Research Article Open Access

Michael Baumgartner* and Christoph Falk

Quantifying the Quality of Configurational Causal Models

DOI: DOI, Received May 24, 2023; revised October 16, 2023; accepted April 12, 2024

Abstract: There is a growing number of studies benchmarking the performance of Configurational Comparative Methods (CCMs) of causal data analysis. A core benchmark criterion used in these studies is a dichotomous (i.e., non-quantitative) correctness criterion, which measures whether all causal claims entailed by a model are true of the data-generating causal structure or not. To date, Arel-Bundock [1] is the only one who has proposed a measure quantifying correctness. That measure, however, as this paper argues, is problematic because it tends to overcount errors in models. Moreover, we show that all available correctness measures are unsuited to assess relations of indirect causation. We therefore introduce a new correctness measure that adequately quantifies errors and does justice to indirect causation. We also offer a new completeness measure quantifying the informativeness of CCM models. Together, these new measures broaden and sharpen the resources for CCM benchmarking.

Keywords: Configurational Causal Modeling, INUS causation, method benchmarking, Coincidence Analysis, Qualitative Comparative Analysis

MSC: 62D20, 03A10

1 Introduction

Configurational Comparative Methods (CCMs) constitute a family of methods of causal learning that track causal complexity by grouping multiple causes into bundles (conjunctions) that only become operative when all of their components are properly co-instantiated and by placing these bundles on alternative (disjunctive) causal paths that can bring about corresponding outcomes independently of one another. CCMs are custom-built to deal with causal structures featuring complex interactions, threshold effects, equifinality, or component causation, which tend to pose challenges for standard methods (e.g. Bayes nets methods or regression methods) because these structures often violate linearity and feature causes and effects that are not correlated in the data, giving rise to violations of causal faithfulness [2]. To this end, CCMs trace causation as defined by modern regularity theories of causation—which define causation in terms of Boolean difference-making and, unlike most other theories, do not entail that pairwise correlation is necessary for causation (cf. [3, 4]).

The two main members of the CCM family are *Qualitative Comparative Analysis* (QCA; [5, 6]) and *Coincidence Analysis* (CNA; [7, 8]). They differ in various aspects, for example, in search targets and implemented algorithms, or in domains of applicability (see [9]). While QCA has been widely used in the social and political sciences, in business administration, or in management, CNA has seen a signifiant uptick in applications in public health in recent years.

Accompanying the increasing dispersion of CCMs, there is a growing body of literature benchmarking the performance of QCA and CNA (e.g. [1, 9–16]). These benchmarking studies conduct inverse searches by, first, (randomly) building data-generating structures—ground truths—, second, simulating data from those structures featuring various deficiencies such as noise or fragmentation, and third, processing the data

***Corresponding Author: Michael Baumgartner:** University of Bergen, Norway; E-mail: michael.baumgartner@uib.no
Christoph Falk: University of Bergen; E-mail: chri.falk@gmail.com

with CCMs to measure the degree to which the produced outputs comply with different quality criteria. One such criterion used in many studies (but not all, cf. [12]) is a dichotomous correctness criterion, which classifies a model as correct if all of its causal implications correspond to causal properties of the ground truth, meaning if it is a submodel of the ground truth, and as incorrect otherwise [11]. That is, a correct model is a model that does not commit a false positive error. But CCM models may have numerous causal implications: they can identify an array of causes, group these causes conjunctively and disjunctively, and they may feature multiple outcomes, for each of which they exhibit disjunctions of conjunctions of causes. Accordingly, only checking whether a model is a submodel of the ground truth amounts to a coarse-grained benchmark. Models that are not submodels of the ground truth can be further compared with respect to how many false implications they have. After all, a model with many true implications and one false positive error is still preferable over a model with many such errors. Hence, measuring correctness not just dichotomously but quantitatively is a natural further development of CCM benchmarking.

However, as this paper will show, adequately quantifying errors in CCM models is an intricate problem. The only solution proposed so far is Arel-Bundock's [1] *wrongness* measure, which counts implications of a model in terms of the number of its submodels and then quantifies wrongness as the proportion of its submodels that are not also submodels of the ground truth. In the first part of this paper, we will argue that this approach is inadequate because it tends to disproportionally overcount errors. In addition, it will be shown that the notion of a submodel, which is at the heart of much of CCM benchmarking to date, is only suited to assess the correctness of models exclusively making claims about relations of *direct* causation, but it cannot handle models expressing *indirect* causation. In consequence, both the standard dichotomous correctness measure and Arel-Bundock's wrongness measure are prone to misjudge the quality of CCM models when the ground truth is a causal chain.

The second part of the paper sets out to rectify these shortcomings. In general terms, correctness of a model relative to a ground truth is the ratio of causal information contained in the model that is true of the ground truth. The problems of overcounting errors and of indirect causation show that sets of submodels do not adequately measure that information content. As an alternative, we introduce the notion of a *causal exposition*, which, in a nutshell, refers to a list of all types of all causal ascriptions, including ascriptions of direct and indirect relevance, made by models and ground truths. These causal expositions can then be intersected and the correctness of the model quantified in terms of the ratio of the complexity of these intersections to the complexity of the model's causal exposition.

Correctness is not the only quality measure, as it exclusively rewards error avoidance and is insensitive to a model's informativeness. Benchmarking studies—in many methodological traditions—, therefore, complement correctness (a.k.a. precision) by a completeness (a.k.a. recall) criterion measuring how much of the ground truth is revealed by a model, that is, how informative a model is [17–20]. In CCM benchmarking, different completeness criteria are in use, some dichotomous, some quantitative [1, 10, 11, 15]. But they typically rely on the notion of a submodel that gives rise to problems when the ground truth is a causal chain. For that reason, we complement our new correctness criterion by an analogous new completeness criterion, which quantifies completeness with exclusive recourse to the tools developed in this paper. To quantify a model's overall quality, we then aggregate its correctness and completeness using the F_β-measure, which is standard in classification theory and machine learning [17]. All new measures and tests are implemented as explicit R functions, which are available in the paper's supplementary material, which moreover provides a script that allows for replicating all calculations of the article.

2 The basics of CCMs

To learn structures featuring causal complexity from data, CCMs draw on the so-called *(M)INUS theory of causation* [3, 4, 21], which is especially suited for the analysis of complexity dimensions that give rise to linearity and faithfulness violations.[1] Contrary to most other theories of causation, the (M)INUS theory does not define causation with recourse to a pairwise dependence between causes and effects. Rather, it defines the relation of causal relevance (i.e., type-level causation) between a factor A taking some value α, A=α, and a factor B taking a value β, B=β, in terms of A=α being a *Boolean difference-maker* of B=β, which, roughly put, amounts to A=α being part of a complex but redundancy-free Boolean function accounting for B=β [3].

Factors in such Boolean functions can either be *crisp-set* (binary), taking two possible values 0 and 1, *fuzzy-set*, taking real values from the unit interval $[0, 1]$, or *multi-value*, taking an open (but finite) number of non-negative integers as possible values. For simplicity, we subsequently focus on crisp-set factors, which allows for abbreviating the "Factor=value" notation. As is conventional in Boolean algebra, we will use "A" as shorthand for A=1 and "a" for A=0.[2] The (M)INUS theory borrows much of its formal machinery from Boolean algebra, in particular, the operations of *negation*, $\neg A$ (expressing "NOT A=1"), *conjunction*, $A{*}B$ ("A=1 AND B=1"), *disjunction*, $A + B$ ("A=1 OR B=1"), *implication*, $A \rightarrow B$ (" IF A=1, THEN B=1"), and *equivalence* $A \leftrightarrow B$ ("A=1 IF, AND ONLY IF, B=1").[3] In case of crisp-set (and multi-value) factors, Boolean operations are given a rendering in classical logic, which we do not reiterate here (see e.g. [23] for a canonical introduction). Based on the implication operator the notions of *sufficiency* and *necessity* are defined, which are the two core dependence relations exploited by the (M)INUS theory: a conjunction $A{*}C{*}E$, for example, is sufficient for B iff (i.e., if, and only if) $A{*}C{*}E \rightarrow B$ (i.e., whenever A AND C AND E are true, B is true); and a disjunction $A + C + E$ is necessary for B iff $B \rightarrow A + C + E$ (i.e., whenever B is true, A OR C OR E is true).

Most sufficiency and necessity relations do not reflect causation, but some of them do, namely the ones that are rigorously freed of redundancies. As shown by Baumgartner and Falk [3], there exists a tight connection between difference-making and redundancy-freeness: A is a Boolean difference-maker of B iff A is a non-redundant part of a minimally sufficient condition Φ_1 (e.g., $A{*}Z_1{*} \ldots {*}Z_n$) of B, such that Φ_1, in turn, is a non-redundant part of a minimally necessary condition $\Phi_1 + \Phi_2 + \ldots + \Phi_n$ of B—where sufficient and necessary conditions are said to be *minimal* iff they do not have proper parts that are, respectively, sufficient and necessary on their own. Correspondingly, CCMs infer minimally necessary disjunctions of minimally sufficient conditions of scrutinized outcomes in disjunctive normal form (DNF),[4] so-called *atomic MINUS-formulas*, from data, which represent causal structures with one outcome. Such one-outcome structures can then be combined to *complex MINUS-formulas* representing multi-outcome structures.[5] (1) is an atomic and (2) a complex exemplar:

$$A{*}b \,+\, c{*}D \;\leftrightarrow\; E \tag{1}$$

$$(H{*}K \,+\, I \;\leftrightarrow\; A) * (A{*}b \,+\, c{*}D \;\leftrightarrow\; E) \tag{2}$$

When causally interpreted, (1) entails that A and b jointly cause E on one path and that c and D jointly cause E on another path. The same also follows from a causal interpretation of (2), but (2) additionally entails that $H{*}K$ and I are two alternative direct causes of A, making them indirect causes of E.

Of course, as deterministic dependencies are rare in (messy) real-life data, strictly sufficient and necessary conditions for an outcome often do not exist. In order to nonetheless distill causal information from

[1] The acronym "INUS" refers to *I*nsufficient but *N*on-redundant parts of *U*nnecessary but *S*ufficient conditions [4, p. 62]. As there are more elegant ways to capture the idea expressed by that expansion, "INUS" is often used as a mere name for a theoretical framework today—void of its original meaning. Accordingly, "MINUS" is a name, without an expansion, locating the corresponding theory in the INUS tradition.

[2] Note that italicization carries meaning: "A" designates the factor and "A" stands for A taking the value 1.

[3] The symbols "$*$" and "$+$" are used as in Boolean algebra here, which means, in particular, that they do not represent the linear algebraic (arithmetic) operations of multiplication and addition (notational variants of Boolean "$*$" and "$+$" are "$\wedge$" and "$\vee$"). For a standard introduction to Boolean algebra see [22].

[4] An expression is in disjunctive normal form iff it is a disjunction of one or more conjunctions of one or more factor values [22, p. 13]

[5] Combining atomic to complex MINUS-formulas requires extra redundancy elimination, which makes it computationally demanding (see [3]). Only CNA builds complex MINUS-formulas.

such data, CCMs approximate deterministic dependency structures by suitably fitting their models to the data. The two core fit measures used for that purpose are *consistency* and *coverage* (for formal definitions, see [24]).[6] Consistency measures the degree to which the behavior of an outcome obeys a corresponding sufficiency or necessity relationship or a whole model; coverage measures the degree to which a sufficiency or necessity relationship or a whole model accounts for the behavior of the corresponding outcome. What counts as acceptable scores on these fit parameters (with values in the unit interval) is defined in thresholds that can either be set by the analyst prior to the analysis or chosen through the robustness protocol recently introduced by Parkkinen and Baumgartner [15]. They determine how closely a dependence in the data must approximate the deterministic ideal in order to pass as one of sufficiency or necessity.

Given their embedding in the (M)INUS theory, CCMs—unlike standard methods—do not infer their outputs from associations (e.g., effect sizes) observed in the data as a whole, rather they exploit difference-making evidence at the level of individual cases (units of observations) in the data. For example, if two cases σ_i and σ_j coincide in all measured factors except for A and B, such that σ_i features A and B and σ_j features a and b, this is evidence—assuming the homogeneity of the unmeasured causal background (for details, see [8])—that there exists a context, *viz.* the one of σ_i and σ_j, in which A makes a difference to B. It follows that A must be part of some conjunction causally relevant for B.

In order to establish, along these lines, that A and C jointly or alternatively cause B, all four logically possible configurations of A and C, namely $A{*}C$, $A{*}c$, $a{*}C$, and $a{*}c$, must be observed in combination with corresponding values of B. In general, the amount of different configurations needed to unambiguously group causes conjunctively or disjunctively increases exponentially with the number of exogenous factors in the analysis. It follows that unambiguously and completely uncovering causal structures by means of CCMs poses very high demands on data diversity; ideally, the behavior patterns of outcomes are observed under all logically possible configurations of exogenous factors. But CCMs are often applied in discovery contexts where such high data diversity is not given. Hence, as CCMs are designed to find all models that equally fit the data, CCM analyses tend to be affected by model ambiguity, meaning that they generate more than one model. Moreover, these models typically are incomplete, that is, they only represent proper parts of data-generating structures.

This has ramifications for the interpretation of CCM models. First, if multiple models $\mathbf{m}_1$ to $\mathbf{m}_n$ are inferred from data, the latter underdetermine their own causal modeling, that is, based on the evidence in the data alone, all of $\mathbf{m}_1$ to $\mathbf{m}_n$ are equally good candidates for being truthful representations of the data-generating structure. Therefore, a CCM output consisting of multiple models is to be interpreted disjunctively: $\mathbf{m}_1$ or $\mathbf{m}_2$ or …or $\mathbf{m}_n$ is true; but the data are insufficient to determine which one(s) exactly.

Second, a model as (1), inferred from data, must be interpreted to be open for later expansions, that is, it must be read with implicit placeholders for additional conjuncts X_i, disjuncts Y_i, and other CCM models Ψ_i [4, see p. 66]:

$$(A{*}b{*}X_1 \;+\; c{*}D{*}X_2 \;+\; Y_1 \;\leftrightarrow\; E)\,{*}\,(\Psi_1) \tag{3}$$

So the fact that, say, G does not appear in model (1) does not entail that G is causally irrelevant to E; it merely means that the data from which (1) was inferred do not contain evidence for the causal relevance of G. By contrast, (1) is committed to all its ascriptions of causal relevance as well as all its ascriptions of conjunctive and disjunctive grouping being true of the complete causal structure regulating the behavior of E—whichever that might be. In other words, the set of causal ascriptions made by a model inferred from data shall be a subset of the causal ascriptions made by the model representing the complete ground truth. In an attempt to define a precise criterion determining when such a subset relation obtains, Baumgartner and Thiem [11] introduced the notion of a *submodel* (we generalize the original definition here):

[6] Consistency is the ratio of true positives to the sum of true and false positives; it is called *positive predictive value* in confusion matrix terminology. Coverage is the ratio of true positives to the sum of true positives and false negatives; it corresponds to *sensitivity* in the confusion matrix.

Submodel. A CCM model $\mathbf{m}_i$ is a *submodel* of another CCM model $\mathbf{m}_j$ iff

(i) if $\mathbf{m}_i$ is an atomic MINUS-formula $\Omega \leftrightarrow Z$, there exists an atomic MINUS-formula $\Gamma \leftrightarrow Z$ in $\mathbf{m}_j$ such that either $\Gamma = \Omega$ or Γ can be transformed into Ω by mere elimination of conjuncts or disjuncts;

(ii) if $\mathbf{m}_i$ is a complex MINUS-formula, all atomic MINUS-formulas in $\mathbf{m}_i$ have counterparts in $\mathbf{m}_j$ for which (i) is satisfied.

For example, $A{*}B \leftrightarrow C$ is a submodel of $A{*}B{*}D \leftrightarrow C$ and of $A{*}B + D \leftrightarrow C$ because $A{*}B{*}D$ and $A{*}B + D$ can be transformed into $A{*}B$ merely by eliminating conjuncts or disjuncts, but not of $A + B \leftrightarrow C$ because $A + B$ cannot be transformed into $A{*}B$ in that way.

If $\mathbf{m}_i$ is a submodel of $\mathbf{m}_j$, $\mathbf{m}_j$ is called a *supermodel* of $\mathbf{m}_i$. The submodel relation is reflexive: every model is a submodel (and supermodel) of itself. Put differently, if $\mathbf{m}_i$ and $\mathbf{m}_j$ are submodels of one another, then $\mathbf{m}_i$ and $\mathbf{m}_j$ are identical. Although the submodel relation, strictly speaking, can only be said to obtain between CCM models, we will subsequently also say, for convenience, that a model $\mathbf{m}_i$ is a submodel of the ground truth (instead of $\mathbf{m}_i$ being a submodel of the model representing the ground truth).

3 Assessing model quality by submodel criteria

3.1 The state-of-the-art in CCM benchmarking

Even though the output a CCM infers from limitedly diverse data often contains more than one model and even though these models cannot be expected to reflect the complete ground truth, the output as a whole can and should be expected to truthfully reflect the data-generating structure. This is satisfied if at least one output model $\mathbf{m}_i$ is a submodel of the model representing the complete ground truth. Against that backdrop, the following is a *qualitative correctness criterion* frequently used in CCM benchmarking [see e.g. 8–11, 15, 16]:[7]

Qualitative Correctness (LCR). A model $\mathbf{m}$ is a correct representation of a ground truth Δ iff $\mathbf{m}$ is a submodel of Δ.

While being important in current CCM benchmarking, (LCR) is clearly insufficient to assess the overall quality of models. For one, (LCR) does not take model complexity or informativeness into account. Models can be very sparse or very complex submodels of the ground truth, yet equally satisfy (LCR). Hence, correctness needs to be complemented by a *completeness criterion* suitably rewarding informativeness.[8] There are various completeness criteria on offer, some qualitative [8], some quantitative [1, 10, 15, 16], but they all measure completeness by drawing on the submodel relation.

Another reason why (LCR) does not suffice for assessing model quality is that it is merely qualitative, meaning it can only be passed or not. As a result, (LCR) cannot capture important differences. To illustrate, let Δ_1 be the ground truth and let models (4) and (5) be inferred from data simulated from that ground

[7] Dusa [12] rejects this criterion and, instead, requires disjuncts in a model $\mathbf{m}$ to be complete in order for $\mathbf{m}$ to count as correct. We take Dusa's position to entail that hardly any CCM models of real-world systems can be correct and, hence, do not discuss it further here (see [25] for a detailed discussion of Dusa's proposal).

[8] Dusa [12] does not conceive of correctness and completeness as independent quality measures. In our view, however, clearly distinguishing between correctness and completeness is crucial for a balanced assessment of model quality. It allows for rewarding models for the true claims they make independently of punishing them for the true claims they fail to make, and it is commonplace in method benchmarking across a wide range of fields (where correctness is often called *precision* and completeness is known as *recall*).

truth in a benchmark test:

$$A{*}b \,+\, c{*}D \;\leftrightarrow\; E \tag{Δ_1}$$

$$A{*}B \,+\, D \;\leftrightarrow\; E \tag{4}$$

$$A{*}B{*}D \;\leftrightarrow\; E \tag{5}$$

As neither (4) nor (5) are submodels of Δ_1, they are both incorrect according to (LCR). But there is a clear sense in which (4) is not equally incorrect as (5). While (4) correctly entails that A and D are causally relevant and places these causes in alternative disjuncts, it erroneously ascribes causal relevance to B (instead of b). (5) makes that same mistake and, in addition, erroneously combines D conjunctively with A. That is, (5) commits one error more than (4). It should count as a worse representation of Δ_1 than (4). However, (LCR), being a mere qualitative criterion, is insensitive to such differences in numbers of errors.

To date, Arel-Bundock [1] is the only one who has proposed a measure that is sensitive to such differences by expressing correctness quantitatively. Strictly speaking, Arel-Bundock does not define a measure for model correctness but for model wrongness: "I measure the level of wrongness by counting the proportion of solution submodels that are not submodels of the truth" [1, p. 7]. But to adjust this proposal to our preferred terminology (which is also standard in the benchmarking literature) we transform Arel-Bundock's wrongness measure into a *quantitative correctness measure* (by negating it):

Quantitative Correctness (NCR). The correctness of a model $\mathbf{m}$ for a ground truth Δ is the proportion of $\mathbf{m}$'s submodels that are also submodels of Δ.

To illustrate, we apply (NCR) to models (4) and (5). Table 1 lists all submodels of (4) and (5), respectively, and indicates whether they are submodels of the ground truth Δ_1. Three of the seven submodels of (4) are also submodels of Δ_1, yielding a (NCR)-score of 0.43. With only two of its seven submodels being submodels of Δ_1, (5) gets a (NCR)-score of 0.29. That these scores are below 1 and above 0 reflects the fact that neither (4) nor (5) are fully correct representations of Δ_1 while still making some true claims. Furthermore, (4) receives a higher score than (5) because it makes less errors. On the face of it, (NCR) thus seems to capture exactly those differences that (LCR) is insensitive to. However, the next two sections will show that (NCR) does not adequately score model correctness in all cases.

$A{*}B + D \leftrightarrow E$	sub(Δ_1)	$A{*}B{*}D \leftrightarrow E$	sub(Δ_1)
$A \leftrightarrow E$	✓	$A \leftrightarrow E$	✓
$B \leftrightarrow E$	✗	$B \leftrightarrow E$	✗
$D \leftrightarrow E$	✓	$D \leftrightarrow E$	✓
$A{*}B \leftrightarrow E$	✗	$A{*}B \leftrightarrow E$	✗
$A + D \leftrightarrow E$	✓	$A{*}D \leftrightarrow E$	✗
$B + D \leftrightarrow E$	✗	$B{*}D \leftrightarrow E$	✗
$A{*}B + D \leftrightarrow E$	✗	$A{*}B{*}D \leftrightarrow E$	✗
	$3/7 = 0.43$		$2/7 = 0.29$

Table 1. All submodels of models (4) and (5), respectively, with marks indicating whether a submodel is also a submodel of the ground truth Δ_1 and resulting (NCR)-scores.

3.2 The problem of overcounting errors

The first problem of (NCR) is best introduced with another concrete example. Thus, let Δ_2 be the ground truth and let models (6) to (8) be inferred from data simulated from Δ_2:

$$A{*}b{*}D{*}F \;+\; a{*}B{*}C{*}D \;\leftrightarrow\; E \tag{Δ_2}$$

$$A \;+\; C \;+\; D \;\leftrightarrow\; E \tag{6}$$

$$A{*}b \;+\; C \;+\; D \;\leftrightarrow\; E \tag{7}$$

$$A{*}b{*}F \;+\; C \;+\; D \;\leftrightarrow\; E \tag{8}$$

The important feature of candidate models (6) to (8) is that they all contain the same error: instead of adding D to the first or second disjunct, they place D into a third disjunct, thus, claiming that D brings about E independently of the other factors. Apart from that mistake, all other causal claims entailed by (6) to (8) are true of Δ_2. More specifically, the difference between (6) and (7) is that the latter truthfully identifies $A{*}b$ as a cause of E, while in the former b is not part of the first disjunct. That is, (7) makes the same mistake as (6) and contains more true information. Analogously, (8) features the same error as (7) (and (6)) in combination with the true conjunctive addition of F to $A{*}b$.

Clearly, an adequate correctness measure must not punish models (7) and (8) for containing more true elements than (6) while committing the same error as (6). More generally, an adequate correctness measure should respect the following *model expansion principle*:

Model Expansion Principle (MEP). Expanding a model by truthfully located elements from the ground truth cannot reduce correctness.

(NCR), however, does not respect (MEP). It assigns the highest correctness score to (6) and the lowest to (8). Model (6) has a total of seven submodels, six of which are also submodels of Δ_2, yielding an (NCR)-score of $6/7 = 0.86$, whereas (7) and (8) only reach (NCR)-scores of $12/15 = 0.80$ and $24/31 = 0.77$, respectively.[9] The reason for this inadequate scoring, in a nutshell, is that (NCR) counts both true and false claims made by models multiple times, in a possibly disproportional manner, which leads to an overcounting of false claims, that is, of errors in case of models (7) and (8).

To bring this out more clearly, we take a closer look at (7). Table 2 lists all of (7)'s 15 submodels. The truthful submodels, that is, the ones that are submodels of Δ_2, are in the left half of the table, the

#	$A{*}b+C+D \leftrightarrow E$	sub(6)	subΔ_2	#	$A{*}b+C+D \leftrightarrow E$	sub(6)	subΔ_2
sm$_1$	$A \leftrightarrow E$	✓	✓	**sm**$_{13}$	$A+C+D \leftrightarrow E$	✓	✗
sm$_2$	$C \leftrightarrow E$	✓	✓	**sm**$_{14}$	$b+C+D \leftrightarrow E$	✗	✗
sm$_3$	$D \leftrightarrow E$	✓	✓	**sm**$_{15}$	$A{*}b+C+D \leftrightarrow E$	✗	✗
sm$_4$	$A+C \leftrightarrow E$	✓	✓				
sm$_5$	$A+D \leftrightarrow E$	✓	✓				
sm$_6$	$C+D \leftrightarrow E$	✓	✓				
sm$_7$	$b \leftrightarrow E$	✗	✓				
sm$_8$	$A{*}b \leftrightarrow E$	✗	✓				
sm$_9$	$b+C \leftrightarrow E$	✗	✓				
sm$_{10}$	$b+D \leftrightarrow E$	✗	✓				
sm$_{11}$	$A{*}b+C \leftrightarrow E$	✗	✓				
sm$_{12}$	$A{*}b+D \leftrightarrow E$	✗	✓				

Table 2. The 15 submodels of (7) with indications of whether they are submodels of (6) and Δ_2 as well.

[9] To see all relevant submodels, consult the paper's replication script at https://github.com/m-baum/quantifyQuality.

false ones in the right. For each submodel, the table moreover indicates whether it is also a submodel of (6). The first thing to highlight is the tendency of (NCR) to count false and true claims made by (7) and its submodels multiple times. For instance, according to Δ_2 it is false to say, as does $\mathbf{sm}_{15}$, that C and D are parts of alternative causes of E, rather they are causally relevant in conjunction. This entails that submodels $\mathbf{sm}_{13}$ and $\mathbf{sm}_{14}$ are also false, as they result from $\mathbf{sm}_{15}$ by mere elimination of a conjunct. The error contained in $\mathbf{sm}_{13}$ and $\mathbf{sm}_{14}$ is the same as the error in $\mathbf{sm}_{15}$. Analogously, given that $\mathbf{sm}_{11}$ is a submodel of Δ_2, and thus only makes true causal claims, it follows that all submodels of $\mathbf{sm}_{11}$, as $\mathbf{sm}_7$ to $\mathbf{sm}_9$, are also submodels of Δ_2, and thus true of Δ_2. That is, models $\mathbf{sm}_7$ to $\mathbf{sm}_9$ do not reveal any truths about Δ_2 not revealed by $\mathbf{sm}_{11}$. Although many submodels of (7) commit the same errors or reveal the same truths, (NCR) counts all of them separately in its correctness calculation.

Now, observe what proportions of true and false submodels are added when model (6) is expanded to (7). Model (6) has seven submodels, which are also marked in Table 2. Six of these submodels are true, one is false. When b is truthfully integrated into (6) to yield model (7), six true and two false submodels are added to the count. That is, the proportion of false submodels increases by a factor of $3/1 = 3$, whereas the proportion of true submodels only multiplies by $12/6 = 2$. In other words, even though (7) results from (6) by integrating true elements only, disproportionally more false than true submodels are thereby introduced. The same happens when (7) is further expanded to (8). It follows that measuring correctness in terms of proportions of true submodels, as done by (NCR), cannot possibly do justice to (MEP).

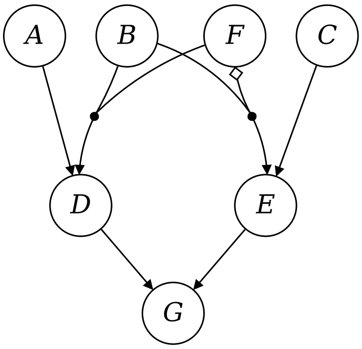
We take this to show, not only that (NCR) does not adequately quantify model correctness, but that any attempt to quantify correctness based on proportions of true or false submodels faces a risk of miscounting errors, because those proportions can be twisted under model expansion and thus are not guaranteed to respect (MEP).

3.3 The problem of indirect causation

To date, CCM benchmarking has predominantly focused on QCA's or CNA's success in recovering single-outcome models, that is, atomic MINUS-formulas. Correspondingly, both (LCR) and (NCR) are custom-built for correctness assessment in single-outcome recovery. This section argues that (LCR) and (NCR) are in fact inadequate when the data are generated by multi-outcome structures with causally related outcomes, that is, by causal chains. In a nutshell, the reason is that models leaving out intermediate links on causal paths to an ultimate outcome may be perfectly correct without being a submodel of the ground truth or even containing such a submodel.

To see this, consider the causal chains in the hypergraph of Figure 1. This graph has two non-standard elements that require introduction: arrows merged by "•" symbolize conjunctive relevance, and "⋄" expresses that the negation of the factor at the tail of the arrow is relevant. Another notable feature of that structure, which will become important in section 4.1, is the switching factor F: its positive value F determines that the impact of B on G is transmitted via D and its negative value f causes that impact to be mediated by E (for more details see [3]). The complex MINUS-formula in Δ_3 expresses that switching structure. Let us assume that Δ_3 is the ground truth used to simulate data in some benchmark test in which the examined method returns model (9). When causally interpreted, (9) claims that A and B are causally relevant for G and that they are parts of alternative causes producing G independently of one another. Both of these claims are indeed true of Δ_3, according to which A and B are alternative causes of D and D is a cause of G, making A and B indirect alternative causes of G. Model (9) is just incomplete. It leaves out the middle link mediating the causal influence of A and B to G—as well as numerous other causes of G. But, as we have seen before, incompleteness does not make a model incorrect.

One might be inclined to respond that, despite making numerous true claims about Δ_3, (9) also falsely claims that A and B are direct causes of G, where in truth they are indirect causes. This response presupposes that there is an objective fact of the matter as to whether a cause—in truth—directly or indirectly brings about its effect. In light of the (widely assumed) continuity of spacetime, however, it is possible to interpolate (suitably defined) intermediate factors on virtually any causal path between



$$(A + B{*}F \leftrightarrow D){*}(C + B{*}f \leftrightarrow E){*}(D + E \leftrightarrow G) \qquad (\Delta_3)$$

$$A + B \leftrightarrow G \qquad (9)$$

Figure 1. A causal chain with switching factor F, the corresponding complex MINUS-formula Δ_3, and a candidate model (9). Arrows merged by "•" symbolize conjunctive relevance and "⋄" expresses that the negation of the factor at the tail of the arrow is relevant.

two factors. Only extremely fine-grained models representing causal structures on the level of objectively fundamental particles—if such exist at all—, could conceivably trace direct causation. Such a view would entail that all macro-level models are incorrect (to some degree) because they represent causal dependencies as direct, which in fact are mediated by intermediate links.

To avoid that consequence, it is standard to view the distinction between direct and indirect causation as inherently relative to the factors contained in a given model [26, 27]. That means that a causal relation can be truthfully represented as a direct one in a first model and as an indirect one in a second. Relative to the factors in model (9), A and B are indeed direct causes of G, because D and E are not contained in (9). But as D and E are contained in Δ_3, the relevance of A and B for G becomes mediated and thus indirect. But Δ_3 might likewise be expandable by further intermediate links, whereby relations represented as direct ones in Δ_3 would be turned into indirect ones. There is no need to stipulate that Δ_3 is an objectively fundamental representation of a causal structure; rather, it truthfully depicts a segment of reality relative to a set of factors suited for that purpose. But the same segment might also be truthfully represented on another level of granularity using other factors.[10]

Against that backdrop, model (9) is incomplete but does not commit an error. An adequate correctness measure should thus reward it with a maximal score. However, both (LCR) and (NCR) fail to do so. The only atomic MINUS-formula for outcome G (i.e., the outcome of (9)) contained in Δ_3 is this one:

$$D + E \leftrightarrow G \qquad (10)$$

But (9) is neither itself a submodel of (10), and thereby of Δ_3, nor does it contain a submodel that would be a submodel of (10), that is, of Δ_3. It follows that (9) does not pass (LCR) and that it receives an (NCR)-score of $^0/_3 = 0$.

4 A new approach to correctness assessment

In the most general terms, correctness of a model $\mathbf{m}$ relative to a ground truth Δ is the ratio of causal information contained in $\mathbf{m}$ that is true of Δ to the totality of causal information contained in $\mathbf{m}$. In other words, it is the ratio of true positives entailed by $\mathbf{m}$ to the sum of true positives and false positives entailed by $\mathbf{m}$—which is also known as *precision* in many fields [17, 20]. The problems of overcounting

[10] Note that viewing the distinction between direct and indirect causation to be model relative neither entails that causation itself is model relative nor that there is no objective fact of the matter whether claims as "X causes Y" are true or false (see [27, section 7] for an extended discussion).

errors and of indirect causation show that sets of submodels of **m** and Δ are not reliable indicators of the information content, or the amounts of true and false positives, relevant for correctness assessments. As an alternative, we propose to identify that content by unpacking all different types of causal ascriptions implied by MINUS-formulas in what we will call *causal expositions*. The causal expositions of **m** and Δ can then be intersected and the correctness of **m** quantified in terms of the ratio of the complexity of these intersections to the complexity of the causal exposition of **m**. The remainder of this section renders that basic idea more precise.

4.1 Building causal expositions

MINUS-formulas contain four types of causal information: ascriptions of causal relevance (i) to individual factor values (or literals), (ii) to conjunctions, (iii) to disjunctions, and (iv) sequential orderings of causal relations in causal paths. For brevity, we refer to these types as *literal*, *conjunctive*, *disjunctive*, and *sequential* ascriptions, respectively. To illustrate, reconsider Δ_3, which represents the switching structure in Figure 1:

$$(A + B{*}F \leftrightarrow D){*}(C + B{*}f \leftrightarrow E){*}(D + E \leftrightarrow G) \qquad (\Delta_3)$$

Among many others, Δ_3 makes the literal ascription that A is causally relevant to G, the conjunctive ascription that $B{*}F$ is relevant to D, the disjunctive ascription that $D{+}E$ is relevant to G, or the sequential ascription that there exists a causal path from A via D to G, expressible as ordered sequence $\langle A, D, G\rangle$. We call the compilation of all causal information contained in a MINUS-formula its *causal exposition*:

Causal Exposition. The causal exposition of a MINUS-formula **m** is the list of all literal, conjunctive, disjunctive, and sequential ascriptions entailed by **m**.

One lesson to learn from the problem of indirect causation is that causal expositions cannot simply be read off the syntax of a MINUS-formula (or of its submodels), because MINUS-formulas only represent direct causation (relative to the factors in the formula) and lack a syntactic expression of indirect causation. But information about indirect causation, and thus causal expositions, can be recovered from MINUS-formulas by syntactic transformations standard in Boolean algebra.

Viewed as a mere Boolean expression, the first atomic MINUS-formula in Δ_3, *viz.* $A + B{*}F \leftrightarrow D$, states that $A + B{*}F$ and D are equivalent, which entails that they are substitutable for one another without breach of Boolean dependence relations of sufficiency and necessity. This substitutability principle allows for replacing D in the third atomic formula in Δ_3, *viz.* in $D + E \leftrightarrow G$, by $A + B{*}F$:

$$A + B{*}F + E \leftrightarrow G \qquad (11)$$

(11) is automatically in disjunctive normal form (DNF). In other examples, additional transformations—for instance, factoring out—may be required to bring expressions resulting from such substitutions into DNF; but any Boolean expression can easily be brought into DNF. The substitutability principle ensures that, if $D + E \leftrightarrow G$ truthfully expresses Boolean dependence relations, then so does (11). But the principle does not guarantee that the expression resulting from the substitution remains redundancy-free and thus causally interpretable. And indeed, (11) contains a redundancy: in the set of all configurations compatible with Δ_3, that is, in so-called *ideal data* (i.e., noise-free and unfragmented data) generated from Δ_3, F does not make a difference to G. The factor F is a mere switch in Δ_3; its positive value F determines that the causal impact of B is transmitted via D to G and its negative value f causes that impact to be mediated by E; but whichever value F takes, B itself is sufficient for G.[11] Hence, $B{*}F$ in (11) is only sufficient for G but not minimally so.

[11] This shows that causation as defined by the (M)INUS-theory is not a transitive relation. F is a cause of D, which is a cause of G, but F is no cause of G (see [3]).

If we additionally minimize sufficient and necessary conditions in (11) relative to ideal data on Δ_3 (e.g. by means of Quine-McCluskey optimization [28]), we get this expression:

$$A + B + E \leftrightarrow G \tag{12}$$

(12) has the form of an atomic MINUS-formula. As it results from syntactic transformations of Δ_3, it can be seen as representing relations of indirect causation entailed by Δ_3. It states that A and B are causally relevant for G, which, relative to the set of factors in Δ_3, amounts to indirect relevance. For brevity, we call it an *indirect* MINUS-formula relative to Δ_3. Indirect MINUS-formulas are recoverable from complex MINUS-formulas by substitution of equivalents, DNF transformation, if needed, and Boolean minimization.

Two further indirect MINUS-formulas can be recovered from Δ_3 in the same way. (13) is built by substituting $C + B{*}f$ for E in $D + E \leftrightarrow G$, and (14) is the result of replacing both D and E by their equivalents in Δ_3 and subsequent minimization:

$$B + C + D \leftrightarrow G \tag{13}$$

$$A + B + C \leftrightarrow G \tag{14}$$

(12), (13), and (14) are all the indirect MINUS-formulas recoverable from Δ_3. We will call the union of all atomic (direct) MINUS-formulas in Δ_3 and all indirect MINUS-formulas recoverable from it the *chain-expansion* of Δ_3. But before we can explicitly define that notion, we have to consider the case where the complex MINUS-formula to be chain-expanded is not a ground truth but a model inferred from data.

Hence, suppose that the following multi-outcome model is inferred from data simulated from ground truth Δ_3:

$$(A + B{*}F \leftrightarrow D){*}(D + E \leftrightarrow G) \tag{15}$$

If we substitute D in $D + E \leftrightarrow G$ by its equivalent $A + B{*}F$ and then minimize relative to ideal data on (15), we do not end up with (12) but with (11). That is, if indirect MINUS-formulas are recovered from (15) through Boolean minimization relative to ideal data on (15), F appears to make a difference to G because F is no switching factor in (15). However, (15) is not inferred from ideal data generated from itself but from data simulated from Δ_3, and according to Δ_3, F does not make a difference to G. That means that the data from which (15) is inferred do not contain evidence for the indirect relevance of F for E. It would therefore not be adequate to recover an indirect relevance ascription from (15) for which there is no evidential basis in the discovery context of that model. We thus submit that when indirect causation is recovered from models inferred from data, Boolean minimization should be conducted relative to that actual data and not, as in case of chain-expanding ground truths, relative to ideal data. In sum, the following is our definition of the notion of a chain-expansion:

Chain-Expansion. The chain-expansion of a MINUS-formula **m** is the union of the atomic (direct) MINUS-formulas contained in **m** and the indirect MINUS-formulas recoverable from **m** by substitution of equivalents, DNF transformation, and Boolean minimization, either relative to the data from which **m** is inferred or, if **m** is not inferred from data, relative to ideal data on **m**.

The important feature of chain-expansions for quantifying model quality is that they syntactically represent all types of causal ascriptions entailed by a MINUS-formula **m**. The literal, conjunctive, and disjunctive ascriptions of **m** for an outcome Z are simply the sets of all factor values and all maximally long conjunctions and disjunctions—freed of duplicates—that appear on the left of "↔" in the atomic MINUS-formulas for Z in **m**'s chain-expansion. The sequential ascriptions for outcome Z are the maximally long ordered sequences of factor values $\langle X_1, \ldots, X_n \rangle$ satisfying the following *path-rule*: for all X_i and X_j with $i < j$ in $\langle X_1, \ldots, X_n \rangle$, there is a MINUS-formula in **m**'s chain-expansion with X_i on the left and X_j on the right of "↔", and Z is the last element of the sequence (i.e., $X_n = Z$).

Table 3 lists the chain-expansion of Δ_3 in the left-most column and the causal exposition, subdivided by outcomes, in the other columns. The literal ascriptions for each outcome in Δ_3 can be recovered from the chain-expansion by removing conjunctors "*", disjunctors "+", and duplicates from the expressions

chain-expansion	**causal exposition**				
		literals	conjunctions	disjunctions	sequences
$A + B{*}F \leftrightarrow D$	D:	$\{A, B, F\}$	$\{A, B{*}F\}$	$\{A + B{*}F\}$	$\{\langle F, D\rangle, \langle B, D\rangle, \langle A, D\rangle\}$
$C + B{*}f \leftrightarrow E$	E:	$\{C, B, f\}$	$\{C, B{*}f\}$	$\{C + B{*}f\}$	$\{\langle f, E\rangle, \langle B, E\rangle, \langle C, E\rangle\}$
$D + E \leftrightarrow G$ $A + B + E \leftrightarrow G$ $B + C + D \leftrightarrow G$ $A + B + C \leftrightarrow G$	G:	$\{A, B, D,$ $C, E\}$	$\{A, B, D,$ $C, E\}$	$\{D + E,$ $A + B + E,$ $B + C + D,$ $A + B + C\}$	$\{\langle A, D, G\rangle,$ $\langle B, D, G\rangle,$ $\langle B, E, G\rangle,$ $\langle C, E, G\rangle\}$

Table 3. Chain-expansion and causal exposition of ground truth Δ_3.

on the left of "↔". The conjunctive ascriptions are obtained by removing "+" and duplicates, and the disjunctive ascriptions are simply the expressions on the left of "↔". Note that, in case of outcome G, conjunctive ascriptions are identical to literal ones because none of G's MINUS-formulas actually features a conjunctor, and a single factor value formally counts as a trivial conjunction (and disjunction). Finally, sequential ascriptions are built by combining as many factor values as possible from the literal ascriptions following the path-rule for every outcome. In case of outcome G, this amounts to combining the factor values on the left-hand sides of D's and E's MINUS-formulas with D and E and adding G if, and only if, the first element of the sequence also appears on the left-hand side of a MINUS-formula of G.

4.2 Intersecting causal expositions

To quantify the correctness of a model $\mathbf{m}$ relative to a ground truth Δ, we propose to intersect the literal, conjunctive, disjunctive, and sequential ascriptions rendered transparent by the causal expositions of $\mathbf{m}$ and Δ. The ratios of the complexities of these intersections to the complexities of $\mathbf{m}$'s literal, conjunctive, disjunctive, and sequential ascriptions then yield measures for *literal*, *conjunctive*, *disjunctive*, and *sequential correctness*.

To make that concrete, assume that the following model is inferred from data generated by ground truth Δ_3.

$$(A{*}B \ \leftrightarrow \ D){*}(D \ + \ B{*}C \ \leftrightarrow \ G) \tag{16}$$

Model (16), which contains no information about outcome E, makes two false claims about Δ_3: first, it erroneously places A and B in the same conjunctive cause of D, and second, B and C appear in the same conjunctive cause of G, which in truth are alternative indirect causes of G. But all literal ascriptions and the placement of D in a separate disjunct leading to G are true of Δ_3. To quantify the correctness of (16), we first chain-expand that model by replacing D in the atomic MINUS-formula of G by $A{*}B$ and then build its causal exposition. The result is in Table 4.

Intersecting the literal, conjunctive, and disjunctive ascriptions of (16) and Δ_3 (cf. Table 3) for each outcome is straightforward. The *literal intersection* is the set of factor values contained in the literal ascriptions of both (16) and Δ_3. The *conjunctive intersection* is the set of all conjunctions with a maximal amount of conjuncts that can be reached from the conjunctive ascriptions of both (16) and Δ_3 by mere

chain-expansion	**causal exposition**				
		literals	conjunctions	disjunctions	sequences
$A{*}B \leftrightarrow D$	D:	$\{A, B\}$	$\{A{*}B\}$	$\{A{*}B\}$	$\{(B, D),\ (A, D)\}$
$D + B{*}C \leftrightarrow G$ $A{*}B + B{*}C \leftrightarrow G$	G:	$\{A, B,$ $C, D\}$	$\{D, B{*}C,$ $A{*}B\}$	$\{D + B{*}C$ $A{*}B + B{*}C\}$	$\{(A, D, G),\ (B, D, G),$ $(B, G),\ (C, G)\}$

Table 4. Chain-expansion and causal exposition of model (16).

	out.	model (16)	ground truth Δ_3	intersection	ratio	weight	correctness
lit.	D:	$\{A, B\}$	$\{A, B, F\}$	$\{A, B\}$	**2/2**	2/6	1
	E:		$\{C, B, f\}$				
	G:	$\{A, B, C, D\}$	$\{A, B, C, D, E\}$	$\{A, B, C, D\}$	**4/4**	4/6	
conj.	D:	$\{A * B\}$	$\{A, B{*}F\}$	$\{A, B\}$	**1/2**	2/7	0.57
	E:		$\{C, B{*}f\}$				
	G:	$\{D, B * C, A * B\}$	$\{D, B, A, C, E\}$	$\{D, B, A, C, E\}$	**3/5**	5/7	
dis.	D:	$\{A * B\}$	$\{A + B{*}F\}$	$\{A, B\}$	**1/2**	2/9	0.56
	E:		$\{C + B{*}f\}$				
	G:	$\{D + B * C,$ $A * B + B * C\}$	$\{C + E,$ $A + B + E,$ $B + C + D,$ $A + B + C\}$	$\{D + B, D + C,$ $A + B, A + C,$ $B + C\}$	**4/7**	7/9	
seq.	D:	$\{\langle B, D\rangle, \langle A, D\rangle\}$	$\{\langle F, D\rangle, \langle B, D\rangle,$ $\langle A, D\rangle\}$	$\{\langle B, D\rangle, \langle A, D\rangle\}$	**2/2**	2/6	1
	E:		$\{\langle f, E\rangle, \langle B, E\rangle,$ $\langle C, E\rangle\}$				
	G:	$\{\langle A, D, G\rangle,$ $\langle B, D, G\rangle,$ $\langle B, G\rangle, \langle C, G\rangle\}$	$\{\langle A, D, G\rangle,$ $\langle B, D, G),$ $\langle B, \cancel{E}, G\rangle,$ $\langle C, \cancel{E}, G\rangle\}$	$\{\langle A, D, G\rangle,$ $\langle B, D, G\rangle,$ $\langle B, G\rangle, \langle C, G\rangle\}$	**4/4**	4/6	

overall correctness (Corr): $6/28 \cdot 1 \ + \ 7/28 \cdot 0.57 \ + \ 9/28 \cdot 0.56 \ + \ 6/28 \cdot 1 \ = \ \boxed{\mathbf{0.75}}$

Table 5. Intersections and correctness scoring for model (16) relative to ground truth Δ_3. Grey shading indicates the expressions used for calculating the correctness ratios. "$\cancel{E}$" represents the removal of E in order to harmonize the factor sets of (16) and Δ_3.

elimination of conjuncts. For example, the (trivial) conjunction B can be reached from (16)'s conjunctive ascription $A{*}B$ for outcome D by elimination of A as well as from Δ_3's conjunctive ascription $B{*}F$ for the same outcome by elimination of F, and there are no conjunctions reachable in that manner with more conjuncts, meaning that B has a maximal amount of conjuncts. The *disjunctive intersection* is the set of all disjunctions with a maximal amount of disjuncts that can be reached from the disjunctive ascriptions of both (16) and Δ_3 by elimination of disjuncts and conjuncts. For example, the disjunctive ascription $D + B$ can be reached by elimination of C from (16)'s disjunctive ascription $D + B{*}C$ for outcome G as well as from Δ_3's disjunctive ascription $B + C + D$ for the same outcome, and there are no longer disjunctions reachable in that manner. Table 5 lists all intersections of (16) and Δ_3. As can easily be seen from that table, conjunctive and disjunctive intersections tend to contain multiple elements; that is, multiple conjunctions and disjunctions with maximal amounts of conjuncts and disjuncts can be reached from both (16) and Δ_3.

As the difference between direct and indirect causal relevance is relative to a set of modeled factors, the correctness of the sequential ascriptions of a model must be assessed relative to its set of factors. This, in turn, requires that the sequential ascriptions of the corresponding ground truth be pruned to the model's factor set before intersecting. More concretely, model (16) is correct to entail that B is a direct cause of G, despite B being an indirect cause of G in Δ_3. The reason is that (16) does not feature the factor E, which is the intermediate link between B and G in Δ_3, meaning that relative to the factors in (16) B indeed is a direct cause of G. Hence, before intersecting sequential ascriptions, factor E must be removed from the sequential ascriptions of Δ_3, which is represented by "$\cancel{E}$" in Table 5. After such harmonizing, the sequential intersection of (16) and Δ_3 simply comes down to the set of sequential ascriptions made by both (16) and Δ_3.

4.3 Quantifying correctness

An intersection expresses the amount of causal information of a particular type shared by the model and ground truth, in other words, it expresses the causal claims made by the model that are true of the ground truth, that is, the model's true positives. As correctness is a measure for the ratio of true information in a model, the next step towards putting a number on the correctness of (16), is to quantify the complexities of intersections and corresponding ascriptions. For literals, conjunctions, and disjunctions we quantify complexities in terms of numbers of factor values. For instance, the set $\{A, B\}$ of (16)'s literal ascriptions for outcome D has complexity 2 because it contains 2 factor values; or the set $\{D + B{*}C, A{*}B + B{*}C\}$ of its disjunctive ascriptions for outcome G has complexity 7 because it contains 7 factor values. The ratio of true information, then, is the ratio of factor values in these ascriptions that have counterparts in the corresponding intersections. Thus, since all factor values in $\{A, B\}$ have counterparts in the literal intersection (see the first row of Table 5), the correctness ratio of $\{A, B\}$ is $^2/_2$. By contrast, the disjunctive ascriptions for outcome G are not completely represented in the disjunctive intersection (row 9 of Table 5). The first disjunction $D + B{*}C$ can be paired with either $D + B$ or $D + C$ in the corresponding intersection, and, since both of the latter have equal complexity, it does not matter which of them is chosen as counterpart. The same holds for the second disjunction $A{*}B + B{*}C$: it can be paired with either $A + B$ or $A + C$ or $B + C$ in the intersection. Whichever elements of the intersection are chosen as counterparts, a total of 4 of the 7 factor values in the set of disjunctive ascriptions for outcome G have counterparts in the corresponding intersection, yielding a correctness ratio of $^4/_7$. In Table 5, the factor values used for calculating the correctness ratios are highlighted with grey shading.

For sequences, we aim to avoid unnecessary double-counting by quantifying complexities not in terms of the number of factor values but in terms of the number of paths. That is, correctness ratios for the sequential ascriptions of a model are ratios of the model's paths that are contained in the sequential intersection with the ground truth. For example, as both paths in the set of sequential ascriptions $\{\langle B, D\rangle, \langle A, D\rangle\}$ for outcome D are also contained in the sequential intersection for that outcome, that set receives a correctness ratio of $^2/_2$.

The next step to a correctness quantification of (16) consists in aggregating these correctness ratios of the component ascriptions. We choose a weighted mean for that purpose, where weights are the complexity shares of component ascriptions. For literal, conjunctive, and disjunctive correctness, weights are calculated based on the number of factor values in a corresponding ascription. In the case of sequential correctness, they are based on the number of paths. For instance, for both outcomes combined, the conjunctive ascriptions of (16) have a total complexity of 7 factor values, with 2 pertaining to outcome D and 5 to outcome G. That is, the weights for the component ascriptions $\{A{*}B\}$ and $\{D, B{*}C, A{*}B\}$ are $^2/_7$ and $^5/_7$, respectively. Weighing the components' ratios by these weights yields a conjunctive correctness score of 0.57. Or, the sequential ascription of (16) contains a total of 6 paths, with 2 leading to outcome D and 4 to outcome G, resulting in the weights $^2/_6$ and $^4/_6$, respectively, and an overall sequential correctness score of 1. Table 5 provides an overview of all weights and resulting correctness scores.

Finally, the four correctness scores must be aggregated into one overall score, which we again do with a weighted mean. The weights are based on the complexity shares of a model's whole causal exposition covered by a corresponding correctness score. The total complexity of the causal exposition is the sum of the complexities of the four types of causal ascriptions. For model (16), it is $6 + 7 + 9 + 6 = 28$, resulting in the weights indicated in the bottom row of Table 5. Overall, the correctness score of model (16) relative to ground truth Δ_3 is 0.75.

Here, then, is our quantitative correctness measure in condensed form. Let $\mathbf{m}$ be a CCM model inferred from data generated from a ground truth Δ. Let $l_{O_i}(\mathbf{m})$, $c_{O_i}(\mathbf{m})$, $d_{O_i}(\mathbf{m})$, and $s_{O_i}(\mathbf{m})$ be $\mathbf{m}$'s literal, conjunctive, disjunctive, and sequential ascriptions for outcome O_i, $i = 1, ..., n$, and analogously for $l_{O_i}(\Delta)$, $c_{O_i}(\Delta)$, $d_{O_i}(\Delta)$, and $s_{O_i}(\Delta)$. Moreover, let $\big|\ldots\big|$ denote the complexity of the enclosed expression, and let $w^x_{O_i}$, $x \in \{l, c, d, s\}$, be the weight associated with the corresponding causal ascriptions of the i^{th} outcome O_i. Then, literal, conjunctive, disjunctive, and sequential correctness, (Corr_l), (Corr_c), (Corr_d), and (Corr_s) are defined as follows:

$$Corr_l = \sum_{i=1}^{n} \frac{\mid l_{O_i}(\mathbf{m}) \, \cap \, l_{O_i}(\Delta) \mid}{\mid l_{O_i}(\mathbf{m}) \mid} \cdot w^l_{O_i} \qquad Corr_c = \sum_{i=1}^{n} \frac{\mid c_{O_i}(\mathbf{m}) \, \cap \, c_{O_i}(\Delta) \mid}{\mid c_{O_i}(\mathbf{m}) \mid} \cdot w^c_{O_i}$$

$$Corr_d = \sum_{i=1}^{n} \frac{\mid d_{O_i}(\mathbf{m}) \, \cap \, d_{O_i}(\Delta) \mid}{\mid d_{O_i}(\mathbf{m}) \mid} \cdot w^d_{O_i} \qquad Corr_s = \sum_{i=1}^{n} \frac{\mid s_{O_i}(\mathbf{m}) \, \cap \, s_{O_i}(\Delta) \mid}{\mid s_{O_i}(\mathbf{m}) \mid} \cdot w^s_{O_i}$$

Aggregating these measures yields the following measure for overall correctness:

Correctness (Corr). The overall correctness of $\mathbf{m}$ for Δ is the weighted mean of $\mathbf{m}$'s (Corr$_l$), (Corr$_c$), (Corr$_d$), and (Corr$_s$) scores, or formally, where w_x are the corresponding weights:

$$Corr(\mathbf{m}, \Delta) = \sum_{x \in \{l,c,d,s\}} Corr_x \cdot w_x$$

An isolated (Corr)-score as 0.75 for (16) is not very informative; it merely says that (16) is neither entirely correct nor entirely incorrect. How (in)correct it is becomes clear only if its correctness score is compared with the scores of other models inferred from the same data. Table 6a thus lists the (Corr)-scores of further model candidates assumed to be inferred from the same data simulated from Δ_3 as (16).[12] The first model, $\mathbf{m}_1$, coincides with (16), except that it does not include D as cause of G. By leaving out D, $\mathbf{m}_1$ leaves out a correct alternative cause of G. Contrary to (16), though, $\mathbf{m}_1$ is not a chain, meaning that $A{*}B$, to which (16) erroneously ascribes causal relevance for both D and G, is not entailed to be causally relevant for G by $\mathbf{m}_1$, which thereby avoids a false conjunctive ascription. Overall, $\mathbf{m}_1$ receives the same (Corr)-score as (16). In model $\mathbf{m}_2$, the incorrect conjunction $A{*}B$ of (16) is replaced by a correct disjunction $A + B \leftrightarrow D$, and model $\mathbf{m}_3$ even gets $B + C \leftrightarrow G$ right. Correspondingly, the (Corr)-score of $\mathbf{m}_2$ is higher than (16)'s and lower than $\mathbf{m}_3$'s. As $\mathbf{m}_3$ contains no error, it receives a perfect (Corr)-score. Likewise, $\mathbf{m}_4$, which was used to illustrate the problem of indirect causation in section 3.3, is error-free and scores perfectly. The same holds for $\mathbf{m}_5$, because it is true that A and B are direct causes of G relative to the set $\{A, B, E, G\}$. That is not true for model $\mathbf{m}_6$, which additionally contains the link D mediating the causal impact of A and B on G in the ground truth Δ_3. It follows that $\mathbf{m}_6$ erroneously entails that A and that B are direct causes of G and alternatives to D relative to the set $\{A, B, D, E, G\}$. Model $\mathbf{m}_6$ reaches a disjunctive correctness of 0.75 and a sequential correctness of 0.5, which, with the perfect literal and conjunctive scores, aggregate to 0.81. Finally, while $\mathbf{m}_6$ makes no incorrect literal and conjunctive ascriptions, $\mathbf{m}_7$, by falsely ascribing

#	model	(Corr)-score
$\mathbf{m}_1$	$(A{*}B \leftrightarrow D){*}(B{*}C \leftrightarrow G)$	0.75
$\mathbf{m}_2$	$(A + B \leftrightarrow D){*}(B{*}C \leftrightarrow G)$	0.88
$\mathbf{m}_3$	$(A + B{*}F \leftrightarrow D){*}(B + C \leftrightarrow G)$	1
$\mathbf{m}_4$	$A + B \leftrightarrow G$	1
$\mathbf{m}_5$	$A + B + E \leftrightarrow G$	1
$\mathbf{m}_6$	$A + B + E + D \leftrightarrow G$	0.81
$\mathbf{m}_7$	$A + B + E + D + F \leftrightarrow G$	0.65

(a)

#	model	(Corr)-score
Δ_2	$A{*}b{*}D{*}F \, + \, a{*}B{*}C{*}D \, \leftrightarrow \, E$	
(6)	$A \, + \, C \, + \, D \, \leftrightarrow \, E$	0.92
(7)	$A{*}b \, + \, C \, + \, D \, \leftrightarrow \, E$	0.94
(8)	$A{*}b{*}F \, + \, C \, + \, D \, \leftrightarrow \, E$	0.95

(b)

Table 6. Table (a) contains additional CCM models and their (Corr)-scores relative to Δ_3 to be contrasted with (16) and its score. Table (b) exhibits the (Corr)-scores of models (6) to (8) for ground truth Δ_2.

[12] For details on how the scores in Table 6 are generated, see the replication script in the supplementary material.

causal relevance for G to F, commits errors in all types of ascriptions. Correspondingly, its (Corr)-score is the lowest.

Lastly, Table 6b exhibits the (Corr)-scores of the examples demonstrating the shortcomings of (NCR) in section 3.2. Model (8) has the highest and (6) the lowest score. That is, contrary to (NCR), (Corr) does not punish (8) for containing more true information than (6), while committing the same mistake as (6). This result generalizes. Adding truthfully located elements from the ground truth to a model increases the complexities of the literal, conjunctive, disjunctive, and sequential intersections and of the model's corresponding causal ascriptions by the same amount, meaning that numerators and denominators of (Corr_l), (Corr_c), (Corr_d), and (Corr_s) increase by the same amount as well. Hence, truthfully expanding models while keeping errors constant increases the (Corr)-score or keeps it constant. By contrast, adding errors to a model while keeping the true information constant only increases the complexities of a model's causal ascriptions but not of their intersections with the ground truth's causal ascriptions. In consequence, the numerators of (Corr_l), (Corr_c), (Corr_d), and (Corr_s) stay the same and the denominators increase, inducing the (Corr)-score to drop or to stay at the minimum of 0. In sum, (Corr) does neither overcount false nor true information in models. It does justice to the model expansion principle (MEP).

5 Completeness

Table 6a also shows that correctness cannot be the only measure of model quality. Models $\mathbf{m}_3$, $\mathbf{m}_4$, and $\mathbf{m}_5$ are all error-free and thus receive (Corr)-scores of 1 each, but they obviously differ in how much detail about the ground truth they reveal. The quality of a model does not only depend on error avoidance, which is what correctness measures, but also on the model's informativeness. To measure that quality aspect, correctness must be complemented by another measure called *completeness*, or *recall* in many fields [17, 20]. As indicated in section 1, there are various completeness measures in use in CCM benchmarking, but they all rely on contrasting submodel sets of models and ground truths. This approach inevitably leads to the problem of indirect causation. For that reason, we now proceed to pair our correctness measure with a completeness measure that builds on the tools developed in this paper.

In the most general terms, completeness of a model $\mathbf{m}$ relative to a ground truth Δ is the ratio of causal information contained in Δ that is revealed by $\mathbf{m}$ to the totality of causal information contained in Δ. As in case of correctness, we propose to break the causal information in $\mathbf{m}$ and Δ down into literal, conjunctive, disjunctive, and sequential ascriptions, as rendered transparent in causal expositions of $\mathbf{m}$ and Δ, respectively. The amount of literal ascriptions of Δ for Outcome O_j, $l_{O_j}(\Delta)$, that is revealed by $\mathbf{m}$ is cashed out in terms of the ratio of the complexity of the intersection between $l_{O_j}(\mathbf{m})$ and $l_{O_j}(\Delta)$ to the complexity of $l_{O_j}(\Delta)$—and analogously for the other types of ascriptions. These ratios are then aggregated for each of the m outcomes in Δ to literal, conjunctive, disjunctive, and sequential correctness measures, (Comp_l), (Comp_c), (Comp_d), and (Comp_s), using a weighted mean with weights, $v^x_{O_j}$, $x \in \{l, c, d, s\}$, corresponding to the complexity share of $\left| l_{O_j}(\Delta) \right|$, $\left| c_{O_j}(\Delta) \right|$, $\left| d_{O_j}(\Delta) \right|$, and $\left| s_{O_j}(\Delta) \right|$:

$$Comp_l = \sum_{j=1}^{m} \frac{|\, l_{O_j}(\mathbf{m}) \,\cap\, l_{O_j}(\Delta) \,|}{|\, l_{O_j}(\Delta) \,|} \cdot v^l_{O_j} \qquad Comp_c = \sum_{j=1}^{m} \frac{|\, c_{O_j}(\mathbf{m}) \,\cap\, c_{O_j}(\Delta) \,|}{|\, c_{O_j}(\Delta) \,|} \cdot v^c_{O_j}$$

$$Comp_d = \sum_{j=1}^{m} \frac{|\, d_{O_j}(\mathbf{m}) \,\cap\, d_{O_j}(\Delta) \,|}{|\, d_{O_j}(\Delta) \,|} \cdot v^d_{O_j} \qquad Comp_s = \sum_{j=1}^{m} \frac{|\, s_{O_j}(\mathbf{m}) \,\cap\, s_{O_j}(\Delta) \,|}{|\, s_{O_j}(\Delta) \,|} \cdot v^s_{O_j}$$

(Comp_l), (Comp_c), (Comp_d), and (Comp_s) are formulated in parallel to the corresponding correctness measures. The only difference is that denumerators in the latter feature complexities of $\mathbf{m}$'s causal ascriptions, while completeness measures contain complexities of Δ's ascriptions in the denumerators.

	out.	ground truth Δ_3	model (16)	intersection	ratio	weight	completeness
lit.	D:	$\{A, B, F\}$	$\{A, B\}$	$\{A, B\}$	$\mathbf{2/3}$	$3/11$	0.55
	E:	$\{C, B, f\}$			$\mathbf{0/3}$	$3/11$	
	G:	$\{A, B, C, D, E\}$	$\{A, B, C, D\}$	$\{A, B, C, D\}$	$\mathbf{4/5}$	$5/11$	
conj.	D:	$\{A, B*F\}$	$\{A*B\}$	$\{A, B\}$	$\mathbf{2/3}$	$3/11$	0.55
	E:	$\{C, B*f\}$			$\mathbf{0/3}$	$3/11$	
	G:	$\{D, B, A, C, E\}$	$\{D, B*C, A*B\}$	$\{D, B, A, C, E\}$	$\mathbf{4/5}$	$5/11$	
dis.	D:	$\{A + B*F\}$	$\{A*B\}$	$\{A, B\}$	$\mathbf{1/3}$	$3/17$	0.47
	E:	$\{C + B*f\}$			$\mathbf{0/3}$	$3/17$	
	G:	$\{C + E,$ $A + B + E,$ $B + C + D,$ $A + B + C\}$	$\{D + B*C,$ $A*B + B*C\}$	$\{D + B, D + C,$ $A + B, A + C,$ $B + C\}$	$\mathbf{7/11}$	$11/17$	
seq.	D:	$\{\langle F, D\rangle, \langle B, D\rangle,$ $\langle A, D\rangle\}$	$\{\langle B, D\rangle, \langle A, D\rangle\}$	$\{\langle B, D\rangle, \langle A, D\rangle\}$	$\mathbf{2/3}$	$3/10$	0.4
	E:	$\{\langle f, E\rangle, \langle B, E\rangle,$ $\langle C, E\rangle\}$			$\mathbf{0/3}$	$3/10$	
	G:	$\{\langle A, D, G\rangle,$ $\langle B, D, G\rangle,$ $\langle B, E, G\rangle,$ $\langle C, E, G\rangle\}$	$\{\langle A, D, G\rangle,$ $\langle B, D, G\rangle,$ $\langle B, G), \langle C, G\rangle\}$	$\{\langle A, D, G\rangle,$ $\langle B, D, G\rangle,$ $\langle B, G\rangle, \langle C, G\rangle\}$	$\mathbf{2/4}$	$4/10$	
		overall completeness (Comp):	$11/49 \cdot 0.55 \; + \; 11/49 \cdot 0.55 \; + \; 17/49 \cdot 0.47 \; + \; 10/49 \cdot 0.4 \; = \; \boxed{\mathbf{0.49}}$				

Table 7. Intersections and completeness scoring for model (16) relative to ground truth Δ_3. Grey shading indicates the expressions used for calculating the completeness ratios.

We aggregate (Comp$_l$), (Comp$_c$), (Comp$_d$), and (Comp$_s$) into an overall completeness measure using a weighted mean where the weights, v_x, correspond to the complexity shares of Δ's whole causal exposition covered by a corresponding completeness measure.

Completeness (Comp). The overall completeness of $\mathbf{m}$ for Δ is the weighted mean of $\mathbf{m}$'s (Comp$_l$), (Comp$_c$), (Comp$_d$), and (Comp$_s$) scores, or formally, where v_x are the corresponding weights:

$$Comp(\mathbf{m}, \Delta) = \sum_{x \in \{l,c,d,s\}} Comp_x \cdot v_x$$

To illustrate, we reconsider model (16) from the previous section and calculate its (Comp)-score relative to ground truth Δ_3. Table 7 reiterates the relevant causal ascriptions from Table 5, but now the ascriptions of Δ_3 are the point of reference and we determine how many of them are reproduced by (16), which are the ones in the intersection. This requires that as many causal ascriptions of Δ_3 as possible are covered by causal ascriptions of (16); and since the former are more numerous than the latter, some ascriptions of Δ_3 may be covered by the same ascription of (16). But each ascription of Δ_3 may only be covered by one ascription of (16). Sometimes the intersections contain multiple elements that can be chosen as counterparts of the elements of Δ_3's causal ascriptions. The ones that enter the completeness calculations in Table 7 are highlighted with grey shading.

As in case of correctness scores, completeness scores are easiest to interpret when contrasting multiple models inferred from the same data. For that reason, let us compare the (Comp)-score of (16) with the scores of the models in Table 8, all of which are assumed to be inferred from the same data simulated from Δ_3 and

#	model	(Comp)-score
$\mathbf{m}_1$	$(A*B \leftrightarrow D)*(B*C \leftrightarrow G)$	0.29
$\mathbf{m}_2$	$(A + B \leftrightarrow D)*(B*C \leftrightarrow G)$	0.31
$\mathbf{m}_3$	$(A + B*F \leftrightarrow D)*(B + C \leftrightarrow G)$	0.43
$\mathbf{m}_4$	$A + B \leftrightarrow G$	0.18
$\mathbf{m}_5$	$A + B + E \leftrightarrow G$	0.29
$\mathbf{m}_6$	$A + B + E + D \leftrightarrow G$	0.40
$\mathbf{m}_7$	$A + B + E + D + F \leftrightarrow G$	0.40
$\mathbf{m}_8$	$(A + B*F \leftrightarrow D)*(B*f + C \leftrightarrow E)*(D + E \leftrightarrow G)$	1
$\mathbf{m}_9$	$(A + B*F + H \leftrightarrow D)*(B*f + C + K \leftrightarrow E)*(D + E \leftrightarrow G)$	1

Table 8. Additional CCM models and their (Comp)-scores relative to Δ_3 to be contrasted with (16) and its score.

the first seven of which were already evaluated for correctness in Table 6a.[13] That comparison highlights two important features of (Comp). First, (Comp) is sensitive to the differences in informativeness to which (Corr) is insensitive. While models $\mathbf{m}_3$, $\mathbf{m}_4$, and $\mathbf{m}_5$ are error-free and thus get perfect (Corr)-scores, they differ in informativeness, which is reflected in their differing (Comp)-scores. Second, contrary to (Corr), (Comp) does not punish for errors in models. To see this, compare models $\mathbf{m}_6$ and $\mathbf{m}_7$: the latter contains one error more than the former, yet they both score the same on (Comp). Or, contrast models $\mathbf{m}_8$ and $\mathbf{m}_9$: the former is error-free while the latter falsely ascribes causal relevance to H and K, still they both receive perfect (Comp)-scores because they contain all the causal information in Δ_3.

6 Aggregating correctness and completeness

In order to assess the overall quality of models in CCM benchmarking, correctness and completeness scores need to be suitably aggregated. Ideally, both scores are 1. In that case, the ground truth is correctly and completely recovered, meaning that the inferred model is identical to the ground truth. It is uncontroversial that this is the optimal result of a benchmark test. It means that the tested method successfully recovers the very structure used to simulate the data. Unfortunately, this ideal scenario often does not obtain when the data are non-ideal, that is, when they feature fragmentation or noise. We cannot expect a method to find the complete ground truth if the evidence in the data is incomplete, and we cannot expect a method to avoid mistakes entirely if some of the evidence is not faithful to the ground truth. But of course, even in non-ideal data scenarios we want the quality of the models to be as high as possible. Methods outputting models of higher quality, on average, are preferable to methods with lower quality outputs. Hence, we need an account of overall model quality that suitably aggregates (Corr)- and (Comp)-scores.

Unfortunately, it is not uncontroversial among CCM methodologists how correctness and completeness should be aggregated. Hasebrouck and Thomann [29, p. 1874] distinguish between two approaches to evaluating models: the *SI-approach* prioritizes the substantive interpretability of models and the *RF-approach* prioritizes the redundancy-freeness of the models. According to the SI-approach, the consistency (see footnote 6) of each disjunct in a model should be as high as possible, even if a disjunct contains conjuncts that are not causes of the outcome. The idea is that each disjunct should constitute a complete recipe—possibly with redundant ingredients—to actualize the outcome. By contrast, the RF-approach demands that each disjunct in a model be exclusively composed of true causes of the outcome, even if the disjunct as a whole does not reach optimal consistency and is only an incomplete recipe for the outcome. It follows that the SI-approach puts more weight on completeness, whereas the RF-approach takes correctness to be more important. A majority of representatives of the QCA method adhere to the SI-approach, while a minor-

[13] For detailed breakdowns of these completness scores, see the replication script in the supplementary material.

#	model	Corr	Comp	$F_{0.5}$	F_2
(16)	$(A{*}B \leftrightarrow D){*}(D + B{*}C \leftrightarrow G)$	0.75	0.49	0.68	0.53
$\mathbf{m}_1$	$(A{*}B \leftrightarrow D){*}(B{*}C \leftrightarrow G)$	0.75	0.29	0.57	0.33
$\mathbf{m}_2$	$(A + B \leftrightarrow D){*}(B{*}C \leftrightarrow G)$	0.88	0.31	0.64	0.35
$\mathbf{m}_3$	$(A + B{*}F \leftrightarrow D){*}(B + C \leftrightarrow G)$	1	0.43	0.79	0.48
$\mathbf{m}_4$	$A + B \leftrightarrow G$	1	0.18	0.53	0.22
$\mathbf{m}_5$	$A + B + E \leftrightarrow G$	1	0.29	0.67	0.34
$\mathbf{m}_6$	$A + B + E + D \leftrightarrow G$	0.81	0.40	0.67	0.45
$\mathbf{m}_7$	$A + B + E + D + F \leftrightarrow G$	0.65	0.40	0.58	0.43
$\mathbf{m}_8$	$(A + B{*}F \leftrightarrow D){*}(B{*}f + C \leftrightarrow E){*}(D + E \leftrightarrow G)$	1	1	1	1
$\mathbf{m}_9$	$(A + B{*}F + H \leftrightarrow D){*}(B{*}f + C + K \leftrightarrow E){*}(D + E \leftrightarrow G)$	0.73	1	0.77	0.93

Table 9. Comparing the overall quality of model (16) relative to ground truth Δ_3 with the quality of other models inferred from the same data at $\beta = 0.5$ and $\beta = 2$.

ity (i.e., those that advocate so-called *parsimonious* QCA solutions) and all representatives of the CNA method adhere to the RF-approach.

We do not want to take a stance, here, on whether correctness or completeness should be preferred when measuring overall model quality. An aggregation that is standard in binary classification and that can easily accommodate either preference is a weighted harmonic mean with a positive real weight β, the so-called *F_β-score* [30]:

Overall Quality. Let $\mathbf{m}$ be a CCM model inferred from data generated from a ground truth Δ. The overall quality of $\mathbf{m}$ for Δ is

$$F_\beta = (1 + \beta^2) \cdot \frac{\mathrm{Corr}(\mathbf{m}, \Delta) \cdot \mathrm{Comp}(\mathbf{m}, \Delta)}{(\beta^2 \cdot \mathrm{Corr}(\mathbf{m}, \Delta)) + \mathrm{Comp}(\mathbf{m}, \Delta)}$$

By assigning a value to β any prioritization of correctness and completeness can be obtained: the completeness of $\mathbf{m}$ relative to Δ, $\mathrm{Comp}(\mathbf{m}, \Delta)$, is β times as important as the correctness $\mathrm{Corr}(\mathbf{m}, \Delta)$. For example, at $\beta = 2$ completeness is twice as important as correctness, and at $\beta = 0.5$ completeness is half as important as correctness. At $\beta = 1$, F_β reduces to the harmonic mean of correctness and completeness.

The harmonic mean is preferred over the arithmetic mean because, contrary to the latter, it requires that a high-quality model strike a balance between correctness and completeness. More specifically, if correctness and completeness scores are balanced at moderate values, the harmonic mean is higher than if the two scores are at opposite extremes, whereas the arithmetic mean is insensitive to such imbalances.

To illustrate F_β-aggregations of (Corr) and (Comp), Table 9 exhibits the F_β-scores of model (16) relative to Δ_3 at $\beta = 0.5$ and $\beta = 2$, respectively, and contrasts them with the corresponding scores of the other model candidates considered in the previous section. Regardless of the value assigned to β, the best model is $\mathbf{m}_8$, which is identical to the ground truth Δ_3. Beyond that clear winner, however, Table 9 shows that different β values not only change the absolute quality scores but also the relative quality ranking among the models. At $\beta = 0.5$, the second best model is $\mathbf{m}_3$, followed by $\mathbf{m}_9$ and (16). At $\beta = 2$, the second best model is $\mathbf{m}_9$, followed by (16) and $\mathbf{m}_3$.

This demonstrates that the relative importance assigned to (Corr) and (Comp) in a CCM benchmark test may have a great influence on the results. Any such test must hence be accompanied by an argument justifying the chosen β. For example, if a test aims to scrutinize a method's reliability in recovering (M)INUS causes from fragmented and noisy data, false positives must be punished more than incomplete ground truth recovery, meaning that β should be lower than 1. By contrast, if a test wants to determine how successfully a method recovers recipes for the outcome, possibly including redundant ingredients, incomplete ground truth recovery should be punished more than false positives, meaning that β should be higher than 1.

7 Conclusion

This paper developed quantitative correctness (precision) and completeness (recall) measures, (Corr) and (Comp), to be used in benchmarking of Configurational Comparative Methods as QCA or CNA. Contrary to the benchmarking criteria currently employed, these new measures do not rely on comparing sets of submodels of candidate models and ground truths. Instead, (Corr) and (Comp), first, unpack the different types of causal ascriptions implied by models and ground truths in causal expositions, second, intersect those expositions, and third, quantify correctness and completeness in terms of the complexities of these intersections. In this manner, (Corr) and (Comp) avoid the problems of overcounting errors and of indirect causation, which affect current benchmarking criteria. The paper ended by accounting for overall model quality in terms of a weighted harmonic mean of (Corr) and (Comp). That account is easily fine-tuned to accommodate any preference ordering of correctness and completeness that may be relevant in a given benchmarking context. Taken jointly, these new measures not only avoid problems of current benchmarking criteria, but they broaden and sharpen the resources for CCM benchmarking more generally.

Acknowledgment: We are grateful to Luna de Souter and Veli-Pekka Parkkinen for extended discussions and their helpful comments on an earlier draft. Moreover, we thank Roel Rutten, Adrian Dusa, the audience at 9th International QCA Expert Workshop, and three anonymous referees.

Funding: The research behind this paper was supported by the Toppforsk program of the University of Bergen and the Trond Mohn Foundation (grant number 811866) and by the Norwegian Research Council (project number 326215).

Conflict of Interest: The authors have no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data availability statement: The datasets generated and analyzed during the current study can be reproduced with the R scripts available at https://github.com/m-baum/quantifyQuality.

References

[1] Arel-Bundock V. The double bind of Qualitative Comparative Analysis. Sociological Methods & Research. 2022;51(3):963–982.

[2] Spirtes P, Glymour C, Scheines R. Causation, prediction, and search (second edition). Cambridge: MIT Press; 2000.

[3] Baumgartner M, Falk C. Boolean difference-making: a modern regularity theory of causation. The British Journal for the Philosophy of Science. 2023;74(1):171–197.

[4] Mackie JL. The cement of the universe: a study of causation. Oxford: Clarendon Press; 1974.

[5] Ragin CC. The comparative method. Berkeley: University of California Press; 1987.

[6] Rihoux B, Ragin CC (Ed.). Configurational comparative methods: Qualitative Comparative Analysis (QCA) and related techniques. Thousand Oaks: Sage; 2009.

[7] Baumgartner M. Uncovering deterministic causal structures: a Boolean approach. Synthese. 2009;170:71–96.

[8] Baumgartner M, Ambühl M. Causal modeling with multi-value and fuzzy-set Coincidence Analysis. Political Science Research and Methods. 2020;8(3):526–542.

[9] Swiatczak MD. Different algorithms, different models. Quality & Quantity. 2022;56:1913–1937.

[10] Baumgartner M, Falk C. Configurational causal modeling and logic regression. Multivariate Behavioral Research. 2023;58(2):292–310.

[11] Baumgartner M, Thiem A. Often trusted but never (properly) tested: evaluating Qualitative Comparative Analysis. Sociological Methods & Research. 2020;49(2):279–311.

[12] Dușa A. Critical tension: sufficiency and parsimony in QCA. Sociological Methods & Research. 2022 Nov;51(2):541––565.

[13] Krogslund C, Choi DD, Poertner M. Fuzzy sets on shaky ground: parameter sensitivity and confirmation bias in fsQCA. Political Analysis. 2015;23(1):21–41.

[14] Lucas SR, Szatrowski A. Qualitative Comparative Analysis in critical perspective. Sociological Methodology. 2014;44(1):1–79.

[15] Parkkinen VP, Baumgartner M. Robustness and model selection in configurational causal modeling. *Sociological Methods & Research*. 2023;52(1):176–208.

[16] Swiatczak MD, Baumgartner M. Data imbalances in Coincidence Analysis: a simulation study. Sociological Methods & Research. 2024 March;doi:10.1177/00491241241227039.

[17] Cheng L, Guo R, Moraffah R, Sheth P, Candan KS, Liu H. Evaluation methods and measures for causal learning algorithms. IEEE Transactions on Artificial Intelligence. 2022;3(6):924–943.

[18] Maier M, Taylor B, Oktay H, Jensen D. Learning causal models of relational domains. In Proceedings of the AAAI Conference on Artificial Intelligence. 2010;24:531–538. https://doi.org/10.1609/aaai.v24i1.7695

[19] Meek C. Causal inference and causal explanation with background knowledge. In Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence. 1995;403–410. https://arxiv.org/ftp/arxiv/papers/1302/1302.4972.pdf

[20] Tabib Mahmoudi F, Samadzadegan F, Reinartz P. Object recognition based on the context aware decision level fusion in multi views imagery. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 2015 Jan;8(1):12–22.

[21] Beirlaen M, Leuridan B, Van De Putte F. A logic for the discovery of deterministic causal regularities. Synthese. 2018;195(1):367–399.

[22] Bowran AP. A Boolean algebra: abstract and concrete. London: Macmillan; 1965.

[23] Lemmon EJ. Beginning logic. London: Chapman & Hall; 1965.

[24] Ragin CC. Set relations in social research: evaluating their consistency and coverage. Political Analysis. 2006;14(3):291–310.

[25] Baumgartner M. Qualitative Comparative Analysis and robust sufficiency. Quality & Quantity. 2022;56:1939–1963.

[26] Parkkinen VP. Variable relativity of causation is good. Synthese. 2022;200:194. https://doi.org/10.1007/s11229-022-03676-0

[27] Woodward J. Response to Strevens. Philosophy and Phenomenological Research. 2008;LXXVII:193–212,

[28] McCluskey EJ. Minimization of Boolean functions. Bell Systems Technical Journal. 1956;35:1417–1444.

[29] Haesebrouck T, Thomann E. Introduction: causation, inferences, and solution types in configurational comparative methods. Quality & Quantity. 2022;56(4):1867–1888.

[30] Chinchor N. Muc-4 evaluation metrics. In Proceedings of the 4th Conference on Message Understanding. 1992 June;22–29. https://aclanthology.org/M92-1002.pdf