Automated Meta-Analysis in Medical Research: A Causal Learning Perspective

Lu Cheng[1,2], Dmitriy A. Katz-Rogozhnikov[1], Kush R. Varshney[1], Ioana Baldini[1]
[1] IBM Research – Thomas J. Watson Research Center, Yorktown Heights, NY, USA
[2] Computer Science and Engineering, Arizona State University, AZ, USA
lcheng35@asu.edu,{dkatzrog,krvarshn,ioana}@us.ibm.com

ABSTRACT

Meta-analysis is a systematic approach for understanding a phenomenon by analyzing the results of many previously published experimental studies related to the same treatment and outcome measurement. It is an important tool for medical researchers and clinicians to derive reliable conclusions regarding the overall effect of treatments and interventions (e.g., drugs) on a certain outcome (e.g., the severity of a disease). Unfortunately, conventional meta-analysis involves great human effort, i.e., it is constructed by hand and is extremely time-consuming and labor-intensive, rendering a process that is inefficient in practice and vulnerable to human bias. To overcome these challenges, we work toward automating meta-analysis with a focus on controlling for the potential biases. Automating meta-analysis consists of two major steps: (1) extracting information from scientific publications written in natural language, which is different and noisier than what humans typically extract when conducting a meta-analysis; and (2) modeling meta-analysis, from a novel *causal-inference* perspective, to control for the potential biases and summarize the treatment effect from the outputs of the first step. Since sufficient prior work exists for the first step, this study focuses on the second step. The core contribution of this work is a multiple causal inference algorithm tailored to the potentially noisy and biased information automatically extracted by current natural language processing systems. Empirical evaluations on both synthetic and semi-synthetic data show that the proposed approach for automated meta-analysis yields high-quality performance.

KEYWORDS

Meta-Analysis, Randomized Clinical Trial, Multiple Causal Inference, Latent-Model

ACM Reference Format:
Lu Cheng[1,2], Dmitriy A. Katz-Rogozhnikov[1], Kush R. Varshney[1], Ioana Baldini[1]. 2021. Automated Meta-Analysis in Medical Research: A Causal Learning Perspective. In *ACM CHIL '21 Workshop: ACM Conference on Health, Inference, and Learning, April 08–10, 2021, Online Virtual Event*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Meta-analysis refers to a tool to amplify statistical power by aggregating weighted information from multiple similar studies [28]. It has been increasingly used in various fields of medical research such as clinical medicine, public health, and psychology, in order to understand a variety of pharmaceutical and non-pharmaceutical interventions. The goal of meta-analysis is to reveal common trends regarding the same treatment and outcome measurement by combining a sufficient number of related studies, even if those individual studies contain sources of error [44]. In medical research, the results of many small studies regarding an issue are diverse and conflicting, rendering clinical decision-making rather challenging [11]. In evidence-based medicine, multiple similar randomized clinical trials (RCTs) are gathered and a meta-analysis is performed before treatments enter clinical practice. Such a study could refer to questions such as: What is the therapeutic association between Vitamin C and breast cancer (negative effect, no effect, or positive effect)? or What is the relationship between the risk factor phosphodiesterase type 5 (PDE5) inhibitor and the cardiac morphology? To answer these clinical questions, medical researchers often have to rely on meta-analysis that pools results across independent studies to increase the population size, mitigate experimental bias or inconsistent results from a single clinical study [27].

Despite its importance, meta-analysis has limited coverage and practicality because the process is extremely time-consuming and labor-intensive. In a standard meta-analysis, a person does a comprehensive literature search, identifies relevant studies using carefully designed inclusion/exclusion criteria, manually extracts data from related publications, and conducts statistical analysis. Several undesired consequences can surface when construct meta-analysis by hand. First, many important topics could be left unexplored due to different research interests or limited time and funding to support such tasks. Second, it is rather challenging for researchers to keep up-to-date with the latest RCTs results [27]. Using out-of-date RCTs can lead to unreliable conclusions that eventually influence decision making, e.g., clinical guidelines. Third, manual meta-analysis is prone to different sources of biases. For example, preconceived domain knowledge can influence the inclusion criteria designed by a meta-analyst, and individual studies with statistically non-significant results may never be made public, thereby skewing the meta-analytic samples [3]. The alternative is an automated meta-analysis that can lead to scalable procedures in which human effort is largely reduced and various potential biases are accounted for.

Toward the overarching goal of automating meta-analysis, we propose a framework composed of two major steps: (1) automatic data extraction from scientific publications (i.e., RCTs) related to the same treatment and outcome using existing natural language

ACM CHIL '21 Workshop, April 08–10, 2021, Virtual Event
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

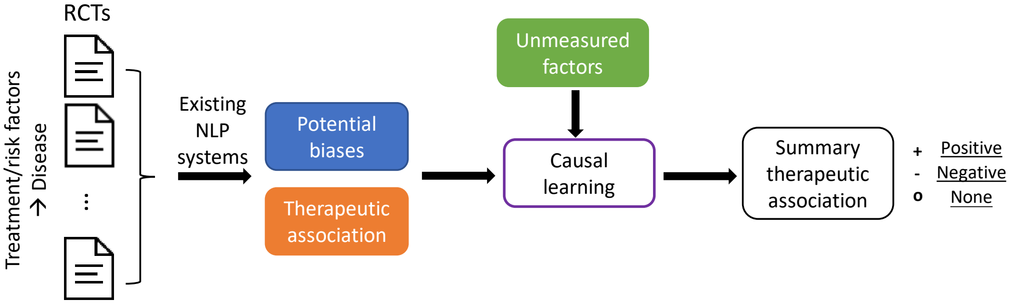


Figure 1: Block diagram of the proposed framework for automating meta-analysis. We first input collected scientific publications into existing NLP systems, which then output risks of bias and therapeutic associations of individual RCTs. A causal inference model is then employed to learn the causal relationship between the risks of bias and the therapeutic association. The resulting model allows us to intervene the risks of bias, i.e., we can set their values to zeros. Here we assume that summary therapeutic association is the predicted therapeutic association with no risks of bias. '+' denotes positive effect, '-' negative effect, and 'o' no effect.

processing (NLP) systems, and (2) automatic inference of the overall therapeutic association or treatment effect with a focus on controlling for the potential biases in the extracted data. The first step has seen recent advances alongside advances in NLP [21, 24, 44]. However, these NLP methods do not extract the *numerical* effect size and variance from the individual studies, which are the standard input to conventional meta-analysis models (e.g., random effect model [2]). Instead, they output a *categorical* therapeutic association (e.g., negative or positive effect) and *categorical* risks of biases such as selective reporting and lack of blinding in the individual RCTs [24] (e.g., high or low risk).

Since conventional meta-analysis models are not applicable to the outputs of these NLP systems, automated meta-analysis requires a new approach for the second step to infer the overall therapeutic association, i.e., *summary therapeutic association* (e.g.,. the overall effect of a drug on a disease), based on the risks of bias and therapeutic association extracted from each individual RCT, with little manual effort. Given that the inference might be further influenced by unmeasured factors, e.g., potential errors from the NLP system and publication bias (many unpublished but relevant work may not be included in the meta-analysis), it is necessary to account for these uncertainties to obtain reliable estimation results for summary therapeutic associations.

To tackle these challenges, in this work, we propose to formulate this task as a causal-inference problem. Since causal inference exploits knowledge of the data generating process, it can uncover the causal relations between variables even in the presence of potential biases. In automated meta-analysis, the risks of bias are *multiple treatments*, the therapeutic associations we observe in different RCTs are the *outcomes*, and the unmeasured factors are *hidden confounders*. Hidden confounders may cause spurious relationships between the treatment and outcome, rendering biased estimation of the summary therapeutic association. From this causal perspective, we seek to understand and uncover the underlying data generating mechanism involving risks of bias, therapeutic association, and unmeasured factors that simultaneously influence the risks of bias and therapeutic association. The direct result is that we can reduce the problem of inferring the summary therapeutic association to questions related to intervention in causal inference — *what the therapeutic association will be if there is no observed risk of biases in the RCTs*. The power of intervention, which ranks higher than correlation, is that it "involves not just seeing what is, but changing what we see" [29]. In automated meta-analysis, this allows us to measure the therapeutic association when we set the risks of bias to zero, i.e., the summary therapeutic association. Fig. 1 depicts the overview of the proposed framework. Our contributions are:

- **Problem.** We study a practical problem of automating meta-analysis to complement conventional meta-analysis often constructed by hand. Despite its importance, automated meta-analysis has not been well-studied in the literature. Here, we pay special attention to the influence of the risks of bias and the hidden confounders on estimating the summary therapeutic association.
- **Algorithm.** We provide an innovative causality perspective to formulate the problem and propose a novel causal learning module for inferring the summary therapeutic association. This is based on therapeutic associations and risks of bias of individual RCTs classified by existing NLP systems.
- **Evaluation.** We evaluate the proposed model on both synthetic and semi-synthetic data. The data generating process of the latter is guided by meta-analyses gathered from real-world clinical studies. We show that the causal perspective can indeed improve the precision of the estimated summary therapeutic association.

2 RELATED WORK

In this section, we briefly review two lines of research that are closely related to our work – meta-analysis in medical research and multiple causal inference.

2.1 Meta-Analysis in Medical Research

The need to make trustworthy clinical decision-making has fostered the momentum toward evidence-based medicine [37]. Typically, an important medical question needs to be studied multiple times, by different research groups in different locations. Central to evidence-based medicine is meta-analysis that aims to integrate results of these conducted studies. There are two core elements of a standard meta-analysis: (1) effect size that reflects the magnitude of the treatment effect of an intervention, and (2) study weight that indicates the importance of the individual clinical trial [2]. One objective of meta-analysis is to determine whether an effect exists; the other objective is to determine whether the effect is positive or negative.

A major challenge of meta-analysis is it is currently constructed by hand, which is extremely time-consuming and labor-intensive. A standard process involves manual screen of related scientific studies, inclusion/exclusion criteria design, manual data extraction, and a statistical analysis of the inconsistencies among these studies. Consequently, there have been a few efforts aimed at envisioning and developing automated meta-analysis [3, 27, 44]. For example, Michelson [27] envisioned an automatic process for creating meta-analyses for any treatment and disease, and keeping them up-to-date automatically. The proposed framework consists of three steps: paper extraction, paper clustering, and standard meta-analysis models. Later, Xiong et al. [44] developed an NLP model seeking to efficiently identify relevant publications. Specifically, it first clustered publications based on common text features using K-means algorithm. Supervised machine learning model was then used to identify clusters of articles most similar to the content of the training set labeled based on a subset of articles from the initial search. It was shown that the machine learning-assisted screening can identify the same set of articles for meta-analysis as those manually screened.

In addition to the efficiency and scalability, another primary limitation of conventional meta-analysis is the potential biases that have long been threats to the validity of meta-analytic results [3]. One such source of bias at the meta-analysis level is the inclusion and exclusion criteria of an RCT. For example, important publications can be left out because researchers are merely interested in those related to their domain expertise. At the level of individual studies, bias can be introduced by the loss of trials and subjects [11], publication bias (studies with significant, positive results are more likely to be published), and poor concealment of treatment allocation or no blinding in studies. All these biases can potentially influence the estimated treatment effects and the summary therapeutic associations [31].

The RobotReviewer of Marshall et al. [24] consists of NLP techniques and pre-trained classifiers to predict the therapeutic association and risks of bias of a single RCT from its written text. Its output includes six types of risks of bias and the therapeutic association. RobotReviewer constitutes the first step before our proposed causal modeling. As the NLP outputs are not standard inputs of the statistical analysis in manual meta-analysis, our contributions herein complement the earlier work. In particular, we explore and demonstrate the inference of the summary therapeutic association *in the presence of potential biases and unmeasured factors from the outputs of NLP techniques such as RobotReviewer* by leveraging a unique multiple causal inference perspective. This enables us to combine an advanced NLP system with causal learning models to accomplish an automated meta-analysis that can directly output a reliable summary therapeutic association.

2.2 Multiple Causal Inference.

Causal inference is central to the goal of a variety of disciplines such as psychology [6], public health [8] and computer science [29]. One particular type of task in causal inference seeks to identify the effects of *multiple* treatment on the outcome in many scientific endeavors. A common approach for causal inference with multiple treatments is the generalized propensity scores (GPS) that extends standard propensity score [36] from a binary treatment to the multiple treatments setting [15, 16]. Different from propensity score – a single value – with binary treatment, conditioning with multiple treatments needs a vector of GPS's, with each element denoting the propensity score w.r.t. a single treatment. Therefore, unbiased effect estimation can only be obtained when all the propensity scores are matched. GPS has been widely used in many causal inference models such as inverse probability of treatment weighted [25], propensity score matching [5, 33], subclassification [34], and imputations [10]. For example, Zanutto et al. [46] extended previous work on evaluating a national anti-drug media campaign into multiple-treatment doses using propensity score subclassification. Genome-wide association studies (GWAS) are typical settings of multiple causal inference, where the multiple treatments are in the form of genetic variations across multiple sites [32, 45]. With ordinal treatments, e.g., doses (low - medium - high), we can compute a scalar balancing score in place of a vector of propensity scores. One such method estimates the treatment assignment as a function of covariates using the proportional odds model [26]. The learned balancing score can then be used to match or subclassify subjects assigned to different levels of an ordinal exposure [22]. Multiple causal inference also requires new kind of assumptions, cf. the discussions of the assumptions for effect identification and estimation by Lechner [20].

The aforementioned approaches, however, assume away the presence of hidden confounding bias, which is often inapplicable in practice. There are many recent efforts aimed at correcting for confounding. For example, one can build a latent variable using the proxies of confounders instead of the confounders themselves [23]. Another common method advocates to learn representations that penalize the differences in the confounder distributions between the treated and control [18, 39]. In multiple causal inference, a variety of methods have been proposed to account for hidden confounders in GWAS [40, 45]. For instance, to estimate kinship matrices between individuals, it was suggested using a subset of the genetic variations, followed by a fitted mixed-model where the kinship provides the covariance for random effect [45]. Factor analysis is another common approach for adjusting for hidden confounding [40]. Tran and Blei [42] described an implicit causal model that leverages neural architectures with an implicit density for GWAS in the presence of hidden confounders. Under the assumptions of shared confounding and of independence given shared confounders, [32] introduced a mutual information based regularizer to balance these two assumptions and also use latent variables to estimate confounders. The recently proposed Deconfounder [43] used the dependencies among multiple treatments as the indirect evidence to find a substitute confounder. It is attractive because of its simplicity and amenability to predictive validation. Our work leverages the advantages of Deconfounder to tackle the challenges in automated

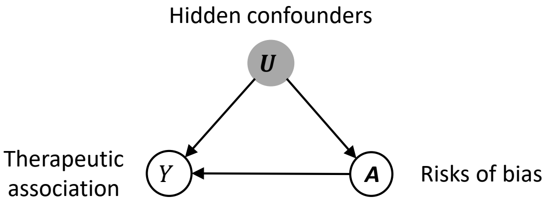


Figure 2: Illustration of multiple causal inference in the presence of hidden confounders for automated meta-analysis. The risks of bias are the multiple treatments, therapeutic association is the outcome. We further assume there is hidden confounding that causes spurious relationships between the treatment and outcome.

meta-analysis, for which we provide unique understandings from the multiple-causal-inference perspective.

3 PROBLEM DESCRIPTION

With the therapeutic association and risks of bias extracted from RCT publications using NLP techniques such as RobotReviewer, what we seek to answer is the question how can one effectively estimate the summary therapeutic association based on the outputs of these NLP systems? The major obstacle is that neither therapeutic association nor risks of bias, both of which are discrete variables, belong to the standard input of conventional meta-analysis models: effect size and variance within each RCT. In addition, the potential biases in RCTs, the uncertainty in NLP systems, and other variables that we cannot measure or even we do not realize their existence further induce undesired factors into the procedure of predicting summary therapeutic association.

These unique challenges in achieving automated meta-analysis motivate us to formulate the problem from a causal inference perspective. In particular, we aim to uncover the underlying causal mechanisms among the risks of bias, therapeutic association, and unmeasured factors. The anchor we leverage here is the rationale that conflicting therapeutic associations we observe across different RCTs are, in part, *caused* by the measured risks of bias. The causal relationship can be further confounded by unmeasured variables, or hidden confounders. For example, analysts' domain expertise can cause them to selectively report the outcome and the summary therapeutic association; uncertainty (e.g., potential errors) in NLP systems can also influence the results of extracted risks of bias and therapeutic association. With *causality-centered* meta-analysis, we answer the question: *What will the therapeutic association be if all risks of bias become zeros?* Or formally, we seek to infer the summary therapeutic association by *intervention*: summary therapeutic association can surface when we intervene on the risks of bias so that their influence on therapeutic association is eliminated.

4 PROPOSED FRAMEWORK

In this section, we introduce the proposed multiple causal inference for automated meta-analysis (MCMA). MCMA first employs existing NLP systems to extract the risks of bias and therapeutic association from collected RCTs regarding the same treatment and outcome measurement. It then builds a multiple causal inference model to uncover the causal relationship between the risks of bias and therapeutic association. The summary therapeutic association across several RCTs can be estimated when the risks of bias are intervened, i.e., set to zeros.

4.1 Automatic Data Extraction

The NLP system we use for extracting therapeutic association and risks of bias assessment is the RobotReviewer pretrained on 12,808 RCTs with ground-truth labels [24]. The RobotReviewer takes a (PDF-formatted) RCT publication as input and labels it with its estimated therapeutic association $Y \in \{0, 1, 2\}$ (negative association, no association, positive association). Another important task it performs is estimating RCT's risks of bias in $D = 6$ domains predefined by the Cochrane risk of bias tool [12]:

- Random sequence generation;
- Allocation concealment;
- Blinding of participants and personnel;
- Blinding of outcome assessment;
- Incomplete outcome data;
- Selective outcome reporting.

For each domain, RobotReviewer determines whether an RCT is at low or high risk of bias and identifies text from the RCT report that supports these bias assessment. What we need for the downstream summary therapeutic association estimation is the predicted therapeutic association Y and a matrix $\mathbf{A} \in \{0, 1\}^D$, where 0 denotes a low risk of bias and 1 denotes a high risk of bias. Note that we recognize the importance of observational studies, however, due to the fact that RobotReviewer is designed for RCTs, our work here focuses on meta-analysis based on RCTs.

4.2 Multiple Causal Inference for Meta-Analysis

Essentially, the summary therapeutic association refers to the therapeutic association when no risks of bias are observed in any of the six domains. Underpinning this is the assumption that the risks of bias are the causes of the inconsistent therapeutic associations across multiple RCTs related to the same treatment and outcome. One may directly employs an off-the-shelf multi-class classifier (e.g., multinomial logistic regression) and predict the summary therapeutic association by setting risks of bias to zeros. However, this method has a number of flaws:

1. *Out-of-Sample Prediction.* In the training data, we do not have access to RCTs where there is no risks of bias. Therefore, directly using trained multi-class classifiers to predict therapeutic association with zero risks of bias can result in unsatisfactory performance.
2. *Unmeasured Factors.* In addition to the risks of bias, other unmeasured factors can also lead to conflicting therapeutic associations we observe in different RCTs, e.g., the medical researcher's intention to assign the treatment to certain type of patients in order to amplify the treatment effect. These factors might also influence how we measure risks of bias, e.g., domain expertise causes the bias of selective outcome reporting.

(3) *Uncertainty in NLP Systems.* NLP systems can make mistakes, especially when they are applied to data collected from specialized domains. For example, drug repurposing in medical research explores the "old" drugs on treating rare and common diseases.

We propose to address these issues by leveraging the power of multiple causal inference. In contrast to multi-class classifiers aimed at prediction, causal learning models seek to explain and understand the underlying data generating process [9, 29]. They have superior robustness or adaptability compared to machine learning models [29]. The goal is now to uncover the causal relationship between risks of bias, i.e., the multiple treatments, and therapeutic association, i.e., the outcome, in the presence of hidden confounders $\mathbf{U}$. Fig. 2 characterizes the causal mechanisms among $\mathbf{A}$, Y and $\mathbf{U}$.

Given $\mathbf{A}$ and Y extracted from a corpus of N RCTs C, we infer the summary therapeutic association with $\boldsymbol{a} = \mathbf{0}$, i.e., $Y(do(\boldsymbol{a} = \mathbf{0}))$, where $do(\cdot)$ represents the intervention [17, 30]. We first infer the mapping function $f(\cdot)$ that characterizes $\mathbb{E}[Y_i(\boldsymbol{a})]$ for $i \in \{1, 2, ..., N\}$ and for each configuration of $\boldsymbol{a} \in \mathcal{A}$, where $\mathcal{A}$ denotes the space of all possible risks of bias:

$$p(Y|do(\boldsymbol{a})) \Leftrightarrow f : \{0,1\}^D \xrightarrow{do(\cdot)} \{0,1,2\} \quad \forall \boldsymbol{a} \in \mathcal{A}. \tag{1}$$

An important notion is that $p(Y|\boldsymbol{a})$ indicates the *correlation* whereas $p(Y|do(\boldsymbol{a}))$ indicates that the change of Y is the result of the change of $\boldsymbol{a}$, i.e., the *causation.*

4.3 Inference Algorithm

We reduce our research question to predicting therapeutic association under intervention in the setting of multiple causal inference. Toward this end, we leverage the Deconfounder [43], a recently proposed multiple causal inference model that combines latent-variable models with predictive model checking to control for hidden confounding. At its core, Deconfounder discovers the substitute confounder $\mathbf{Z}$ from the dependencies among multiple treatments to approximate the hidden confounder $\mathbf{U}$.

Suppose that each RCT is represented by a vector $\boldsymbol{a} = \{a_1, a_2, ..., a_D\}$. A potential outcome function $y_i(\boldsymbol{a}) : \{0,1\}^D \rightarrow \{0,1,2\}$ maps configurations of risks of bias to the outcome therapeutic association for each RCT i. The goal of multiple causal inference is to characterize the sampling distribution of the potential outcome, i.e., $\mathbb{E}[Y_i(\boldsymbol{a})]$. We do not have access to the full distribution of $Y_i(\boldsymbol{a})$ for any $\boldsymbol{a}$ due to the fundamental problem of causal inference [14] that we can only observe one potential outcome (either outcome under treated or control) for each subject. Basic classifiers directly estimate the conditional distribution $\mathbb{E}[Y_i(\boldsymbol{a})|\mathbf{A} = \boldsymbol{a}]$. However, in the presence of hidden confounders, $\mathbb{E}[Y_i(\boldsymbol{a})] \neq \mathbb{E}[Y_i(\boldsymbol{a})|\mathbf{A} = \boldsymbol{a}]$. We first elaborate the required assumptions:

Assumption.

(1) *Stable Unit Treatment Value Assumption (SUTVA) [35, 36]. There is no interference between individuals and there is one single version of each cause.*

(2) *Positivity (Overlap). The substitute confounder* $\mathbf{Z}_i$ *satisfies:*

$$p(A_{ij} \in \mathcal{A}|\mathbf{Z}_i) > 0, \quad p(\mathcal{A}) > 0, \tag{2}$$

where $\mathcal{A}$ *is the set of* $A_{ij}, i = 1, 2..., N, j = 1, 2, ..., D$.

(3) *No unobserved single-cause confounders. Formally,*

$$A_{ij} \perp Y_i(\boldsymbol{a})|\mathbf{X}_i, \quad j = 1, ..., D, \tag{3}$$

where $\mathbf{X}_i$ *is some observed background variable.*

The SUTVA assumption implies that the risks of bias of one RCT do not affect risks of bias and therapeutic association of any other RCT. The positivity assumption indicates that given the substitute confounder, the probability of risks of bias being high should be nonzero in each domain. The last assumption, also referred to as *single-ignorability*, implies that there are no such hidden confounders that exclusively influence one single domain of risk of bias. While this assumption is also untestable in practice, it is weaker than assuming there is no hidden confounder.

Given $\mathbf{A}$ and Y, MCMA consists of three major steps: substitute confounder inference, predictive check of the latent-variable model, and outcome model inference. First, we examine the correlations of all pairs of risks of bias, and remove highly correlated ones to better satisfy the single-ignorability assumption. Next, we define and fit a latent-variable model of the risks of bias: $p(\mathbf{z}, a_1, a_2, ..., a_D)$, where $\mathbf{z} \in \mathbf{Z}$. Specifically, the model is characterized as

$$\begin{aligned} \mathbf{Z}_i &\sim p(\cdot|\alpha) \quad i = 1, ..., N, \\ A_{ij}|\mathbf{Z}_i &\sim p(\cdot|\mathbf{z}_i, \theta_j) \quad j = 1, ..., D, \end{aligned} \tag{4}$$

where α and θ_j parameterize the distribution of $\mathbf{Z}_i$ and the per-cause distribution of A_{ij}, respectively. The latent-variable model for the experimentation is the probabilistic PCA (PPCA) [41]. Other factor models are left to explore in future work.

Theorem. *If the distribution of the observed risks of bias* $p(\boldsymbol{a})$ *can be written as the factor model* $p(\theta_{1:D}, z, \boldsymbol{a})$*, we obtain the following ignorable treatment assignment:*

$$Y_i(\boldsymbol{a}) \perp (\mathbf{A}_{i1}, ..., \mathbf{A}_{iD})|Z_i \quad \forall i \in \{1, 2..., N\}. \tag{5}$$

This theorem is proved by the fact that the substitute confounder Z is inferred without knowledge of the potential outcome Y and the fact that the risks of bias $(\mathbf{A}_{i1}, ..., \mathbf{A}_{iD})$ are jointly independent given Z. The results indicate that Z captures all the dependencies between the therapeutic association and the risks of bias. The identifiability holds if all the required assumptions are satisfied. When the effects are not identifiable, the estimates will present large variance and uncertainty in the potential outcomes can be used to quantify the reliability of the inference algorithm.

In the second step of predictive check, we randomly hold out a subset of the risks of bias for each RCT i, denoted as $\boldsymbol{a}_{i,held}$ and the rest $\boldsymbol{a}_{i,obs}$. We then fit the latent model to $\{\boldsymbol{a}_{i,obs}\}_{i=1}^N$ and perform a predictive check on the held-out dataset. A predictive check compares the given risk of bias with the risk of bias drawn from the model's predictive distribution [43]. If the predictive check score $p \in (0, 1)$ is larger than 0.5, we conclude that the latent model generates values of the held-out causes that give similar log-likelihoods to their real values [43].

In the final step of inferring outcome model, we use the fitted factor model L to infer the substitute confounder for each RCT $p(\mathbf{z}_i|\boldsymbol{a}_i)$, i.e., $\mathbf{z}_i = \mathbb{E}_L[\mathbf{Z}_i|\mathbf{A}_i = \boldsymbol{a}_i]$. We then augment the risks of bias $\boldsymbol{a}_i$ with $\mathbf{z}_i$. The outcome model can be estimated as follows:

$$\mathbb{E}[Y_i(\mathbf{A}_i)|\mathbf{Z}_i = \mathbf{z}_i, \mathbf{A}_i = \boldsymbol{a}_i] = f(\boldsymbol{a}, \mathbf{z}) \tag{6}$$

with augmented data $\{\boldsymbol{a}_i, \mathbf{z}_i, y_i(\boldsymbol{a}_i)\}$ via a multi-class classification model $f(\boldsymbol{a}, \mathbf{z})$, e.g., multinomial logistic regression[1]. At last, we infer the summary therapeutic association by setting $\boldsymbol{a} = \mathbf{0}$ in the fitted outcome model.

5 EMPIRICAL EVALUATION

The fundamental problem in causal inference – we can only measure one potential outcome for the same subject at the same study – makes it almost impossible to validate causal approaches using observational data. In addition, empirical evaluation for automated meta-analysis on real-world data is challenging because: (1) the number of RCTs studying the same treatment/risk factor and disease is limited due to the ethical and financial considerations when conducting RCTs, and (2) we need to follow the standard meta-analysis process, which is extremely time-consuming, to obtain the ground truth for the summary therapeutic association. Consequently, to examine the different aspects of MCMA, here, we follow standard causal inference evaluation method and focus on synthetic data for which we know the ground truth. In addition, we will also present a case study using semi-synthetic data generated based on the statistics collected from a real-world meta-analysis.

In this section, we begin by showing the experimental setup and then discuss the experimental results/findings in detail. The evaluation centers on the following four perspectives:
P1. How does MCMA fare against standard multi-class classifiers with respect to the precision of predicting therapeutic association in individual RCTs?
P2. What is the effect of varying the level of introduced confounding noise on model performance with respect to the precision of predicting the therapeutic association in individual RCTs?
P3. How does MCMA fare against standard multi-class classifiers with respect to the precision of predicting the summary therapeutic association?
P4. Will the answer to *P1* remain the same when applying MCMA to semi-synthetic data generated based on real-world RCTs?

5.1 Experimental Setup

We are not aware of any similar work in the literature to automate meta-analysis with risks of bias and hidden confounders. Inaccessible to the background variables $\mathbf{X}$, conventional multiple causal inference models (e.g., [13]) under Unconfoundedness assumption[2] [35] revert to basic classification models. Thus, given the studied problem inherently relates to multi-class classification, we compare MCMA with widely-used classifiers, including k-Nearest Neighbors (k-NN), Multinomial Logistic Regression (MNLogit), Gaussian Naïve Bayes (Gaussian NB), Multilayer Perceptron (MLP), and XGBoost [4]. Note that for each baseline, there is a corresponding MCMA that uses the same classification model but integrated with the introduced causal knowledge. The key difference is that the input of standard classification model is risk of bias whereas that of MCMA is the augmented data – risk of bias and substitute confounder. We consider three evaluation metrics: AUC and F1 scores for predicting therapeutic association in each RCT, and absolute error of predicting the probabilities of each class being the summary therapeutic association. For the Deconfounder implementation, we used Tensorflow [1] and Statsmodels [38]. The dimension of the substitute confounder $\mathbf{Z}$ is set low (e.g., 1 in our implementation) to keep the estimation variance small [43]. To enforce positivity, it is suggested that the dimension of $\mathbf{Z}$ is set to be smaller than the number of treatments in practice [43]. The latent-variable model PPCA is optimized by Adamax [19] with a learning rate of 0.01. The implementation code will be released in due course.

5.2 Synthetic Experiments

We begin by simulating the causal mechanism in Fig. 2. Specifically, we define $\mathbf{A}$, Y and U as:

$$\begin{aligned} u_i &\sim \text{Unif}(0,1)\\ \boldsymbol{a}_i|u_i &\sim \text{Bern}(0.75u_i + 0.25(1-u_i));\\ y_i &\sim \text{Multinomial}(p_1, p_2, p_3),\\ \text{where } p_1 = \frac{\mathbf{w}_1^\intercal \boldsymbol{a}_i u_i + 4u_i}{l}, p_2 = \frac{\mathbf{w}_2^\intercal \boldsymbol{a}_i}{l}, p_3 &= \frac{\mathbf{w}_3^\intercal \boldsymbol{a}_i + w_u u_i}{l};\\ l = \mathbf{w}_1^\intercal \boldsymbol{a}_i u_i + 4u_i &+ \mathbf{w}_2^\intercal \boldsymbol{a}_i + \mathbf{w}_3^\intercal \boldsymbol{a}_i + w_u u_i,\\ \mathbf{w}_1 \sim \text{Possion}(3), \mathbf{w}_2 \sim &\text{Possion}(2), \mathbf{w}_3 \sim \text{Possion}(1). \end{aligned} \tag{7}$$

Here, $w_u \in \mathbb{R}$ indicates the level of the induced confounding noise through the hidden confounder u_i, which is modeled as a one-dimensional continuous variable following the uniform distribution. We set $w_u = 2$ in *P1* and *P3*, and vary w_u in *P2* to examine its impact on model performance. The number of domains for risks of bias is set to $D = 10$. We also ensure that at least one domain of risks of bias is high. This data generating process explicitly introduces confounding bias between $\mathbf{A}$ and Y as they both hinge on the assignment of "unobserved" u_i. Additionally, we parameterize the distribution of Y in a way that $Y = 0$ is more likely to be the summary therapeutic association. By setting $\boldsymbol{a}$ to $\mathbf{0}$, we obtain the ground truth for the probabilities of each class being the summary therapeutic association, i.e., (p_1^s, p_2^s, p_3^s). All the experimental results are averaged over 10 replications.

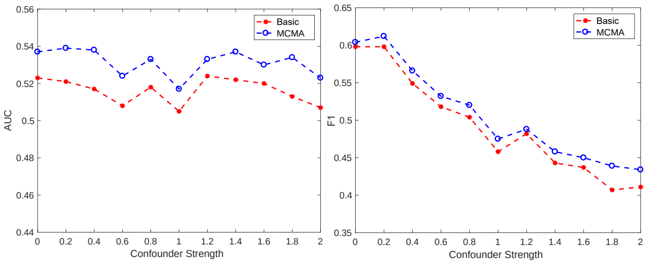
5.2.1 Predicting Therapeutic Association in Each RCT (P1). We first investigate the models' performance on predicting therapeutic associations in different RCTs. These are the therapeutic associations potentially influenced by risks of bias and hidden confounders. We set the sample size $N = 1000$ in this experiment and report mean AUC and F1 scores along with standard deviations in Table 1. We denote the baselines as Basic and ours as MCMA. We observe that the causality-guided multi-class classifiers, i.e., MCMA, can mostly achieve the best performance with respect to both AUC and F1 scores. For example, MCMA improves AUC and F1 scores over standard MNLogit by 31.6% and 56.0%, respectively. Standard deviations are small overall.

5.2.2 Varying the Level of Induced Confounding Noise (P2). We vary w_u from 0 to 2 with an increment of 0.2 to further examine the effect of the induced confounding noise on models' performance in predicting the therapeutic association. Intuitively, as w_u becomes larger, the increased confounding noise will exacerbate the performance of all models. The generated risks of bias and therapeutic

[1] Note that MCMA can also be used for continuous outcomes, e.g., effect size, after we replace $f(\cdot)$ with regression models.
[2] All the confounders can be captured by $\mathbf{X}$, i.e., no hidden confounders.

Table 1: Performance comparisons w.r.t. predicting observed therapeutic associations. Each set of columns presents the results with a standard classifier (Basic) and our method (MCMA).

AUC scores									
MNLogit		k-NN		MLP		Gaussian NB		XGBoost	
Basic	MCMA	Basic	MCMA	Basic	MCMA	Basic	MCMA	Basic	MCMA
.507±.014	**.523±.022**	.519±.018	**.529±.024**	.535±.020	**.539±.019**	.573±.026	**.581±.025**	**.539±.022**	.536±.033
F1 scores									
Basic	MCMA	Basic	MCMA	Basic	MCMA	Basic	MCMA	Basic	MCMA
.411±.050	**.434±.038**	.421±.028	**.433±.039**	.436±.032	**.437±.035**	**.424±.037**	.421±.037	.423±.035	**0.443±0.043**



(a) AUC score.

(b) F1 score.

Figure 3: Performance comparisons using synthetic data and MNLogit with varied levels of confounding noise.

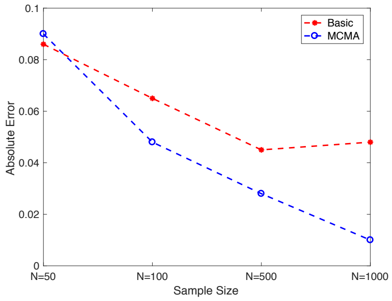


Figure 4: How the performance of basic XGBoost and MCMA changes as N increases.

association are then fed into the Basic MNLogit and the corresponding MCMA. Here, we use MNLogit as a working example; similar results are found using other classifiers. Fig. 3 shows that as expected, the prediction performance of both standard MNLogit and MCMA tend to worsen as we introduce more confounding noise in the data generation process. The trend is especially apparent for F1 score. This is mainly because the synthetic data is simulated to be imbalanced to ensure there are more samples with $Y = 0$. As such, the probability of $Y = 0$ being the summary therapeutic association will be the largest when $\boldsymbol{a} = \mathbf{0}$. Compared to AUC, F1 is more sensitive to the effect of the unevenness among classes. Another important finding is that MCMA consistently outperforms Basic MNLogit when varying the confounder strength w_u.

5.2.3 Predicting the Summary Therapeutic Association (P3). Ultimately, we are interested in inferring the summary therapeutic association between the treatment/risk factor and the disease of interest. With the known data generating process, we can thus compare the estimated probabilities of each class being the summary therapeutic association $\hat{p}_k^s$ ($k \in \{1, 2, 3\}$) with the ground-truth probabilities when $\boldsymbol{a} = \mathbf{0}$. In the real world, the number of available RCTs varies. Thus, we evaluate across sample sizes $N \in \{50, 100, 500, 1000\}$. Absolute error is used to measure the discrepancies, defined as:

$$\text{Absolute Error} = |p_k^s - \hat{p}_k^s| \quad k \in \{1, 2, 3\}. \tag{8}$$

The results can be seen in Fig. 5-8.

We begin by observing that indeed the embedded causal perspective confers an advantage to using the substitute confounder, cf., the results for XGBoost with $N = 500$. Additionally, the best performance w.r.t. each probability prediction is achieved by various models whereas MCMA mostly presents at least two lowest estimation errors among the three classes (p_1^s, p_2^s, p_3^s), e.g., MCMA based on MNLogit, k-NN, and Gaussian NB with $N = 500, 1000$. The improvement is more significant in terms of $Y = 0$ (i.e., the ground-truth summary therapeutic association) when the number of sample size N is larger. To further validate this finding, we also present the trend of models' performance against sample size in Fig. 4 using XGBoost as an example outcome model. These results suggest that (1) it is challenging to simultaneously optimize probabilities prediction for all three classes; (2) the overall performance improvement of MCMA over Basic multi-class classifiers corroborates the effectiveness of the embedded causal mechanisms in MCMA; and (3) in the era of big data, MCMA presents great potentials to accurately predict the summary therapeutic associations.

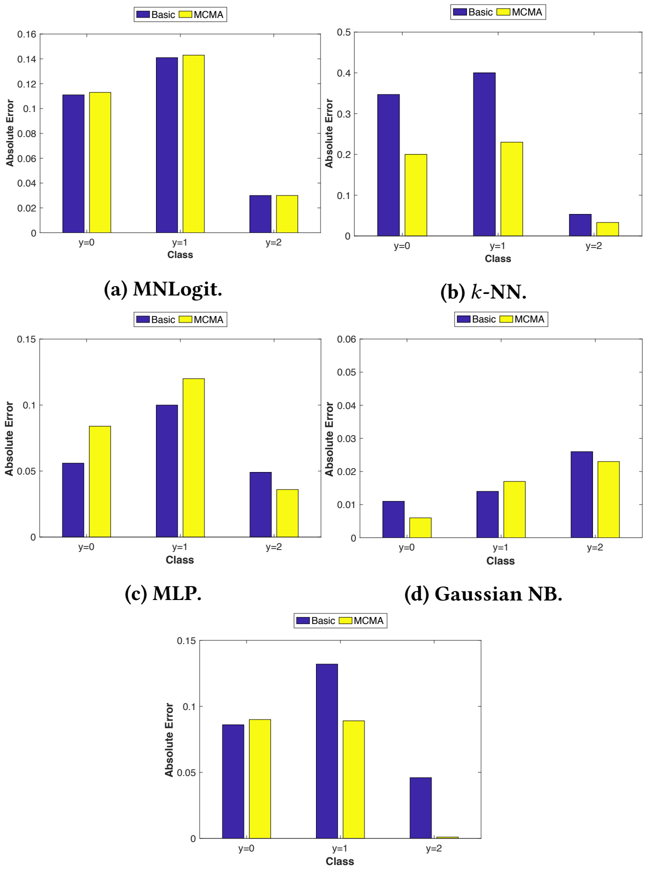
In summary, solutions to *P1-3* based on synthetic experiment lead to the conclusion that compared to standard classification models, the unique perspective of multiple causal inference in the presence of hidden confounders can help uncover the underlying data generating process involving risks of bias, therapeutic association, and unmeasured factors. This improves the precision of predicting both therapeutic association and summary therapeutic association.

5.3 Semi-Synthetic Experiments (P4)

Due to the difficulty of empirical evaluation using real-world data, here, we introduce an alternative using semi-synthetic data generated from real-world meta-analyses. To get the ground truth for the summary therapeutic association, we collect the scientific papers of meta-analysis for specific treatment/risk factor-disease pairs from

Table 2: Performance comparisons w.r.t. predicting therapeutic association between PDE5 inhibitor and the cardiac morphology in different RCTs. Each set of columns presents the results with a standard classifier (Basic) and our method (MCMA).

AUC scores									
MNLogit		*k*-NN		MLP		Gaussian NB		XGBoost	
Basic	MCMA	Basic	MCMA	Basic	MCMA	Basic	MCMA	Basic	MCMA
.500±.032	**.504±.031**	.490±.071	**.491±.050**	.513±.070	**.543±.078**	.512±.052	**.521±.082**	**.532±.046**	.520±.065
F1 scores									
Basic	MCMA	Basic	MCMA	Basic	MCMA	Basic	MCMA	Basic	MCMA
.389±.071	**.402±.100**	.365±.099	**.384±.062**	.420±.078	**.458±.089**	**.410±.070**	.409±.085	.427±.081	**.434±.077**

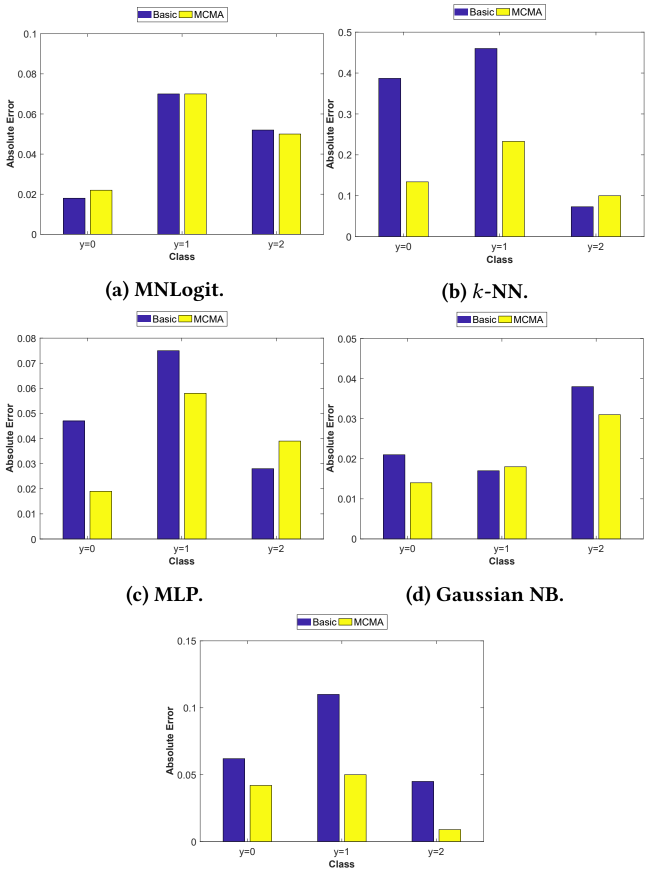


(a) MNLogit. **(b) *k*-NN.**

(c) MLP. **(d) Gaussian NB.**

(e) XGBoost.

Figure 5: Absolute Errors of the estimated probabilities of being the summary therapeutic association with $N = 50$.



(a) MNLogit. **(b) *k*-NN.**

(c) MLP. **(d) Gaussian NB.**

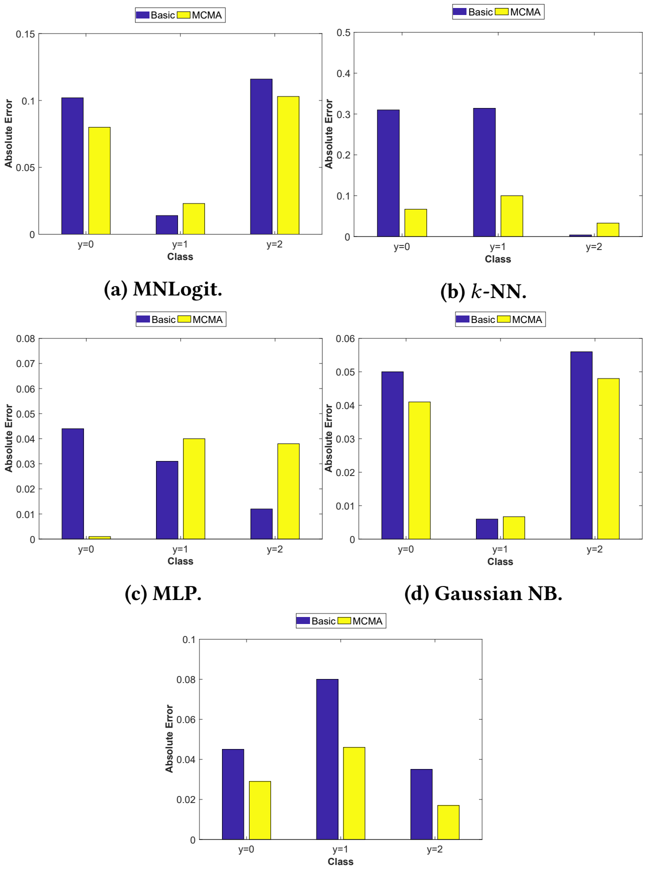
(e) XGBoost.

Figure 6: Absolute Errors of the estimated probabilities of being the summary therapeutic association with $N = 100$.

PubMed and Cochrane Library databases. In doing so, we assume that all the published papers are selected based on strict reviewing processes, the results thereby should be in high quality. The ground truth for summary therapeutic associations is obtained from these meta-analysis papers. The next step identifies and collects the primary studies (i.e., RCTs in PDF format) included in each meta-analysis. The RobotReviewer then takes these primary studies as input and outputs the risks of biases matrix **A** and therapeutic association Y. The last step of MCMA predicts the summary therapeutic association using the fitted outcome model in the Deconfounder.

The number of existing meta-analysis papers consisting exclusively of RCTs is limited.[3] We eventually managed to collect 54 such meta-analysis papers. Preliminary data analysis shows that

[3] Many meta-studies in medical research use both RCTs and observational studies for ethical and financial considerations.

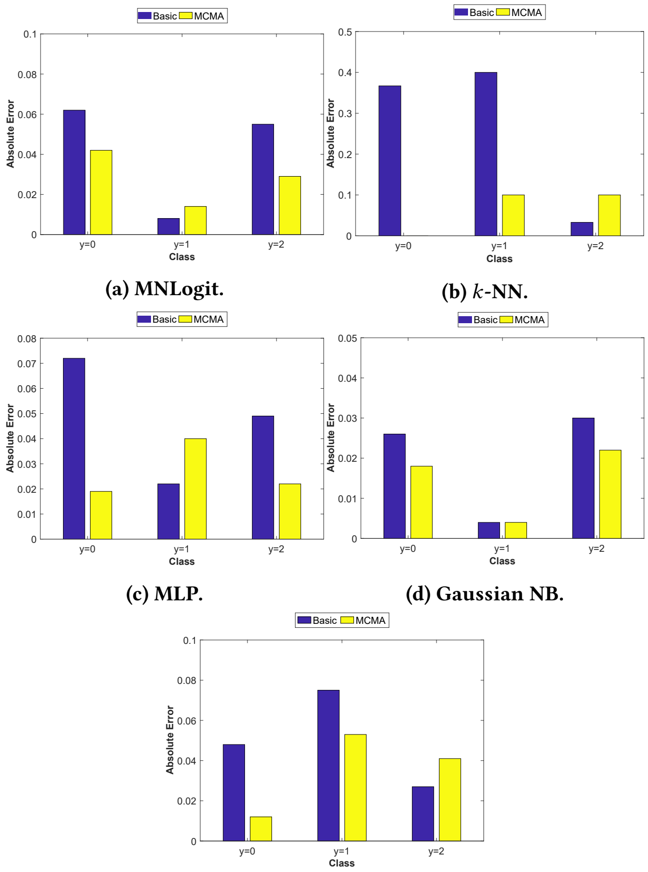


(a) MNLogit. **(b) k-NN.**

(c) MLP. **(d) Gaussian NB.**

(e) XGBoost.

Figure 7: Absolute Errors of the estimated probabilities of being the summary therapeutic association with $N = 500$.



(a) MNLogit. **(b) k-NN.**

(c) MLP. **(d) Gaussian NB.**

(e) XGBoost.

Figure 8: Absolute Errors of the estimated probabilities of being the summary therapeutic association with $N = 1000$.

the average number of RCTs included in each meta-analysis is 13, which is insufficient to train a classification model. To augment the data, we first gather the statistics obtained from these real-world RCTs and then use them to parameterize the simulation process. For example, the meta-analysis of PDE5 inhibitor and cardiac morphology in [7] includes 18 RCTs, based on which the authors suggest positive summary therapeutic association between PDE5 inhibitor and cardiac morphology, i.e., $Y = 2$. The simulation process starts with the risks of bias and therapeutic association of these 18 RCTs extracted from the RobotReviewer. We parameterize the Bernoulli distribution and the Multinomial distribution with the estimates from the real-world data to generate risks of bias $\mathbf{A}$ and therapeutic association Y, respectively. The sample size in this experiment is set to 100.

With unknown probabilities of each category being the summary therapeutic association, we evaluate the models' performance on predicting therapeutic association in the RCTs and the summary therapeutic association. Results in Table 2 suggest findings that are similar to those we get from synthetic data. For instance, MCMA often outperforms the basic classification models on both AUC and F1 scores. All models correctly predict the summary therapeutic association, but the overall performance is unsatisfactory partly because of the limited training sample size.

6 DISCUSSION

The empirical results shown in this work demonstrate the efficacy of causal learning models to process the outputs of NLP-based data extraction and estimate summary therapeutic association in the presence of hidden confounding bias. While there are many other potential biases not captured by existing NLP systems, currently identified biases can provide supporting evidence for existing conjectures and explore new sources of biases for further investigation.

The observation that causality-guided multi-class classifiers often outperform basic classifiers in predicting therapeutic association and the probabilities of each class being summary therapeutic associations suggests that via the lens of multiple causal inference, we have a better understanding of the underlying data generating process involving risks of bias and therapeutic association. In contrast to correlation, causation implies that bias and therapeutic association have a cause-and-effect relationship with one another. Furthermore, controlling for hidden confounders presents an opportunity to alleviate the influence of both the "known unknown"

(known but unmeasured confounders) and "unknown unknown" (unknown confounders) in the process of automating meta-analysis. Incorporating these factors will ensure more reliable and consistent prediction of summary therapeutic association, as suggested by both synthetic and semi-synthetic experiments.

The results here are not without limitations. As with other causal learning models relying on assumptions that may be violated in practice, the single-ignorability assumption in MCMA may limit its utility. There might be hidden confounders that only influence one domain of risk of bias. Furthermore, limited by the functionality of RobotReviewer, the current implementation focuses solely on RCTs, resulting in insufficient data for training classifiers, and leaving numerous observational studies unexplored. While the meta-analysis results in scientific publications are considered high-quality, there are inherent issues of publication and selection biases [3]. Simulation has limitations in general: real-world scenarios are certainly more complex than the presumed data generating process. Moreover, MCMA relies on meta-analysts for literature search and paper relevance identification. Therefore, to achieve the ultimate goal of automating meta-analysis, tools for automatic paper inclusion and paper clustering/classification given the treatment and outcome are still needed.

Nevertheless, this work advances collaborative research efforts in the automation of meta-analysis, which is currently a manual and complicated process. In this study, we provide initial solutions, but much work remains to be done in order to bring meta-analysis to the scale of medical research. With our work, we can potentially discover effective and personalized therapies, improve the timeliness of clinical guidelines and even develop new directions for medical research and other domains of social good. We hope to bring to the forefront concerns and broaden the discussions about the potential research directions of automated meta-analysis.

ACKNOWLEDGEMENTS

This work was conducted under the auspices of the IBM Science for Social Good initiative. The authors thank Iain Marshall and Byron Wallace for help with RobotReviewer and the staff of Reboot Rx for discussions.

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467* (2016).

[2] Michael Borenstein, Larry V Hedges, Julian PT Higgins, and Hannah R Rothstein. 2011. *Introduction to meta-analysis*. John Wiley & Sons.

[3] Evan Carter. 2019. Deep learning for robust meta-analytic estimation. (2019).

[4] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *KDD*. 785–794.

[5] Rajeev H Dehejia and Sadek Wahba. 2002. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and statistics* 84, 1 (2002), 151–161.

[6] E Michael Foster. 2010. Causal inference and developmental psychology. *Developmental psychology* 46, 6 (2010), 1454.

[7] Elisa Giannetta, Tiziana Feola, Daniele Gianfrilli, Riccardo Pofi, Valentina Dall'Armi, Roberto Badagliacca, Federica Barbagallo, Andrea Lenzi, and Andrea M Isidori. 2014. Is chronic inhibition of phosphodiesterase type 5 cardioprotective and safe? A meta-analysis of randomized controlled trials. *BMC medicine* 12, 1 (2014), 185.

[8] Thomas A Glass, Steven N Goodman, Miguel A Hernán, and Jonathan M Samet. 2013. Causal inference in public health. *Annu Rev Public Health* 34 (2013), 61–75.

[9] Ruocheng Guo, Lu Cheng, Jundong Li, P Richard Hahn, and Huan Liu. 2020. A survey of learning causality with data: Problems and methods. *CSUR* 53, 4 (2020), 1–37.

[10] Roee Gutman and Donald B Rubin. 2015. Estimation of causal effects of binary treatments in unconfounded studies. *Statistics in medicine* 34, 26 (2015), 3381–3398.

[11] Anna-Bettina Haidich. 2010. Meta-analysis in medical research. *Hippokratia* 14, Suppl 1 (2010), 29.

[12] Julian PT Higgins, Douglas G Altman, Peter C Gøtzsche, Peter Jüni, David Moher, Andrew D Oxman, Jelena Savović, Kenneth F Schulz, Laura Weeks, and Jonathan AC Sterne. 2011. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *Bmj* 343 (2011), d5928.

[13] Jennifer L Hill. 2011. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* 20, 1 (2011), 217–240.

[14] Paul W Holland. 1986. Statistics and causal inference. *JASA* 81, 396 (1986), 945–960.

[15] Kosuke Imai and David A Van Dyk. 2004. Causal inference with general treatment regimes: Generalizing the propensity score. *JASA* 99, 467 (2004), 854–866.

[16] Guido W Imbens. 2000. The role of the propensity score in estimating dose-response functions. *Biometrika* 87, 3 (2000), 706–710.

[17] Guido W Imbens and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

[18] Fredrik Johansson, Uri Shalit, and David Sontag. 2016. Learning representations for counterfactual inference. In *ICML*. 3020–3029.

[19] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[20] Michael Lechner. 2001. Identification and estimation of causal effects of multiple treatments under the conditional independence assumption. In *Econometric evaluation of labour market policies*. Springer, 43–58.

[21] Eric W Lee, Byron C Wallace, Karla I Galaviz, and Joyce C Ho. 2020. MMiDaS-AE: multi-modal missing data aware stacked autoencoder for biomedical abstract screening. In *ACM CHIL*. 139–150.

[22] Michael J Lopez, Roee Gutman, et al. 2017. Estimation of causal effects with multiple treatments: a review and new ideas. *Statist. Sci.* 32, 3 (2017), 432–454.

[23] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. 2017. Causal effect inference with deep latent-variable models. *NeurIPS* 30 (2017), 6446–6456.

[24] Iain J Marshall, Joël Kuiper, and Byron C Wallace. 2016. RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. *JAMIA* 23, 1 (2016), 193–201.

[25] Daniel F McCaffrey, Beth Ann Griffin, Daniel Almirall, Mary Ellen Slaughter, Rajeev Ramchand, and Lane F Burgette. 2013. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in medicine* 32, 19 (2013), 3388–3414.

[26] Peter McCullagh. 1980. Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)* 42, 2 (1980), 109–127.

[27] Matthew Michelson. 2014. Automating meta-analyses of randomized clinical trials: a first look. In *2014 AAAI Fall Symposium Series*.

[28] David Moher, Alessandro Liberati, Jennifer Tetzlaff, Douglas G Altman, Prisma Group, et al. 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS med* 6, 7 (2009), e1000097.

[29] Judea Pearl. 2019. The seven tools of causal inference, with reflections on machine learning. *Commun. ACM* 62, 3 (2019), 54–60.

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

[31] J Pildal, Asbjørn Hrobjartsson, KJ Jørgensen, Jørgen Hilden, DG Altman, and PC Gøtzsche. 2007. Impact of allocation concealment on conclusions drawn from meta-analyses of randomized trials. *Int J Epidemiol* 36, 4 (2007), 847–857.

[32] Rajesh Ranganath and Adler Perotte. 2018. Multiple causal inference with latent confounding. *arXiv preprint arXiv:1805.08273* (2018).

[33] Jeremy A Rassen, Daniel H Solomon, Robert J Glynn, and Sebastian Schneeweiss. 2011. Simultaneously assessing intended and unintended treatment effects of multiple treatment options: a pragmatic "matrix design". *Pharmacoepidemiology and drug safety* 20, 7 (2011), 675–683.

[34] Paul R Rosenbaum and Donald B Rubin. 1984. Reducing bias in observational studies using subclassification on the propensity score. *JASA* 79, 387 (1984), 516–524.

[35] Donald B Rubin. 1980. Randomization analysis of experimental data: The Fisher randomization test comment. *JASA* 75, 371 (1980), 591–593.

[36] Donald B Rubin. 1990. Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statist. Sci.* 5, 4 (1990), 472–480.

[37] David L Sackett, William MC Rosenberg, JA Muir Gray, R Brian Haynes, and W Scott Richardson. 1996. Evidence based medicine: what it is and what it isn't.

[38] Skipper Seabold and Josef Perktold. 2010. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.

[39] Uri Shalit, Fredrik D Johansson, and David Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *ICML*. PMLR, 3076–3085.

[40] Minsun Song, Wei Hao, and John D Storey. 2015. Testing for genetic associations in arbitrarily structured populations. *Nature genetics* 47, 5 (2015), 550–554.

[41] Michael E Tipping and Christopher M Bishop. 1999. Probabilistic principal component analysis. *Statistical Methodology* 61, 3 (1999), 611–622.

[42] Dustin Tran and David M Blei. 2017. Implicit causal models for genome-wide association studies. *arXiv preprint arXiv:1710.10742* (2017).

[43] Yixin Wang and David M Blei. 2019. The blessings of multiple causes. *JASA* 114, 528 (2019), 1574–1596.

[44] Zhaohan Xiong, Tong Liu, Gary Tse, Mengqi Gong, Patrick A Gladding, Bruce H Smaill, Martin K Stiles, Anne M Gillis, and Jichao Zhao. 2018. A machine learning aided systematic review and meta-analysis of the relative risk of atrial fibrillation in patients with diabetes mellitus. *Frontiers in physiology* 9 (2018), 835.

[45] Jianming Yu, Gael Pressoir, William H Briggs, Irie Vroh Bi, Masanori Yamasaki, John F Doebley, Michael D McMullen, Brandon S Gaut, Dahlia M Nielsen, James B Holland, et al. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature genetics* 38, 2 (2006), 203–208.

[46] Elaine Zanutto, Bo Lu, and Robert Hornik. 2005. Using propensity score subclassification for multiple treatment doses to evaluate a national antidrug media campaign. *JEBS* 30, 1 (2005), 59–73.