SAND2019-14320J

A causal perspective on reliability assessment

Lauren Hund, Benjamin Schroeder

Sandia National Laboratories, 1515 Eubank SE, Albuquerque NM 87123

Abstract

Causality in an engineered system pertains to how a system output changes due to a controlled change or intervention on the system or system environment. Engineered systems designs reflect a causal theory regarding how a system will work, and predicting the reliability of such systems typically requires knowledge of this underlying causal structure. The aim of this work is to introduce causal modeling tools that inform reliability predictions based on biased data sources. We present a novel application of the popular structural causal modeling (SCM) framework to reliability estimation in an engineering application, illustrating how this framework can inform whether reliability is estimable and how to estimate reliability given a set of data and assumptions about the subject matter and data generating mechanism. When data are insufficient for estimation, sensitivity studies based on problem-specific knowledge can inform how much reliability estimates can change due to biases in the data and what information should be collected next to provide the most additional information. We apply the approach to a pedagogical example related to a real, but proprietary, engineering application, considering how two types of biases in data can influence a reliability calculation.

Highlights:

- Understanding reliability often requires causal, rather than predictive, inferences.
- Structural causal modeling is a framework for thinking causally about reliability.
- Sensitivity studies investigating data bias can improve decisions about reliability.

Keywords: Causality, Reliability, Margin, Risk assessment

1. Introduction

Reliability analysis aims to quantify the likelihood that a system will meet its functional requirements over time [24]. Predicting the reliability of engineered systems is difficult, as system testing is often expensive and modeling the performance of engineered systems is a challenge. Modern reliability assessments often rely on ‘reliability assurance,’ defined as combining information from various sources to learn about reliability, instead of traditional ‘reliability demonstration,’ defined as asserting a reliability requirement is met based on statistical modeling of pass/fail or time-to-failure test data [24]. When the failure event is not directly observed frequently enough to demonstrate reliability, reliability assurance activities often supplement test data with auxiliary information, including physics-based modeling and expert judgment [4].

Reliability assurance activities involve describing and testing causal theories about a system, using explanatory models that describe a causal mechanism rather than predictive models that forecast based on observed data alone [37]. As noted in [1], there is a need to establish and to develop methods for inferring causality in risk and reliability assessment. Heuristically, causality in an engineered system pertains to how a system output changes due to a controlled, possibly hypothetical, change or intervention on the system or system environment [6]. Understanding causal relationships allows generalizations to be made in the absence of test data, deemed ‘extrapolative prediction’. As an example, computer simulations of engineering systems are typically causal models, because they predict how outputs change as a function of unobserved inputs through describing the underlying physics. In the absence of strictly physics-based models for predicting reliability, statistical models based on observational data are often used for reliability assessment. As in physics-based modeling, these statistical models must accurately reflect the underlying causal mechanisms to produce accurate reliability predictions.

The aim of this work is to introduce causal modeling tools that can support reliability assurance activities, particularly statistical modeling, when data sources contain bias. In the ideal setting, we can collect data from a large designed physical or computer experiment, using randomization of inputs to infer causal relationships about

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy’s National Nuclear Security Administration under contract DE-NA0003525.

how outputs change across inputs. In practice, data are often ‘observational,’ where the analyst does not have control over how the data are collected. Observational data often contain biases. For instance, the data may not be a representative sample from the target population (selection bias) or spurious associations may exist between inputs and outputs due to omitted variables (confounding bias). With imperfect data, equating predictive models with explanatory models is identical to inferring causation from correlation; causal modeling facilitates moving from correlation to causation and thus produces more generalizable results than prediction alone [8].

To address data insufficiencies, we present a novel application of the popular structural causal modeling (SCM) framework [2, 27] to engineering reliability assessment based on biased data. While others have investigated causality in reliability analysis and quality [6, 10, 20, 23], we are unaware of any previous work that implements the SCM framework, which is a popular framework for causal inference in human health and social science applications [27]. In related work, Broniatowski and Tucker (2017) described high-level notions of validity that can be used to assess data-driven causal claims about engineering systems [6]. The authors emphasize that many methodologies that are successful for prediction (e.g. machine learning) are not effective for drawing causal inferences. Previous work has also considered how to learn causal networks in engineered systems from manufacturing data [20, 23]; methods for learning causal networks depend heavily on having ample, bias-free data from which to learn the network. Deviating from this work, we are more concerned with reliability assurance applications where expert judgment is the primary source of information for building the causal network because data are biased and often sparse, which is a common situation in practice [2, 13, 14, 33]. We are targeting the question of: given a causal network, when can we accurately estimate reliability given the available data? Unlike previous risk assessment literature that only alludes to the need to adjust for biases in data to infer causality, we give concrete steps for implementing methods to address this bias.

The paper is structured as follows. In Section 2, we discuss a motivating engineering example. In Section 3, we provide a high-level review of causality in risk assessment to further motivate the presented methods. In Section 4, we describe how the SCM framework can be applied to reliability assessment. Then, we illustrate how the SCM framework can inform sensitivity studies to assess the implications of assumption violations in reliability estimation (Section 5). The paper concludes with a discussion (Section 6).

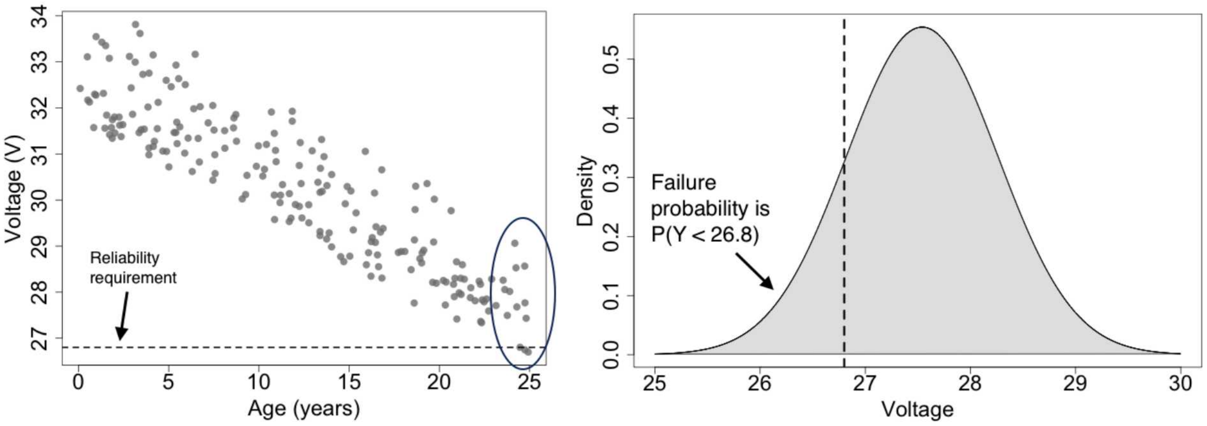
2. Motivating example: Predicting battery performance

This work was motivated by applications that we have encountered concerning the issue of causality from the perspective of design and manufacturing of complex and expensive engineered systems; specifically, in our case, we are concerned with production of nuclear weapon components (which are then integrated into a larger system). A frequent challenge is to demonstrate that components are meeting their requirements with limited test or computational data; the ‘quantifications of margins and uncertainties framework’ (QMU) was developed in the early 2000s to address risk assessment in nuclear weapon applications with limited data and high consequence decision-making [26, 31, 36]. QMU is essentially risk assessment for nuclear weapons, with an emphasis on the credibility of the results for informing decision-making [31].

Most QMU analyses rely heavily on observational data, modeling, and subject matter knowledge, as collecting large amounts of data from designed experiments is often cost prohibitive. In other engineering applications, it is sometimes feasible to conduct structured designs of experiments [25] to answer causal questions about how inputs relate to outputs. Most QMU applications rely on data that are primarily ‘observational,’ where we cannot specifically control all aspects of the design of experiments due to limited resources and/or testing constraints.

Consider the following hypothetical example motivated by a real QMU application: a component must meet performance requirements across a broad range of environments (e.g. temperature, mechanical shock and vibration, etc.) and over a specified product lifetime. However, data are limited; for instance, testing combined environments is infeasible and the amount of hardware for testing is less than the desired level. The question of interest then becomes: with the imperfect data that are available, what can be said about the likelihood of a component meeting its performance requirements?

To illustrate how the SCM framework can be applied to predict reliability using statistical models based on observational data, we use an example motivated by a proprietary engineering application, where the objective is to predict component performance and estimate reliability in an aging system. Specifically, we consider change in voltage of a thermal battery design over time. The batteries must meet a functional voltage requirement (26.8V) throughout their lifespan for different inputs and environments with 98% reliability. The batteries were tested in surveillance for 25 years. In Figure 1, battery voltages are displayed as a function of battery age; from simply plotting the data, the downward trend in voltage over time suggests that the reliability at the current time point, 25 years, may fall below the 98% reliability requirement. In fact, fitting a linear regression to these data and predicting the voltage distribution at 25 years (Figure 1), the best-estimate of reliability at 25 years is 0.850, with 95% confidence interval (.775, .910), suggesting the requirement is no longer met.



(a)

(b)

Figure 1: (a) Battery voltage as a function of age. (b) The estimated prediction distribution for voltage at age 25 based on the linear regression model. The estimated failure probability is the area under the curve to the left of the requirement, 26.8V.

A key challenge is that the data contain certain biases. To achieve the goal of making inference about battery performance at the current time point (25 years after production), biases in the data collection mechanism must be addressed. Specifically, the tested batteries are not a representative sample from the population of fielded batteries; in the surveillance data, load is higher, on average, than what would be expected in normal use conditions. Further, some important measurements were not collected in the surveillance data, namely battery load. Considering the underlying causal model behind battery performance and its relationship to the collected data is required to build an appropriate statistical model for predicting current battery performance. While all data used in this manuscript are simulated, the example is motivated by real but proprietary case.

3. Causality in risk assessment

The most common definition of causality in risk analysis pertains to the effects of hypothetical interventions on a system. Specifically, "causality in risk analysis is ... about how changes, if made, would propagate through systems" [10]. The system can be an environment, e.g. how much would pollution levels change under a proposed regulatory policy; a population of individuals, e.g. how much would disease incidence change given a vaccine campaign in a population; or, as considered in this paper, manufactured goods, e.g. how much would the reliability of a batch of components change given a specific design choice? Notions of causality are typically framed in terms of counterfactuals, which consider what would happen to an outcome in a system under a well-defined *hypothetical* intervention on that system.

Without the ability to actually implement the intervention (as would happen in a design of experiments or clinical trial), estimating counterfactual quantities (and, subsequently, causal effects), requires modeling the system using available data and subject matter expertise and then using this model to make causal inferences. Note that, when we can implement the intervention in the population of interest, we do not necessarily need this system model; rather, causal inferences are drawn using empirical observations from a sample of the population. On the other hand, when the intervention is not actually implemented in the population, then a model of the system is required for inferring causality; stated otherwise, inferences about counterfactual quantities from observatinoal data are conditional on the assumed model, which introduces a layer of approximation into these causal inferences, since, of course, all models are approximations of the true system.

There are many forms of 'causal modeling' in risk assessment and therefore it is worth clarifying what is meant by this phrase herein. We are concerned with in quantitative causal inferences arising from probabilistic models about a system. In typical risk assessment applications, these causal probabilistic models are represented using Bayesian networks [9]. Bayesian networks are graphical models used to probabilistically represent information in an uncertain domain [3]; when arrows between nodes in a Bayesian network represent causal dependencies, then the probabilistic models underlying the Bayesian network can be used to make inferences about causal dependencies in the underlying system [30]. By fully specifying a causal Bayesian network, an analyst can query causal relationships of interest.

For instance, in a probabilistic risk assessment (PRA) analysis, one could ask how a proposed risk management intervention (e.g. a vaccine campaign) would change the ultimate health consequences for a population (e.g. the disease incidence) [9]. First, we must posit an underlying model for this system (represented via a Bayesian network) and then consider the impact of intervening on this system to increase the vaccine rate. A hypothetical Bayesian network for this example is shown in Figure 2. Since we aim to understand how the disease incidence of the flu (Y) would change if we could intervene on the vaccine rate (X_3), we can consider the probability distribution of Y fixing X_3 to the anticipated vaccine rate under a population-level intervention. Other factors also impact the disease incidence and the vaccine rate, such as population demographics (X_1), geographic location (X_2), and infectiousness of the flu strain (X_4). To evaluate the causal effects of intervening on the vaccine rate, the analyst fixes X_3 in the Bayesian network to reflect the intervention, and then propagates uncertainty over the network, *conditioning on the intervention.*

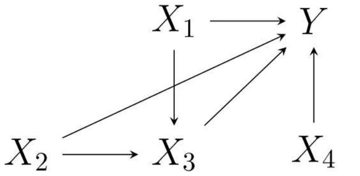


Figure 2: Bayesian network for vaccine intervention example.

In risk assessment and reliability analysis, causal reasoning via Bayesian networks has gained in popularity. However, the practical implementation of Bayesian networks in PRA applications is often inhibited by difficulties associated with specifying the full probability model induced by the Bayesian network; stated more simply, drawing the diagram is often simpler than learning all the relationships between variables in the diagram [19]. In the vaccine example, under the posited model, we would need to predict the disease incidence as a function of the geographic location, population demographics, vaccine rate, and disease infectiousness; accurate prediction would of course require ample data that are relevant to the problem at hand. In practical applications, specifying these relationships is often difficult due to lack of relevant data; that is, either the available data is too sparse or too biased to infer conditional dependencies between nodes in a Bayesian network. It is exactly this problem that we address in this paper by leveraging the structural causal modeling approach [29]. While [9] alludes to the impact of biases in data on assertions of causality and [10] alludes to the potential utility of counterfactual modeling methods (like SCM) for risk assessment, there remains a gap in the literature regarding how to apply these modern counterfactual-based causal inference frameworks (e.g. [29]) to infer causality under biases in data.

4. Applying the SCM framework to reliability applications

In the presence of observational data that contain some biases, SCM provides a framework for defining causality and estimating causal effects. First, causality is predicated on the existence of some causal model of the world (the structural causal model); and second, causal effects are counterfactual quantities, defined as the results of a hypothetical intervention on this causal model. In this section, we review the SCM framework and demonstrate how the framework can be applied to reliability applications. Our review draws from the existing literature; the novelty of the work is applying this framework to improve reliability assessment. The notation and concepts used herein follow [2].

The basic premise of the SCM framework is that there exists some true model, e.g., a system of mathematical equations, that predicts outputs using a set of inputs and relationships between those inputs; however, this true model is often unknown, particularly when systems are too complex to characterize. With enough knowledge about the system (i.e., data and assumptions), we can still predict outputs of interest without directly collecting data on these outputs. More formally, a structural model $\mathcal{M}$ consists of three elements: (1) the model, defined as a set of functions describing the relationship between a set of inputs and outputs; (2) input variables or, in causal terminology, exogenous variables; and (3) output variables whose values are uniquely determined given the input variables and model; in causal terminology, these variables are called endogenous.

In reliability applications, it is useful to think of the underlying structural model as a physics-based system of equations predicting the output of a widget based on design parameters and environments. For example, the voltage of the thermal battery in Section 2 is the output, which could hypothetically be predicted using a model and inputs to the 'battery system,' such as design parameters and environmental exposures. In this work, we consider the case where we do not know the model exactly, but there are some observations of the inputs and outputs (i.e., data), and some knowledge of the system that can be used to learn about the model and subsequently predict a target of estimation. For instance, in the battery example, we have some measurements of voltage (the output) and inputs age and load, and are aiming to predict voltage for unobserved combinations of load and age.

The SCM framework can provide a generalized process for assessing whether the available data are sufficient for estimating reliability and for identifying the set of assumptions under which reliability estimation is feasible. The 'causal inference' question pertains to: is the available data sufficient to "make up for our ignorance" of the true causal model linking the inputs and outputs [2]? The available data may consist of a single dataset or multiple datasets of different types. In engineering applications, 'data' may constitute actual experimental data, computational simulations, or expert judgment. A critical aspect of the SCM framework is integration of engineering judgment and expert knowledge, emphasizing that data alone cannot replace "substantive knowledge in practical decision making and scientific explorations" [2].

In the following sections, we overview the steps of estimating causal effects (Section 4.1); and apply the framework to the motivating example (Section 4.2).

4.1. Steps of structural causal modeling

To implement the SCM framework to estimate causal effects, the following steps are often used:

1. Define the causal effect to be estimated (Section 4.1.1).
2. Identify assumptions required for causal estimation given the available data and knowledge about the problem (Section 4.1.2).
3. Estimate causal effect from the available data under certain modeling assumptions (Section 4.1.3).

In the following sections, we describe each of these steps in more detail.

The following notation is used throughout. The output of the causal model is denoted Y, which is a function of inputs X. Capital letters (X, Y) denote random variables, whereas lower case letters (x, y) denote realizations of random variables. $P()$ denotes a probability density function for a random variable.

4.1.1. Defining the target of estimation

In this section, we consider how to define the target of estimation, reliability, using counterfactual notation used to represent causal quantities. In the SCM framework, the target of estimation Q is called a 'causal query,' reflecting the idea that 'causal' estimands reflect an intervention on a structural model $\mathcal{M}$ [2]. Causal queries are functions of counterfactuals, which are outcomes under hypothetical interventions on a system. Mathematically, counterfactuals on a system model are conditional probability distributions deduced from an intervention on a system model. To denote such an intervention, [2] uses the 'do'-notation. Specifically, the counterfactual $Y|do(X = x)$ is the random variable representing the output Y if we intervene on the structural model to fix input X to x [2, 14, 27]. That is, $P(Y|do(X = x))$ is the distribution of Y if we could fix X to the value of x, keeping everything else the same when propagating uncertainty over the structural model. Causal queries Q typically concern functions of distributions of counterfactuals, e.g., $Q = g\{P(Y|do(X = x))\}$.

In reliability applications, reliability is the most obvious target of estimation Q, where reliability is defined as a quantified measure of uncertainty about a probability of system success [4]; other quantities such as margin or safety factors may also be reasonable targets, but we focus herein on reliability. When reliabilty is defined in terms of meeting a performance requirement, a causal query for reliability can be written as $Q = P(Y < \tau|do(X = x))$, where X is a variable representing specific conditions at which reliability is estimated and τ is a required performance threshold on Y. To define an intervention on X, we could intervene on product tolerances or environmental exposures (e.g. different mechanical or thermal environments, ages, etc.). We consider τ fixed and known for simplicity, though in practice τ could also be a random variable.

4.1.2. Assumptions for causal estimation from observational data

Causal estimation is the act of linking observed data to a causal query, e.g., expressing a counterfactual quantity $P(Y|do(X = x))$ as a function of observable quantities $P(Y|X, S = 1)$. The observable quantity represents the association between Y and X *in the observed data* and is thus empirically estimable from the data alone. The random variable S is a sample selection indicator reflecting how data were sampled for inclusion in the dataset [2, 28]; $S = 1$ denotes data sampled according to the data generating mechanism for the observed data, which may be different from the data generating mechanism for the population of interest. Unlike the observable quantity $P(Y|X, S = 1)$, the counterfactual quantity $P(Y|do(X = x))$ is only estimable with knowledge of how the data were generated to disentangle the causal effect of X on Y from other associations in the data. Biases present in the data, such as selection and confounding bias, will cause these two quantities to differ, often substantially, in practice.

For instance, in the battery example (Section 2), biases in the data preclude direct estimation of the causal query using associations in the observed data. The counterfactual of interest is the voltage distribution at a specific age, i.e., $P(Y|do(A = a))$. Simply estimating the empirical association between age and voltage in the data ($P(Y|A = a, S = 1)$) is insufficient to estimate the counterfactual, because load is higher in the observed data than in the fielded batteries, which leads to an over-estimation of voltage and under-estimation of reliability in the observed data relative to the causal estimands.

The causal literature includes many different methods for estimating counterfactual quantities from observed data [14]. Herein, we apply the popular back-door adjustment formula [2], because this estimator is conducive to reliability applications and is frequently used in practice. The back-door adjustment formula is:

$$P(Y|do(X=x)) = \sum_z P(Y|X=x, z, S=1)P(Z=z) \tag{1}$$

where the terms on the right hand side of the equation can be estimated from the available data [2]. *The objective is to find a reasonable set of assumptions and set of variables Z such that the equality in Equation 1 holds*, as demonstrated in the following sections.

Estimating Q requires two assumptions: (1) we can identify a set of variables Z to control for biases present in the data and (2) we have enough data or information to accurately estimate the terms in Equation 1, namely $P(Y|X, Z, S=1)$ and $P(Z=z)$. We distinguish between these two types of assumptions using the terminology structural versus functional assumptions, where structural assumptions pertain to biases in the data generating mechanism and functional assumptions (Section 4.1.3) pertain to biases in the statistical model form, assuming the structural assumptions are correct.

First, we discuss how to identify structural assumptions required for causal estimation, i.e. how to determine a sufficient set of variables Z to condition on in Equation 1. To identify these assumptions, we must consider the types of structural biases that are present in the data [2]:

- Confounding bias occurs when variables are omitted in a statistical model that are present in $\mathcal{M}$ and confound the relationship between X and Y.

  Confounding bias is formally defined in terms of conditional independence between variables; specifically, no confounding bias is equivalent to $Y|do(X=x) \perp\!\!\!\perp X|Z$ (where $\perp\!\!\!\perp$ represents independence), i.e. assignment to level $X=x$ in the data does not impact the value of the counterfactual given Z (called 'ignorability' of the assignment mechanism to X). Practically, no confounding implies that the distribution of the counterfactual $Y|do(X=x)$ is the same across different values of X; that is, there are no imbalances in predictors of Y between levels of X that would change the distribution of the counterfactuals across levels of X. In practice, controlling for confounding bias is achieved by conditioning on variables Z that are associated with both X and Y.

  Returning to the battery example, confounding bias could arise if, *due to a tester error*, battery load increased over time in battery tests, but load was not measured during testing. In this case, the random variable representing voltage fixing age to a certain value (a counterfactual quantity that is not observed) is correlated with age, because this counterfactual is depends on load and the distribution of load changes with age. Stated more simply, load confounds the relationship between age and voltage and will bias the reliability estimate.

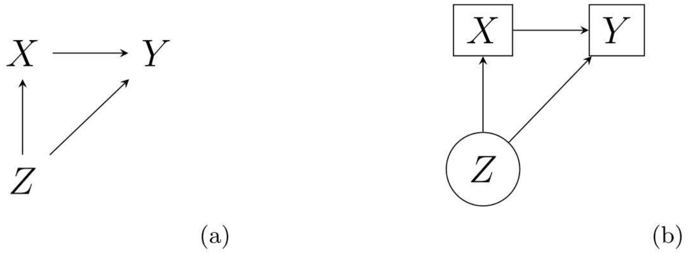
- Selection bias occurs when the available data are not representative of the target population.

  Selection bias arises due to non-random sampling of observations in the data; specifically, the probability of inclusion in the dataset varies as a function of some variables Z that are associated with the outcome Y. Without appropriately controlling for selection variables Z, the distribution of $Y|do(X=x)$ will be specific to the selection mechanism S. By adjusting for Z, 'ignorability' of the selection mechanism (defined as $Y \perp\!\!\!\perp S|X,Z$) can be achieved [2]. (Note: alternative methods for adjusting for selection bias are beyond the scope of this paper; see [2] for details.)

  Returning to the battery example, if battery load increases over time in the battery tests due to a tester error, then the tests do not represent a 'random sample' of battery loads. Rather, load will be higher, on average, in the battery tests than in the use-conditions of interest (that is, selection into the dataset is associate with battery load). Therefore, selection bias on load will bias reliability estimation.

A set of rules has been established for determining what set of variables Z is sufficient to control for selection and confounding bias. Causal diagrams are a graphical tool for identifying these variables Z by visually representing the causal structure of $\mathcal{M}$. That is, causal diagrams inform the identifiability of a query Q by determining whether the available data $\mathcal{D}$ are sufficient for statistical modeling given the true model $\mathcal{M}$. To make this determination, the analyst encodes in a causal diagram a set of assumptions about how variables in $\mathcal{M}$ are related and how the data were generated. Expert knowledge is typically used to build the causal diagram, which is simply "a graphical tool to represent our qualitative expert knowledge and a priori assumptions about the causal structure of interest" [14].

Causal diagrams are a type of Bayesian network, which (as noted in Section 3) is a commonly used tool for graphically representing models in risk assessment [5, 18, 38, 40] as well as computational simulation modeling [21, 22, 35]. Causal diagrams are directed acyclic graphs (DAGs), defined as graphs with directional arrows between nodes (directed) and with no cycles (loops); in causal diagrams, arrows between nodes represent potential *causal* dependencies between variables in the structural model $\mathcal{M}$. An arrow connecting X and Y states that we cannot rule out a causal relationship between X and Y; in counterfactual language, intervening on X could change the outcome for Y. An example of a simple DAG is shown in Figure 3a. An outcome Y is a function of variables X and Z. Rectangles and circles can be used to depict whether variables in a DAG are observed or unobserved, respectively. For instance, in Figure 3b, the output Y and variable X are observed, while Z is unobserved.



(a)

(b)

Figure 3: DAG concepts: (a) Simple DAG and (b) DAG with observed status.

Explicit rules have been derived to determine identifiability of a query from a causal diagram, using the rules of 'd-separation'; a detailed review of all rules for determining the set of variables Z to condition on using causal diagrams is beyond the scope of this work, and we refer the reader to [14, 27]. To summarize, the rules of 'd-separation' detect potential biases in causal queries due to biases in the data by looking for conditional dependences between variables. Causal estimation is theoretically feasible when two conditional independence assumptions are met: $Y|do(X=x) \perp\!\!\!\perp X|Z$ (no unmeasured confounding) and $Y \perp\!\!\!\perp S|X, Z$ (ignorable selection bias). Hence, to determine the identifiability of a causal estimand, we must check these conditional independencies. The rules of d-separation are used to determine whether two variables in a DAG are independent, conditional on a set of variables Z. Hernan *et. al* (2002) concisely summarize the rules of d-separation as:

> Two variables are d-separated if all paths between them are blocked (otherwise they are d-connected). A path is blocked if and only if it contains a non-collider that has been conditioned, or it contains a collider that has not been conditioned on and has no descendants that have been conditioned on [13].

A collider is a node in the graph where two arrows collide; for instance, in the diagram $A \rightarrow C \leftarrow B$, C is a collider.

To check for confounding bias, we can look for unblocked paths between X and Y in a DAG where the direct causal association between X and Y has been removed. If there is no confounding bias, then no information should flow between X and Y in a graph where the direct causal connection between X and Y has been removed. To check for selection bias, we introduce a node S into the graph that denotes selection in the dataset and include arrows in the graph between S and variables that are related to S. If the selection mechanism is ignorable, then no information should flow between Y and S after conditioning on X and Z, i.e., there are no unblocked paths between Y and S in a DAG.

As an example, in Figure 4a, if Z is unobserved, then there remains a path between X and Y in the causal diagram, even after we remove the direct causal link between X and Y. In this case, we cannot estimate $P(Y|do(X=x))$ because the relationship between Y and X is confounded by the unobserved variable Z. If Z was observed (Figure 4b), Equation 1 could be applied to estimate $P(Y|do(X=x))$.

4.1.3. Estimate causal effects under functional modeling assumptions

After determining how to apply Equation 1 to estimate reliability using causal diagram, we must have enough information available to estimate the terms in Equation 1, $P(Y|X, Z, S=1)$ and $P(Z=z)$, from the data. In the absence of ample data to infer these relationships, we make modeling assumptions about these functional relationships. For instance, we can specify a statistical regression model for $Y|X, Z, S=1$, assuming Y varies linearly with X and Z. Interactions and nonlinearities can be added to the statistical model for additional flexibility. Naturally, the accuracy with which we can specify $Y|X, Z, S=1$ depends on how much data and *a priori* knowledge are available. When data are sparse, inferences will more heavily depend on assumptions based on subject matter knowledge; with ample data, inferences will be more data-driven. Standard statistical rules of thumb and power analyses apply to determining whether enough data are available.

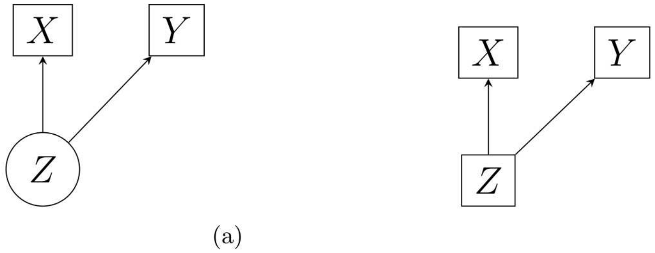


Figure 4: (a) DAGs with direct causal association between X and Y removed. If information can still flow between X and Y, then confounding bias is present; (b) if this path is blocked, then no bias is present.

Further, to apply the adjustment formula in the presence of selection bias, we must specify a probability distribution for Z, $P(Z = z)$; note that, in the absence of selection bias, the empirical distribution of Z in the collected data can be used to marginalize over Z in Equation 1 and there is no need to specify this distribution. In the presence of selection bias, we need auxiliary information about the distribution of Z to apply the adjustment formula; this information can come from expert knowledge or other data sources, but of course cannot be learned from the dataset with selection bias on Z present.

To summarize the process for reliability estimation, we specify a set of variables Z such that, under the data generating mechanism, structural biases are not present after controlling for Z in the analysis via Equation 1 (Section 4.1.2). Then, we determine how to statistically model Y as a function of X and Z (Section 4.1.3). By assembling and assessing this set of assumptions, we have a means to assess the credibility of a data-driven reliability assessment within a causal framework.

4.2. Example: Predicting battery performance

To illustrate how the SCM framework can be applied in practice to explicate the assumptions behind a reliability assessment, we return to the motivating example from Section 2 and consider predicting battery voltage at the current time point, defined as $A = a$. The goal is to use a statistical model to predict the frequency of battery failure due to insufficient voltage, denoted $Q = P(Y < \tau | do(A = a))$. Because we are aiming to illustrate how to apply the SCM concepts, we use a simple statistical model in this example, but note that the concepts can be extended to arbitrarily complicated statistical models. Specifically, to model the observed data, we assume a linear relationship between battery voltage and age:

$$Y|A, S = 1 \quad = \quad \alpha_0 + \alpha_1 A + \epsilon, \epsilon \sim N(0, \sigma) \tag{2}$$

The output voltage Y is a function of the age A, where this function is parameterized by unknown parameters $\alpha = \{\alpha_0, \alpha_1\}$ and random noise $\epsilon \sim N(0, \sigma)$. Both age and voltage are measured in the collected surveillance data.

Equation 2 is the fitted statistical model and can be used to predict functions of Y in the observed data. However, it remains unclear as to whether Equation 2 can be used to estimate the causal query of interest, Q. That is, under what assumptions does it hold that $P(Y|A) = P(Y|do(A = a))$? Recall that there is an additional variable 'battery load', denoted L, which also impacts battery voltage. However, L was not recorded in the data. In this case, a causal diagram could be used to determine *whether* the effect of load must be considered to estimate reliability. The causal diagram is determined by knowledge of the data generating mechanism. For instance, confounding and selection bias could arise if, due to an error in the testing procedure, the input load to the tester L was artificially increased over time, applying a higher load to the battery as time increased (DAG shown in Figure 6b).

Figures 5a-5d show simplified examples of how different biases in the data could impact the identifiability of the causal query.

- In Figure 5a, Q is estimable without knowledge of L. While L was not observed, the causal diagram implies that knowledge of L is not needed, because L is neither a confounder nor a variable associated with selection (in the diagram, this result follows from no unblocked paths between A and Y through L and no selection indicator S present).

- *Unmeasured confounding.* In Figure 5b, confounding bias exists, because L is associated with both Y and A, but is not observed. Hence, information can 'flow' between Y and A in the causal diagram even after removing the direct arrow between A and Y, and the relationship is confounded.

- *Adjustment for confounding.* On the other hand, if L was measured (Figure 5c), conditioning on L controls for confounding bias in the analysis by blocking the flow of information between A and Y in the diagram; hence, per Equation 1, statistical models predicting voltage Y would need to include both A and L (including A alone is not sufficient).

- *Unmeasured selection.* In Figure 5d, selection bias precludes estimation because the terms in Equation 1 cannot be estimated from the data, since L is associated with S but L is not observed. In the causal diagram, information can flow between Y and S and thus the selection bias cannot be ignored in estimation.
- *Adjustment for selection.* On the other hand, if L was measured (Figure 5e), selection bias can be corrected for using the adjustment formula in Equation 1, given auxiliary information about the distribution $P(L = l)$ in the target population. The path between Y and S is now blocked by conditioning on L.

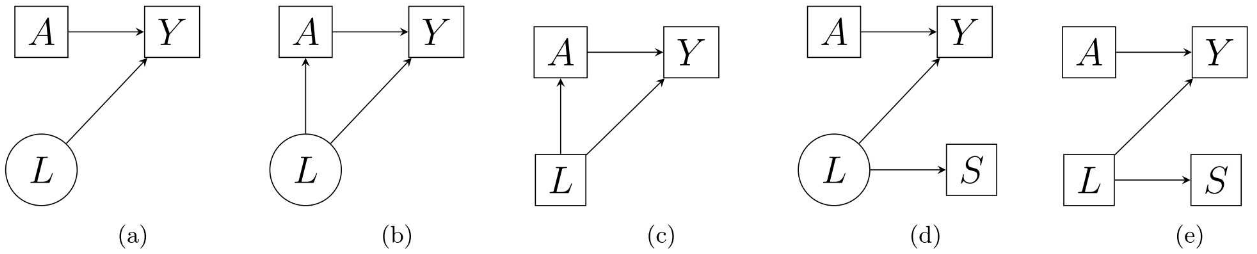


Figure 5: Different SCM identifiability cases: (a) Q estimable, (b) Q not estimable due to confounding, (c) Q estimable conditioned on L, (d) Q not estimable due to selection bias, and (e) Q estimable with auxiliary information.

Hence, whether Q is estimable depends on the true data generating mechanism and the available data. Note the difference between the DAGs in Figure 5 and Figure 6a. In Figure 6a, L is along the causal pathway between A and Y. This type of relationship could happen, for instance, if aging mechanisms caused an increase in load, which then caused an increase in voltage. On the other hand, if the load association is driven by a tester error (e.g Figure 6b), L is not on the causal pathway. If confounding is driven by a tester error, selection bias is also likely to be present. If L is along the causal pathway between A and Y, then we should not control for L in our analysis per the rules of d-separation (Section 4.1.2).

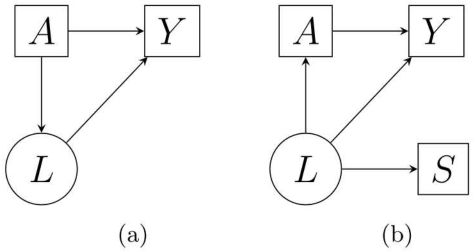


Figure 6: (a) L is on the causal pathway between A and Y. (b) L is not on the causal pathway between A and Y. Note, selection bias is also present in this case, as might be observed with a tester error.

The structural and functional assumptions required to use Equation 2 and the data $\mathcal{D}$ to make inference about Q are presented in Table 1. Structurally, we must assume the DAG in Figure 5a is correct. Under these assumptions, we can equate the counterfactual quantity $P(Y|do(A = 25))$ to $P(Y|A = 25, S = 1)$, which is estimable using Equation 2; stated differently, if our assumptions are correct, $P(Y|A = 25, S = 1)$ is an estimator for $P(Y|do(A = 25))$. Note that, in this example, we assumed the form of the input-output relationship was linear with normally distributed residual variability (Equation 2), which is a reasonable assumption per Figure 1. With sufficient data, we could theoretically learn about this relationship from the data. Without sufficient data, these functional assumptions are required. Determining what constitutes 'sufficient data' is not necessarily straightforward, but statistical model selection and power analyses may help inform whether certain types of effects are detectable with the available data.

In this case, the implications of the functional assumption regarding normality of Y are quite subtle and important to consider. Specifically, we must consider how the residual variability in Y, denoted ϵ, arises. Per the causal diagram, ϵ may contain variability due to production variability, measurement error, *and the impact of* L *on* Y. If the effect of L on Y changes with A, then ϵ will no longer be 'constant variance' over A and the statistical model must be updated to account for this heteroskedasticity. Hence, assuming ϵ has a constant variance also implies no interaction between A and L on Y.

Structural assumptions	Functional assumptions
• No confounding of the age-voltage relationship. • No selection bias in the observed data.	• Linearity: linear relationship between $E(Y)$ and A. • Normality: $Y\|A$ normally distributed with constant variance over A (and no interaction between A and L on Y).

Table 1: Structural and functional assumptions needed to estimate Q using Equation 2.

5. Sensitivity studies

Using sensitivity studies, we consider the implications of violations of assumptions required for causal inference. When assumptions are not supported by evidence, sensitivity studies are commonly used to determine the quantitative impact of violating an assumption on the final results [11, 39]. Such studies typically require eliciting information from experts concerning which assumptions could be wrong, *how* an assumption might be wrong, and the impact of a violation of the assumption. In this section, we explore what information is required to conduct sensitivity studies to examine the credibility of results using the battery data as an exemplar, specifically focusing on how to extend Equation 2 to account for violations of structural assumptions.

Specifically, we consider how the distribution of voltage at the current time point changes under different structural assumptions. Changes in the distribution of voltage will subsequently change the estimated reliability (Figure 1). Because the data are simulated, we know the true reliability at the current time point (99%) and can compare this true reliability to estimates under varying assumptions. The sensitivity studies illustrate how bias-corrected statistical models can improve reliability estimation with good prior information about the data generating mechanism. These sensitivity studies can be constructed for arbitrary types of assumption violations, but we consider two specific examples: selection bias (Section 5.1) and confounding bias (Section 5.2).

5.1. Example: sensitivity study to address selection bias

First, we consider the effect of selection bias on the reliability estimate Q. Selection bias on battery load L could arise as an artifact of the test plan, where the battery was tested at higher loads than would be representative of use conditions (the target population). In the presence of selection bias (5d), we cannot estimate Q using Equation 2. However, we can adjust for the bias with additional assumptions about:

- the true distribution of load in the target population of batteries $P(L = l)$,
- the selection distribution on load in the surveillance data $P(L = l|S = 1)$, and
- the load-voltage association, $P(Y|L, A, S = 1)$.

We first consider the hypothetical case where we have exact knowledge of these additional quantities. To adjust for selection bias, we can apply the adjustment formula in Equation 1, with knowledge about the true distribution of L and the distribution of $Y|A, L, S = 1$:

$$\begin{aligned} L_i &\sim TN(.5, .25, 0, 1) \text{ (true load distribution in target population)} \\ P(Y|do(A = a)) &= \int_l P(Y|A = a, L = l, S = 1)\underbrace{P(L = l)}_{true\,dist.}\,dl \end{aligned} \tag{3}$$

Then, we fit a statistical model to estimate the distribution of $Y|L, A, S = 1$:

$$\begin{aligned} Y_i|A_i, L_i, S = 1 &= \beta_0 + \beta_1 A_i + \beta_2 L_i + \epsilon_i \\ \epsilon_i &\sim N(0, \sigma) \\ L_i|(S = 1) &\sim TN(1, .25, 0, 1) \text{ (true load distribution in the observed data, i.e. the selection distribution)} \\ \beta_2 &= -5 \text{ (true load-voltage association)} \end{aligned} \tag{4}$$

where TN(μ, σ, *min*, *max*) denotes the truncated normal distribution with mean μ, standard deviation σ, and bounds *min* to *max*. Lines 3-4 in Equation 4 are the additional assumptions used to address selection bias. Note that, in line 1, load is included in the statistical model though it is not explicitly measured in the data. Rather, load is treated as missing data that is estimated in the statistical model under the assumptions. The error term ϵ now has a completely different interpretation from Equation 2 and no longer represents variability due to load, only uncertainty due to production variability and measurement uncertainty.

To illustrate this approach, we implement this model in a simulated dataset ($n = 200$), where the data were simulated from the model in Equation 4, with data collected uniformly over time from 0 to 25 years. We fit the model in Equation 4 using Bayesian inference [12] with improper flat priors on the unknown parameters β_0, β_1 and σ. Then, we compare the estimate of $P(Y|do(A = 25))$ under Equation 3 to a 'naive estimate' under Equation 2 which does not address selection bias. All statistical models were fit in using the R Statistical Software package [32], interfaced with STAN for Bayesian inference [7]. Code for duplicating the sensitivity study examples is provided in the online supplementary material.

The adjusted estimator for $P(Y|do(A = 25))$ aligns well with the truth in this simulated example (Figure 7a). This analysis is clearly overly optimistic, since the selection mechanism and load-voltage association are typically not known exactly. However, considering this analysis suggests an approach to a sensitivity study.

- Specify probability distributions reflecting uncertainty about the true load distributions, selection distributions, and load-voltage associations. That is, Equations 3 and 4 can be extended to add 'prior distributions' on the sensitivity study parameters.
- Fit the statistical model with the additional prior distributions on the uncertain sensitivity parameters.
- Compare inferences on Q from the original statistical model and the sensitivity study.

Because the data typically will not contain information about structural biases in the data, the end results will be sensitive to selection of the prior distributions on the sensitivity study parameters.

For instance, we can extend Equation 4 to allow for uncertainty in the load-voltage association and in the mean of the selection distribution (lines 3-5 below):

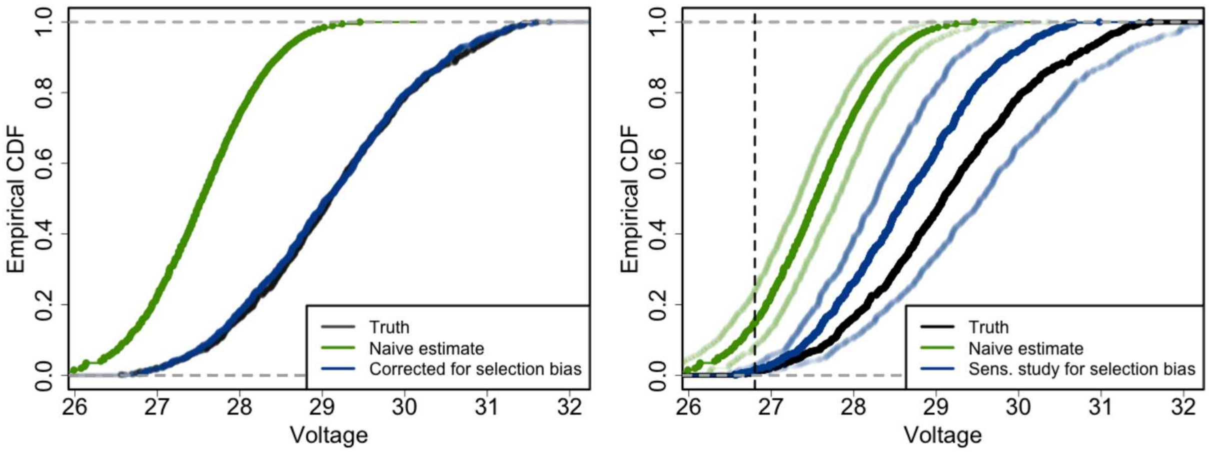
$$\begin{aligned} Y_i|A_i, L_i, S = 1 &= \beta_0 + \beta_1 A_i + \beta_2 L_i + \epsilon_i \\ \epsilon_i &\sim N(0, \sigma) \\ L_i|S = 1 &\sim TN(\mu_l, .25, 0, 1) \text{ assumed selection distribution} \\ \mu_l &\sim N(.9, .2) \text{ assumed selection distribution} \\ \beta_2 &\sim N(-4, 2) \text{ assumed load-voltage association} \end{aligned} \tag{5}$$

We can apply the adjustment formula in Equation 3 under this revised model to again estimate Q. Now, we use the sample of size $n = 200$ to calculate a best-estimate and 95% pointwise confidence intervals on the voltage distribution, as well as estimate reliability and a corresponding 95% CI. Results from this analysis compared to the naive analysis based on Equation 2 are shown in Figure 7b. Because the prior on the sensitivity study parameters in Equation 5 is reasonably aligned with the truth (Equation 4), the 95% pointwise confidence intervals also contain the truth. If the priors were bounded away from the truth, then the sensitivity study results will be inaccurate. The estimated reliability in the sensitivity study is .984, with 95% CI (.975, .993); hence, the change in reliability from the original analysis, where the reliability estimate was .850, is substantial and the sensitivity study suggests that selection bias should likely not be ignored when estimating reliability.

The results of sensitivity studies can be used to help determine what actions to take next. For instance, if results are exceptionally sensitive to certain dubious assumptions, then more information might need to be collected prior to decision-making. For example, under selection bias, we can consider collecting different types of information to improve our estimates. For instance, we could conduct more surveillance tests, learn the load-voltage association more precisely, or learn about the load selection distribution more precisely. This leads to a decision question is: what type of information would be most beneficial to collect next? In Appendix A, we provide an example of how to use a decision-analytic framework to determine what information to collect next based on the sensitivity study results.

5.2. Example: Sensitivity study to address confounding bias

Suppose now that we want to design a sensitivity study to address unmeasured confounding by load. That is, load L is associated with both age A and voltage Y, but is not along the causal pathway between A and Y (Figure 6a). As noted in the Section 4.2, this type of confounding could arise due to a tester error, where load was accidentally increased over time. However, note that this confounding would also be accompanied by selection bias, since the observed loads in the surveillance data are no longer representative of the expected load distribution in the target



(a)

(b)

Figure 7: Selection bias sensitivity study examining the empirical distributions of $Y|A = 25$. (a) The selection mechanism is known. The truth is in black, naive estimator in green, and corrected estimate in blue. The truth is covered by the corrected estimate. (b) The selection mechanism is uncertain. The best-estimate (pointwise median) is in a darker color than the 95% point-wise confidence intervals. The true distribution lies within the 95% confidence intervals from the sensitivity study.

population. When predicting Y, we should predict to the 'correct' distribution of L, depending on whether selection bias is present. In this example, we assume confounding bias is due to tester error, such that the distribution of load in the target population is the same as in Equation 3 in Section 5.1. A causal diagram for this case is shown in Figure 6b.

To apply the adjustment formula (Equation 1) to adjust for confounding bias, we need a statistical model for the distribution of $Y|L, A, S = 1$, but load L is not measured. To address this confounding bias, we use auxiliary information about:

- the load-age association $P(L|A, S = 1)$, and
- the load-voltage association $P(Y|L, A, S = 1)$.

In the presence of selection bias, we also need information about the distribution of load in the target population to apply the adjustment formula in Equation 1. With knowledge of these quantities, we can adjust for confounding via a statistical model. For example, suppose the confounding mechanism is known exactly to be:

$$\begin{array}{rcl} Y_i|A_i, L_i, S = 1 & = & \beta_0 + \beta_1 A_i + \beta_2 L_i + \epsilon_i \\ \epsilon_i & \sim & N(0, \sigma) \\ L_i|A_i, S = 1 & = & TN(.5 + .02A_i, .25, 0, 1) \text{ (true load-age association)} \\ \beta_2 & = & -5 \text{ (true load-voltage association)} \end{array} \tag{6}$$

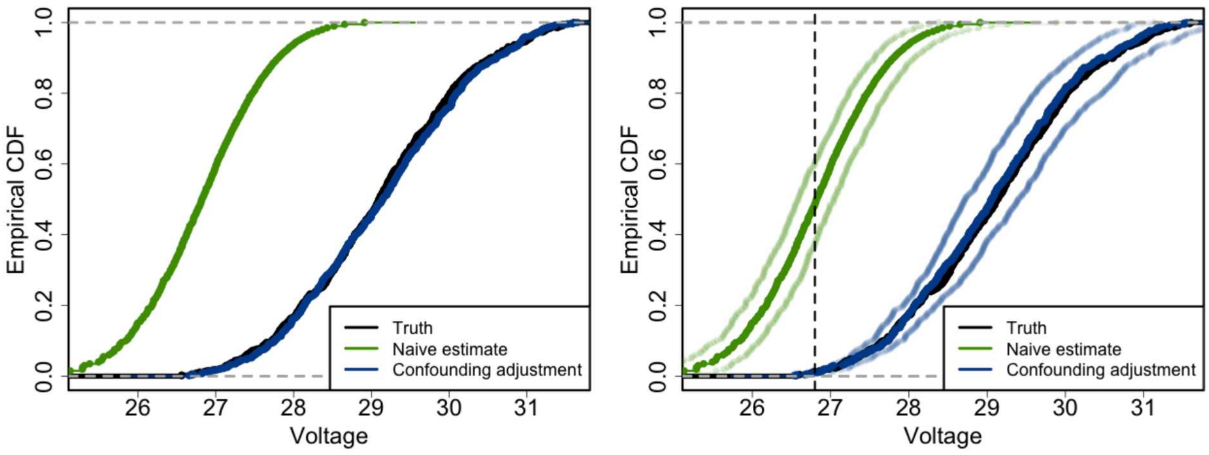
Lines 3-4 model the confounding mechanism given exact knowledge of the load-age and load-voltage associations. Given the distribution of $Y|L, A, S = 1$, as well as the distribution of load in the target population $P(L = l)$ (if selection bias is present), we can apply the adjustment formula in Equation 3.

Results from fitting the confounding-adjusted model are shown in Figure 8a. As in Section 5.1, the adjusted estimator aligns well with the truth in this simulated example, though this analysis is optimistic, since we plugged in the exact confounding mechanism. In practice, the exact load-age and load-voltage associations are unknown. Sensitivity studies can be used to determine how the results would change under different assumptions about the relationships. For instance, a model with uncertain load-age and load-voltage relationships might be:

$$\begin{array}{rcl} Y_i|A_i, L_i, S = 1 & = & \beta_0 + \beta_1 A_i + \beta_2 L_i + \epsilon_i \\ \epsilon_i & \sim & N(0, \sigma) \\ L_i|A_i, S = 1 & = & TN(.5 + \gamma_1 A_i, .25, 0, 1) \text{ (assumed load-age association)} \\ \gamma_1 & \sim & N(.01, .02) \text{ (assumed load-age association)} \\ \beta_2 & \sim & N(-4, 2) \text{ (assumed load-voltage association)} \end{array} \tag{7}$$

Note that a simple way to specify the load-age association is by specifying the distribution of load at two points and then interpolating.

Figure 8b shows the results of fitting this hierarchical model to the $n = 200$ samples, with 95% pointwise confidence intervals included. While uncertainty in the voltage distribution is higher due to the uncertainty about the confounding mechanism, inferences about the voltage distribution remain quite accurate relative to the truth and the truth is contained within the confidence intervals. The reliability estimates change substantially after addressing confounding in the sensitivity study. Specifically, without addressing confounding, the reliability estimate is .515 with 95% CI (.400, .630); in the sensitivity study with adjustment, the reliability estimate is .992 with 95% CI (.988, .994). Again, these inferences are sensitive to the prior on the sensitivity study parameters, and incorrect prior specification would produce inaccurate results. Regardless, the results of the sensitivity study indicate the potential high sensitivity of the reliability estimate to confounding bias.



(a)

(b)

Figure 8: Confounding bias sensitivity study examining the empirical distributions of $Y|A = 25$. (Left) Confounding mechanism is known. The truth is in black, naive estimator in green, and corrected estimate in blue. The truth is covered by the corrected estimate. (Right) Confounding mechanism is uncertain. The best-estimate (median) is in a darker color than the 95% pointwise confidence intervals. The true distribution lies within the 95% confidence intervals from the sensitivity study.

6. Discussion

In this paper, we link reliability analysis to SCM and illustrate how this framework can improve inferences and drive sensitivity studies. Connecting reliability to causality will ideally allow the reliability and risk analysis community to leverage the existing large body of work on causal analysis in other areas of application. This paper is only intended to frame how SCM and reliability are connected, and therefore contains very limited details about SCM. Further details can be found in review books such as [14, 27, 29].

Unlike SCM, existing approaches for reliability assessment uncertainty quantification often do not systematically consider statistical model credibility, essentially ‘conditioning’ on the statistical model being correct. This SCM framework puts a direct emphasis on the need to systematically assess statistical model credibility, i.e. do we have enough information to quantify reliability? Ignoring statistical model uncertainty can naturally lead to underestimation of uncertainty and inefficient allocation of resources for improving reliability. An important element of the SCM framework is the idea that data alone are typically not sufficient to learn causal relationships; rather, reliance on expert judgment and a fundamental understanding of how the data were generated is critical. Expert judgment is used to identify the structure of the causal network; to specify a statistical model for the data (when data are sparse); and to specify parameters in sensitivity studies assessing potential violations of assumptions.

Sensitivity studies examining structural modeling assumptions about how the data were generated can inform modeling results. The sensitivity studies in this paper only address violation of structural assumptions. The analysis results still rely on strong functional assumptions, namely that the conditional distribution of $Y|L, A$ is correctly specified. The statistical models could be extended to relax certain assumptions; however, as assumptions are relaxed, more parameter uncertainties must be specified in the sensitivity study, adding to the complexity of the analysis. Note that sensitivity studies are conceptually different from sensitivity analysis, which refers to quantifying how much variance in a model's outputs is explained by each model input [34]. On the other hand, sensitivity studies is a broader term that refers to analyzing how violations of certain assumptions impact the credibility of the modeling results.

Herein, we only discussed two types of structural biases (confounding and selection bias); Bareinboim and Pearl (2016) identify a third type of bias, transportability bias [2], where results from the available data must be 'transported' to a different target population. Confounding and selection bias compromise the internal validity of studies, while transportability bias compromises external validity [6]. Methods for adjusting for transportability and selection bias are similar, and therefore we focused on selection bias herein, but note that the results can be shown to generalize to transportability.

A primary limitation of our analysis is the simplicity of the example, namely modeling a linear trend over time from a single data source to predict reliability. One of the strengths of the SCM framework is the ability to use expert judgment and auxiliary data sources to improve estimation; future work will explore how to conduct data integration via the SCM framework to improve reliability estimation. Further, our analysis relied on the assumption that the functional model for battery voltage was correctly specified; in practical reliability applications, statistical models for reliability are difficult to specify, particularly due to the need to extrapolate about the tail behavior of the statistical model [16]. In non-engineering applications, such as health policy, a common goal is to compare alternatives, and Q is often a difference in conditional means, e.g. $E(Y|do(X = x_1)) - E(Y|do(X = x_2))$. In engineering applications, mean estimation is typically of less interest than reliability or margin estimation, which requires estimation of the full distribution of a counterfactual, rather than just a mean estimate. Estimation of this full distribution requires more modeling assumptions or more data than simply estimating a mean.

One of the strengths of the SCM framework is that the concepts are sufficiently general to apply across a broad array of settings. In future work, we aim to explore how SCM can improve risk and reliability assessments via combining different data sources, deemed data fusion [2]. Additionally, this paper focused on statistical, rather than computational simulation, models for estimating reliability. Another interesting area of research in risk analysis is how SCM can inform calibration and forward propagation of uncertainty over mis-specified computer models.

Appendix A. Assessing value of additional information

Using the selection bias example from Section 5.1, we apply a value of information decision framework (see, for instance, [15, 17, 41] for more details) to decide whether and what information to collect next by maximizing our expected utility (minimizing expected cost). Value of information (VoI) considers the cost relative to information gain and is defined as the difference in the expected cost of a certain decision with and without the new information:

$$VoI = E(C|\mathcal{D}) - E(C|\mathcal{D}, \mathcal{D}^*) \tag{A.1}$$

where C is the total cost, $\mathcal{D}$ is the current information, and $\mathcal{D}^*$ represents the new information. If collecting the information substantively decreases the expected cost *relative to the cost of the new information*, then the information has high value.

To calculate VoI, we need the following information: a statistical model for the outcome, the decision being made based on the outcome model, the cost structure for the decision, and the new information that can be collected. To illustrate the concept, we consider VoI in the context of selection bias (Section 5.1). The statistical model for the outcome is in Equation 5. The decision being made is whether or not the 98% reliability requirement is met for the 26.8 V requirement. We consider a linear cost structure, where we can either reject now and improve reliability at cost C_2 or approve the product and impose a linear penalty on the failure probability p for any probability exceeding $p_r = .01$. With knowledge of p, this cost structure is:

$$C(p) = min(C_1 pI(p > p_r), C_2) \tag{A.2}$$

Since p is uncertain, we consider the expected cost of approving the product. Then, the overall expected cost is:

$$C = min(E_{p|\mathcal{D}}[C_1 pI(p > p_r)], C_2) \tag{A.3}$$

where the expectation is taken over the posterior distribution of p given the data $\mathcal{D}$. Now, suppose we obtain additional information that induces a distribution on the model space, denoted $\mathcal{D}^*$. The expected cost given additional information $\mathcal{D}^*$ is defined as:

$$C^* = E_{\mathcal{D}^*} min(E_{p|\mathcal{D},d^*}[C_1 pI(p > p_r)], C_2) \tag{A.4}$$

where $p|\mathcal{D}, d^*$ is the posterior distribution of p given data $\mathcal{D}$ and a realization of $\mathcal{D}^*$. That is, the expectation in Equation A.4 is taken over the prior predictive distribution over $\mathcal{D}^*$. Decisions about whether to collect new data $\mathcal{D}^*$

are driven by the relative cost of C_1 to C_2, the expected reliability, and the change in uncertainty about reliability given the new information.

We consider collecting the following additional types of information:

- Conduct n^* additional tests at the current time point $A = 25$. $\mathcal{D}^*$ is a vector of n^* samples from the posterior predictive distribution of Y.
- Obtain perfect information about the load-voltage association, β_2. $\mathcal{D}^*$ is a vector of samples from the prior distribution on β_2.
- Obtain perfect information about the mean of the load selection distribution, μ_l. $\mathcal{D}^*$ is a vector of samples from the prior distribution μ_l.

To estimate the value of information, we compare the expected cost under each of these three different scenarios to the expected cost given the current information only. (Note that we could also consider the value of jointly conducting multiple actions, such as both collecting new samples and learning about the load-voltage association; however, such joint actions are not considered herein.)

Returning to the selection bias example (Section 5.1), we now consider a data-limited scenario where only $n = 50$ tests were conducted, rather than $n = 200$ tests. First, we examine the value of information for the case where the sample size increases by conducting n^* tests at the current time point. We consider 4 different cases: $n^* = \{1, 5, 10, 100\}$. To estimate VoI, we sample n^* new data points from the posterior predictive distribution of the voltage Y after fitting the model in Equation 5. Then, we fit the model in Equation 5 again to the revised sample with the n^* additional points; repeating this procedure 100 times, we then obtain the approximate prior predictive distribution of p given $\mathcal{D}^*$, from which we can calculate the expected cost given the n^* new samples using Equation A.4. Results of the VoI analysis are show in Figure A.9. The value of information peaks for a C_1/C_2 ratio near 40; that is, collecting new information is most valuable when the linear penalty on reliability (C_1) is approximately 40 times as large as the cost of initially rejecting and approving reliability up front (C_2). The VoI increases with the number of new samples; the confidence interval width for reliability also shrinks as n^* increases. Whether to collect more samples depends on the cost of the additional data collection relative to the VoI.

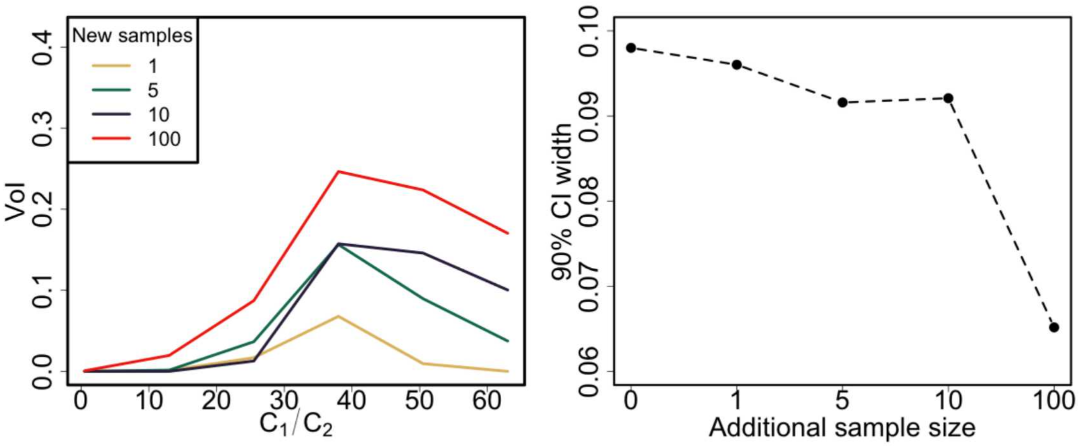


Figure A.9: (Left) Value of information for increasing the sample size as a function of the ratio of C_2 to C_1. (Right) Confidence interval width on reliability as a function of the added sample size.

Next, we calculate the value of information under the three different scenarios as a function of the ratio of C_1/C_2 (Figure A.10). For the case where the sample size increases, we use $n^* = 100$ (since VoI was highest at $n^* = 100$). The average confidence interval width under the different scenarios is also plotted in Figure A.10. Collecting the new information does decrease uncertainty about reliability and therefore tends to decrease expected cost; but, the value of information is of course sensitive to the ratio of C_1/C_2. Obtaining perfect information about the load-voltage association provides the most additional value once the ratio of C_1/C_2 exceeds 20. However, the decision maker must also consider the cost of obtaining the additional information. Therefore, whether to collect additional information and what information to collect depends on the difference between the VoI and the cost of the additional information.

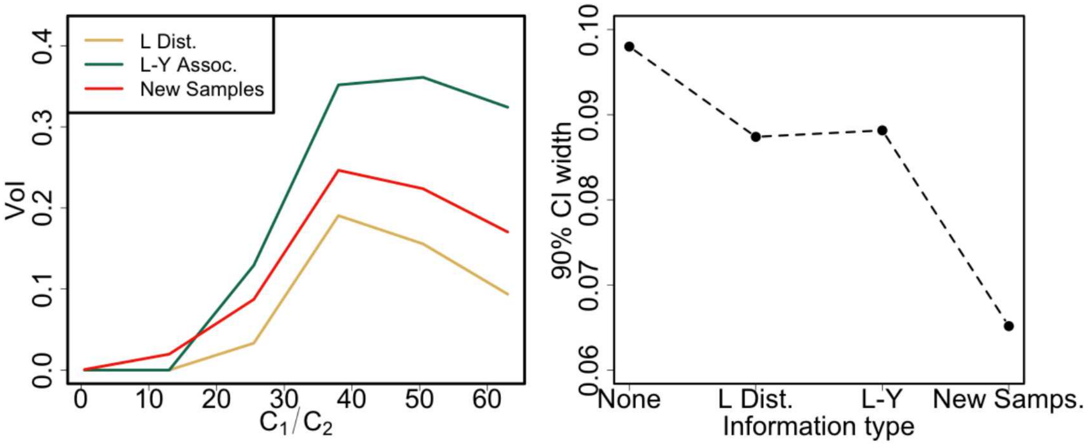


Figure A.10: (Left) Value of information for different types of additional information as a function of the ratio of C_2 to C_1. (Right) Confidence interval width on reliability for different types of information.

Acknowledgements

Funding: This work was supported by a Sandia National Laboratories Laboratory Directed Research and Development (LDRD) grant. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525. Approved for unlimited release, SANDXXXX.

This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government

[1] Aven, T. "Foundational issues in risk assessment and risk management." *Risk Analysis*, 34(7) (2014).

[2] Bareinboim, E. and Pearl, J. "Causal inference and the data-fusion problem." *Proceedings of the National Academy of Sciences*, 113(27):7345–7352 (2016).

[3] Ben-Gal, I. "Bayesian networks." In Kenett, R. and Faltin, F. (eds.), *Encyclopedia of statistics in quality and reliability*, volume 1. New York: Wiley Online Library (2008).

[4] Bennett, T. R., Booker, J. M., Keller-McNulty, S., and Singpurwalla, N. D. "Testing the untestable: Reliability in the 21st century." *IEEE Transactions on Reliability*, 52(1):118–124 (2003).

[5] Bobbio, A., Portinale, L., Minichino, M., and Ciancamerla, E. "Improving the analysis of dependable systems by mapping fault trees into Bayesian networks." *Reliability Engineering & System Safety*, 71(3):249–260 (2001).

[6] Broniatowski, D. A. and Tucker, C. "Assessing causal claims about complex engineered systems with quantitative data: internal, external, and construct validity." *Systems Engineering*, 20(6):483–496 (2017).

[7] Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. "Stan: A probabilistic programming language." *Journal of Statistical Software*, 76(1) (2017).

[8] Cox, L. A., Jr. "Clarifying types of uncertainty: when are models accurate, and uncertainties small?" *Risk Analysis: An International Journal*, 31(10):1530–1533 (2011).

[9] Cox Jr, L. and Ricci, P. "Causation in risk assessment and management: models, inference, biases, and a microbial risk–benefit case study." *Environment international*, 31(3):377–397 (2005).

[10] Cox Jr, L. A. "Improving causal inferences in risk analysis." *Risk Analysis*, 33(10):1762–1771 (2013).

[11] EricksonKirk, M. et al. "Sensitivity studies of the probabilistic fracture mechanics model used in FAVOR version 03.1." *NUREG-1808, US Nuclear Regulatory Commission, ADAMS ML*, 61580349 (2004).

[12] Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., and Rubin, D. B. *Bayesian Data Analysis*. Chapman and Hall/CRC (2013).

[13] Hernán, M. A., Hernández-Díaz, S., Werler, M. M., and Mitchell, A. A. "Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology." *American Journal of Epidemiology*, 155(2):176–184 (2002).

[14] Hernán, M. A. and Robins, J. M. *Causal Inference*. Chapman & Hall, CRC: forthcoming (2018). Https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/.

[15] Howard, R. A. "Information value theory." *IEEE Transactions on Systems Science and Cybernetics*, 2(1):22–26 (1966).

[16] Hund, L., Schroeder, B., Rumsey, K., and Huerta, G. "Distinguishing between model- and data-driven inferences for high reliability statistical predictions." *Reliability Engineering & System Safety*, 180:201–210 (2018).

[17] Katz, R. W. and Murphy, A. H. "Quality/value relationship for imperfect information in the umbrella problem." *The American Statistician*, 41(3):187–189 (1987).

[18] Khakzad, N., Khan, F., and Amyotte, P. "Safety analysis in process facilities: Comparison of fault tree and Bayesian network approaches." *Reliability Engineering & System Safety*, 96(8):925–932 (2011).

[19] Lee, C.-J. and Lee, K. J. "Application of Bayesian network to the probabilistic risk assessment of nuclear waste disposal." *Reliability Engineering & System Safety*, 91(5):515–532 (2006).

[20] Li, J. and Shi, J. "Knowledge discovery from observational data for process control using causal Bayesian networks." *IIE Transactions*, 39(6):681–690 (2007).

[21] Ling, Y., Mullins, J., and Mahadevan, S. "Selection of model discrepancy priors in Bayesian calibration." *Journal of Computational Physics*, 276:665–680 (2014).

[22] Mahadevan, S. and Rebba, R. "Validation of reliability computational models using Bayes networks." *Reliability Engineering & System Safety*, 87(2):223–232 (2005).

[23] Marazopoulou, K., Ghosh, R., Lade, P., and Jensen, D. "Causal discovery for manufacturing domains." *arXiv preprint arXiv:1605.04056* (2016).

[24] Meeker, W. Q. and Escobar, L. A. "Reliability: The other dimension of quality." *Quality Technology & Quantitative Management*, 1(1):1–25 (2004).

[25] Montgomery, D. C. *Design and analysis of experiments*. John wiley & sons (2017).

[26] Newcomer, J. T. "A new approach to quantification of margins and uncertainties for physical simulation data." *Sandia National Laboratories, Albuquerque, NM, Technical Report No. SAND2012-7912. https://prod-ng. sandia. gov/techlib-noauth/access-control. cgi/2012/127912. pdf* (2012).

[27] Pearl, J. *Causality*. New York: Cambridge University Press (2009).

[28] Pearl, J. and Bareinboim, E. "External validity: From do-calculus to transportability across populations." *Statistical Science*, 579–595 (2014).

[29] Pearl, J., Glymour, M., and Jewell, N. P. *Causal Inference in Statistics: A Primer*. John Wiley & Sons (2016).

[30] Pearl, J. and Russell, S. "Bayesian networks." In Arbib, M. (ed.), *Handbook of brain theory and neural networks*. Cambridge: MIT Press (2001).

[31] Pilch, M., Trucano, T. G., and Helton, J. C. "Ideas underlying the quantification of margins and uncertainties." *Reliability Engineering & System Safety*, 96(9):965–975 (2011).

[32] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2017).
URL `https://www.R-project.org/`

[33] Robins, J. M. "Data, design, and background knowledge in etiologic inference." *Epidemiology*, 313–320 (2001).

[34] Saltelli, A., Chan, K., Scott, E. M., et al. *Sensitivity Analysis*, volume 1. New York: Wiley (2000).

[35] Sankararaman, S., McLemore, K., Mahadevan, S., Bradford, S. C., and Peterson, L. D. "Test resource allocation in hierarchical systems using Bayesian networks." *AIAA Journal* (2013).

[36] Sharp, D. H. and Wood-Schultz, M. M. "QMU and Nuclear Weapons Certification-What's under the Hood?" *Los Alamos Science*, 28:47–53 (2003).

[37] Shmueli, G. "To explain or to predict?" *Statistical Science*, 289–310 (2010).

[38] Urbina, A., Mahadevan, S., and Paez, T. L. "Quantification of margins and uncertainties of complex systems in the presence of aleatoric and epistemic uncertainty." *Reliability Engineering & System Safety*, 96(9):1114–1125 (2011).

[39] VanderWeele, T. J. and Arah, O. A. "Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders." *Epidemiology*, 22(1):42–52 (2011).

[40] Wallstrom, T. C. "Quantification of margins and uncertainties: A probabilistic framework." *Reliability Engineering & System Safety*, 96(9):1053–1062 (2011).

[41] Yokota, F. and Thompson, K. M. "Value of information analysis in environmental health risk management decisions: past, present, and future." *Risk Analysis*, 24(3):635–650 (2004).