New sufficiency and necessity measures for model building with Coincidence Analysis

Luna De Souter[1*] and Michael Baumgartner[1]

[1*]Dept. of Philosophy, University of Bergen, Sydnesplassen 12-13, 5020 Bergen, Norway.

*Corresponding author(s). E-mail(s): Luna.Souter@uib.no;
Contributing authors: Michael.Baumgartner@uib.no;

Abstract

Background Coincidence Analysis (CNA) is a configurational comparative method of causal learning that has seen a significant uptick in applications in public health in recent years. To build its causal models, CNA searches for redundancy-free relations of sufficiency and necessity in data using a sufficiency measure called consistency and a necessity measure called coverage.

Methods This paper argues that consistency and coverage have severe limitations. In particular, they are not reliable when the relative frequencies of candidate causes and outcomes are at high or low extremes. We propose alternative sufficiency and necessity measures that are not affected by these limitations and benchmark them against standard consistency and coverage in an extended simulation experiment analyzing binary, so-called crisp-set, data.

Results Across a wide range of data scenarios, the overall quality of CNA models built by means of the new measures is more than 20% higher than when models are built using the standard measures.

Conclusion We recommend that the new measures are made available in relevant CNA software and that CNA users transition to building crisp-set models with them.

Keywords: configurational causal modeling, causal complexity, INUS causation, consistency, coverage

1 Introduction

Coincidence Analysis (CNA) belongs to the family of so-called *configurational comparative methods* of causal learning that includes Qualitative Comparative Analysis (QCA) as its

best-known member [1–4]. CNA tracks causal complexity by assembling multiple causes in bundles (conjunctions) that only bring about their effect when all of their components are co-instantiated and by placing these bundles on alternative (disjunctive) causal paths that operate independently of one another. The method is custom-built to deal with causal structures featuring interactions and equifinality, which pose challenges for standard methods because structures with these features often violate linearity and comprise causes and effects that are not pairwise correlated [5, 6]. To this end, CNA identifies redundancy-free (i.e. minimal) sufficiency and necessity relations in data and combines them to causal structures as defined by a modern regularity theory of causation—which, unlike most other theories, does not entail that pairwise correlation is necessary for causation (cf. [7, 8]). CNA is the only configurational comparative method that can process data generated by causal structures with multiple outcomes, for example, causal chain structures.

In recent years, the method has been applied in various fields including the social, political, and behavioral sciences [9–12]. CNA is currently seeing a significant uptick in applications in public health, covering a wide range of topics from safety culture in medical homes, opioid and obesity treatment, to cancer care, surgical site infection reduction, or the impact of firearm laws on suicide and homicide rates [13–18].[1] Simultaneously, the method is continually being developed further [20], its performance benchmarked under varying data scenarios and against other methods [2, 5], and its software implementation updated [21–23].

This paper contributes to these methodological advancements by improving CNA's approach to evaluating whether dependencies in data meet the standard for sufficiency or necessity. Since its first introduction in 2009 [1], CNA has conducted such evaluations using a sufficiency measure called *consistency* and a necessity measure called *coverage*. Both measures were directly imported from QCA, where they had been introduced on common-sense grounds a few years earlier [24]. Consistency and coverage assess how frequently sufficiency and necessity are satisfied in the data. They are equivalent to positive predictive value (aka precision) and sensitivity (aka recall), respectively, which are well-known from fields such as binary classification [25] and information retrieval [26]. Consistency and coverage play a twofold role in CNA: on the one hand, they are key in CNA's model-building algorithm, and on the other, they are used in selecting among multiple model candidates output by that algorithm.

De Souter [27] has recently shown that consistency and coverage do not take into account all the evidence relevant for assessing whether some X is sufficient or necessary for some Y in binary (crisp-set) data. In particular, the measures are insensitive to cases (units of observation) in the data in which both X and Y are absent, despite such cases often containing important information about whether X should be considered sufficient/necessary for Y. In consequence, De Souter introduces a sufficiency measure called *contrapositive consistency* and a necessity measure called *contrapositive coverage*, which are sensitive to the evidence neglected by standard measures. She demonstrates that using these supplementary measures to select among multiple models—built in the usual way with standard measures—succeeds in selecting models of significantly higher quality.

We follow up on these results by investigating the suitability of consistency, coverage, and their contrapositive counterparts for evaluating sufficiency and necessity relationships within CNA's model-building algorithm. After reviewing the basics of CNA, the first part of

[1] For a full overview of the CNA literature see the Zotero Coincidence Analysis Group Library [19].

the paper identifies data scenarios—characterized by very high or low relative frequencies of X and Y—in which the four measures become unreliable. As these scenarios can coincide, sufficiency and necessity relations cannot be reliably evaluated in all possible data scenarios by merely aggregating consistency and coverage with the corresponding contrapositive measures. Instead, new sufficiency and necessity measures are needed.

The second part of the paper introduces two new sufficiency measures and two new necessity measures that mitigate the limitations of consistency, coverage, and their contrapositive counterparts. In essence, they are appropriately weighted versions of existing measures. We then conduct a comprehensive simulation experiment to benchmark the new measures against standard consistency and coverage and to identify the pair of measures that performs best overall. The results show that, across a wide range of binary data scenarios, the overall quality of CNA models built using that best-performing measure pair exceeds that of models built using conventional measures by more than 20%.

2 Preliminaries

We begin by introducing the relevant notation and concepts. CNA builds causal models that conform to the so-called *MINUS theory* of causation [7, 8].[2] That theory 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 part of a complex sufficient and necessary condition of B=β that is rigorously freed of all redundant elements [7]. Factors 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. We will develop our argument focusing on crisp-set data and, thus, abbreviate our notation according to conventions in Boolean algebra: we use "A" as shorthand for A=1 and "a" for A=0.[3]

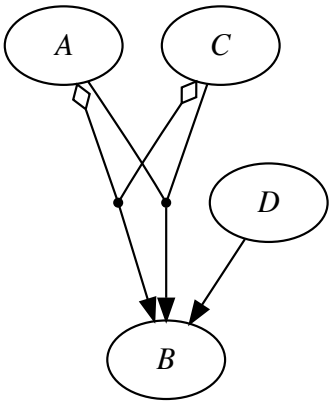
The MINUS theory borrows much of its formal machinery from Boolean algebra, in particular the operations of *negation*, $\neg A$ (expressing "NOT A=1"), *conjunction*, $A{*}B$ ("A=1 AND B=1"), *disjunction*, $A + B$ ("A=1 OR B=1"), *implication*, $A \rightarrow B$ ("IF A=1, THEN B=1"), and *equivalence* $A \leftrightarrow B$ ("A=1 IF, AND ONLY IF, B=1").[4] For crisp-set and multi-value factors, Boolean operations are given a rendering in classical logic, which we do not reiterate here (see e.g. [29]). The relations of *sufficiency* and *necessity* are defined based on the implication operator. A Boolean expression, for example the conjunction $A{*}C$, is sufficient for B iff (i.e., if, and only if) $A{*}C \rightarrow B$, that is, whenever A AND C is true, B is true—or equivalently, it is not the case that A AND C is true and B false. Conversely, an expression, say, $A{*}C + a{*}c + D$ is necessary for B iff $B \rightarrow A{*}C + a{*}c + D$, meaning that whenever B is true, $A{*}C$ OR $a{*}c$ OR D is true—or equivalently, B does not obtain without any of $A{*}C$ OR $a{*}c$ OR D also obtaining.

Most sufficiency and necessity relations do not reflect causation, but those that are minimal do [7]. $A{*}C$ is a *minimally sufficient* condition of B iff $A{*}C \rightarrow B$ holds and no proper

[2] Originally, "INUS" was an acronym referring to *I*nsufficient but *N*on-redundant parts of *U*nnecessary but *S*ufficient conditions [8, 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. Accordingly, "MINUS" is a name, without an expansion, locating the corresponding theory in the INUS tradition.

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

[4] The symbols "*" and "+" are used as in Boolean algebra here (notational variants are "$\wedge$" and "$\vee$"). For a standard introduction to Boolean algebra see [28].



(a)

	A	C	D	B
c_1	1	1	1	1
c_2	0	1	1	1
c_3	1	0	1	1
c_4	0	0	1	1
c_5	1	1	0	1
c_6	0	0	0	1
c_7	0	1	0	0
c_8	1	0	0	0

(b)

	A	C	D	B
c_1	1	1	1	1
c_2	0	1	1	1
c_4	0	0	1	1
c_5	1	1	0	1
c_6	0	0	0	1
c_7	0	1	0	0
c_8	1	0	0	0
c_9	0	0	1	0
c_{10}	0	1	0	1

(c)

Figure/Table 1: Figure (a) is a causal hypergraph where arrows represent direct causal relevance, '$\diamond$' means negation, '$\bullet$' stands for conjunction, and multiple arrows into the same node form a disjunction. It plots the MINUS-formula (1). Table (b) contains ideal data generated by (a)/(1) and Table (c) real data.

part of $A{*}C$ is sufficient for B, that is, neither $A \rightarrow B$ nor $C \rightarrow B$ hold. $A{*}C + a{*}c + D$ is a *minimally necessary* condition of B iff $B \rightarrow A{*}C + a{*}c + D$ holds and no elimination of at least one disjunct from $A{*}C + a{*}c + D$ results in a necessary condition of B. Correspondingly, CNA infers minimally necessary disjunctions of minimally sufficient conjunctions of investigated outcomes in disjunctive normal form (DNF) from data and outputs them as so-called *MINUS-formulas*[5], for instance:

$$A{*}C \; + \; a{*}c \; + \; D \; \leftrightarrow \; B \tag{1}$$

The minimality (redundancy-freeness) of MINUS-formulas guarantees that each factor value X_i on the left-hand side of '$\leftrightarrow$' makes a difference to the right-hand side, because without X_i the left-hand side would not satisfy sufficiency or necessity and, thus, would not account for all variation in the factor on the right-hand side. Accordingly, MINUS-formulas have a straightforward causal interpretation: conjunctions represent complex causes (interactions) and disjunctions stand for alternative causes (equifinality). Hence, (1) entails that A and C jointly cause B on one path, a and c on another, and D on a third path. A concrete interpretation might be that the factors A, C, and D represent positions of electrical switches, say, "up" and "down", and B stands for a lamp being on. In this context, (1) asserts that the lamp is caused to be on when either switches A and C are both in the "up" or "down" position or switch D is in the "up" position. The causal hypergraph in Figure 1a provides a graphical representation of this structure.

Table 1b features ideal data generated by structure (1), that is, data comprising all $2^3 = 8$ possible configurations of the 3 mutually independent exogenous factors A, C, D such that the values of B are assigned in accordance with (1). For brevity, we will henceforth refer to a causal structure generating data as the *ground truth* of these data; thus, (1) is the ground

[5]Two remarks: first, an expression is in DNF iff it is a disjunction of one or more conjunctions of one or more factor values [28, p. 13]. Second, there are *atomic* and *complex* MINUS-formulas, the former representing causal structures with one outcome, the later structures with multiple outcomes. That distinction will not be relevant in this paper.

truth of Table 1b. These data can be seen as obtained when experimentally investigating the structure connecting the switches to the lamp under ideal laboratory conditions. Note that the positions of the switches A and C are uncorrelated with the lamp being on or off. In consequence, any method that takes pairwise correlation to be necessary for causation cannot learn the structure in Figure 1a from the ideal data in 1b.

CNA, by contrast, has no problem recovering (1) from Table 1b. To this end, it proceeds in two algorithmic phases (see [2] for more details). In phase (I), it searches for minimally sufficient conditions of B. This is done by, first, testing for each individual value of A, C, and D whether it is sufficient for B. Factor values that are not sufficient by themselves are then gradually combined to increasingly complex conjunctions until sufficiency for B is satisfied. By building these conjunctions from the bottom up and never adding an extra element to a conjunction that is already sufficient for B, the algorithm ensures that conjunctions passing a sufficiency test are minimally sufficient. The result of CNA's first algorithmic phase applied to Table 1b is this set of minimally sufficient conditions of B: $\{A{*}C,\ a{*}c,\ D\}$.

In phase (II), CNA first tests whether individual elements of the set $\{A{*}C,\ a{*}c,\ D\}$ are necessary for B. It then proceeds to building increasingly complex disjunctions of the non-necessary elements and determining, at each complexity level, whether necessity for B is satisfied. By building these disjunctions from the bottom up and never adding extra elements to a disjunction that is already necessary, CNA ensures that resulting disjunctions are minimally necessary. For Table 1b, the result of phase (II) is the singleton set of minimally necessary conditions $\{A{*}C + a{*}c + D\}$,[6] which is easily completed to the MINUS-formula (1) by adding the outcome on the right-hand side of an equivalence operator. For ideal data, CNA is guaranteed to find the ground truth.[7]

But of course, ideal data are rare in real causal learning contexts. Real data tend to be affected by various deficiencies, particularly fragmentation and noise. *Fragmentation* refers to the ratio of configurations of exogenous factors that are compatible with the ground truth but missing from the data, due to practical limitations of data collection. To illustrate, consider the data in Table 1c, again assumed to be generated by ground truth (1). The configuration instantiated by case c_3, with switches A and D in "up" and C in "down" position, is compatible with (1) but missing from those data. As it is the only missing configuration of a total of 8 configurations compatible with (1) (see Table 1b), Table 1c has a fragmentation of $1/8 = 0.125$. The higher the fragmentation, the less information about the ground truth is contained in the data, and thus the more incomplete the resulting CNA models, on average.

By *noise* we mean the ratio of cases in the data that are incompatible with the ground truth, due to, for example, measurement error or confounding. An incompatible case either features an outcome without its causes or a cause without its outcome(s). To illustrate, two of the 9 cases in Table 1c, specifically c_9 and c_{10}, are incompatible with (1). In c_9, a sufficient cause of B, $a{*}c$, is given without B; and in c_{10}, B is given without any of its causes. The noise level of Table 1c, hence, is $2/9 = 0.22$. Data with zero fragmentation and zero noise are ideal data.

B cannot be modeled as a (strict) Boolean function of the other factors in Table 1c: the value of B is not determined given the values of the other factors, as there are pairs of cases, (c_4, c_9) and (c_7, c_{10}), in which A, C, and D take constant values but the value of B changes.

[6] There often exists more than one minimally necessary condition, giving rise to model ambiguities.
[7] To replicate the CNA analyses of Tables 1b and 1c see the paper's online replication materials.

So, the former factors cannot be combined to a minimally necessary disjunction of minimally sufficient conditions of B, meaning that no MINUS-formula of B is inferrable from Table 1c—provided that we require *strict* satisfaction of sufficiency and necessity. Nonetheless, most instances of the true causes of B—$A{*}C$, $a{*}c$, and D—are associated with B, and most instances of B are covered by an instance of at least one of its causes. Sufficiency and necessity are *almost* (strictly) satisfied. There is only one case, c_9, that violates sufficiency and one case, c_{10}, that violates necessity. Hence, if we lower the bar for an association to pass the sufficiency or necessity test, we can extract information about the ground truth from Table 1c despite the presence of fragmentation and noise.

Plainly though, causal learning from fragmented and noisy data comes with an inherent error risk. Any conjunction of exogenous factor values may appear to be sufficient for the outcome because counter-evidence is missing from the data due to fragmentation, and not because it actually is causally relevant. And any conjunction that really is causally relevant may appear to be insufficient for the outcome because of noise.

This is where measures evaluating the evidence for and against the satisfaction of sufficiency and necessity in the data become crucial. When running the CNA algorithm on real data δ with an outcome Y, we need to assess for each conjunction Φ_i of exogenous factor values that is tested for sufficiency for Y in phase (I) whether the evidence in δ warrants accepting $\Phi_i \rightarrow Y$. Analogously, we have to assess, for each disjunction $\Phi_i + \ldots + \Phi_j$ of minimally sufficient conditions tested for necessity in phase (II), whether the evidence in δ warrants accepting $Y \rightarrow \Phi_i + \ldots + \Phi_j$. An association that does not meet the strict sufficiency or necessity standard in δ should be accepted as passing the sufficiency or necessity test if, and only if, we can be reasonably confident that the association would be one of strict sufficiency or necessity, were the same ground truth investigated under ideal laboratory circumstances, that is, if (hypothetically) we were analyzing ideal data.

The evaluation measures utilized by CNA for this purpose were introduced to configurational comparative methods by Ragin in 2006 [24]. He labeled the sufficiency measure *consistency* and the necessity measure *coverage*. In essence—formal definitions are given in the next section—consistency and coverage, which both take values from the unit interval, measure how often a tested sufficiency or necessity relation is satisfied in the data. Before processing data with CNA, the analyst sets a consistency and a coverage threshold, typically between 0.7 and 1.[8] These thresholds express the degree to which a sufficiency or necessity relationship needs to be satisfied in the real data for the analyst to accept that it holds in the (hypothetical) ideal data. In phase (I), the algorithm then assesses each tested dependency against the consistency threshold and collects all minimal conjunctions reaching this threshold. In phase (II), disjunctions of these minimally sufficient conjunctions are tested for necessity by determining whether they meet the coverage threshold. The result is a set of minimal disjunctions satisfying the coverage threshold where each disjunct is a minimal conjunction that meets the consistency threshold.

If the thresholds for consistency and coverage are set to 0.7 and 0.8, respectively, CNA infers the following MINUS-formula from Table 1c:

$$C + D \leftrightarrow B \tag{2}$$

[8] Setting suitable thresholds is an intricate task: the higher the thresholds, the higher the risk that resulting models are overfitted; the lower the thresholds, the higher the risk that the models are uninformative or spurious. CNA provides a robustness protocol that balances these two risks (see [20]).

When causally interpreted according to the MINUS theory, model (2) claims that C and D are causally relevant to B on two separate paths. Both of these claims are true according to the ground truth (1) of Table 1c. As (2) only makes true causal claims, it is a correct model. But it is incomplete, meaning that it does not fully represent the ground truth. (2) is a proper *submodel* of (1): all causal claims entailed by (2) are also entailed by (1), but not vice versa (see [2] for more on the submodel relation).

Incomplete ground truth recovery is a commonplace result when CNA analyzes real data. The higher the fragmentation and noise levels, the sparser the evidence about the ground truth, and therefore, the less completely that ground truth can be recovered. As CNA cannot be expected to completely recover ground truths from fragmented and noisy data, sufficiency and necessity measures fulfill their purpose if they enable CNA to reliably infer submodels of the ground truth that are as complete as possible under non-ideal discovery circumstances. There may be substantive differences in the quality of CNA outputs when the algorithm is run with different evaluation measures. Of two alternative measures, preference should be given to the one that, on average, recovers more correct and complete models. But how building models based on alternative measures—of which there are many in other methodological frameworks—influences CNA's performance has not been investigated to date.

3 Current sufficiency and necessity measures of CNA

Before we turn to filling that gap, this section reviews the formal details of the evaluation measures currently utilized in CNA and highlights their limitations. For our ensuing discussion, we use Φ as a placeholder for a Boolean expression in DNF (see footnote 5), for example, A, $A{*}C$, or $A{*}C + a{*}c$, while ϕ represents the negation of that DNF. Analogously, Y and y shall be placeholders for single factor values and their negations, for example, A and a. Φ is called the *antecedent* of the implication, and Y is the *consequent*. Moreover, we use cardinality bars $|\ldots|$ to refer to the number of cases in the analyzed data δ satisfying the enclosed condition. For example, $|\Phi{*}Y|$ designates the number of cases in δ instantiating $\Phi{*}Y$.

3.1 Consistency

Consistency—the standard measure for evaluating the sufficiency of Φ for Y—is equivalent to positive predictive value (PPV) in binary classification [25] and to precision in information retrieval [26]. It determines whether Φ is likely to be sufficient for Y in ideal data by assessing whether $\Phi \rightarrow Y$ is satisfied often enough in the analyzed data δ. Consistency considers all cases in δ featuring Φ and measures the proportion of them that satisfy $\Phi \rightarrow Y$, which are those that instantiate both Φ and Y; or formally:[9]

$$consistency(\Phi \rightarrow Y) = \frac{|\Phi{*}Y|}{|\Phi|} = \frac{|\Phi{*}Y|}{|\Phi{*}Y| + |\Phi{*}y|} \tag{3}$$

By measuring the proportion of cases with Φ that satisfy $\Phi \rightarrow Y$, consistency penalizes the cases with Φ that violate $\Phi \rightarrow Y$, that is, cases with $\Phi{*}y$: for a given $|\Phi|$, or a given $|\Phi{*}Y|$, consistency decreases as $|\Phi{*}y|$ increases.

[9] The evaluation measures discussed in this section contain arithmetic sums. We symbolize sums with script-style "+", as opposed to the "+" used for Boolean disjunction.

As an implication can only be violated by cases instantiating the antecedent but not the consequent, this penalization scheme makes good intuitive sense. That is why it was adopted in configurational comparative methods without much exploration of alternatives.[10] Still, some issues have been noted in the literature. In particular, as cases exhibiting Y cannot violate $\Phi \rightarrow Y$, consistency tends to be high in data δ with a high proportion of cases featuring Y, meaning with high outcome prevalence. But of course, the mere fact that most cases in δ instantiate Y is not evidence in favor of $\Phi \rightarrow Y$ holding in the (hypothetical) ideal version of δ. This is typically viewed as a problem stemming from an inadequate calibration of Y giving rise to a skewed distribution of Y's values [3]. But De Souter [27] locates the source of this problem in the consistency measure itself: consistency is an unsuited sufficiency measure for high-prevalence data.

To see this, consider an antecedent Φ that is entirely independent of an outcome Y in data δ. That is, the occurrence of Y is neither more nor less likely when Φ happens compared to when ϕ happens. The subset of cases for which Φ is true does not contain a higher or lower proportion of cases featuring Y than all cases in δ (or than the subset of cases with ϕ). In other words, the proportion of cases with Y among the cases with Φ, which amounts to the consistency of $\Phi \rightarrow Y$, is equal to the proportion of cases with Y among all cases in δ, which amounts to prevalence. In brief, consistency is equal to prevalence.

That consistency is equal to prevalence when antecedent and outcome are independent exposes the limitation of consistency. When prevalence is high, say 95%, any expression $\Phi \rightarrow Y$ such that Φ and Y are independent in the data has consistency 0.95. For example, let Y represent employees (self-)reporting as being competent at work, and Φ shall stand for having a surname of less than 6 characters. Most people report as being competent, meaning that Y is highly prevalent. But clearly, such reports are not influenced by name length. Nevertheless, having a short name turns out to be sufficient for competence with an almost perfect consistency of 0.95. The reason is that only 5% of the cases considered by consistency (i.e. the few people reporting as incompetent), could possibly violate $\Phi \rightarrow Y$. In high-prevalence scenarios, therefore, high consistency does not indicate strong evidence that $\Phi \rightarrow Y$ holds in corresponding ideal data. However, a high score on an adequate sufficiency measure should always signal strong evidence for $\Phi \rightarrow Y$.

A further limitation of consistency is its susceptibility to noise when prevalence is low. If cases with Y are rare, there are only few cases that could possibly corroborate that $\Phi \rightarrow Y$ holds (i.e. the few cases with Y), and if some of them are affected by noise, consistency plummets. In consequence, the chances that consistency can detect sufficiency satisfaction are low. That is, even when $\Phi \rightarrow Y$ does in fact reflect a causal relation between Φ and Y, a small number of noisy cases with $\Phi{*}y$ are enough to yield a low consistency score, rendering it impossible to find that relation. While consistency is too lenient when prevalence is high, it is overly strict when prevalence is low. Overall, consistency is an unreliable evaluation measure for the sufficiency of Φ for Y when prevalence is at extreme highs or lows.

3.2 Coverage

Coverage—the standard measure for evaluating the necessity of Φ for Y—is equivalent to sensitivity in binary classification and to recall in information retrieval. It determines whether

[10] Some alternative sufficiency measures have been considered in recent years (e.g. [30, 31]), but those are only intended to address alleged weaknesses of standard consistency in fuzzy-set analyses.

Φ is likely to be necessary for Y in (hypothetical) ideal data by assessing whether $Y \rightarrow \Phi$ is satisfied often enough in the analyzed data δ. To this end, coverage considers all cases in δ featuring Y and measures the proportion of them that satisfy $Y \rightarrow \Phi$, which are those that instantiate both Φ and Y; or formally:

$$coverage(Y \rightarrow \Phi) = \frac{|\Phi{*}Y|}{|Y|} = \frac{|\Phi{*}Y|}{|\Phi{*}Y| + |\phi{*}Y|} \tag{4}$$

By measuring the proportion of cases with Y that satisfy $Y \rightarrow \Phi$, coverage penalizes the cases with Y that violate $Y \rightarrow \Phi$, that is, cases with $\phi{*}Y$: for a given $|Y|$, or a given $|\Phi{*}Y|$, coverage decreases as $|\phi{*}Y|$ increases.

This is a sensible penalization scheme, since the necessity of Φ for Y can only be violated by cases instantiating Y but not Φ. Still, some limitations have been highlighted in the literature. Notably, coverage tends to be high in data δ with a high proportion of cases featuring Φ. But the mere fact that most cases in δ instantiate Φ is not evidence in favor of $Y \rightarrow \Phi$ being underwritten by a causal relation. This problem is typically attributed to the trivialness of necessary conditions with high $|\Phi|/N$ [3]—where N is the total number of cases in δ. However, De Souter [27] shows that it is a consequence of coverage's unsuitability as necessity measure when $|\Phi|/N$ is high.

This becomes apparent when considering an outcome Y and some Φ that are statistically independent in data δ, such that Φ is not more or less likely when Y happens than when y happens. Accordingly, the set of cases for which Y holds does not exhibit a higher or lower proportion of cases featuring Φ than all cases in δ (or than the cases with y). Hence, the proportion of cases with Φ among those with Y, which is the coverage of $Y \rightarrow \Phi$, is equal to the proportion of cases with Φ among all cases in δ, which is $|\Phi|/N$. In other words, coverage is equal to $|\Phi|/N$. That, in turn, means that a very frequent Φ is expected to score high on coverage for *any* Y. To illustrate, let Y again stand for people self-reporting as competent at work and Φ for people having names of fewer than 10 characters. As most English last names have fewer than 10 characters, Φ is very frequent. As a result, being short-named almost perfectly covers being competent. Plainly though, scoring high on coverage merely because short names are very frequent does not amount to strong evidence that being short-named is actually necessary for reporting as competent. Coverage does not adequately evaluate necessity when $|\Phi|/N$ is high.

Another weakness of coverage is its high susceptibility to noise when $|\Phi|/N$ is low. Coverage penalizes cases with $\phi{*}Y$ in proportion to cases with $\Phi{*}Y$. Therefore, if there are only a few cases with Φ and thus with $\Phi{*}Y$, coverage can be pulled down significantly by only a few noisy cases with $\phi{*}Y$. This happens even if $Y \rightarrow \Phi$ is actually underwritten by a causal dependence. So, coverage is both too lenient when $|\Phi|/N$ is high and too strict when $|\Phi|/N$ is low, making it an unreliable evaluation measure for the necessity of Φ for Y when $|\Phi|/N$ is at high or low extremes.

3.3 Contrapositive consistency

To mitigate the problems of consistency when prevalence is high, De Souter [27] introduced the new measure of contrapositive consistency, or *C-consistency* for short. C-consistency is equivalent to specificity in binary classification. Its use as sufficiency measure in CNA is

based on the rule of contraposition, which states that $\Phi \rightarrow Y$ is logically equivalent to $y \rightarrow \phi$. Thus, by measuring the proportion of cases with Φ that also feature Y, the original consistency measure does not only assess whether $\Phi \rightarrow Y$ is satisfied often enough, but also whether $y \rightarrow \phi$ is satisfied often enough. Likewise, C-consistency evaluates both $y \rightarrow \phi$ and $\Phi \rightarrow Y$ by measuring the proportion of cases with y that feature ϕ:

$$C\text{-}consistency(\Phi \rightarrow Y) = \frac{|\phi{*}y|}{|y|} = \frac{|\phi{*}y|}{|\phi{*}y| + |\Phi{*}y|} \tag{5}$$

In order for Φ to be sufficient for Y, the cases with y must exhibit ϕ. Accordingly, C-consistency penalizes the cases with y exhibiting Φ, which are the same cases penalized by consistency and exactly the cases violating $\Phi \rightarrow Y$ and $y \rightarrow \phi$. Overall, C-consistency is a sensible additional sufficiency measure.

In section 3.1, we have shown that regular consistency is unreliable for evaluating $\Phi \rightarrow Y$ in high-prevalence data because it is equal to prevalence when Φ and Y are independent. The same does not hold for C-consistency, making it a preferable sufficiency measure in certain scenarios where regular consistency fails. This does not imply, however, that C-consistency is always reliable. Analogously to how we demonstrated that consistency and coverage are equal to prevalence and $|\Phi|/N$, respectively, when Φ and Y are independent, it can be shown that C-consistency is equal to $|\phi|/N$ when Φ and Y are independent. It follows that C-consistency is unreliable as sufficiency measure when $|\Phi|/N$ is low (i.e. when $|\phi|/N$ is high). Furthermore, just as consistency and coverage are too strict when, respectively, $|Y|/N$ and $|\Phi|/N$ are low, C-consistency is too strict when $|\Phi|/N$ is high ($|\phi|/N$ is low), since such scenarios have few cases with $\phi{*}y$ and C-consistency penalizes cases with $\Phi{*}y$ relative to cases with $\phi{*}y$.

Unfortunately, the weakness scenarios of consistency and C-consistency can coincide. For example, if prevalence is very high while $|\Phi|/N$ is very low, consistency and C-consistency are both unreliable. Therefore, simply aggregating these measures does not yield dependable sufficiency evaluation in all scenarios. Instead, a new measure is needed, one that is reliable regardless of $|Y|/N$ and $|\Phi|/N$.

3.4 Contrapositive coverage

Analogously to the introduction of C-consistency as sufficiency measure to complement consistency, De Souter [27] introduced contrapositive coverage, or *C-coverage*, as an alternative to coverage for measuring necessity:

$$C\text{-}coverage(Y \rightarrow \Phi) = \frac{|\phi{*}y|}{|\phi|} = \frac{|\phi{*}y|}{|\phi{*}y| + |\phi{*}Y|} \tag{6}$$

C-coverage is equivalent to negative predictive value (NPV) from binary classification. Like regular coverage, C-coverage is justified by the rule of contraposition: $Y \rightarrow \Phi$ is logically equivalent to $\phi \rightarrow y$. Both expressions are violated if and only if $\phi{*}Y$ holds. Correspondingly, C-coverage penalizes the cases exhibiting $\phi{*}Y$ by measuring the proportion of cases with ϕ that instantiate y. But in contrast to regular coverage, C-coverage can reliably assess whether $Y \rightarrow \Phi$ holds even when $|\Phi|/N$ is at extremes, because contrary to coverage, C-coverage is not equal to $|\Phi|/N$ when Φ and Y are independent.

Still, C-coverage is itself sometimes unreliable as necessity measure. It is too lenient when $|y|/N$ is high, that is, when prevalence is low, and it is too strict when prevalence is high. As the data scenarios in which coverage and C-coverage are unreliable can coincide, a mere aggregation of both measures does not yield dependable necessity evaluation in all scenarios. Therefore, a new necessity measure, which can be trusted regardless of $|Y|/N$ and $|\Phi|/N$, is needed.

4 New sufficiency and necessity measures for CNA

While there is limited research on alternative sufficiency and necessity measures in CNA, other fields, including binary classification [25], association rule learning [32], and Bayesian epistemology [33], have extensively investigated evaluation measures for sufficiency and necessity—though under different labels such as *classification performance*, *association rule interestingness*, or *confirmation*. Some measures proposed in that context are suitable for the purposes of CNA, others are not, for example, because they evaluate sufficiency and necessity simultaneously and not, as is required for CNA model building, independently of one another. Still other measures, which are not directly suitable for CNA, say, because they take values outside of the [0,1] interval, can be adapted to fit CNA through slight modifications, such as transformations into the [0,1] interval.

In our search for better model-building measures, we explored a wide range of candidates from various disciplines and selected those for closer scrutiny that satisfied the conceptual criteria for evaluating sufficiency and necessity in CNA model building. We also adapted some measures to serve CNA's purposes. Among the measures considered were adjusted versions of consistency, C-consistency, coverage, and C-coverage featuring weights attached to the respective penalty terms, which deliver stronger or weaker penalization depending on prevalence and $|\Phi|/N$. We incorporated penalty weight exponents to adjust the influence of the weights on the final score. Moreover, we considered suitable variants of the Z-measure [33], as well as a range of harmonic means—both unweighted and weighted—of consistency and C-consistency for sufficiency evaluation, and of coverage and C-coverage for necessity evaluation.

We conducted small-scale initial simulation experiments with all measures in this pool of candidates to determine whether using them for CNA model building showed promise for improving the quality of CNA's outputs. Many of the candidates could be excluded based on the findings from these initial tests. However, four of them yielded promising results; two sufficiency measures and two necessity measures. This section presents the formal details of these four promising candidates and justifies them theoretically, before the next section reports the results of resource-intensive, large-scale benchmarking experiments we conducted with them.

4.1 Prevalence-adjusted consistency

The first promising sufficiency measure is what we will call *prevalence-adjusted consistency* (PA-consistency). This measure is equivalent to calibrated precision as proposed by Siblini et al. [34]. It addresses the limitations of consistency, which is too lenient when prevalence is high and too strict when prevalence is low, by imposing a stronger penalty when prevalence

is high and a weaker penalty when prevalence is low. It is defined as follows:

$$PA\text{-}consistency(\Phi \to Y) = \frac{|\Phi{*}Y|}{|\Phi{*}Y| + \frac{|Y|}{|y|} \cdot |\Phi{*}y|} \tag{7}$$

The difference between PA-consistency and regular consistency (cf. equation (3)) is that PA-consistency attaches a weight $\frac{|Y|}{|y|}$ to the penalty term $|\Phi{*}y|$ in the denominator. When $|Y| = |y|$, this weight is 1, to the effect that PA-consistency and regular consistency are equal. However, the weight increases as $|Y|$ increases relative to $|y|$, implying that a high weight is assigned to $|\Phi{*}y|$ when prevalence is high. Consequently, every case with $\Phi{*}y$ is penalized more strongly by PA-consistency than by regular consistency, meaning that the former is lower than the latter when prevalence is high (except when $|\Phi{*}y| = 0$ and both are equal to 1).

One crucial upshot of this is that, contrary to consistency, PA-consistency is not equal to prevalence when antecedent and consequent are independent. Instead, it is equal to $1/2$ in such scenarios.

Proof. When Φ and Y are statistically independent, Φ is not more or less likely to occur when Y occurs than when y occurs. This means that the proportion of cases with Φ among those featuring y is equal to the proportion of cases with Φ among those featuring Y. So,

$$\frac{|\Phi{*}y|}{|y|} = \frac{|\Phi{*}Y|}{|Y|} \tag{8}$$

Using this equality, the value of PA-consistency is determined as follows:

$$\begin{aligned} PA\text{-}consistency(\Phi \to Y) &= \frac{|\Phi{*}Y|}{|\Phi{*}Y| + \frac{|Y|}{|y|} \cdot |\Phi{*}y|} = \frac{|\Phi{*}Y|}{|\Phi{*}Y| + |Y| \cdot \frac{|\Phi{*}y|}{|y|}} \\ &= \frac{|\Phi{*}Y|}{|\Phi{*}Y| + |Y| \cdot \frac{|\Phi{*}Y|}{|Y|}} = \frac{|\Phi{*}Y|}{|\Phi{*}Y| + |\Phi{*}Y|} \\ &= \frac{|\Phi{*}Y|}{2 \cdot |\Phi{*}Y|} = \frac{1}{2} \end{aligned}$$

□

This shows that PA-consistency depends neither on prevalence nor on $|\Phi|/N$ when antecedent and consequent are independent. Contrary to consistency, PA-consistency cannot be high when antecedent and consequent are independent. It follows that PA-consistency is more reliable than regular consistency for sufficiency evaluation when prevalence is high.

Furthermore, when prevalence is low, the weight $\frac{|Y|}{|y|}$ is low, implying that PA-consistency assigns a weaker penalty to $|\Phi{*}y|$ than consistency. Consequently, whereas consistency plummets in low prevalence data once there are only a few noisy cases with $\Phi{*}y$, even if $\Phi \to Y$ is actually underwritten by a causal dependence, PA-consistency can still be high in such scenarios. This makes PA-consistency a more reliable sufficiency measure than regular consistency when prevalence is low.

4.2 Antecedent-adjusted C-consistency

The second promising sufficiency measure is defined analogously. We call it *antecedent-adjusted C-consistency* (AAC-consistency). It mitigates C-consistency's leniency when the antecedent Φ of a sufficiency relation is infrequent in the data (i.e. $|\Phi|/N$ is low) and its strictness when Φ is frequent (i.e. $|\Phi|/N$ is high):

$$AAC\text{-}consistency(\Phi \to Y) = \frac{|\phi{*}y|}{|\phi{*}y| + \frac{|\phi|}{|\Phi|} \cdot |\Phi{*}y|} \tag{9}$$

That measure differs from C-consistency by the weight $\frac{|\phi|}{|\Phi|}$ attached to the penalty term $|\Phi{*}y|$. If $|\phi| = |\Phi|$, this weight is 1, rendering AAC-consistency and C-consistency equal. But the weight increases as $|\Phi|$ decreases relative to $|\phi|$, implying that a high weight is assigned to $|\Phi{*}y|$ in datasets with low $|\Phi|/N$, which, in turn, yields that all cases with $\Phi{*}y$ are penalized more strongly by AAC-consistency than by C-consistency. As a result, the former is lower than the latter when $|\Phi|/N$ is low (except when $|\Phi{*}y| = 0$).

Again, a crucial upshot is that AAC-consistency, unlike C-consistency, is not equal to $|\phi|/N$ when antecedent and consequent are independent, rather, it is equal to $1/2$.

Proof. When Φ and Y are statistically independent, y is not more or less likely to occur when Φ occurs than when ϕ occurs. That means that the proportion of cases with y among those featuring Φ is equal to the proportion of cases with y among those featuring ϕ. So,

$$\frac{|\Phi{*}y|}{|\Phi|} = \frac{|\phi{*}y|}{|\phi|}$$

Using this equality, we can calculate the value of AAC-consistency as follows:

$$\begin{aligned} AAC\text{-}consistency(\Phi \to Y) &= \frac{|\phi{*}y|}{|\phi{*}y| + \frac{|\phi|}{|\Phi|} \cdot |\Phi{*}y|} = \frac{|\phi{*}y|}{|\phi{*}y| + |\phi| \cdot \frac{|\Phi{*}y|}{|\Phi|}} \\ &= \frac{|\phi{*}y|}{|\phi{*}y| + |\phi| \cdot \frac{|\phi{*}y|}{|\phi|}} = \frac{|\phi{*}y|}{|\phi{*}y| + |\phi{*}y|} \\ &= \frac{|\phi{*}y|}{2 \cdot |\phi{*}y|} = \frac{1}{2} \end{aligned}$$

□

That is, AAC-consistency neither depends on $|\Phi|/N$ nor on prevalence when antecedent and consequent are independent. Unlike C-consistency, AAC-consistency cannot be high when Φ and Y are independent. This makes AAC-consistency more reliable for evaluating $\Phi \to Y$ when $|\Phi|/N$ is low (i.e. when $|\phi|/N$ is high). Moreover, when $|\Phi|/N$ is high, the weight $\frac{|\phi|}{|\Phi|}$ is low, implying that AAC-consistency assigns a weaker penalty to $\Phi{*}y$ than C-consistency. As C-consistency is too strict when $|\Phi|/N$ is high, assigning a weaker penalty in these scenarios makes AAC-consistency a more reliable sufficiency measure when $|\Phi|/N$ is high.

4.3 Antecedent-adjusted coverage

The promising candidates for necessity evaluation are adjusted versions of coverage and C-coverage that feature appropriate weights attached to the penalty term $|\phi{*}Y|$. A first new necessity measure is *antecedent-adjusted coverage* (AA-coverage):

$$AA\text{-}coverage(Y \to \Phi) = \frac{|\Phi{*}Y|}{|\Phi{*}Y| + \frac{|\Phi|}{|\phi|} \cdot |\phi{*}Y|} \tag{10}$$

As regular coverage (cf. equation (4)) is too lenient when $|\Phi|/N$ is high and too strict when $|\Phi|/N$ is low, AA-coverage adds the weight $\frac{|\Phi|}{|\phi|}$ to the penalty $|\phi{*}Y|$. This makes AA-coverage stricter than coverage when $|\Phi|/N$ is high and more lenient than coverage when $|\Phi|/N$ is low, thereby mitigating the limitations of coverage.

Like PA-consistency and AAC-consistency, AA-coverage has a constant expected value of $1/2$, if antecedent and consequent are independent—which can be proven in close analogy to the constancy proofs given in sections 4.1 and 4.2. This makes AA-coverage more reliable than regular coverage (which has an expected value of $|\Phi|/N$ when antecedent and consequent are independent) for necessity evaluation when $|\Phi|/N$ is at high or low extremes.

4.4 Prevalence-adjusted C-coverage

Since C-coverage (cf. equation (6)) is too lenient when prevalence is low and too strict when prevalence is high, the second promising new necessity measure is *prevalence-adjusted C-coverage* (PAC-coverage):

$$PAC\text{-}coverage(Y \to \Phi) = \frac{|\phi{*}y|}{|\phi{*}y| + \frac{|y|}{|Y|} \cdot |\phi{*}Y|} \tag{11}$$

If $|y| = |Y|$, PAC-coverage and C-coverage are equal because the penalty weight $\frac{|y|}{|Y|}$ is 1. But the weight increases as $|y|$ increases relative to $|Y|$. It follows that, when prevalence is low, cases with $\phi{*}Y$ are penalized more strongly by PAC-coverage than by C-coverage.

Whereas C-coverage is equal to $|y|/N$ when Φ and Y are independent, PAC-coverage is equal to the constant $1/2$. The proof is analogous to the constancy proofs given in sections 4.1 and 4.2. When prevalence is low (i.e. when $|y|/N$ is high), therefore, PAC-coverage is not expected to be high when Φ and Y are independent. This makes PAC-coverage a more reliable necessity measure than C-coverage in low prevalence scenarios. Finally, when prevalence is high (i.e. when $|y|/N$ is low) the weight $\frac{|y|}{|Y|}$ is low, meaning that PAC-coverage penalizes less than C-coverage, which is too strict in such scenarios. The result is a more reliable necessity evaluation also when prevalence is high.

5 Benchmarking

We now have two promising sufficiency measures and two promising necessity measures on the table, but the CNA algorithm only needs one of each. Hence, a selection must be made. We have to identify the combination of adjusted sufficiency and necessity measures that, when

implemented in CNA's model building algorithm, yields the outputs with maximal quality. As this selection cannot be based on theoretical and conceptual considerations alone, we conduct an extended series of large-scale simulation experiments benchmarking the performance of four combinations—⟨PA-consistency, AA-coverage⟩, ⟨AAC-consistency, AA-coverage⟩, ⟨PA-consistency, PAC-coverage⟩, and ⟨AAC-consistency, PAC-coverage⟩—and comparing it to the performance of regular consistency and coverage. This section presents the setup and results of those benchmarking experiments. The code of the test series is available in the paper's online replication materials.

5.1 Test setup and data simulation

We design the tests as inverse search trials. In a nutshell, the trials consist in, first, randomly building data-generating structures (or ground truths), second, simulating data with noise, fragmentation, and prevalence imbalances from those structures, third, processing those data with a combination of evaluation measures, and fourth, determining the degree to which the resulting outputs comply with various benchmark criteria.

In the first step, we generate a stock of 1000 ground truths Δ_1 to Δ_{1000}, from the factor set $\mathbf{F} = \{\text{A, B, C, D, E, F, G}\}$. To avoid excessive runtimes, we restrict the maximal complexity of the ground truths: our Δ_i have one outcome only and a maximum of five alternative paths (i.e. disjuncts), with a maximum of three causes on each path (i.e. conjuncts), producing the outcome. The outcome is fixed to be A. Within these complexity confines, the generation of Δ_1 to Δ_{1000} is random. Some Δ_i are as simple as $C \leftrightarrow A$, while others feature antecedents with 5 disjuncts comprising 3 conjuncts each. In the second step, we simulate nine datasets δ_i^k from each ground truth Δ_i with randomized sample sizes, fragmentation and noise levels, and outcome prevalence systematically varied to nine different ratios. The data are generated in five phases.

(I), we create ideal data δ_i^{id} for Δ_i, such that each configuration compatible with Δ_i is represented by exactly one case in δ_i^{id}. (II), a fragmentation level $\mathcal{F}$ is randomly drawn from the interval $[0.2, 0.5]$, and $N_i^{id} \cdot \mathcal{F}$ randomly selected rows are removed from δ_i^{id}, where N_i^{id} is the sample size of δ_i^{id}. The resulting fragmented data δ_i^{fr} have a small to intermediate sample size (between 32 to 64 cases) and exactly one case per configuration. But of course, real data may have larger sizes and multiple cases may instantiate the same configuration. For that reason, we then, (III), sample a random number of rows (possibly 0) from δ_i^{fr} and combine them with δ_i^{fr} to an augmented dataset δ_i^{au} with a sample size anywhere between the original size of δ_i^{fr} and 200. In phase (IV), we introduce noise into every δ_i^{au}. This is done by drawing a noise level $\mathcal{S}$ from the interval $[0.01, 0.3]$, and then reversing (negating) the outcome value in $N_i^{au} \cdot \mathcal{S}$ randomly selected rows of δ_i^{au}, where N_i^{au} is the sample size of δ_i^{au}. Replacing the outcome value in a configuration that is compatible with Δ_i by its negation yields a configuration that is incompatible with Δ_i. In other words, the rows with reversed outcomes change from being signal to being noise rows.

Finally, in phase (V), we systematically manipulate the prevalence of the outcome, which is A in all datasets, to every value in the following *variation sequence*:

$$\langle 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 \rangle.^{11}$$

[11] Prevalence cannot be set to 0 or 1 because without variation in the outcome data would not contain difference-making evidence, forcing CNA to return nothing.

This is accomplished by purposefully duplicating and eliminating cases δ_i^{no} in such a way that the fragmentation and noise levels of δ_i^{no} remain unchanged. For example, in order to reduce the outcome prevalence to some lower value in the variation sequence, we duplicate noise and signal cases featuring A=0 at the noise ratio of δ_i^{no}, thereby maintaining that noise ratio, and we analogously eliminate cases with A=1, but only those that have duplicates in δ_i^{no}, to avoid increasing fragmentation.[12] The result are the nine test datasets δ_i^k for Δ_i, where k designates the value to which outcome prevalence is set. For example, $\delta_{23}^{0.9}$ refers to the test data simulated from ground truth Δ_{23} in which the prevalence of the outcome is 0.9.

5.2 Data analysis and benchmark criteria

We analyze the test data with five different combinations of sufficiency and necessity measures, yielding five different experiments A to E. Experiment A is the control run that applies CNA with regular consistency and coverage. In experiment B, we run CNA with PA-consistency and AA-coverage, experiment C implements AAC-consistency and AA-coverage, experiment D uses PA-consistency and PAC-coverage, and experiment E applies AAC-consistency combined with PAC-coverage. In each experiment, 9 000 test datasets are processed in nine test arms, $\mathcal{T}^{0.1}$ to $\mathcal{T}^{0.9}$, each corresponding to one prevalence level from the variation sequence and each comprising 1 000 datasets. That is, $\delta_1^{0.1}$ to $\delta_{1000}^{0.1}$ are analyzed in the first test arm, $\delta_1^{0.2}$ to $\delta_{1000}^{0.2}$ in the second, etc.

We use an adapted implementation of the model building algorithm in the **cna** R-library—this implementation is available in the paper's online replication materials—to run CNA with the new evaluation measures, and we employ the robustness analysis protocol developed by Parkkinen and Baumgartner [20]. That means that each δ_i^k is not only analyzed at one designated setting of sufficiency and necessity thresholds for the tested pair of measures but re-analyzed at all combinations of settings that can be built from a whole threshold sequence, in our case $\langle 0.95, 0.85, 0.75, 0.65\rangle$. All MINUS-formulas CNA recovers in that re-analysis series are collected and their robustness and overall model fit are measured and scored. For every δ_i^k, the 95^{th} percentile of top-performing MINUS-formulas is returned as CNA's output set $\mathbf{O}_i^k$ for these data.

The elements of such an output set, which on average contains between 1 and 5 models in our test series, are indistinguishable on the basis of the evidence contained in δ_i^k. Accordingly, if $\mathbf{O}_i^k$ comprises more than one MINUS-formula, CNA cannot determine which of those formulas truthfully represents the ground truth Δ_i; all that it infers is that at least one of them is true of Δ_i. It follows that a set $\mathbf{O}_i^k$ featuring, say, three MINUS-formulas $\mathbf{m}_1$, $\mathbf{m}_2$, and $\mathbf{m}_3$ is to be causally interpreted disjunctively: $\mathbf{m}_1$ OR $\mathbf{m}_2$ OR $\mathbf{m}_3$ is true of Δ_i.[13] Overall, analyzing the data of our entire test series yields 1 000 output sets for each of the test arms $\mathcal{T}^{0.1}$ to $\mathcal{T}^{0.9}$ in each of the experiments A to E.

We assess the quality of the output sets in each $\mathcal{T}^k$ by a summary score that aggregates two complementary benchmark scores averaged over all trials in $\mathcal{T}^k$ with the ratio of trials resulting in empty outputs. The first benchmark is a qualitative *correctness* criterion, which has been repeatedly used in CCM benchmarking before (e.g. [2, 5, 36]). What CNA infers

[12] Note that this procedure to manipulate prevalence while maintaining fragmentation and noise cannot also keep the sample size fixed. In consequence, test data at the lower and upper ends of the variation sequence typically have much higher sample sizes than datasets with mid-range prevalence levels.

[13] On par with Bayesian network methods, but different from typical regression methods, CCMs automatically build all equally data-fitting models (for more on CCM model ambiguities, see, e.g. [35]).

from δ_i^k counts as correct if, and only if, that inference is true of the ground truth Δ_i. As we have seen above, that is the case if, and only if, at least one MINUS-formula $\mathbf{m}_j$ in $\mathbf{O}_i^k$ is true of Δ_i, which, in turn, holds if, and only if, all factor values contained in $\mathbf{m}_j$ are in fact causes of the outcome of Δ_i and all conjunctive and disjunctive groupings in $\mathbf{m}_j$ are in agreement with Δ_i.[14] For example, if formula (1), i.e. $A{*}C + a{*}c + D \leftrightarrow B$, is the ground truth, models as $A{*}C \leftrightarrow B$ or $A + D \leftrightarrow B$ are correct because all factor values contained in these models are actually causes of B and all conjunctive and disjunctive groupings are true of (1). By contrast, a model as $A{*}d \leftrightarrow B$ is incorrect because d is not in fact a cause of B, or $A + C \leftrightarrow B$ is incorrect because A and C are conjunctively and not disjunctively grouped in (1). If CNA does not infer anything from δ_i^k and, thus, $\mathbf{O}_i^k$ is empty—say, because the chosen sufficiency or necessity thresholds cannot be met—we assign 'NA' to the correctness benchmark for that trial.

Making only true claims about Δ_i, as is required to pass the correctness benchmark, can be easily accomplished by models that make only very few causal claims. As more informative models are preferable to less informative ones, the second benchmark, which is called *completeness*, examines how much information about a ground truth is contained in a model. The completeness of a MINUS-formula $\mathbf{m}_j$ with respect to Δ_i is the proportion of Δ_i that is recovered by $\mathbf{m}_j$. To measure this, we divide the complexity of the maximal syntactic intersection of $\mathbf{m}_j$ and Δ_i by the complexity of Δ_i, where the complexity of a MINUS-formula is the number of factor value appearances in its antecedent. For example, if $A{*}C{+}a{*}c{+}D \leftrightarrow B$ is the ground truth, a model such as $A{*}C + A{*}D \leftrightarrow B$ scores $3/5 = 0.6$ on completeness, because the maximal syntactic intersection, $A{*}C + D \leftrightarrow B$, has complexity 3 and the ground truth has complexity 5. When $\mathbf{O}_i^k$ is empty, completeness also gets the value 'NA'.

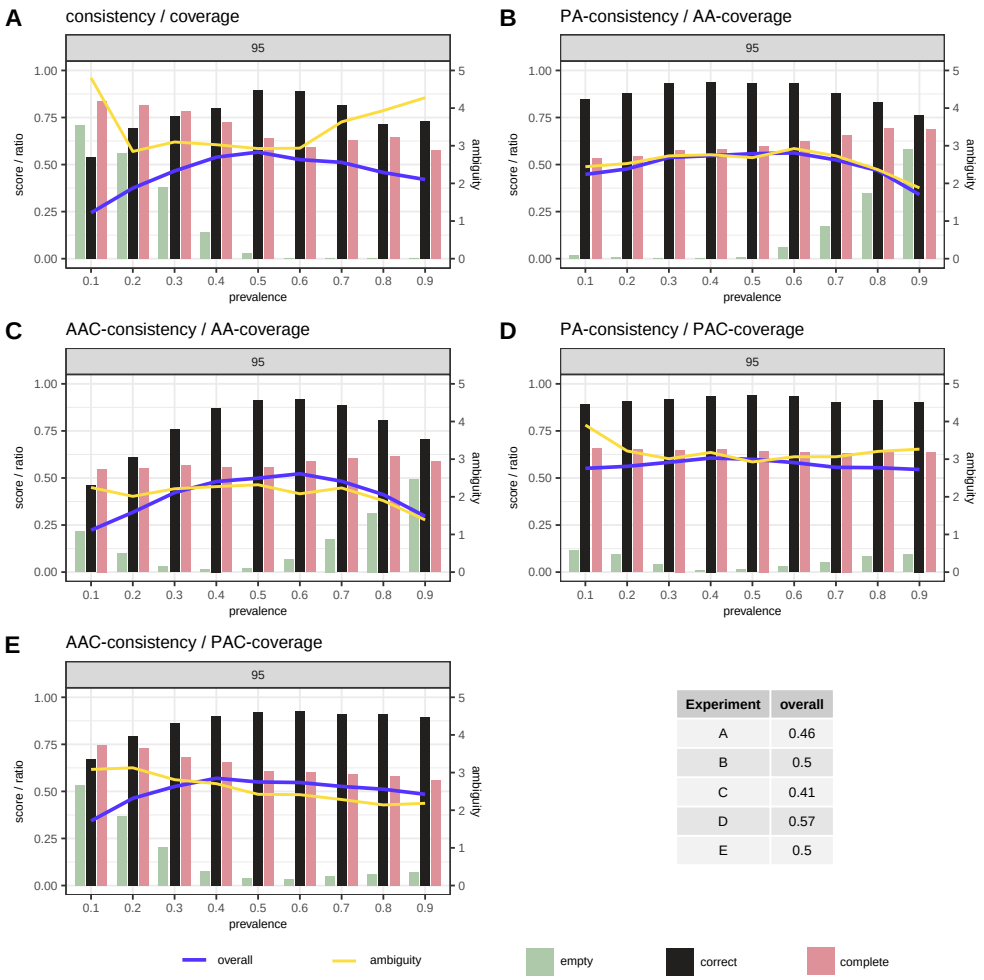
The ratio of trials in each test arm $\mathcal{T}^k$ with an empty output set $\mathbf{O}_i^k$ yields a final quality benchmark called *emptiness*. Although an empty output is suboptimal, it is still preferable to a false output. Issuing a correct and complete output is more important than issuing a non-empty one. Correspondingly, when aggregating correctness, completeness, and emptiness to an overall quality score, we give less weight to emptiness. More specifically, our overall quality score for an individual test arm $\mathcal{T}^k$ is the product of the correctness score averaged over all trials in $\mathcal{T}^k$ (with NAs removed), the average completeness score in $\mathcal{T}^k$ (with NAs removed), and the square root of the ratio of non-empty trials in $\mathcal{T}^k$:

$$overall(\mathcal{T}^k) = correct(\mathcal{T}^k) \cdot complete(\mathcal{T}^k) \cdot \sqrt{1 - empty(\mathcal{T}^k)} \tag{12}$$

5.3 Results

The results are plotted in Figure 2, subdivided by experiments A to E and complemented by a table with mean overall quality scores for the experiments as a whole. In the plots for the individual experiments, the benchmark scores are depicted separately for each test arm $\mathcal{T}^{0.1}$ to $\mathcal{T}^{0.9}$. Correctness, completeness, and emptiness are represented as black, red, and green bars, while the overall score is presented as a blue line. The relevant values for all benchmark scores are on the left y-axis. In addition, the yellow line labeled *ambiguity*, with relevant values on the right y-axis, indicates how many MINUS-formulas the output sets contain on average in each test arm. All represented values are averages over the 1000 trials in the corresponding

[14] These conditions are satisfied if $\mathbf{m}_j$ is a submodel of Δ_i (e.g. [2]).



Experiment	overall
A	0.46
B	0.5
C	0.41
D	0.57
E	0.5

Figure 2: Results of experiments A to E. Prevalence levels are given on the x-axis. The left y-axis indicates average correctness, completeness, and emptiness scores, depicted as bars, and overall quality scores, represented as lines. Average ambiguity ratios are given as lines with corresponding values on the right y-axis. All scores are averages from 1000 analyzed datasets in each test arm. The table in the bottom right corner provides mean overall quality scores for the experiments as a whole.

test arm. For instance, the emptiness ratio of 0.71 depicted by the left-most green bar in the plot for experiment A means that the output of CNA is empty in 71% of the 1000 trials of $\mathcal{T}^{0.1}$. The adjacent black correctness bar expresses that in 54% of the non-empty trials in that same test arm the CNA output contains a correct model.

The first notable finding derives from the control experiment A. It shows that the performance of CNA drops significantly at extreme prevalence levels when CNA builds its models with regular consistency and coverage. Specifically, if 80% or more of the cases in the data feature the outcome, CNA returns a correct model in fewer than 75% of the trials. If prevalence is at low extremes, CNA outputs are empty in over two-thirds of the trials. Among the non-empty outputs, only slightly more than half yield correct models, which, however, reach

the highest completeness scores of the whole test series, meaning that these models are overfitted supermodels of the ground truth.[15] Furthermore, ambiguity ratios increase noticeably at prevalence extremes. In sum, when employing consistency and coverage, reliable and informative model building with CNA requires prevalence to be balanced at mid-range levels. At those mid-range levels, however, the standard sufficiency and necessity measures yield an output quality on par with all other tested measures.

Our results also show that not all tested combinations of the promising measure candidates mitigate the sensitivity of CNA's output quality to prevalence. When models are built with the combination of AAC-consistency and AA-coverage, CNA's performance is even more sensitive to prevalence extremes than when regular consistency and coverage are used. In experiment C, fewer than half of the non-empty outputs contain a correct model at a prevalence of 0.1, and at a prevalence of 0.9, CNA does not find a model in one out of two trials. While the mean overall performance score in experiment A is 0.46, it is only 0.41 in experiment C.

In experiments B and E, CNA reaches overall performance scores of 0.5, which is about 9% better than in the control experiment A, and sensitivity to prevalence is merely one-sided: in experiment B, the quality of CNA's output only plunges when prevalence is very high, while in experiment E, overall performance dives at very low prevalence only. By contrast, in experiment B, there are virtually no empty trials when prevalence is low, while correctness remains above 0.85 and completeness above 0.53. Moreover, whereas the non-empty output sets contain almost 5 models, on average, at the lower prevalence end of experiment A, ambiguity is only at 2.4 models per trial at that same prevalence level of experiment B. When prevalence is high in experiment E, ambiguity ratios are even lower, while correctness and completeness are at 0.89 and 0.56, respectively.

Clearly though, the most notable result is that the combination of PA-consistency and PAC-coverage, which is examined in experiment D, renders the correctness and completeness of CNA's output entirely insensitive to prevalence extremes. The mean overall performance score for that experiment is 0.57, which is 23% higher than in the control experiment. Correctness stays above 0.89 in all test arms of experiment D, and completeness is consistently between 0.63 and 0.65. The only benchmark score that is slightly affected by prevalence imbalances is emptiness. At low and high prevalence extremes, CNA's output is empty in, respectively, 11% and 10% of the trials. This is a mild prevalence sensitivity in comparison to the severe impact prevalence imbalances can have on the number of trials with empty outputs in the other experiments. Averaged over all prevalence levels, the combination of PA-consistency as sufficiency measure and PAC-coverage as necessity measure is the winner of the test series.

6 Discussion

Our simulation experiments confirm much of what we anticipated based on the theoretical considerations in sections 3 and 4. First, the performance of CNA drops substantively

[15] That CNA's overall performance drops significantly at prevalence extremes aligns with recent findings of Swiatczak and Baumgartner [37]. But they use a different completeness benchmark, which measures the complexity of correct models and not, as does our completeness criterion, the complexity of the maximal intersection with the ground truth. Hence, regarding completeness, our results differ from Swiatczak and Baumgartner's.

at prevalence extremes when models are built with standard consistency and coverage. Second, when prevalence is balanced around 0.5, standard measures perform reliably; in fact, any combination of sufficiency and necessity measures works roughly as well as any other. Third, at prevalence extremes, three of the four tested combinations of adjusted sufficiency and necessity measures outperform consistency and coverage.

At the same time, there are unanticipated findings. On the one hand, given the promise that the new sufficiency and necessity measures have shown in section 4, we had reason to expect that all four combinations of them would yield higher quality CNA outputs at prevalence extremes. That does not materialize, as the combination of AAC-consistency and AA-coverage underperforms. On the other hand, the extent to which the combination of PA-consistency and PAC-coverage outperforms standard measures, rendering the correctness and completeness of CNA's output insensitive to prevalence extremes, is very impressive and exceeds our expectations.

The crucial follow-up question now becomes whether this impressive result can be generalized from the specifics of our concrete simulation experiments to model building with CNA in general. Should PA-consistency and PAC-coverage replace consistency and coverage for model building purposes in all research contexts and data scenarios or only in some; and if the latter, in which ones? To answer that question, we must consider to what degree the limitations of our experiments impact generalizability.

A first limitation of our experiments is that, in order to keep the computational demand manageable, the tested ground truths all have only a single outcome and their complexity is restricted. However, we have no reason to suspect that the results for single-outcome ground truths with limited complexity should not transfer to multi-outcome ground truths with higher complexity. Single-outcome models are mere conjunctions of multi-outcome models. Hence, if the former have higher quality when built with the new measures, the latter are also expected to have higher quality. Moreover, analyzing ground truths with complex causes of more than 3 conjuncts and 5 disjuncts increases the computational burden but does not pose any qualitatively different challenges than the analysis of the causal structures in our test series. Also, we see no significant differences in computation times between the five different experiments in our series, implying that the new measures are not expected to impose higher computational demands when building highly complex models than the original measures. That is, the complexity limitation of the ground truths in our experiments should not constrain the generalizability of our results.

A second limitation is that we vary prevalence by purposefully duplicating and eliminating cases, resulting in frequency-induced prevalence variation only. However, prevalence in real-life data can also vary because of the structural properties of the ground truth (see [37] for a detailed discussion). The reason for our focus on frequency-induced prevalence variation is that ground truths with structural properties yielding extremely high or low prevalence are rare. It is thus impractical to generate enough of these ground truths to obtain a sufficiently large sample size for reasonably powered experiments. We also expect prevalence in real-life data to be frequency-induced much more often than structure-induced. Besides, we see no reason why structure-induced prevalence variation would affect sufficiency and necessity measures in a different way than frequency-induced prevalence variation. We conclude that the unilateral focus on frequency-induced prevalence variation, though a limitation of the experiments, does not hinder the generalizability of our results.

A third limitation is that we only manipulate outcome prevalence and not the frequency of candidate causes, or more specifically, $|\Phi|/N$. It might appear that manipulating prevalence rather than $|\Phi|/N$ biases the results in favor of PA-consistency and PAC-coverage, which have weights that depend on prevalence and are designed to handle prevalence variation. However, note that these prevalence-adjusted measures are derived from consistency and C-coverage, which are sensitive to prevalence in the first place. Consequently, as pushing prevalence to extremes results in data that are problematic for the unadjusted counterparts of PA-consistency and PAC-coverage, our test design might also bias the experiments *against* these measures. But even if there is an underlying bias, the extent to which PA-consistency and PAC-coverage outperform standard consistency and coverage is so large that it is implausible to arise exclusively due to bias. Hence, the combination of PA-consistency and PAC-coverage can be expected to outperform standard measures also in contexts without systematic prevalence variation. At the same time, it must be recognized that the performance increase delivered by PA-consistency and PAC-coverage is surprisingly high. This suggests that some of the other pairs of new sufficiency and necessity measures might also yield surprising results if, instead of prevalence, we were to push the frequency of some of the candidate causes to extremes. Follow-up studies will be needed to investigate this question. Until then, at least those three pairs of new measures that outperformed consistency and coverage in our experiments should remain under consideration.

A fourth limitation is that all our experiments exclusively analyze crisp-set data, which allow for the most streamlined discussion of methodological issues related to configurational comparative methods in general, and evaluation measures in particular. But CNA can also analyze multi-value and fuzzy-set data (cf. section 2). Although CNA processes all data types with the same underlying algorithm, each type poses specific challenges. Multi-value data, for instance, tend to have very low prevalence. The reason, in short, is that when factors can take more than two values the total space of possible configurations is much larger, resulting in each value being taken relatively less frequently; and since outcomes are factors taking values, outcomes are less frequent, on average, in multi-value than in crisp-set data. In fact, we are not aware of any multi-value study, using either QCA or CNA, with an outcome reaching a prevalence of 0.8 or above. A possible consequence is that the combination of PA-consistency and AA-coverage, which yields virtually no empty outputs at the lower prevalence spectrum in our crisp-set test series, produces better overall CNA outputs for multi-value data than the combination of PA-consistency and PAC-coverage, with a significantly lower ambiguity ratio at low prevalence. In any case, our crisp-set results cannot easily be generalized to the multi-value case. A separate study will be needed to determine which sufficiency and necessity measures perform best when building models for multi-value data.

Fuzzy-set data have the particularity that the properties represented by fuzzy-set factors are not mutually exclusive. In other words, it is possible for cases in the data to have non-zero membership both in Φ and in ϕ, as well as in Y and in y. It follows that the proofs of section 4—showing that the adjusted sufficiency and necessity measures go to $1/2$ for crisp-set (and multi-value) data when Φ and Y are independent—do not generalize to the fuzzy-set case. Instead, to reach an expected value of $1/2$ at independence, an additional correction must be applied to the penalty terms of the adjusted measures. More concretely, the mean of $min(\Phi, \phi, Y, y)$ must be added to the numerator and subtracted from the denominator

of the penalty weight. The term $min(\Phi, \phi, Y, y)$ is always equal to 0 in crisp-set (or multi-value) data, which is why it does not have to be included in the crisp-set formulations of the measures. In some preliminary simulation experiments with fuzzy-set data, we have seen that this correction often leads to very large or very small penalties, generating a lot of empty outputs. It therefore needs to be relaxed in practice by means of a sensitivity hyperparameter. We have experimented with some parameter settings, obtaining promising first results, which however are not conclusive yet. More research is needed. Clearly though, the results from our crisp-set experiments do not generalize to fuzzy-set data.

Overall, we conclude that the results reported in section 5 can be generalized beyond our specific test design to all crisp-set applications of CNA. Across all possible prevalence levels, crisp-set CNA outputs built by PA-consistency and PAC-coverage can be expected to be of substantively higher quality than outputs built by standard consistency and coverage. When prevalence is low, standard CNA outputs tend to be empty, and when prevalence is high, standard outputs tend to be overfitted. By contrast, in all of these data scenarios, PA-consistency and PAC-coverage enable CNA to reliably find correct and reasonably complete models.

7 Outlook

Based on our results, we recommend that the four new adjusted measures all be made available in relevant CNA software [21] and that CNA users transition to building crisp-set models using PA-consistency and PAC-coverage, especially if their data are affected by prevalence imbalances. More research is needed before analogous recommendations can be made for multi-value and fuzzy-set applications. Still, once all new measures are available in the CNA software, we advise users to integrate the values of the new scores into their data analysis routines also when analyzing multi-value or fuzzy-set data. The new measures can be valuable for model selection or cross-validation purposes, on par with existing solution attributes such as complexity, exhaustiveness, or faithfulness [38]. Furthermore, we strongly encourage follow-up studies analogous to ours that focus on evaluating sufficiency and necessity in multi-value and fuzzy-set analyses.

At the same time, our results should not be taken to imply that standard consistency and coverage would no longer be valuable tools for crisp-set CNA. First, as we saw in section 5, all sufficiency and necessity measures, including consistency and coverage, yield CNA outputs of about equal quality when prevalence is close to balanced, meaning that models might just as well be built with consistency and coverage in these circumstances. Second, consistency and coverage remain valuable for model evaluation and model selection—along with C-consistency and C-coverage [27], case knowledge, and theoretical knowledge [3, p. 172]. Third, running complementary analyses using consistency and coverage in addition to PA-consistency and PAC-coverage may facilitate cross-validation. Although more research will be needed to determine exactly how much more reliable (a part of) a model is when it is returned by more than one pair of sufficiency and necessity measures, we strongly suspect that successful cross-validation delivers a substantive boost in reliability.

However, our results should be taken as a reason to reconsider the growing literature on data imbalances in configurational comparative methods. The problems that prevalence imbalances create for methods such as QCA and CNA have long been noticed and discussed

in the literature [3, 37, 39], but they have typically been attributed to deficiencies in data collection or calibration. That is, their source has been located in pre-analytic phases of QCA or CNA studies. Our results show that it is possible to greatly mitigate the problems posed for CNA by prevalence imbalances through suitable adjustments of the implemented sufficiency and necessity measures. This locates the source of the problems within CNA's analytic phase itself. Consequently, collecting more data or re-calibrating the data are no longer the only available responses when data are affected by prevalence imbalances. Instead, such imbalances can newly also be addressed by choosing the right sufficiency and necessity measures. Much more work will be needed in that area in the future. In particular, studies similar to ours are needed that explore the potential to improve QCA's performance through new sufficiency and necessity measures. Consistency and coverage have a long, unquestioned, and unrivaled status as gold standard for sufficiency and necessity evaluation in configurational comparative methods, especially in crisp-set analyses. That status needs to be questioned. It is time to seriously consider alternatives.

Abbreviations

CNA: Coincidence Analysis
QCA: Qualitative Comparative Analysis
C-consistency: contrapositive consistency
C-coverage: contrapositive coverage
PA-consistency: prevalence-adjusted consistency
PAC-coverage: prevalence-adjusted contrapositive coverage
AAC-consistency: antecedent-adjusted contrapositive consistency
AA-coverage: antecedent-adjusted coverage

Declarations

Ethics approval and consent to participate. Not applicable.

Consent for publication. Not applicable.

Availability of data and materials. The simulated datasets used in the experiments presented in this paper, complete results of the conducted CNA analyses, and scripts to fully replicate the presented examples and simulation experiments are available both on Github: https://github.com/Luna-De-Souter/New-sufficiency-and-necessity-measures-for-model-building-with-Coincidence-Analysis and in the Zenodo repository, https://doi.org/10.5281/zenodo.13259601.

Competing interests. The authors declare that they have no competing interests.

Funding. This work was funded by the Research Council of Norway (grant number 326215) and the Trond Mohn Foundation (grant number 811866).

Authors' contributions. LDS analyzed the limitations of existing sufficiency and necessity measures and proposed candidate new measures. Both authors contributed to the selection of candidate measures to be tested in the simulation experiments and to the design of these simulation experiments. MB conducted the experiments. Both authors contributed to the first

draft of the manuscript and made critical revisions, and both authors read and approved the final manuscript.

Acknowledgements. We are very grateful to Mathias Ambühl for developing implementations of CNA's model-building algorithm to conduct the experiments. Furthermore, we thank Veli-Pekka Parkkinen, Martyna Swiatczak, and the participants in the 3rd International Conference on Current Issues in Coincidence Analysis, the 9th biennial meeting of the European Philosophy of Science Association, and the 11th International QCA Workshops for helpful feedback.

References

[1] Baumgartner M. Inferring causal complexity. Sociological Methods & Research. 2009;38:71–101. https://doi.org/10.1177/0049124109339369

[2] Baumgartner M, Ambühl M. Causal modeling with multi-Value and fuzzy-set Coincidence Analysis. Political Science Research and Methods. 2020;8:526–542. https://doi.org/10.1017/psrm.2018.45.

[3] Oana IE, Schneider CQ, Thomann E. Qualitative Comparative Analysis (QCA) Using R: A Beginner's Guide. Cambridge: Cambridge University Press; 2021.

[4] Ragin CC. Redesigning Social Inquiry: Fuzzy Sets and Beyond. Chicago: University of Chicago Press; 2008.

[5] Baumgartner M, Falk C. Configurational causal modeling and Logic Regression. Multivariate Behavioral Research. 2023;58:292–310. https://doi.org/10.1080/00273171.2021.1971510.

[6] Spirtes P, Glymour C, Scheines R. Causation, Prediction, and Search. 2nd ed. Cambridge: MIT Press; 2000.

[7] Baumgartner M, Falk C. Boolean difference-making: A modern regularity theory of causation. The British Journal for the Philosophy of Science. 2023. https://doi.org/10.1093/bjps/axz047.

[8] Mackie JL. The Cement of the Universe. A Study of Causation. Oxford: Clarendon Press; 1974.

[9] Dodge J, Sullivan K, Miech E, Clomax A, Riviere L, Castro C. Exploring the social determinants of mental health by race and ethnicity in Army wives. Journal of Racial and Ethnic Health Disparities. 2023. https://doi.org/10.1007/s40615-023-01551-3.

[10] Haesebrouck T, Thomann E. Introduction: Causation, inferences, and solution types in configurational comparative methods. Quality & Quantity. 2022;56(4):1867–1888. https://doi.org/10.1007/s11135-021-01209-4

[11] Roczniewska M, Tafvelin S, Nielsen K, von Thiele Schwarz U, Miech EJ, Hasson H, et al. Simple roads to failure, complex paths to success: An evaluation of conditions explaining perceived fit of an organizational occupational health intervention. Applied Psychology. 2023. https://doi.org/10.1111/apps.12502.

[12] Swiatczak MD. Towards a neo-configurational theory of intrinsic motivation. Motivation and Emotion. 2021;45(6):769–789. https://doi.org/10.1007/s11031-021-09906-1.

[13] Adams K, Miech E, Sobieraj D. Factors that distinguish opioid withdrawal during induction with buprenorphine microdosing: A configurational analysis. Addiction Science & Clinical Practice. 2022;17(1). https://doi.org/10.1186/s13722-022-00336-z.

[14] Cragun DL, Hunt PP, Dean M, Weidner A, Shields AK, Tezak A, et al. Applying the framework for developing and evaluating complex interventions to increase family communication about hereditary cancer. PEC Innovation. 2023;2:100133. https://doi.org/10.1016/j.pecinn.2023.100133.

[15] Dy SM, Acton RM, Yuan CT, Hsu YJ, Lai AY, Marsteller J, et al. Association of implementation and social network factors with patient safety culture in medical homes: A Coincidence Analysis. Journal of Patient Safety. 2020. https://doi.org/10.1097/PTS.0000000000000752.

[16] Rich J, Miech E, Semenza D, Corbin T. How combinations of state firearm laws link to low firearm suicide and homicide rates: A configurational analysis. Preventive Medicine. 2022. https://doi.org/10.1016/j.ypmed.2022.107262.

[17] Schlick CJR, Huang R, Brajcich BC, Halverson AL, Yang AD, Kreutzer L, et al. Unbundling bundles: Evaluating the association of individual colorectal surgical site infection reduction bundle elements on infection rates in a statewide collaborative. Diseases of the Colon & Rectum. 2022;65(8):1052–1061. https://doi.org/10.1097/DCR.0000000000002223.

[18] Yakovchenko V, Miech EJ, Chinman MJ, Chartier M, et al S S. Strategy configurations directly linked to higher hepatitis C virus treatment starts: An applied use of Configurational Comparative Methods. Medical Care. 2020;58(5). https://doi.org/10.1097/MLR.0000000000001319.

[19] Zotero Coincidence Analysis Group Library. https://www.zotero.org/groups/4567107. Accessed on 14 August 2024.

[20] Parkkinen VP, Baumgartner M. Robustness and model selection in configurational causal modeling. Sociological Methods & Research. 2021. https://doi.org/10.1177/0049124120986200.

[21] Ambühl M, Baumgartner M.: **cna**: Causal modeling with Coincidence Analysis. Version 3.5.4. https://cran.r-project.org/package=cna.

[22] Falk C, Ambühl M, Baumgartner M.: **causalHyperGraph**: Drawing Causal Hypergraphs. Version 0.1.0. https://cran.r-project.org/package=causalHyperGraph.

[23] Parkkinen VP, Baumgartner M, Ambühl M.: **frscore**: Functions for Calculating Fit-Robustness of CNA-Solutions. Version 0.4.1. https://cran.r-project.org/package=frscore.

[24] Ragin CC. Set relations in social research: Evaluating their consistency and coverage. Political Analysis. 2006;14(3):291–310. https://doi.org/10.1093/pan/mpj019.

[25] Tharwat A. Classification assessment methods. Applied Computing and Informatics. 2021;17(1):168–192. https://doi.org/10.1016/j.aci.2018.08.003.

[26] Manning CD, Raghavan P, Schütze H. Introduction to Information Retrieval. Cambridge: Cambridge University Press; 2008.

[27] De Souter L. Evaluating Boolean relationships in Configurational Comparative Methods. Journal of Causal Inference. 2024. https://doi.org/10.1007/s11229-008-9348-0.

[28] Bowran AP. A Boolean Algebra. Abstract and Concrete. London: Macmillan; 1965.

[29] Lemmon EJ. Beginning Logic. London: Chapman & Hall; 1965.

[30] Haesebrouck T. Pitfalls in QCA's consistency measure. Journal of Comparative Politics. 2015;8(2):65–80.

[31] Veri F. Aggregation bias and ambivalent cases: A new parameter of consistency to understand the significance of set-theoretic sufficiency in fsQCA. Comparative Sociology. 2019;18(2):229–255. https://doi.org/10.1163/15691330-12341496.

[32] Glass DH. Confirmation measures of association rule interestingness. Knowledge-based Systems. 2013;44:65–77. https://doi.org/10.1016/j.knosys.2013.01.021

[33] Crupi V, Tentori K, Gonzalez M. On Bayesian measures of evidential support: Theoretical and empirical issues. Philosophy of Science. 2007;74(2):229–252. https://doi.org/10.1086/520779.

[34] Siblini W, Fréry J, He-Guelton L, Oblé F, Wang YQ. Master your metrics with calibration. In: Berthold MR, Feelders A, Krempl G, editors. Advances in Intelligent Data Analysis XVIII. Cham: Springer; 2020. p. 457–469.

[35] Baumgartner M, Thiem A. Model ambiguities in configurational comparative research. Sociological Methods & Research. 2017;46(4):954–987. https://doi.org/10.1177/0049124115610351.

[36] Baumgartner M, Thiem A. Often trusted but never (properly) tested: Evaluating Qualitative Comparative Analysis. Sociological Methods & Research. 2020;49:279–311. https://doi.org/10.1177/0049124117701487.

[37] Swiatczak MD, Baumgartner M. Data imbalances in Coincidence Analysis: A simulation study. Sociological Methods & Research. 2024. https://doi.org/10.1177/00491241241227039.

[38] Baumgartner M, Ambuehl M. **cna**: An R Package for Configurational Causal Inference and Modeling. Vignette included in R package cna, version 362. 2024. https://CRAN.R-Project.org/package=cna.

[39] Schneider CQ, Wagemann C. Set-Theoretic Methods: A User's Guide for Qualitative Comparative Analysis (QCA) and Fuzzy-Sets in the Social Sciences. Cambridge: Cambridge University Press; 2012.