Access to this work was provided by the University of Maryland, Baltimore County (UMBC) ScholarWorks@UMBC digital repository on the Maryland Shared Open Access (MD-SOAR) platform.

Please provide feedback

CKH: Causal Knowledge Hierarchy for Estimating Structural Causal Models from Data and Priors

Riddhiman Adib   RIDDHIMAN.ADIB@MARQUETTE.EDU
Department of Computer Science, Marquette University

Md Mobasshir Arshed Naved   NAVED@PURDUE.EDU
Department of Computer Science, Purdue University

Chih-Hao Fang   FANG150@PURDUE.EDU
Department of Computer Science, Purdue University

Md Osman Gani   MOGANI@UMBC.EDU
Department of Information Systems, University of Maryland, Baltimore County

Ananth Grama   AYG@CS.PURDUE.EDU
Department of Computer Science, Marquette University

Paul Griffin   PMG14@PSU.EDU
Department of Industrial Engineering, Pennsylvania State University

Sheikh Iqbal Ahamed   SHEIKH.AHAMED@MARQUETTE.EDU
Department of Computer Science, Marquette University

Mohammad Adibuzzaman   ADIBUZZA@OHSU.EDU
Oregon Clinical and Translational Research Institute, Oregon Health & Science University

Abstract

Structural causal models (SCMs) provide a principled approach to identifying causation from observational and experimental data in disciplines ranging from economics to medicine. SCMs, however, require domain knowledge, which is typically represented as graphical models. A key challenge in this context is the absence of a methodological framework for encoding priors (background knowledge) into causal models in a systematic manner. We propose an abstraction called causal knowledge hierarchy (CKH) for encoding priors into causal models. Our approach is based on the foundation of "levels of evidence" in medicine, with a focus on confidence in causal information. Using CKH, we present a methodological framework for encoding causal priors from various data sources and combining them to derive an SCM. We evaluate our approach on a simulated dataset and demonstrate overall performance compared to the ground truth causal model with sensitivity analysis.

1. Introduction

Causal knowledge discovery or identifying cause-and-effect relationships between variables is a fundamental objective in various domains such as epidemiology and medicine (Rothman and Greenland, 2005), sociology (Gangl, 2010), and economics (Gow et al., 2016). Without understanding causal relationships, scientists rely on correlations, which do not allow for estimation of intervention effect (i.e., doing) of a variable on a model outcome. While randomized controlled trials remain the gold standard for exploring causation (Kovesdy and

Kalantar-Zadeh, 2012), they are often infeasible because of cost, time, and/or ethical reasons (Glass et al., 2013). Thus, causal discovery from observational data that is complementary to experimental studies is of significant interest (Hernán, 2018; Listl et al., 2016; Nichols, 2007).

Recent developments in the theory of causal inference under the Pearl causal hierarchy (PCH) (Bareinboim et al., 2020; Pearl et al., 2016, 2009; Bareinboim and Pearl, 2016), also known as structural theory of causation (within the potential outcome framework) provides the methodologies to estimate causal effects from observational data. Within this, a causal model is expressed through structural causal models (SCMs) (Pearl, 2010). SCMs represent variables of interest (exogenous and endogenous), causal relationships between the variables, and underlying probability distributions. SCMs use a graphical representation of the causal model, formalize the identification of causal effects from observational and experimental data, estimate the interventional distribution ($P(y|do(x))$) through do-calculus (Pearl, 2019), and assess hypothetical scenarios from available data and model with necessary assumptions explicitly encoded into the model.

PCH is grounded in the three layers of causation – (L_1) seeing, (L_2) doing, and (L_3) imagining. Recent work on the causal hierarchy theorem (CHT) proved that discovery in a higher layer of causation using only information from a lower layer is not feasible (Bareinboim et al., 2020). In other words, estimating the effect of experimentation (i.e., "doing (L_2)") is not feasible based only on observational data (i.e., "seeing (L_1)") (Bareinboim et al., 2020). Hence it is critical to augment observational data (L_1) with other sources of information such as expert knowledge to derive the effect of intervention. Expert knowledge can come in different forms such as expert opinions, established causal relationships and, peer-reviewed literature (Druzdzel and Díez, 2003). Within each of these forms, confidence in knowledge can vary. However, no methodological framework exists for incorporating domain expertise with data-driven causal discovery from observational data in a systematic way (Spirtes and Zhang, 2016).

We develop a methodological framework to augment data-driven causal discovery tools with human in the loop (HTL) models. The additional causal knowledge sources include different tiers of knowledge, but is not limited to: background knowledge, expert opinion, and literature. For this purpose, we have broadly categorized possible causal knowledge sources into three tiers and proposed a causal knowledge hierarchy (CKH) between the tiers. Using this causal knowledge hierarchy, we develop an associated standardized methodology to curate necessary causal information and merge them to derive the structural causal model (SCM). We also provide both theoretical and simulated results of the framework along with algorithmic pseudo code detailing the implementation.

We make the following specific contributions in this work:

1. We propose a causal knowledge hierarchy (CKH) on the foundation of levels of evidence in medicine based on the confidence in the causal information.

2. We present a standard methodological pipeline, based on CKH, to capture causal knowledge from different sources and combine them to derive the SCM.

3. We show the effectiveness of our proposed method in a simulated experiment, detailing the implementation, along with evaluation.

2. Related Work

The need for causal inference from sources other than data has been explored in the literature (Spirtes and Zhang, 2016). From a theoretical perspective, research has been conducted on integrating causal information from varying sources. Lee et al. (Lee and Bareinboim, 2020) proposed GID-PO that identifies the causal effect from partially-observed distributions. In another result (Lee et al., 2020), an algorithmic approach that combines data collected under multiple, disparate regimes (observational and interventional) to identify specific causal effects was presented. However, the focus of these experiments was not on prior knowledge or varying knowledge sources. Borboudakis et al. (Borboudakis and Tsamardinos, 2012) used path-constraints to incorporate prior causal knowledge, without explicitly discussing the impact of causal knowledge sources. For tiered knowledge, Andrews et al. (Andrews et al., 2020) proposed tiered background knowledge where each tier consists of sets of variables with causal relationships, preceding another set of variables (aka, tier), and demonstrated that FCI (Fast Causal Inference) is a sound and complete causal structure elarning algorithm with and incorporation of this knowledge. From an empirical viewpoint, (Nordon et al., 2019) developed a causal model from medical literature and electronic medical record (EMR) data, by generating two independent graphs, one based on the literature and one from the EMR data, and merged them. The method did not consider other sources of knowledge and did not compare knowledge sources as well as their confidence of information. In a related result, a prior-knowledge-based causal discovery algorithm (Wang et al., 2020) has been proposed to discover the underlying causal mechanism between bone mineral density and its factors from the clinical data, where prior knowledge was handpicked manually and added to the algorithm as a whitelist between edges. Finally, from a software application perspective, a graphically similar software application to help researchers navigate published findings has been proposed by the software "ResearchMaps" (Landreth and Silva, 2013). However, this is primarily a visualization tool to illustrate interconnected features with a graph. In summary, although previous research has attempted to resolve causal information from varying sources, a unified and principled approach to build a generic SCM is still needed.

3. Expert Augmented Causal Model with Knowledge Hierarchy

We propose the concept of tiers of causal knowledge and the generalized algorithm for causal model learning through knowledge hierarchy. Specifically, we propose *Causal Knowledge Hierarchy (CKH)* that uses three tiers of knowledge analogous to the "hierarchy of evidence" (Ackley et al., 2008) and the associated weight for each tier. We discuss the assumptions within the CKH required for our proposed methodology. Finally, we establish step-by-step actions for the method. An overview of our method is presented in Figure 1.

3.1. Causal Knowledge Hierarchy

Levels of evidence (Ackley et al., 2008), is a well established knowledge hierarchy based on the study design, data collection, and sample size. Based on this concept we propose a "Causal Knowledge Hierarchy (CKH)" to incorporate causal information from different sources. CKH is a multi-level descriptor between types of knowledge and their contribution
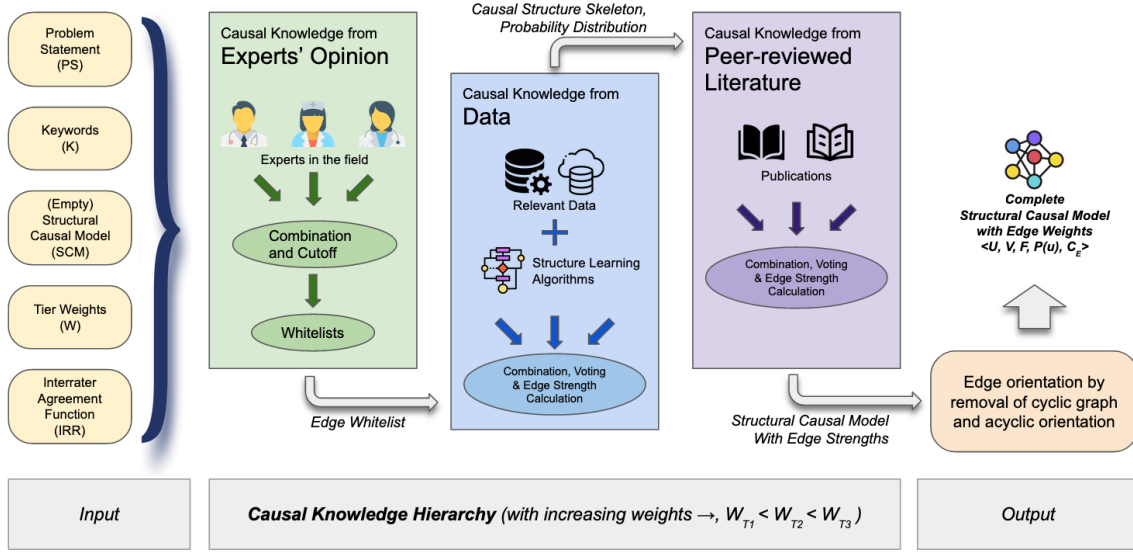
Figure 1: Our approach starts with the inputs: problem statement (PS), defined keywords (K), (empty) structural causal model (SCM), pre-defined tier weights (W), and Inter-rater agreement function (IRR) (*discussed in supplementary document*). For each tier, a general series of steps is described. Finally, the SCM goes through an edge orientation phase to produce the fully specified SCM with individual edge weights ($\langle U, V, F, P(u), C_E \rangle$)

to the overall causal structure in a problem domain. We initially define three common sources of causal information and propose a hierarchy between them. We define necessary assumptions to make our proposed framework effective and generalizable. We categorize sources of causal knowledge for scientific studies into three distinct classes and define a hierarchy (**CKH**) based on the statistical confidence in the causal information they hold.

Definition 1 *The three tiers of the CKH are: (1) **Tier 1: Causal knowledge from expert opinion / expertise (**CK_E**), (2) **Tier 2: Causal knowledge from data (**CK_D**), and (3) **Tier 3: Causal knowledge from peer-reviewed literature (**CK_L**). The target structural causal model is a function of convex combination of causal knowledges from the three tiers of sources,*

$$SCM \leftarrow f(CK_E, CK_D, CK_L)$$

3.1.1. TIER 1: CAUSAL KNOWLEDGE FROM EXPERT OPINION

Tier 1 of CKH incorporates causal knowledge based on the expertise and opinions (CK_E) from researchers, scientists, and, subject matter experts (SME). This includes, but is not limited to, inputs from physicians, discussion with application users and intervention participants and, by researchers working in a specific problem domain; and excludes any knowledge
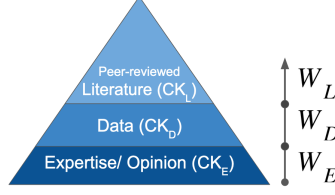
Figure 2: Tiers of Causal Knowledge Source

directly from peer-reviewed literature. Causal knowledge is generally captured through surveys or structured communications' methods-driven group discussions (e.g., Delphi method (Linstone et al., 1975)). This collaborative knowledge requires further validation through scientific studies, and is prone to high levels of bias due to variation in the expert's training and experience. We classify this causal knowledge as Tier 1 ($\mathbf{CK_E}$), and assign a lower weight ($\mathbf{W_E}$) since it contributes diverse information with lesser confidence.

3.1.2. Tier 2: Causal Knowledge from Data

Tier 2 encodes causal knowledge generated from data sources ($\mathbf{CK_D}$). Data can be from various study designs such as an experimental study (e.g., data from randomized controlled trials), an observational study (e.g., text mining data from social media), or in between (e.g., data from pragmatic clinical trial). Depending on the data generation mechanism, different structural causal models are used to explain the causal relationships between the variables. However, there may be bias from selection, confounding, or other experimental design features. We associate the causal knowledge gathered from data ($\mathbf{CK_D}$) at Tier 2, with a relatively higher weight ($\mathbf{W_D}$) than that of Tier 1 ($\mathbf{CK_E}$). The rationale for this is: (a) data can be collected from different study designs, locations, and corroborated over time, (b) data can be analyzed further with newer methodologies and models, and (c) data can convey the effect of causal relationships between covariates for scientific studies. Because of the higher weights, it contributes more to that conjoined causal model, and can even alter directions of certain causal relationships defined from Tier 1.

3.1.3. Tier 3: Causal Knowledge from Literature

Tier 3 is causal information from peer-reviewed literature (CK_L) and has the highest weight in the CKH. It excludes any knowledge from opinions of experts, without references. Examples of Tier 3 include causal knowledge from peer-reviewed and published literature, systematic reviews, meta-analyses, evidence syntheses, article synopses, and causal effect of interventions published as studies. Within Tier 3, there may be different levels of evidence. The data extraction process additionally falls under the domain of text-mining and natural language processing (NLP). Causal knowledge from literature ($\mathbf{CK_L}$) may have its own biases such as selection bias for inclusion exclusion criteria or transportability bias for differences in population.

3.1.4. TIER WEIGHTS FOR CAUSAL KNOWLEDGE HIERARCHY

Axiom 2 *Each tier of causal knowledge hierarchy (CKH) has individual weights (W_E, W_D, W_L), signifying the confidence of the causal information. A higher tier holds a higher weight and provides more robust causal information compared to that from lower tiers. By definition,*

$$CK_E \propto W_E, \; CK_D \propto W_D, \; CK_L \propto W_L$$

Based on the causal information hierarchy proposed, we define three weights $W = \{W_E, W_D, W_L\}$ for each tier of the causal knowledge hierarchy (*refer to Axiom 1*). The weights are defined such that:

1. $\sum_{\mathbf{i}} \mathbf{W_i} = \mathbf{1}$: By definition of convex combination, the sum of all three tier weights $(\sum W = W_E + W_D + W_L)$ is 1.0. A full agreement for a specific edge connection and direction from all three tiers of the CKH results in maximum edge confidence of 1.0.

2. $\mathbf{W_E} < \mathbf{W_D} < \mathbf{W_L}$: The weights are defined in an increasing order. At any time, causal information from one tier can only contribute a maximum of their tier weight. Thus, this increasing score ensures a hierarchy between each tier.

3.1.5. ASSUMPTIONS

For our proposed methodological framework, there are two associated assumptions.

Assumption 1 *Knowledge within the same tier of CKH does not override one another.*

For conflict resolutions with contradictory causal information within the same tier (such as, $A \rightarrow B$ from dataset 1 and $A \leftarrow B$ from dataset 2, both from CK_D),we find the strength of the causal relationship based on all the information within the same tier. Selection of knowledge sources within a tier is subjective and depends on the experimenter. Consequently, we do not propose any hierarchy within a tier, rather, we compute the conjoined strength of the causal connections. A similar direction in causal connections and edges increase the confidence, whereas contradictory causal connections and edges reduce the confidence of the edge.

Assumption 2 *Within CKH, knowledge from upper tier (or, in special case, tier with more weight) can reverse/ override knowledge from lower tier.*

Unlike the earlier assumption, when we have contradictory causal information from different tiers (such as, $A \rightarrow B$ from Tier 1 (CK_E) and $A \leftarrow B$ from Tier 2 (CK_D)), the direction of causal relationship from an upper tier can override that from a lower tier.

3.2. Algorithm

Our algorithm works with the following inputs: problem statement (**PS**), defined keywords (**K**), (empty) structural causal model (**SCM**), pre-defined tier weights (**W**), and inter-rater agreement function (**IRR**). For each of the three tiers of CKH, we encode specific and relevant causal knowledge within the tier, and systematically update the knowledge base to derive the causal structure.

We start with an empty SCM, with no values assigned to **U**, **V**, **F** and **P(u)** (*introduced and discussed in the supplementary document*). For each individual tier of causal knowledge hierarchy, we follow the steps described below:

1. **Encode tier-specific information:** For each tier, we encode all the information in the causal graph specific to that tier. Specifically, we encode:

   (a) experts' opinion and background knowledge as individual edges, and their confidence score (between 0 and 1.0) in those edges (with directions), in Tier 1 ($\mathbf{CK_E}$),

   (b) edges generated by the causal structure learning algorithms run on the data sets, in Tier 2 ($\mathbf{CK_D}$), and,

   (c) causal relationships extracted from literature as directional edges in Tier 3 ($\mathbf{CK_L}$).

2. **Develop a scoring matrix from encoded information:** From the information encoded for each tier, we build a causal information-based scoring matrix P for the specific tier, with a dimension of $m \times n$. Here m is the number of rows equal to the count of unique pairs of nodes (variables). For q number of total variables ($U \cup V$) in an individual tier, we have $m = \binom{q}{2} = \frac{q(q-1)}{2}$. n is the number of columns in the scoring matrix P. For a specific row in P with the nodes (or variables) A and B, we have four columns signifying the type of causal connection between them: i) $A \rightarrow B$, ii) $A \leftarrow B$, iii) no causal connection between A and B and, iv) no causal information available between A and B. The complete matrix represents the causal knowledge in the specific tier.

3. **Compute individual edge confidence based on agreement from scoring matrix**: Next we calculate individual edge confidence from each row in the scoring matrix P, through plurality voting. For each pair of nodes A, B in P, we iterate through rows i of P and use the equation of edge confidence of causal connections between variables A and B:

$$q_i = [P_i(n)] \text{ for } n = 0, 1, 2 \tag{1}$$

$$e_{A,B} = \frac{max(q_i)}{\sum q_i} \tag{2}$$

Here, $e_{A,B}$ signifies the confidence of the causal connection or on the directional edge between variables A & B, q_i represents the first three values of row i of scoring matrix P with variables A and B.

4. **Estimate agreement score from the scoring matrix:** Using the inter-rater agreement function (IRR), we calculate the agreement score (α) (i.e., Fleiss' kappa, varies from 0 to 1) from the generated scoring matrix P (*explained further under background in supplementary document*).

5. **Compute individual weighted edge confidence:** Next, we calculate the weighted edge strength for all edges within a tier, using the equation:

$$weighted\_e_{A,B} = e_{A,B} \times \alpha \times W_i \tag{3}$$

Here, $weighted\_e_{A,B}$ signifies the weighted edge confidence of the causal connection between variables A & B, α is the agreement score calculated previously, and W_i

represents the weight of the specific tier (W_E / W_D / W_L). Within a specific tier, only $e_{A,B}$ is different for individual edges, whereas α and W_i remains the same. In the best case, where edge weight for a specific edge and agreement score are both 1.0, the weighted edge strength can be the maximum (the weight of the Tier).

6. **Extract tier-specific insights:** From the generated weighted edge confidences $weighted\_e_{A,B}$, we extract tier-specific causal insights and carry them forward to the next tier. Specifically, in Tier 1 ($\mathbf{CK_E}$), we set a predefined confidence threshold to select edge whitelist. Any edge with edge confidence over the threshold is put in the whitelist and used in the structure learning algorithms in Tier 2 ($\mathbf{CK_D}$). Similarly, in Tier 2 ($\mathbf{CK_D}$), we extract an incomplete causal structure skeleton $\langle U, V, F \rangle$ and probability distribution $\langle P(u) \rangle$, to carry over to the next tier. Finally, in Tier 3 ($\mathbf{CK_L}$), we extract the complete SCM $\langle U, V, F, P(u) \rangle$, with encoded information from all tiers of the CKH.

7. **Move extracted insights to next step:** Next we get the output of the specific tier, the extracted insights and use them in the next tier as inputs.

Edge Orientation Step: In this step, we check for any potential cycles between variables, and re-orient them prioritizing the weighted edge strengths, and generate a complete directed acyclic graph. For this task, we implement the edge orientation process from PC algorithm (Scutari et al., 2019). For each triplet of nodes, $A - B - C$, we recursively set edge directions using the two rules: (a) if A is adjacent to B and there is a strictly directed path from A to B, we then replace $A - B$ with $A \to B$ (to avoid introducing cycle), and (b) if A and C are not adjacent and we have $A \to B$ & $B - C$, we replace $B - C$ with $B \to C$ (to avoid introducing new v-structures). The complete directed acyclic graph, along with the computed $\langle U, V, F, P(u) \rangle$ is the resultant SCM for the problem domain. The algorithmic pseudocode is presented in 1.

4. Experimental Results

For our experiments, we rely on *a simulation with a standard causal model.* We use predefined default values for hyper-parameters, and validate our results with the initial ground truth causal model and provide sensitivity analysis of the SCM.

Ground Truth Causal Model We start the simulation with a causal model with a ground truth SCM (Figure 3), (a). For this, we refer to the "clgaussian" dataset from 'bnlearn' library (Scutari, 2021). The dataset has 5000 data-points and is generated from a causal model with one normal (Gaussian) variable, four discrete variables and three conditional Gaussian variables. For validation, we assume this initially defined causal model to be the Ground Truth Directed Acyclic Graph (GTDAG).

Optimization Function Keywords are defined with a complete set of variables $\{A, B, C, D, E, F, G, H\}$. We define the optimization problem *to identify the causal effect and relations between variables D and G,* along with all associated variables $\{A, B, C, E, F\}$ with the best-fitting GTDAG for the SCM. We also hypothesize that the domain is well-explored *(with sufficient experts, data and literature)* and set the values of weights as $W_{T1} = 0.2, W_{T2} = 0.3, W_{T3} = 0.5$, as well as confidence threshold for Tier 1 $threshold = 0.8$.

Algorithm 1: Structural Causal Model Estimation using Causal Knowledge Hierarchy (CKH)

Input: PS, K, $SCM = NULL$, $W = [W_E, W_D, W_L]$, IRR
Output: $SCM_m = \langle U, V, F, P(u), CONF \rangle$
Initialize empty confidence for all edges: $CONF \leftarrow \phi$;
Update modified SCM for output: $SCM_m \leftarrow SCM + CONF$;
/* (a) Tier 1: Causal Knowledge from Experts' opinion (CK_E):   */
Select group of experts: $Exp \leftarrow [exp_1, exp_2, exp_3, ...]$;
while *expert exp_i in Exp* **do**
$\quad \mid \quad CR'_{CK_E}[i] \leftarrow extract\_causal\_relationships(exp_i, PS, K, U \cup V = \phi)$;
end
$U, V \leftarrow get\_vars(CR'_{CK_E})$;
while *expert exp_i in Exp* **do**
$\quad \mid \quad CR''_{CK_E}[i] \leftarrow extract\_causal\_relationships(exp_i, PS, K, U \cup V)$;
end
$U, V \leftarrow get\_vars(CR'_{CK_E} \cup CR''_{CK_E})$;
$P \leftarrow create\_grading\_tuple(CR'_{CK_E} \cup CR''_{CK_E})$;
$F, CONF \leftarrow compute\_weighted\_confidence(W_E, P, IRR, SCM_m)$;
$SCM_m \leftarrow update\_scm(SCM_m, [U, V, F, CONF])$;
/* (b) Tier 2: Causal Knowledge from Data (CK_D):   */
Using $U \cup V$ and K, gather relelvant datasets: $D = [d_1, d_2, d_3, ...]$;
Select different causal structure learning algorithms: $SLA = [sla_1, sla_2, sla_3]$;
while *output model in $SLA \times D$* **do**
$\quad \mid \quad CR'_{CK_D}[i] \leftarrow extract\_causal\_relationships(model)$;
end
$U, V \leftarrow get\_vars(CR'_{CK_D})$;
$P(u) \leftarrow get\_probability\_distribution(D)$;
$P \leftarrow create\_grading\_tuple(CR'_{CK_D})$;
$F, CONF \leftarrow compute\_weighted\_confidence(W_D, P, IRR, SCM_m)$;
$SCM_m \leftarrow update\_scm(SCM_m, [U, V, F, P(u), CONF])$;
/* (c) Tier 3: Causal Knowledge from Literature (CK_L):   */
Using PS and K, gather relevant literature: $L = [l_1, l_2, l_3, ...]$;
$U, V \leftarrow get\_vars(L)$;
while *literature l in L* **do**
$\quad \mid \quad CR'_{CK_L}[i] \leftarrow extract\_causal\_relationships(L)$;
end
$P \leftarrow create\_grading\_tuple(CR'_{CK_L})$;
$F, CONF \leftarrow compute\_weighted\_confidence(W_L, P, IRR, SCM_m)$;
$SCM_m \leftarrow update\_scm(SCM_m, [U, V, F, CONF])$;
/* (d) Edge orientation: */
$SCM_m \leftarrow orient\_edges(SCM_m)$;
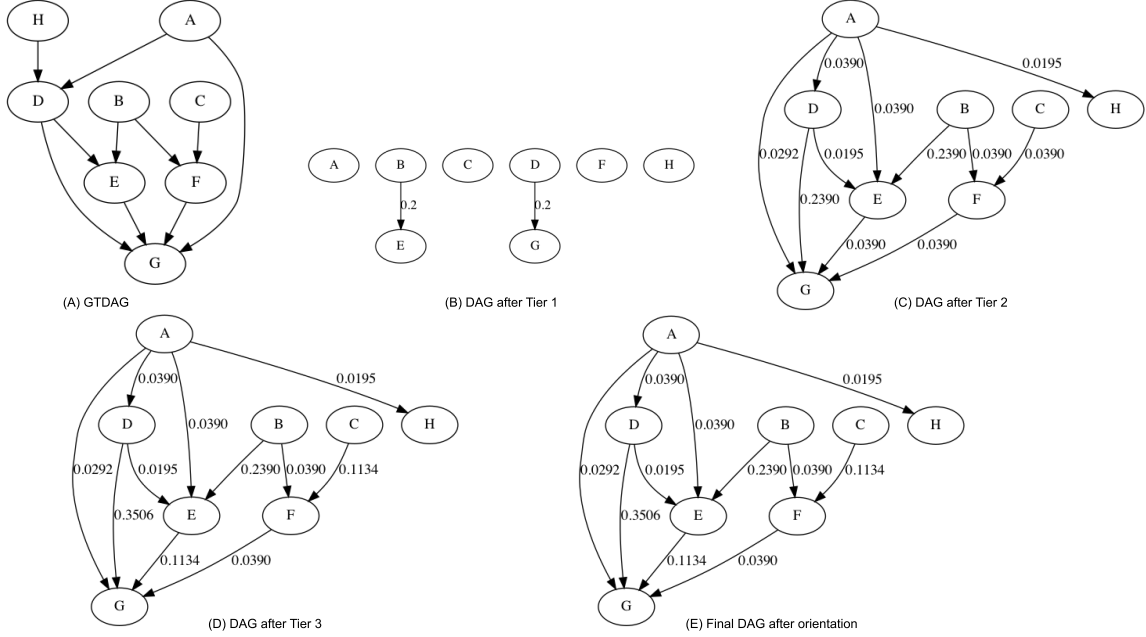**return** SCM_m

---

Figure 3: Ground-truth Causal Model versus Structural Causal Models with edge confidences at individual tiers of CKH

Tier 1: Causal Knowledge from Experts' Opinion For Tier 1, we consider only one expert $Exp = \{exp_1\}$, and use their knowledge, expressed through causal graphs *(although in reality, we might have multiple experts)*. We go through this step in two phases, the first causal graphs are encoded from each of the experts by describing the problem statements (PS) and keywords (K). We integrate all the variables and create a super-set of nodes, $U \cup V$. The second causal graph is derived after providing $U \cup V$, along with both PS and K to the experts. This generates $2x$ causal graphs from experts' opinions based on the problem statement. From these graphs, we generate simplified causal connections:

- $A \rightarrow D$, confidence: 0.5
- $D \rightarrow G$, confidence: 1.0
- $B \rightarrow E$, confidence: 1.0

Based on the pre-defined confidence threshold for Tier 1, we derive a combined causal graph from them. In the combined causal graph in Tier 1, we have $M = < U, V, F, P(u) = \phi, CONF = \phi >$, where U and V comes from the super-set of variables suggested by the experts, and F is defined from the causal relationships suggested by the experts. We calculate the agreement score (using Kappam' fleiss), and measure total confidence values using the agreement score and $W_{T1} = 0.2$. Since we use one expert in this experiment, the agreement score for this tier is 1.0, and we use this value in Equation 3 to compute the weighted edge confidence. We build our scoring matrix P based on the weighted edge

confidence higher than the confidence threshold. The resulting causal graph after Tier 1 is presented in Figure 3, (b).

Tier 2: Causal Knowledge from Data For Tier 2, we generate three separate datasets, $D = \{d_1, d_2, d_3\}$ based on the GTDAG. To simulate a varying number of datasets in real-world, we generate and use three different datasets for which the underlying causal relations are invariant. d_1, is generated from variables $\{A, D, G\}$, d_2, is generated from only variables $\{B, D, E, G, H\}$, and d_3, is generated from all the variables. We use two specific structure learning algorithms, $SLA = \{sla_1, sla_2\}$, with sla_1 being **PC algorithm** and sla_2 being **MMHC algorithm**. Additionally, we use edge whitelists from Tier 1 for structure learning. We show outputs of all SLAs on datasets in the supplementary document, Figure 3.

We update the scoring matrix P for this tier and compute weighted edge confidence. We resolve contradictory edges between Tier 1 and Tier 2 by selecting the highest weighted edge confidence of the two. We sum up the weighted edge confidences and the resulting causal graph after Tier 2 is shown in Figure 3, (c).

Tier 3: Causal Knowledge from Literature In Tier 3, we go through peer-reviewed literature for the problem domain. For this simulation, we assume three causal information sets, $L = \{l1, l2, l3\}$, each of which is extracted from individually published literature. The summary of causal relationships extracted from the literature L is:

1. Literature 1 ($l1$): $D \rightarrow G$, $A \rightarrow D$
2. Literature 2 ($l2$): $F \rightarrow G$, $E \rightarrow G$, $D \rightarrow G$, $C \rightarrow F$
3. Literature 3 ($l3$): $B \rightarrow E$, $C \rightarrow F$, $D \rightarrow E$, $D \rightarrow G$, $E \rightarrow G$

Similar to the previous tiers, we update the scoring matrix P and compute weighted edge confidence as well. We also resolve conflicting edges, *(if any)*, between new causal graph in Tier 3 and original causal graph after Tier 2, depending on edge confidence. Finally, we sum up the weighted edge confidences with previous tiers and the resulting causal graph after Tier 3 is shown in Figure 3, (d). A summary of the edges with edge orientations and combined weighted edge confidence is shown in the supplementary document, Table 3.

Edge Orientation In the last stage of edge orientation, we see whether any cycles were created in the process. In case one is found, we follow edge orientation process *(as described in method section)* and derive the updated SCM. We present the eventual output in Figure 3, (e).

Evaluation For evaluating our proposed method, we compare the output SCM of the algorithm with the GTDAG. Specifically, we compare edges with directions from our proposed method with that of the GTDAG, and report the average accuracy, along with precision, recall and F1 score. An edge-by-edge comparison of generated output DAG with that of GTDAG is considered as a classification problem (Tsamardinos et al., 2006; Scutari et al., 2019). For the simulation, with a node number of $q = 8$, we check for $\binom{n}{2} = 28$ edges' causal directions. On average, our proposed method achieved an average accuracy of **89.29**% (precision: **86.80**%, recall: **89.29**%, f1-score: **87.70**%).

It is possible to incorporate incorrect causal knowledge due to biased opinion, dataset, or publication. For this, we additionally perform sensitivity analysis with perturbed edges within individual tiers. We randomly select three edges and alter their directions. Specifically, we add the following information:

Table 1: Iterations of simulations with incorrect edge directions injected for each tier (one edge incorrect (A1), two edges incorrect (A1+A2), and three edges incorrect (A1+A2+A3)). Corresponding agreement scores and performance metrics are reported.

		Agreement scores			Metrics			
Change in edge		Tier 1	Tier 2	Tier 3	Accuracy	Precision	Recall	f1-score
No alteration		1.0	0.13	0.372	0.8929	0.8680	0.8929	0.8770
Tier 1	A1	1.0	0.13	0.372	0.8571	0.8438	0.8571	0.8427
	A1+A2	1.0	0.13	0.372	0.8214	0.8364	0.8214	0.8244
	A1+A2+A3	1.0	0.13	0.372	0.8214	0.8364	0.8214	0.8244
Tier 2	A1	1.0	0.124	0.28	0.8929	0.8680	0.8929	0.8770
	A1+A2	1.0	0.115	0.28	0.8929	0.8680	0.8929	0.8770
	A1+A2+A3	1.0	0.107	0.28	0.8929	0.8680	0.8929	0.8770
Tier 3	A1	1.0	0.13	0.319	0.8929	0.8680	0.8929	0.8770
	A1+A2	1.0	0.13	0.28	0.8929	0.8680	0.8929	0.8770
	A1+A2+A3	1.0	0.13	0.28	0.8929	0.8680	0.8929	0.8770

Table 2: Alteration in causal edges in the simulation

Id	Change in simulation	Ground truth	Altered information
A1	Add false causal edge	$C \ldots D$	$C \to D$
A2	Alter true causal edge	$E \to G$	$E \leftarrow G$
A3	Remove true causal edge	$B \to F$	$B \ldots F$

We run the simulation initially without any alteration, and then with multiple perturbed edge directions. For each tier, we perturb one edge (A1), two edges (A1+A2), and three edges (A1+A2+A3), and report the general accuracy, precision, recall, and F1-score, along with the change in agreement scores in each case. Table 1 shows the reported outcomes for each simulations.

In general, with gradual perturbation, performance metrics as well as agreement scores decrease, however this decrease is not drastic, due to the weights of tiers of CKH. For each tier, agreement scores decrease, however the decrease in agreement score does not necessarily alter the outcome.

5. Discussion

Identifying causality is a critical part of many analyses, specifically in clinical research where trust in models is low, and safety and efficacy of clinical decisions is essential. In that context, SCM provides the theoretical foundation for identifying causation from large datasets. However, the lack of methodologies to derive an SCM for estimating causal effect

is a fundamental research problem. We have proposed a novel methodology to combine causal knowledge from various sources such as experts' opinion, data, and literature to derive domain-specific accurate SCMs. We discuss the importance of causal information from sources other than just data, and present the rationale behind using hierarchy of causal knowledge. As demonstrated by our experiments, our proposed method (CKH) effectively identifies the most compatible causal models, with higher accuracy and F1-score, from opinions of experts working in the field, outputs of SLAs on existing data, and reported information in peer-reviewed publications. Further discussion is addressed in the appendix.

The CKH-driven algorithm relies on availability and abundance of causal knowledge sources, making it unreliable when there is a lack of causal knowledge from multiple tiers of sources (e.g., CKH-generated causal model from only data will have low confidence compared to that generated from all three sources). Similar to other open research problems in causal inference, it cannot verify a ground-truth SCM. An alternate to CKH is to not make strong inferences about causal directions to build a DAG and rather derive a Markov equivalence class. Within individual tiers under CKH for a specific problem domain, there is a challenge in finding and selecting experts. Similar difficulty arises within Tier 3 (literature, CK_L), since extraction of causal knowledge from literature is itself a research problem under Natural Language Processing (NLP) (Wood-Doughty et al., 2018; Keith et al., 2020) and is currently being tackled by NLP researchers.

Acknowledgments

References

B Ackley, G Ladwig, BA Swan, S Tucker, et al. Evidence based nursing care guidelines. *Medical Surgical Interventions. Mosby Elsevier, syf*, 15, 2008.

Anthony K Akobeng. Principles of evidence based medicine. *Archives of disease in childhood*, 90(8):837–840, 2005.

Bryan Andrews, Peter Spirtes, and Gregory F Cooper. On the completeness of causal discovery in the presence of latent confounding with tiered background knowledge. In *International Conference on Artificial Intelligence and Statistics*, pages 4002–4011. PMLR, 2020.

Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.

Elias Bareinboim, JD Correa, Duligur Ibeling, and Thomas Icard. On pearl's hierarchy and the foundations of causal inference. *ACM Special Volume in Honor of Judea Pearl (provisional title)*, 2020.

M Bikak, R Adib, W Ingram, P Griffin, and M Adibuzzaman. Outcomes of use of antipsychotic for delirium in the icu: A big data approach. In *D50. CRITICAL CARE: THE METAMORPHOSIS-PAIN, SEDATION, DELIRIUM, ICU-ACQUIRED WEAKNESS, AND PALLIATIVE CARE*, pages A6672–A6672. American Thoracic Society, 2019.

Giorgos Borboudakis and Ioannis Tsamardinos. Incorporating causal prior knowledge as path-constraints in bayesian networks and maximal ancestral graphs. *arXiv preprint arXiv:1206.6390*, 2012.

Mathias Drton and Marloes H Maathuis. Structure learning in graphical modeling. *Annual Review of Statistics and Its Application*, 4:365–393, 2017.

Marek J Druzdzel and Francisco J Díez. Combining knowledge from different sources in causal probabilistic models. *The Journal of Machine Learning Research*, 4:295–316, 2003.

David Freedman and Paul Humphreys. Are there algorithms that discover causal structure? *Synthese*, 121(1):29–54, 1999.

Markus Gangl. Causal inference in sociological research. *Annual review of sociology*, 36, 2010.

Thomas A Glass, Steven N Goodman, Miguel A Hernán, and Jonathan M Samet. Causal inference in public health. *Annual review of public health*, 34:61–75, 2013.

Ian D Gow, David F Larcker, and Peter C Reiss. Causal inference in accounting research. *Journal of Accounting Research*, 54(2):477–523, 2016.

Christina Heinze-Deml, Marloes H Maathuis, and Nicolai Meinshausen. Causal structure learning. *Annual Review of Statistics and Its Application*, 5:371–391, 2018.

Miguel A Hernán. The c-word: scientific euphemisms do not improve causal inference from observational data. *American journal of public health*, 108(5):616–619, 2018.

Katherine A Keith, David Jensen, and Brendan O'Connor. Text and causal inference: A review of using text to remove confounding from causal estimates. *arXiv preprint arXiv:2005.00649*, 2020.

Csaba P Kovesdy and Kamyar Kalantar-Zadeh. Observational studies versus randomized controlled trials: avenues to causal inference in nephrology. *Advances in chronic kidney disease*, 19(1):11–18, 2012.

Anthony Landreth and Alcino J Silva. The need for research maps to navigate published work and inform experiment planning. *Neuron*, 79(3):411–415, 2013.

Sanghack Lee and Elias Bareinboim. Causal effect identifiability under partial-observability. In *International Conference on Machine Learning*, pages 5692–5701. PMLR, 2020.

Sanghack Lee, Juan D Correa, and Elias Bareinboim. General identifiability with arbitrary surrogate experiments. In *Uncertainty in Artificial Intelligence*, pages 389–398. PMLR, 2020.

Harold A Linstone, Murray Turoff, et al. *The delphi method*. Addison-Wesley Reading, MA, 1975.

Stefan Listl, Hendrik Jürges, and Richard G Watt. Causal inference from observational data. *Community dentistry and oral epidemiology*, 44(5):409–415, 2016.

Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3): 276–282, 2012.

Austin Nichols. Causal inference with observational data. *The Stata Journal*, 7(4):507–541, 2007.

Galia Nordon, Gideon Koren, Varda Shalev, Benny Kimelfeld, Uri Shalit, and Kira Radinsky. Building causal graphs from medical literature and electronic medical records. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1102–1109, 2019.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Judea Pearl. Causal inference. *Causality: Objectives and Assessment*, pages 39–58, 2010.

Judea Pearl. The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62(3):54–60, 2019.

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

Judea Pearl et al. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.

Kenneth J Rothman and Sander Greenland. Causation and causal inference in epidemiology. *American journal of public health*, 95(S1):S144–S150, 2005.

Frank E Saal, Ronald G Downey, and Mary A Lahey. Rating the ratings: Assessing the psychometric quality of rating data. *Psychological bulletin*, 88(2):413, 1980.

Marco Scutari. Clgaussian-test.html, 2021. Last accessed May 2, 2022from https://www.bnlearn.com/documentation/man/clgaussian-test.html.

Marco Scutari, Catharina Elisabeth Graafland, and José Manuel Gutiérrez. Who learns better bayesian network structures: Accuracy and speed of structure learning algorithms. *International Journal of Approximate Reasoning*, 115:235–253, 2019.

Shohei Shimizu. Lingam: Non-gaussian methods for estimating causal structures. *Behaviormetrika*, 41(1):65–98, 2014.

Pater Spirtes, Clark Glymour, Richard Scheines, Stuart Kauffman, Valerio Aimale, and Frank Wimberly. Constructing bayesian network models of gene expression networks from microarray data. 2000.

Peter Spirtes. Introduction to causal inference. *Journal of Machine Learning Research*, 11 (5), 2010.

Peter Spirtes and Kun Zhang. Causal discovery and inference: concepts and recent methodological advances. In *Applied informatics*, volume 3, pages 1–28. SpringerOpen, 2016.

Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning*, 65(1):31–78, 2006.

Wei Wang, Gangqiang Hu, Bo Yuan, Shandong Ye, Chao Chen, Yayun Cui, Xi Zhang, and Liting Qian. Prior-knowledge-driven local causal structure learning and its application on causal discovery between type 2 diabetes and bone mineral density. *IEEE Access*, 8: 108798–108810, 2020.

Zach Wood-Doughty, Ilya Shpitser, and Mark Dredze. Challenges of using text classifiers for causal inference. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2018, page 4586. NIH Public Access, 2018.

Appendix A. Background

Before presenting our proposed methodology we briefly discuss the relevant scientific concepts needed to explain the individual steps as well as the rationale behind them. The objective of our method is to generate an expert augmented **Structural Causal Model (SCM)** for a specific hypothesis, where part of the causal information used to generate the graphical representation comes from the application of **Structure learning algorithms** on datasets. The building blocks of our proposed methodology includes structural causal models (SCM), **Inter-rater agreement functions** for generating aggregated information and a Causal Knowledge Hierarchy inspired from **Hierarchy of Evidence** in evidence-based health research. These concepts are introduced in this section with details.

A.1. Structural Causal Models

Developed on the foundations of probabilistic graphical models, SCMs are graphical representations of the causal relationships between variables, and are used to draw causal inferences. An SCM is often expressed by a causal graph G. Each node V in G represents an observed or unobserved variable, and each directed edge E represents the causal relationships between them.

An SCM M is a 4-tuple $\langle U, V, f, P(u) \rangle$ (Pearl, 2009) where,

- U is a set of background (exogenous) variables that are determined by factors outside of the model,

- V is a set of observable (endogenous) variables that are determined by variables in the model,

- F is a set of functions such that each $f_i \subseteq F$ is a mapping from the respective domains of $U_i \cup PA_i$ to V_i, where $U_i \subseteq U$ and

- $PA_i \subseteq V \setminus V_i$ and the entire set F forms a mapping from U to V, and $P(u)$ is a probability distribution over the exogenous variables.
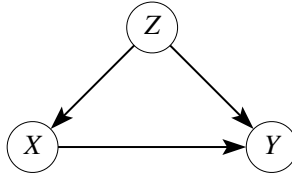
Figure 4: A simple graphical model representing observational study with three variables, with treatment X, outcome Y and confounder Z, expressed using causal directed acyclic graph *(nodes are the variables, edges portray causal relationships between variables)*

To find the causal effect of variable X on variable Y, do-calculus is introduced (Pearl et al., 2016). Do-calculus comes with its own set of strong mathematical tools, such as,

rules of do-calculus, backdoor criterion, that is used to map the observational reality to the corresponding experimental reality with the identifiability equation by adjusting for different kinds of biases (e.g., confounding bias), if it exists.

Figure 4 represents a simple graphical model for an SCM. Although an SCM is used to represent the underlying causal model, in reality, the ground truth causal model in social sciences or medical sciences is never fully known (Freedman and Humphreys, 1999). The causal graph usually represents a set of assumptions explicitly in the problem domain of interest. Given a data set, an SCM can be any of the causal model for a Markov equivalence class (Spirtes, 2010), meaning multiple causal models can be true within a Markov equivalence class for a given data set. Consequently, the validation of a causal model is one of the fundamental challenges in causal inference research. The state of the art generates the most-fitting SCM from datasets using structure learning algorithms *(described in next subsection)* using the properties of conditional probability distributions.

A.2. Structure Learning Algorithms

Other than domain expertise, observational or experimental data can be used to generate a causal graph. Data are the result or snapshot of the underlying causal mechanisms between variables. To recover the causal relationships from data, a rich set of algorithms have been developed over the past thirty years (Spirtes et al., 2000; Tsamardinos et al., 2006; Shimizu, 2014). Causal structured learning is where we try to learn the causal graph or aspects of the causal mechanism. The problem is fundamentally a model selection problem, and these algorithms are called structured learning algorithms (SLA) (Heinze-Deml et al., 2018; Drton and Maathuis, 2017), where a graph is learned or estimated that best describes the dependence structure in a given data set. The learning process includes relying on necessary assumptions (i.e., causal sufficiency, causal faithfulness, linearity), finding conditional dependencies between variables (i.e., Bayes' theorem) and differentiating between different causal structures (i.e., chains, forks, colliders).

Specifically, learning an SCM (or, Bayesian Networks) with a directed acyclic graph (DAG) G and parameters θ from a dataset D with n observations is completed in two steps (Scutari et al., 2019): (1) finding the DAG G which encodes the dependence structure of data D, called structured learning, and (2) estimating the parameters θ, given the obtained G from structured learning, called parameter learning:

$$P(G, \theta|D) = P(G|D) \cdot P(\theta|G, D)$$

Consequently, SLAs are a key component in estimating causal effects within a dataset.

Several algorithms have been proposed in the literature for SLAs (Heinze-Deml et al., 2018), however they differ in their approaches, assumptions, and graphical objects generated. This makes their outcomes varying (even based on the same data source) and difficult to compare.

The main classes of existing SLAs (Heinze-Deml et al., 2018) are:

1. **Constraint-based methods:** Peter-Clark (PC), rankPC , fast causal inference (FCI), and rankFCI

2. **Score-based methods:** greedy equivalence search (GES), rankGES, greedy interventional equivalence search (GIES), and rankGIES

3. **Hybrid methods:** Max-min hill climbing (MMHC)

4. **Structural equation models with additional restrictions:** linear non-Gaussian acyclic models(LINGAM)

An overview of their generated graphical models and assumptions required are summarised in Table 3.

Table 3: Summary of different SLAs and their outputs

	PC	FCI	GES	GIES	MMHC	LINGAM
Causal sufficiency	Yes	*No*	Yes	Yes	Yes	Yes
Causal faithfulness	Yes	Yes	Yes	Yes	Yes	*No*
Acyclicity	Yes	Yes	Yes	Yes	Yes	Yes
Non-gaussian errors	*No*	*No*	*No*	*No*	*No*	Yes
Known do-intervention	*No*	*No*	*No*	Yes	*No*	Yes
Output	CPDAG	PAG	CPDAG	PDAG	DAG	DAG

Since different SLAs can generate different SCMs from the same datasets, there is a need for a principled approach for combining information, which is also correlated with the agreement between them. For this purpose, we leverage inter-rater agreement functions *(described in next section)* to generate an aggregated graphical model that best represents the data along with other sources of causal information (e.g., output SCMs of SLAs, expert opinion or peer-reviewed literature).

A.3. Inter-rater Agreement

Inter-rater agreement (McHugh, 2012) is the degree of agreement among raters, which generates a score on homogeneity, or consensus, in the ratings given by judges or raters. In causal inference we frequently arrive at multiple ratings on causal relationships (by experts' opinion, or from outputs of SLAs), and this is a mechanism to mitigate the discrepancy. To the best of our knowledge, this mechanism has not been previously used in the context of causal graph generation.

Inter-rater agreement function relies on three operational definitions of agreement (Saal et al., 1980):

1. Reliable raters agree with the "official" rating of a performance.
2. Reliable raters agree with each other about the exact ratings to be awarded.
3. Reliable raters agree about which performance is better and which is worse.

In addition, reliable raters are assumed to behave as independent witnesses to the model where they express their independence by disagreeing slightly. In our proposed methodology, we assume the expert opinion, literature, or SLAs are independent raters of causal relationships who capture and express their judgements based on their individual knowledge sources.

Different types of inter-rater agreement functions and scores have been proposed, each with their unique features and strengths. We present a brief overview (McHugh, 2012) of a few of them here.

1. Percent agreement
2. Cohen's kappa coefficient
3. Fleiss kappa (adaptation of Cohen's kappa for 3 or more raters)
4. Joint probability of agreement
5. Pearson r coefficient
6. Krippendorff's alpha (useful when there are multiple raters and multiple possible ratings)

Along with inter-rater agreement function applied on rated causal relationships between variables by raters or algorithms, we propose to incorporate the well established **Hierarchy of Evidence** in evidence-based health research in our methodology.

A.4. Hierarchy of Evidence

A hierarchy of evidence is needed when there are multiple results or inferences from similar scientific studies (sometimes even contradictory) and one has to choose one or combine them. Hierarchy of evidence (or, levels of evidence) is a scoring that quantifies the rank or strength of the results or outcomes from scientific and experimental studies. The hierarchy relies on the study design, validity and applicability to patient care, and quality of data (Ackley et al., 2008). When choosing between multiple findings from experimental studies, hierarchy of evidence is critically important. For example, in healthcare professionals are required to decide on clinical actions based on the best evidence available. One of the most significant reasons behind using a hierarchy of evidence is to upgrade quality of care, by identifying and promoting practice that is effective and by eliminating those who are ineffective or harmful (Akobeng, 2005).

Different hierarchy of evidence have been proposed in the literature, based on design of studies and the endpoints measured. A commonly accepted level of effectiveness rating scheme (Ackley et al., 2008) is presented in Figure 5.
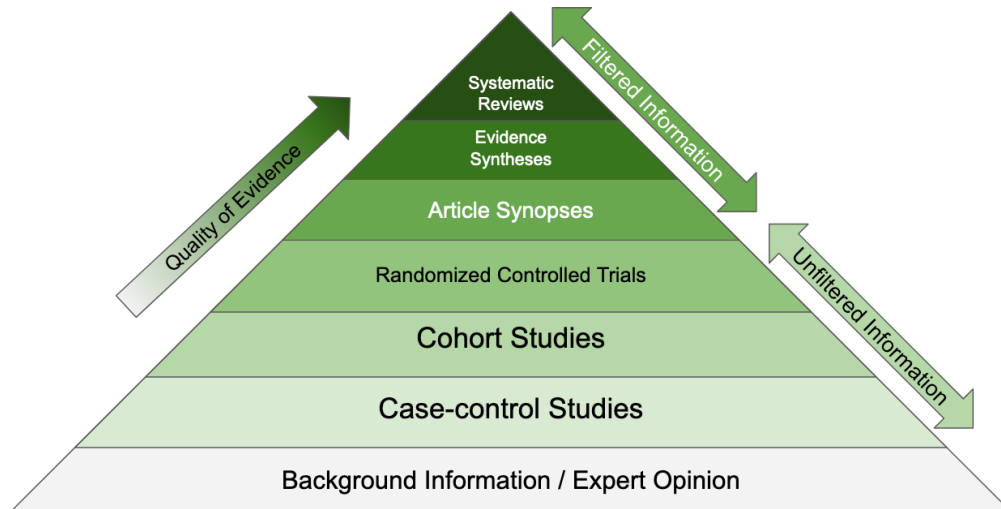

Figure 5: Evidence-Based Medicine (EBM) Resources

Appendix B. Method

The following table discusses a potential dictionary for translating causal knowledge from literature to the syntax of structural theory of causation. The table is not comprehensive and part of an ongoing solution to this research problem.

Table 4: Causal knowledge (CK_L) extraction from peer-reviewed literature

Section in Publication	Information	Translation to Causal Discovery	do-Calculus
Objective / Outcome measures	Objective	Causal Query	$P(y\|do(x))$
Objective / Interventions	Effect of __	Treatment	X
Objective / Outcome measures	Effect on __	Outcome	Y
Methods	Patient demographic	Covariates (potential confounders)	Z
Methods	Study type	Backdoor from confounders to treatment	$Z \rightarrow X$
Methods / Results	Inclusion & Exclusion criteria	Target group	$S = 1$
Results / Conclusion	Findings	causal effect	Value of $P(y\|do(x))$

Appendix C. Algorithm

The following algorithmic pseudocode discusses the process of computing weighted edge confidence for specific edge within a tier. The pseudocode is a part of the complete algorithm discussed in the paper.

Appendix D. Results of Simulation

Here we present each individual structural causal models generated through application of structure learning algorithms on the datasets. We have applied $SLA = \{sla_1, sla_2\}$ on dataset $D = \{d_1, d_2, d_3\}$.

The following table described the edge confidences, weighted edge confidences within each tier of CKH for this simulation.

Algorithm 2: Computation of Weighted Confidence for Individual Tier

---

Input: W, P, IRR, SCM

Output: $F, CONF$

Measure agreement score: $\alpha \leftarrow IRR(P)$;

$CONF \leftarrow \phi$;

for *each edge in SCM* **do**

    Equation 1 and 2 to find edge confidence: C_{edge};

    Equation 3 to find weighted edge confidence: $weighted\_C_{edge}$;

    $CONF.append(weighted\_C_{edge})$;

end

$F \leftarrow extract\_edge\_connections(P)$;
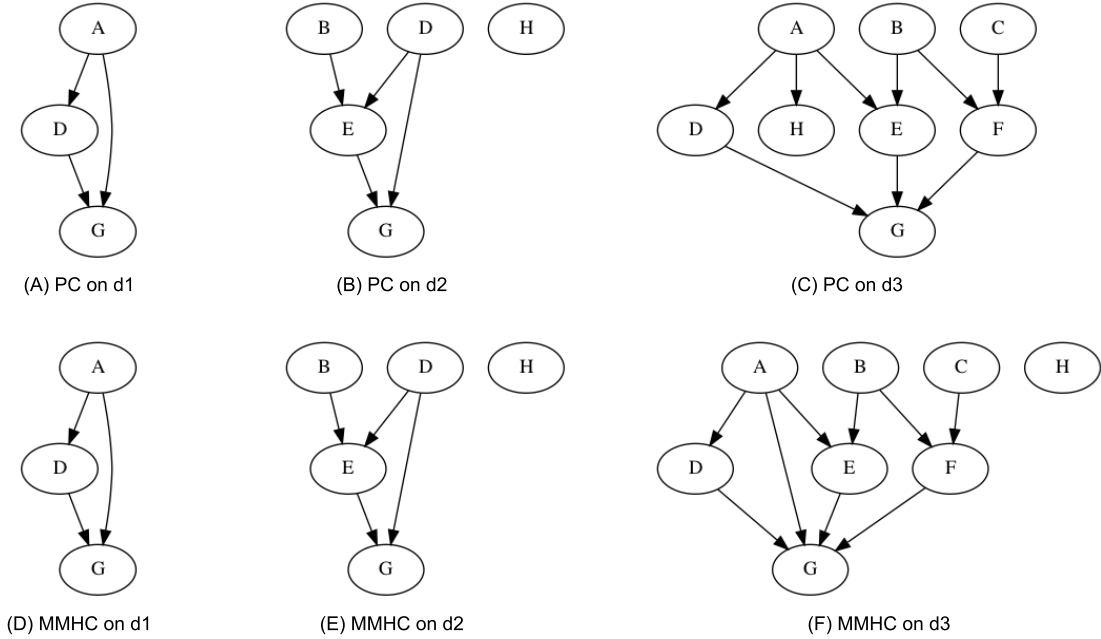
return $F, CONF$

Figure 6: Structural Causal Models as outputs of Tier 2 in CKH

22

Table 5: Summary of simulation results. Each tier has edge directions, edge confidence and weighted edge confidence. A positive value represents direction from first node to second node, and a negative value represents the opposite edge direction. Summarised output results are presented in the last two columns with edge directions and combined weighted edge confidences

Vars	Tier 1 ($W_E = 0.2$)			Tier 2 ($W_D = 0.3$)			Tier 3 ($W_L = 0.5$)			Final Direction	Combined Conf.
	Edge Direction	Edge Conf.	Weighted Edge Conf.	Edge Direction	Edge Conf.	Weighted Edge Conf.	Edge Direction	Edge Conf.	Weighted Edge Conf.		
A, B	$A \ldots B$	0.0	0.0	$A \ldots B$	0.0	0.0	$A \ldots B$	0.0	0.0	$A \ldots B$	0.0
A, C	$A \ldots C$	0.0	0.0	$A \ldots C$	0.0	0.0	$A \ldots C$	0.0	0.0	$A \ldots C$	0.0
A, D	$A \ldots D$	0.0	0.0	$A \rightarrow D$	1.0	0.1116	$A \ldots D$	0.0	0.0	$A \rightarrow D$	0.1116
A, E	$A \ldots E$	0.0	0.0	$A \rightarrow E$	1.0	0.1116	$A \ldots E$	0.0	0.0	$A \rightarrow E$	0.1116
A, F	$A \ldots F$	0.0	0.0	$A \ldots F$	0.0	0.0	$A \ldots F$	0.0	0.0	$A \ldots F$	0.0
A, G	$A \ldots G$	0.0	0.0	$A \rightarrow G$	0.75	0.0837	$A \ldots G$	0.0	0.0	$A \rightarrow G$	0.0837
A, H	$A \ldots H$	0.0	0.0	$A \rightarrow H$	0.5	0.0558	$A \ldots H$	0.0	0.0	$A \rightarrow H$	0.0558
B, C	$B \ldots C$	0.0	0.0	$B \ldots C$	0.0	0.0	$B \ldots C$	0.0	0.0	$B \ldots C$	0.0
B, D	$B \ldots D$	0.0	0.0	$B \ldots D$	0.0	0.0	$B \ldots D$	0.0	0.0	$B \ldots D$	0.0
B, E	$B \rightarrow E$	1.0	0.2	$B \rightarrow E$	1.0	0.1116	$B \ldots E$	0.0	0.0	$B \rightarrow E$	0.3116
B, F	$B \ldots F$	0.0	0.0	$B \rightarrow F$	1.0	0.1116	$B \ldots F$	0.0	0.0	$B \rightarrow F$	0.1116
B, G	$B \ldots G$	0.0	0.0	$B \ldots G$	0.0	0.0	$B \ldots G$	0.0	0.0	$B \ldots G$	0.0
B, H	$B \ldots H$	0.0	0.0	$B \ldots H$	0.0	0.0	$B \ldots H$	0.0	0.0	$B \ldots H$	0.0
C, D	$C \ldots D$	0.0	0.0	$C \ldots D$	0.0	0.0	$C \ldots D$	0.0	0.0	$C \ldots D$	0.0
C, E	$C \ldots E$	0.0	0.0	$C \ldots E$	0.0	0.0	$C \ldots E$	0.0	0.0	$C \ldots E$	0.0
C, F	$C \ldots F$	0.0	0.0	$C \rightarrow F$	1.0	0.1116	$C \rightarrow F$	0.6667	0.0744	$C \rightarrow F$	0.186
C, G	$C \ldots G$	0.0	0.0	$C \ldots G$	0.0	0.0	$C \ldots G$	0.0	0.0	$C \ldots G$	0.0
C, H	$C \ldots H$	0.0	0.0	$C \ldots H$	0.0	0.0	$C \ldots H$	0.0	0.0	$C \ldots H$	0.0
D, E	$D \ldots E$	0.0	0.0	$D \rightarrow E$	0.5	0.0558	$D \ldots E$	0.0	0.0	$D \rightarrow E$	0.0558
D, F	$D \ldots F$	0.0	0.0	$D \ldots F$	0.0	0.0	$D \ldots F$	0.0	0.0	$D \ldots F$	0.0
D, G	$D \rightarrow G$	1.0	0.2	$D \rightarrow G$	1.0	0.1116	$D \rightarrow G$	1.0	0.1116	$D \rightarrow G$	0.4232
D, H	$D \ldots H$	0.0	0.0	$D \ldots H$	0.0	0.0	$D \ldots H$	0.0	0.0	$D \ldots H$	0.0
E, F	$E \ldots F$	0.0	0.0	$E \ldots F$	0.0	0.0	$E \ldots F$	0.0	0.0	$E \ldots F$	0.0
E, G	$E \ldots G$	0.0	0.0	$E \rightarrow G$	1.0	0.1116	$E \rightarrow G$	0.6667	0.0744	$E \rightarrow G$	0.186
E, H	$E \ldots H$	0.0	0.0	$E \ldots H$	0.0	0.0	$E \ldots H$	0.0	0.0	$E \ldots H$	0.0
F, G	$F \ldots G$	0.0	0.0	$F \rightarrow G$	1.0	0.1116	$F \ldots G$	0.0	0.0	$F \rightarrow G$	0.1116
F, H	$F \ldots H$	0.0	0.0	$F \ldots H$	0.0	0.0	$F \ldots H$	0.0	0.0	$F \ldots H$	0.0
G, H	$G \ldots H$	0.0	0.0	$G \ldots H$	0.0	0.0	$G \ldots H$	0.0	0.0	$G \ldots H$	0.0

Appendix E. Extended Discussion

Rationale with the Theory of SCM   Finding the right and the most-compatible causal model with the underlying data generating mechanism is critical for estimating causal effects through rules of do-calculus. This is a challenging research problem since data itself cannot differentiate between SCM within a Markov equivalence class. It has been proven in a recent seminal work (Bareinboim and Pearl, 2016) that additional complementary causal knowledge is needed along with data to derive a SCM and the proposed methodology can be a strong tool to aid in that process. Our proposed methodology uses a systematic approach to incrementally derive the SCM with appropriated scoring for different levels of evidence and can generate high accuracy even in the presence of contradictory causal connections. The algorithm would be beneficial for applied causal inference researchers, specially in epidemiology, medicine and social sciences. Proposed CKH-driven algorithm effectively estimates all necessary components of a computed SCM, U, V, F from all tiers of causal knowledge and $P(u)$ from Tier 2 (data) only. However, it also produces weighted edge confidences $CONF$, which is a key contribution of this algorithm. The edge confidence signifies the strength or confidence of information we have on the specific edge, however it does not state anything about the strength of causal relationships between variables or parameters of functions F.

Completeness of the Derived Causal Model   The causal model estimated through CKH algorithm is built through encoding of causal knowledge sources iteratively and thus holds a summary of causal information from all possible and relevant sources. The rationale for its completeness is that, $\langle U, V, F, P(u) \rangle$, values of four (4) components of the structural causal model derived with the proposed CKH-driven algorithm, is curated from all the tiers of CKH *(and thus contains all necessary and relevant information needed to generate a complete outcome)*. This collaborated information is also weighted appropriately based on the significance and the impact of causal knowledge sources. Another key argument for completeness of the derived causal model comes from the tier weights and their usage. Till now, we have used an increasing weight for the three tiers ($W_E < W_D < W_L$) for a well-researched problem domain, with sufficient experts, data and publications available on the domain specific problem of interest. For an evolving problem domain (e.g., COVID-19 crisis), where we do not have an abundance of well-established peer-reviewed literature, we can alter and adjust the tier weights as fit for the problem at hand. For example, in estimating the effect of a specific old drug in treatment of COVID-19 patients, we would have more weight on Tier 2 (data, CK_D) compared to Tier 3 (literature, CK_L), simply because of lack of strong evidence from literature and might use an alternative variation of tier weights hierarchy ($W_E \leq W_D > W_L$). For such reasons, we conjecture that CKH provides a fundamental, necessary, and sufficient building mechanism for constructing structural causal models for a problem domain, given causal knowledge from a varying sources. A rigorous proof for the completeness is still under investigation.

Application in specific fields of Science   Proposed CKH-driven causal model generation has high impact for specific fields of science, and especially in health science. Identifying the cause for an outcome and quantifying the causal effect is of high importance in health science and epidemiology. An ongoing work is aiming to derive a SCM for the treatment of

24

delirium patients in the ICU (Bikak et al., 2019) based on the CKH. Other than that, CKH has implications in other branches of science, where the notion of causality is critical, such as, sociology and finance.