Artificial Intelligence in Test and Evaluation: Test Design and Analysis Using AI-Based Causal Inference

Briefing

Joseph George Caldwell

15 July 2022

(*Briefing*: Microsoft PowerPoint file test-design-and-analysis-using-causal-inference-briefing.pptx, .pdf
Posted at
http://www.foundationwebsite.org/index13-causal-inference-and-matching.htm)

Road Map

1. A brief summary of major methodologies in Operational Test and Evaluation (OT&E): Descriptive Analysis and Causal Inference (Designed Experiments, Observational Studies)

2. Difficulties with Designed Experiments and traditional Observational Data Analysis in OT&E applications

3. Potential roles of Artificial Intelligence (AI) in OT&E: test design, test-data analysis, generalization of field-test results, and automated scenario generation

4. Alternatives to Experimental Design (ED) for causal inference in OT&E: Neyman-Rubin Causal Model (NRCM, the traditional statistical approach to analysis of observational data); Pearl Structured Causal Model (AI-based approach)

5. Comparison of basic technical features of NRCM and Pearl approaches

6. Assessment: Pearl's approach has significant advantages over NRCM in operational settings

7. Recommendations for future development of AI-based causal inference in OT&E

1. Two Basic OT&E Activities: Estimation of Average Performance and Conditional Performance

1. Estimate the **average performance** of a system or component for one or more "standard" situations

   Descriptive analysis of population characteristics: Simple random sampling (infinite populations); (descriptive) sample survey (finite populations)

   Methodology: Determine sample sizes to achieve desired level of precision for estimates; confidence intervals, tests of statistical significance

2. Estimate performance **conditional on** values of system and environmental variables

   Causal inference: Obtain an accurate estimate of the distribution (or distribution feature; "causal effect") of an output variable, Y, conditional on values for input variables, X, in the presence of covariates, Z.

   Methodologies: Experimental design, analytical sample survey design, and observational data analysis (Neyman-Rubin Causal Model (**traditional statistical approach**), Pearl Structured Causal Model (**AI-based approach**))

This presentation will summarize and compare methodologies for **causal inference**, focusing on the application of the **NRCM** and **Pearl's AI-based methodology** in OT&E. (Also relevant to descriptive analysis.)

2. Before We Proceed, Some Observations

1. This presentation addresses inference about the **effects of causes**, not the **causes of effects**.

2. George E. P. Box: ”All models are wrong, but some are useful.” **Agree**, and some are more useful than others!

3. George E. P. Box: “To find out what happens to a system when you interfere with it you have to interfere with it (not just passively observe it).” (“Use and Abuse of Regression,” *Technometrics*, Vol. 8, No. 4, Nov. 1966.) **Yes and no – it depends.**

4. Paul W. Holland: “No causation without manipulation,” (“Statistics and Causal Inference,” *Journal of the American Statistical Association*, Vol. 81, No. 396, Dec. 1986.) **Disagree!** The moon causes tides, but we cannot manipulate it. (Holland was referring to defining a causal effect at the level of an individual experimental unit.)

5. The **causal effect of X on Y is defined** to be the conditional probability distribution, P(Y|X) (or aspects of it, such as E(Y|X), or $E(Y|X_1)$-$E(Y|X_2)$). This distribution **can be estimated** without manipulation, given assumptions.

3. Classical Approaches to Test Design: Designed Experiments and Sample Surveys

Experiment: An investigation to determine cause-effect relationships in which selection of subjects and assignment of treatments (experimental conditions) to subjects is controlled by the experimenter

Observational Study: This control is absent (for any number of reasons, including physical, legal, ethical, behavior)

Elements of Experimental Design (ED) and Sample Survey Design:

> **Randomization** (randomized selection from a population (or distribution) of interest; randomized assignment to treatment levels of interest)
> **Replication** (for estimation of estimate variances, precision, power and significance levels)
> **Local control** (matching, blocking, stratification)
> **Symmetry** (balance, orthogonality)
> **Sample size determination** (statistical precision analysis, statistical power analysis)
> **Sample selection procedures** (simple random sampling, with/without replacement sampling, multistage sampling, stratified sampling, sampling with variable selection probabilities, matching)

Straightforward analysis, matched to design (e.g., analysis of variance, general linear statistical model, sample survey analysis)

4. Designed Experiments Encounter Many Difficulties in OT&E Applications

Expertise (knowledge of experimental design (ED); creative skill)
Strict adherence to design protocol
Many opportunities for departure from the design (resulting in a “broken” ED):

- Inability to apply treatments as intended (e.g., equipment failure)
- Inability to employ randomization as intended (e.g., self-selection of respondent, noncompliance)
- Inability to implement design structure as intended (e.g., lack of orthogonality)
- Inability to control design variables as intended (e.g., weather conditions)
- Missing data (item and unit nonresponse)

Inability to repair broken EDs as EDs (**if randomization is compromised, the data become observational, not experimental**)
Difficulty in assessing the impact of design breakdowns
Inability to combine observational data with experimental data
Low level of external validity (scope of inference): difficulty in extending experimental results (ED often applied to a restricted population and conditions)
Stable unit-treatment-value assumption (SUTVA): a unit’s response is not affected by other units’ responses (no interference, such as competition for resources)

While these difficulties are largely addressable in a laboratory setting, they can represent serious problems in non-laboratory settings such as in evaluation of social and economic programs or in operational test and evaluation of military systems and equipment.

4b. Key Problems for Designed Experiments in OT&E Applications

1. **The ED causal model is extremely simple, but fragile, not robust:**

   Once the key assumptions of **randomized selection** and **randomized assignment to treatment levels** are compromised, the controlled experiment has failed, the model is invalid, the causal-effect estimates are biased, and there is no "fallback" procedure to salvage the situation.

   The simple ED model cannot be adjusted (as an ED) to accommodate the issue. **The data are now observational data, not experimental data.**

2. **The ED causal model has a low level of external validity (restricted scope):**

   Design variables are often specified and levels set **without reference to a complete causal model**.

   Randomized controlled experiments are generally performed on **restricted** portions of a larger **population** of interest, and under **restricted** and highly controlled **conditions**.

   The ED causal-effect estimates are **conditional** on the design variables, restricted population, data, and simple causal model of the ED, and ED methodology does not prescribe **how to generalize the results**.

The designed-experiment approach has a high level of internal validity, but **many OT&E settings require evaluation methodologies that are more flexible than EDs**.

5. Problems with the Traditional Approach to Observational Data Analysis (ODA)

The **Neyman-Rubin Causal Model** (NRCM) (potential-outcomes model, counterfactuals model), which is much used, is very useful in teaching and research, but **has severe shortcomings** in operational settings.

1. The method is based on **untestable assumptions** about the **unobservable** joint distribution of treatment and response. The approach has been characterized as “**metaphysical**.”

2. The method is **not based on a specification of a complete causal model**. It is difficult to assess the estimability of causal effects and difficult to determine formulas for estimating causal effects.

3. The method involves the use of “propensity scores” (probabilities of assignment to treatment), which are **much-used** for matching. **Propensity scores should not be used for matching.** It often increases imbalance, inefficiency, model dependence and bias.

6. A Significant OT&E Issue: How to Generalize Field Test Results (Increase Scope, External Validity)

Specific Issue: How to generalize the **conditional causal-effect** estimates from an Experimental Design (ED) or traditional Observational Data Analysis (ODA) to obtain **average causal-effect** estimates for settings (scenarios) of interest?

In the absence of a complete causal model of the system under test, this cannot reasonably be done.

The ED causal model is essentially useless for this purpose (randomized assignment to treatment severs the causal links from all system variables to treatment – basically, it is a “degenerate” causal model).

Traditional ODA (NRCM) is not based on a complete causal model, but on separate sets of (unconfoundedness) assumptions specific to estimating particular causal effects.

To achieve generalization of test results from EDs and ODA to more general settings (scenarios), a complete causal model is required.

The Pearl Structured Causal Model is a complete causal model, and can be used for generalization.

6b. Example: How to Generalize a Field Test of an Electronic Warfare System Performed at the US Army Electronic Proving Ground (EPG), Fort Huachuca, Arizona

Alternative Test Designs:

Descriptive Analysis: System set up and tested in a standard configuration and environmental conditions.
Result: Estimate of average performance of the system for a "base case."

Experimental Design: Fractional factorial design with the factors: distance to target, terrain type, vegetation type, weather, type of target, density of targets, and deployment/employment of targets.
Result: Estimates of system performance conditional on design factors (all orthogonalized).

Pearl Structured Causal Model: Design based on a complete causal model.
Result: Estimates of system performance conditional on causal-model variables (varying in a realistic fashion).

Generalization Issue: Based on the field-test results, how is it expected that the system would perform in Ukraine, under combat conditions?

Descriptive Analysis: Performance expected to be as in field test, in locations and conditions similar to field test.
Experimental Design: How to extend the conditional, orthogonal-effect estimates to Ukraine situation?
Pearl AI-Based Approach: Configure the causal model match the Ukraine situation, and obtain causal-effect estimates of the system performance in Ukraine from the model. **Relative to the issue of generalization, Pearl's model is more useful than the alternatives.**

7. Artificial Intelligence Can Be Used to Overcome Many of the Problems Encountered by Designed Experiments and Traditional Observational Data Analysis in OT&E Settings

The AI-based methodology of Pearl's Structured Causal Models can be used to address these problems, and overcome them. Specifically, Pearl's methodology can be used:

To assist experimental design (increase scope, relevance and efficiency)

To "fix" broken experiments (in the sense of recovering useful information)

To assist observational test design

To analyze data from observational tests

To generalize the conditional causal-effect estimates of EDs and traditional ODA to other settings

To support automated scenario generation for operational tests

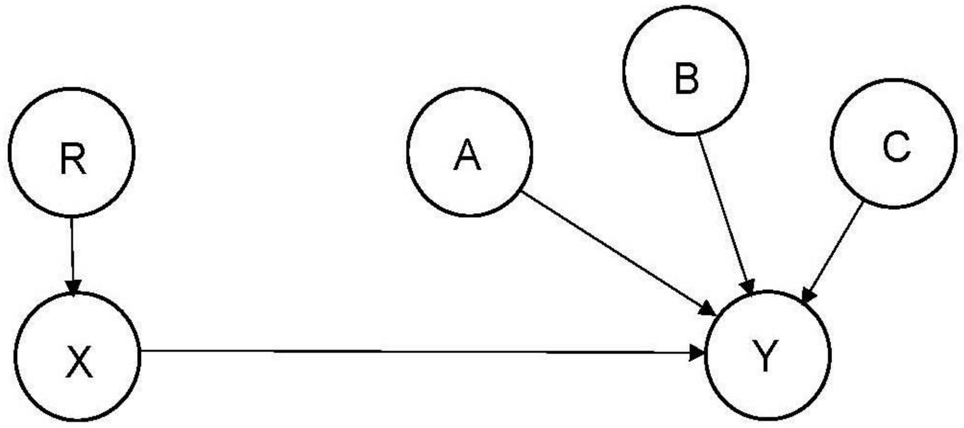
This briefing will summarize how Pearl's methodology addresses these problems and can be used to accomplish the preceding evaluation functions.

(The methodology is useful for assisting **descriptive analysis as well as causal inference**, but that is not a subject of this briefing.)

8. Alternative Methodologies for Causal Inference in Operational Test and Evaluation

- Experimental Tests
  - **Experimental Design** (Randomized Controlled Trials)
  - **Analytical Sample Survey Design** (an amalgam of sample survey and experimental design)
- Observational Tests
  - Traditional Statistical Approach
    - **Neyman-Rubin Causal Model (Counterfactuals Model, Potential-Outcomes Model)**
      - Rosenbaum-Rubin approach (“balancing approach,” “statistical approach”)
      - Heckman approach (“econometric approach”)
  - AI-Based Approach
    - **Judea Pearl Structured Causal Model**

8b. Alternative Methodologies for Causal Inference in OT&E (Cont'd.) – Experimental Design Model and Analytical Sample Survey Model



"R" denotes randomized selection of an experimental unit from a population of interest and randomized assignment to a value of the variable X. Randomization assures that there are **no variables that affect both X and Y**.

The model provides estimates of the **Average Causal Effect** (ACE, or, simply, Causal Effect, CE) of input X on output Y **for each design group (combination of design variables** (X, which is a vector of variables)).

The **ACE for a population (or probability distribution) of interest** is the weighted average of the design-group ACEs, where the weights are the proportions of the population represented by each group.

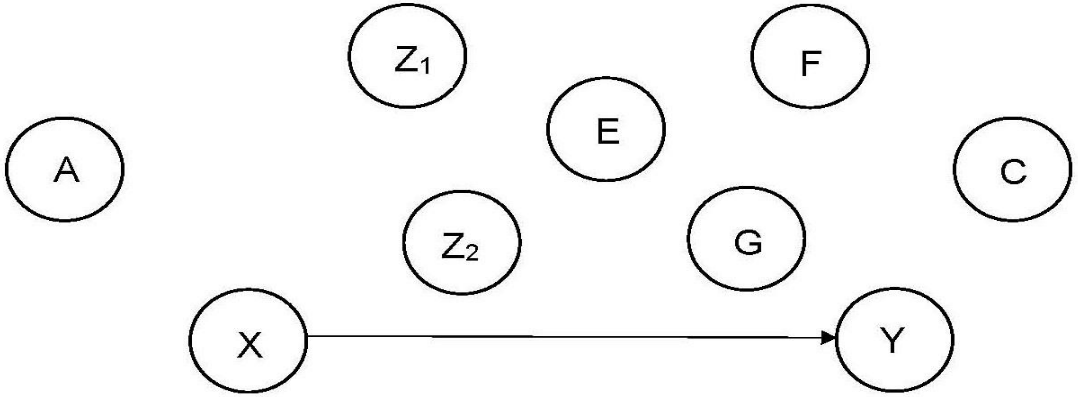
8c. Alternative Methodologies for Causal Inference in OT&E (Cont'd.) – Neyman-Rubin Causal Model

The **Neyman-Rubin Causal Model** (NRCM, Counterfactuals Model, Potential-Outcomes Model) is the traditional statistical approach to analysis of observational data, and the most widely used model for analysis of observational data.

Suppose that we have a sample of inputs and outputs, S = (Y_o,X_o) and covariates, A, B, …, Z.

If we can find a set of covariates Z such that (Y,X) $\perp X_o$ | Z where $0 < \Pr(X_o) < 1$ for all Z (i.e., "**strong ignorability**" of X_o given Z), then

> Average Causal Effect of X on Y = Pr(Y|X) = the weighted average over Z of the Z-conditioned Observed Effect of X on Y, = $\Sigma_Z \Pr_S(Y|X,Z)P(Z)$.



8d. Alternative Methodologies for Causal Inference in OT&E (Cont'd.) – Neyman-Rubin Causal Model (NRCM)

Note: The terminology **"counterfactual"** derives from the case in which X is a binary variable (e.g., treated (X=1) and untreated (X=0)), and each sample unit can be observed for only one of its two values. The observed outcome is called the "actual" (or "factual") and the unobserved outcome is called the "counterfactual." The term **"potential outcomes"** recognizes the fact that, until the unit is observed, either outcome is possible.

Issues:

1. Without specifying a complete causal model (depicting causal relationships among all model variables), how can Z reasonably be identified?

2. A separate "model" (a strong ignorability assumption) is required for each causal-effect estimate of interest.

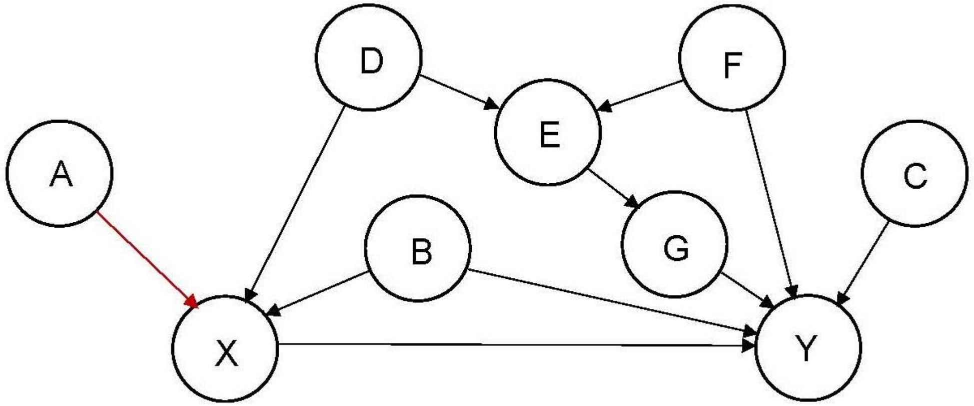
3. What are candidates for Z? Are there alternative choices for Z? If so, how to choose among them? The NRCM methodology provides no guidance about what are good choices or bad choices for Z!

4. It is impossible to observe the joint pair (Y,X), and hence impossible to justify any assertion about the joint distribution of (X,Y), so how is it reasonable to make assumptions about it (outside of research or teaching)?

5. The condition that the joint probability distribution function (PDF) of (Y,X) be independent of X_o given Z is stronger than is needed. We are interested in Y|X, not in (Y,X). So why make assumptions about (Y,X)?

8e. Alternative Methodologies for Causal Inference in OT&E (Cont'd.) – Pearl Structured Causal Model

A **complete causal model** of the system under test is constructed, and displayed as a **directed acyclic graph (DAG, causal model diagram)**. (It is the specification of a **complete causal model** that distinguishes the Pearl approach from the NRCM approach, and qualifies it as an "AI-based" approach.)

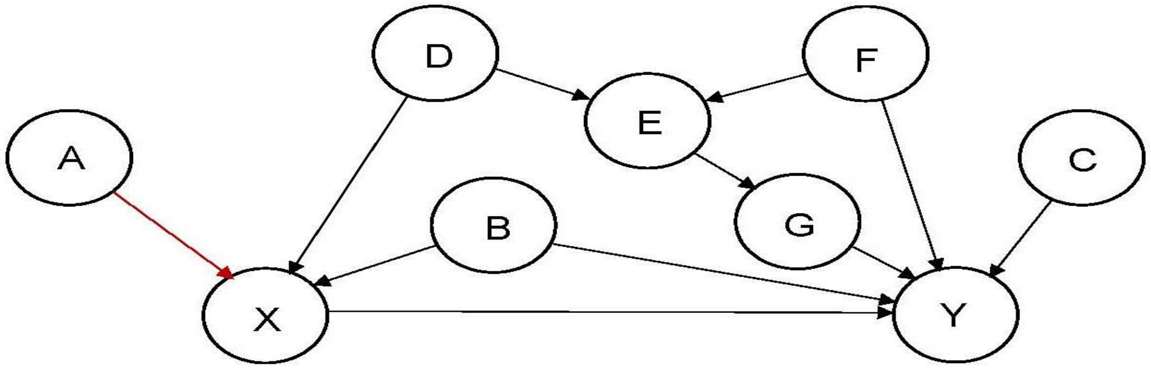


The causal model (a **Markovian Bayesian Network**) is defined in terms of probabilistic relationships among the variables depicted in the graph. (The probability formulas are not presented here – see Pearl's **Causality** for definitions (similar in some respects to specification of economic Structural Equation Models (SEMs)).)

The causal effect of a variable X on a variable Y is defined as the conditional distribution of Y given X, averaged over a population (or probability distribution) of interest.

Estimability of causal effects and the general form of estimates of causal effects are determined from connectedness properties of the graph ("**Back-Door Criterion**," "**Front-Door Criterion**").

9. The Back-Door Criterion for Assessing Estimability of Causal Effects for the Pearl Model



Directional separation (d-separation) (Pearl): A path p is said to be *d-separated* (or *blocked*) by a set of nodes Z if and only if

1. p contains a **chain** i→m→j or a **fork** i ←m→ j such that the middle node m is in Z, or
2. p contains an **inverted fork** (or *collider*) i→m←j such that the middle node m is not in Z and such that no descendent of m is in Z.

A set Z is said to ***d-separate*** X from Y if and only if Z blocks every path from a node in X to a node in Y.

Back-Door Criterion (Pearl): A set of variables Z satisfies the *Back-Door Criterion* relative to an ordered pair of variables (X,Y) in a DAG G if:

1. no node in Z is a descendent of X; and
2. Z blocks every path between X and Y that contains an arrow into Y.

In the graph shown, the set of nodes {B,D} satisfies the Back-Door Criterion relative to the pair {X,Y}.

9b. The Back-Door Criterion: Back-Door Adjustment

The following procedure provides a basis for determining, from the Back-Door Criterion, **whether a causal effect is estimable**, and **provides a formula for estimating the causal effect** of X on Y:

Back-Door Adjustment (Pearl): If a set of variables Z satisfies the Back-Door Criterion relative to (X,Y), then the causal effect on X on Y is identifiable (estimable) and is given by the formula

$$P(y|do\ x) = \sum_{z} P(y|x,z)P(z)\,.$$

If Z contains a single variable, then the expression is the simple weighted average of the distribution P(Y|X,Z) over Z. If Z contains more than one variable, the expression may be evaluated using a Markov Chain Monte Carlo (MCMC) procedure, such as Gibbs sampling, or by using propensity scores. In most applications, interest focuses on estimation of means, not of the distribution function.

Many evaluation applications involve the case in which treatment is binary (treated / untreated). The method is general, and applies to the cases of single, multiple or continuous input (treatment) variables.

(The Front-Door Criterion addresses cases in which the causal diagram includes a covariate along the path from X to Y.)

For a shorter briefing, skip the following sections, 10-14 (retained in presentation for possible use in discussion)

10. Comparison of the Fundamental Assumptions of the Pearl and NRCM Methodologies

11. Pros and Cons of the Neyman-Rubin Causal Model Approach

12. Pros and Cons of the Pearl Structured Causal Model Approach

13. Summary of the Advantages of the Pearl Approach over the NRCM Approach

14. If the Pearl AI-Based Approach to Causal Inference Is So Useful, Then Why Has the Statistical Establishment Not Embraced It?

10. Comparison of the Fundamental Assumptions of the NRCM and Pearl Methodologies

The various approaches to causal inference for observational data are based on similar assumptions, which are referred to variously as **conditional independence (CI)**, ignorability, exogeneity, selection on observables, or unconfoundedness assumptions.

The goal of satisfying these assumptions is accomplished (conceptually) by finding subsets of the sample, defined by values of observed covariates, Z, such that, within each subset, **the distribution of Y|X does not depend on observed treatment, X_o.**

The **Average Causal Effect** can then be estimated simply by averaging the subset causal effect with respect to the distribution of Z (often called **"conditioning on Z"**).

10b. Comparison of the Fundamental Assumptions of the NRCM and Pearl Methodologies (Cont'd.)

Conditional Independence Assumption for the NRCM Approach

The NRCM approach requires "**strong ignorability**": Conditional on observed covariates, Z, the joint probability distribution of response (output), Y, and treatment (input), X, is not dependent on the observed treatment (or assignment to treatment), X_o. Symbolically, this conditional-independence (CI) assumption is written as:

$$(Y,X) \perp X_o | \mathrm{Z} \, , \, 0 < \Pr(X_o=1) < 1 \text{ for all } Z.$$

A difficulty associated with this approach is that, for an individual experimental unit, the response, Y, can be observed for only one level of the treatment variable, X. Since the joint distribution of (Y,X) cannot be observed, **the fundamental assumption is untestable**. Furthermore, since a complete causal model is not specified, the reasonableness of the assumption **cannot even be assessed theoretically**.

A detailed discussion of this difficulty is presented in **Pearl's book, *Causality: Models, Reasoning, and Inference***, 2nd ed., Cambridge University Press (2009 (1st ed. 2000)) and in **A. P. Dawid's paper, "Causal Inference without Counterfactuals"** / Comments (by D. R. Cox, George Casella, Stephen P. Schwarz, Judea Pearl, James M. Robins, Sander Greenland, Donald B. Rubin, Glenn Shafer, and Larry Wasserman) / Rejoinder, *Journal of the American Statistical Association*, June 2000, Vol. 95 No. 450, Theory and Methods. The impossibility of testing the strong ignorability assumption led Dawid to refer to the counterfactuals approach to causal inference as "**metaphysical**."

10c. Comparison of the Fundamental Assumptions of the NRCM and Pearl Methodologies (Cont'd.)

Conditional Independence Assumption for the Pearl Approach

The requirement that the joint variables (Y,X) be independent of observed treatment X_o given covariates Z is very strong, and stronger than necessary. **Pearl uses a weaker (but sufficient) assumption**, that the conditional variable (Y|X) be independent of observed treatment X_o given covariates Z.

Note that, whereas the **joint** random variables (Y,X) are not observable at the level of the individual experimental unit, the **conditional** random variable (Y|X) *is* observable. (The reasonableness of the assumption could be assessed from observed data (Heterogeneous Treatment Effects "subgroup analysis").)

Symbolically, this conditional-independence (CI) assumption is written as:

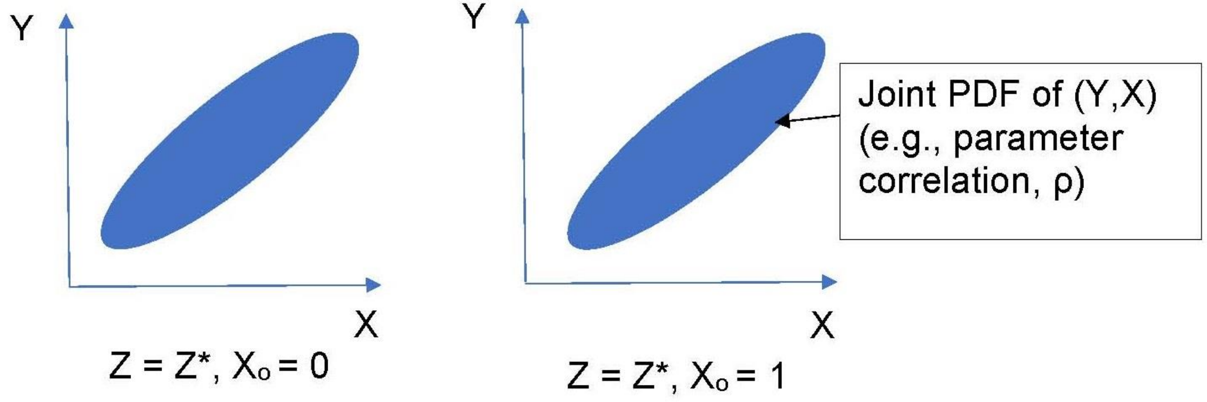
> $Pr_P(Y=y \mid X=x, X_o=x_o, Z=z)$ depends only on covariates Z and treatment X, and not further on observed treatment X_o; that is, $Pr_P(Y=y \mid X=x, X_o= x_o, Z=z) = Pr(Y=y \mid X=x, Z=z)$. (The subscript P refers to the joint probability distribution of all model variables **over a population of interest**.)

Under this assumption, the distribution of Y conditional on X is the same as the distribution of Y conditional on observed treatment, X_o; that is, **given Z, the Average Treatment Effect (ATE) is equal to the Observed Treatment Effect (OTE)**.

Given the causal diagram, the validity of this assumption can be assessed (Back-Door and Front-Door Criteria).

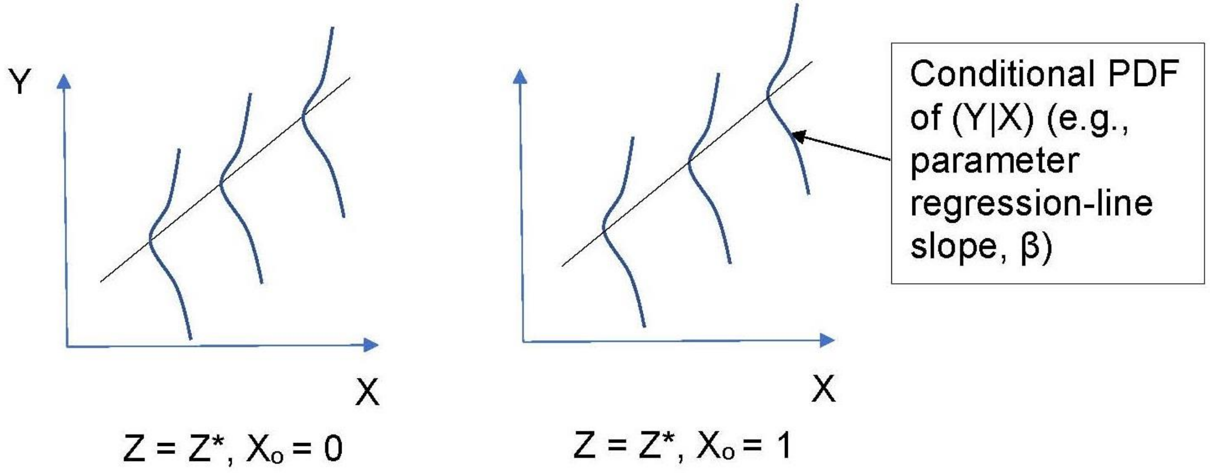
10d. Comparison of the Fundamental Assumptions of the NRCM and Pearl Methodologies (Cont'd.)

NRCM: Conditional Independence of (Y,X) and X_o given Z



Joint PDF of (Y,X) (e.g., parameter correlation, ρ)

Pearl: Conditional Independence of (Y|X) and X_o given Z



Conditional PDF of (Y|X) (e.g., parameter regression-line slope, β)

10e. Comparison of the Fundamental Assumptions of the NRCM and Pearl Methodologies (Cont'd.)

Points of Comparison:

With the NRCM approach, knowledge about the causal nature of a system is embodied in a **collection of strong-ignorability assumptions** about jointly unobservable variables (Y,X). In the Pearl approach, knowledge about the causal system is embodied **in a single graph** displaying system variables, Y|X.

Although the use of counterfactual statements is useful in understanding individual-unit causal concepts, **comprehension** of a complete causal system **is much easier with a simple DAG** that depicts observable conditional relationships among variables, than with a collection of strong-ignorability statements about unobservable joint probability distributions at the individual-unit level.

This fundamental difference has a profound effect on the ease of constructing and describing a causal model, assessing estimability of causal effects, and identifying estimation formulas.

A tremendous advantage of the Pearl approach is that, given the causal model, it is possible to readily identify, from a visual examination of the causal graph, **the estimability of any causal effects of interest, and a formula for estimating the causal effect.**

10f. Comparison of the Fundamental Assumptions of the NRCM and Pearl Methodologies (Cont’d.)

Points of Comparison (cont’d.):

The NRCM requirement of conditional independence (CI) of the joint random variable (Y,X) given Z **is much stronger** than the Pearl requirement of CI of the conditional random variable (Y|X).

In many, if not most, evaluation applications, all that is of interest is differences in group means, not heterogeneity of treatment effects (HTE) at the level of the individual unit, and **the stronger requirement is essentially irrelevant**.

Since **the joint distribution is not observable** at the level of the individual treatment unit, the CI condition of strong ignorability is **impossible to validate**.

This requirement renders the NRCM approach difficult to use and exposes it to severe criticism in operational settings.

10g. Comparison of the Fundamental Assumptions of the NRCM and Pearl Methodologies (Cont'd.)

Of What Use Is the Counterfactual Model?

Since the strong-ignorability assumption of the NRCM approach is impossible to validate, and is unnecessary in many applications, it is reasonable to ask **wherein lies its value**.

In **Donald Rubin's comment on Dawid's paper**, "Causal Inference without Counterfactuals," Rubin pointed out two very useful features of the counterfactuals / strong-ignorability framework.

One is to facilitate ***teaching*** of casual inference – the concept of counterfactuals greatly facilitates understanding of causality.

The second is to assist ***research***: analysis of the features of the joint distribution (Y,X) is very helpful in analysis of otherwise-intractable problems such as noncompliance and missing data.

The counterfactuals / potential-outcomes framework is useful for ***teaching and research***. In those contexts, assumptions are simply asserted, with no need for validation / justification.

For practical applications, the stringent requirement of strong ignorability, and the requirement to represent causal knowledge in the form of strong-ignorability statements, **greatly diminishes the usefulness of the NRCM approach**.

11. Pros and Cons of the Neyman-Rubin Causal Model Approach

Pros:

Simplicity: Through use of the **propensity score** (the probability of selection for treatment), the “curse of dimensionality” of having to average over a multivariate Z distribution (Gibbs sampling, MCMC algorithm)) is obviated. All that is necessary is to average over the propensity score, a single (univariate) variable. (Estimation of the propensity score is straightforward (a logistic regression model based on observed covariates).)

The counterfactuals approach to causal inference is very useful for **teaching and research**.

The NRCM approach is **generally accepted by the statistical establishment**.

Some government agencies now require estimation of propensity scores in analysis of observational data.

For his work on the NRCM, James Heckman was awarded the Nobel Prize in Economics.

The method is presented in many books on econometrics.

Statistical software packages are widely available to perform the analysis.

11b. Pros and Cons of the Neyman-Rubin Causal Model Approach (Cont'd.)

Cons:

A fundamental problem with the NRCM approach is the difficulty in justifying the model assumptions:

> Pearl: "Strong ignorability is a convenient syntactic tool for manipulating counterfactual formulas, as well as a convenient way of formally assuming admissibility (of Z) without having to justify it. …Hardly anyone knows how to apply it in practice because the counterfactual variables are unobservable, and scientific knowledge is not stored in a form that allows reliable judgment about conditional independence of counterfactuals. It is not surprising, therefore, that "strong ignorability" is used almost exclusively as a surrogate for the assumption "Z is admissible" and rarely, if ever, as a criterion to protect us from bad choices of Z."

11c. Pros and Cons of the Neyman-Rubin Causal Model Approach (Cont’d.)

Cons (Cont’d.):

There is no well-defined procedure for assessing model validity (strong ignorability), or for assessing estimability, or for deciding which variables should or should not be conditioned on to obtain unbiased estimates of causal effects. It is impossible to validate the strong ignorability assumption, and easy for a critic to fault.

There is no well-defined procedure for repairing or revising a model if the strong ignorability assumption is invalidated. Invalidation of the assumption of strong ignorability in the NRCM approach has dire consequences, similar to invalidation of the assumption of randomized assignment in an experimental design, and the methodology offers no guide to rectifying the situation.

Although very useful in teaching and research applications, in real-world applications, the NCMR approach is an inflexible, fragile, non-robust “all-or-nothing” approach. **It is an arcane, abstruse, “metaphysical” approach that is difficult to implement and justify in operational settings.**

11d. Pros and Cons of the Neyman-Rubin Causal Model Approach (Cont'd.)

Cons (Cont'd.):

Problems related to the propensity score:

Recall: The **propensity score** is the probability of assignment to treatment, given covariates Z=z: L(z) = P(X=1|Z=z).

Rosenbaum-Rubin result: The causal effect may be estimated by averaging over the **univariate** propensity score given the covariates, instead of averaging over the joint **multivariate** distribution of the covariates.

$$P(y|do\ x) = \sum_z P(y|x,z)P(z) = \sum_l P(y|x,l)P(l)$$

This averaging is accomplished by post-stratifying over categories of the propensity score, or by including it in a regression model, or by inverse-probability weighting with it.

Note that the propensity-score simplification has nothing to do with causality**: it is simply a computational device**.

11e. Pros and Cons of the Neyman-Rubin Causal Model Approach (Cont'd.)

Cons (Cont'd.):

Problems related to the propensity score (cont'd.):

The propensity score is **subject to all of the assumptions about strong ignorability** as the multivariate formula that it replaces. Its only role is to simplify computation.

In widespread use of the propensity score, **awareness of all of the assumptions** on which its use is based **tends to be forgotten**.

The causal-effect estimates are correct **only for the true value of the propensity score, not for estimates of it**.

If the true propensity score depends on **unobserved variables**, then the design must be configured so that these variables "drop out" in the analysis. For example, if selection depends on individual personality characteristics, use a pretest-posttest-comparison-group design in which the same individuals are interviewed in the posttest.

The propensity score cannot be equal to zero or one (the "overlap" condition).

11f. Pros and Cons of the Neyman-Rubin Causal Model Approach (Cont’d.)

Cons (Cont’d.):

Problems related to the propensity score (cont’d.):

Use of the propensity score (PS) complicates the task of estimating the ATE for populations other than the one corresponding to the field test (since groups having equal propensity scores do not correspond to covariate-identified segments of the population).

Propensity scores should not be used for matching. The use of propensity-score matching (PSM) often increases imbalance, inefficiency, model dependence and bias. For more information, see the article, “Why Propensity Scores Should Not Be Used for Matching,” by Gary King and Richard Nielsen, *Political Analysis* (2019), or http://www.foundationwebsite.org/StatCours4&5CausalInferenceAndMatching.pdf.

In the R&R methodology, the PS is used to reduce selection bias by post-stratification on the PS, using the PS as a covariate in a regression model, or by inverse-probability weighting using the PS. **It was never proposed or intended as a basis for matching, and does so very poorly.** Units that match on the PS may not match well at all on variables that are strongly causally related to output variables of interest.

Despite the inappropriateness of the use of propensity-score matching (PSM), it is widely used for this purpose, and the data analysis is usually done incorrectly: only 28% correct, per “A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003,” by Peter C. Austin, *Statistics in Medicine*, 2008, Vol. 27, pp. 2037-2049.

12. Pros and Cons of the Pearl Structured Causal Model Approach

Pros:

Simplicity; transparency; ease of understanding

> High degree of face validity (model does not depend on untestable assumptions about the joint distribution of treatment and response); easy to defend / justify (reference to graphic causal-model diagram); difficult to fault.
>
> Ease of understanding model assumptions and of assessing validity of model assumptions
>
> Ease of assessing estimability of causal effects

Robustness, flexibility. Works for broken EDs or for observational data analysys (ODA). Easy to modify the model to accommodate changes in assumptions.

A more general approach: Incorporates modern AI methodology. Methodology is readily automated.

Broader scope: Based on a complete causal model, the causal-effect estimates can be extended to populations of interest other than the particular one used for a test (by suitably modifying the causal model).

Ease of documentation of model validity, estimability assessment, and estimation procedures (results are conditional on the model validity, which is relatively easy for anyone to assess).

12b. Pros and Cons of the Pearl Structured Causal Model Approach (Cont'd.)

Cons:

The Pearl Structured Causal Model approach is **not widely accepted or used by the statistical establishment**.

The Pearl methodology is **presented in few reference texts**, such as Pearl's *Causality* and Stephen L. Morgan and Christopher Winship's *Counterfactuals and Causal Inference: Methods and Principles for Social Research*, 2nd ed., Cambridge University Press (2015, 1st ed. 2007).

Software is not widely available that automates the process of constructing causal graphs (DAGs) and making inferences from them.

The Pearl model **does not accommodate nonrecursive causal effects** (variables that have a simultaneous (mutual, reciprocal) causal effect on each other).

Effort must be expended in constructing a complete causal model.

12c. Pros and Cons of the Pearl Structured Causal Model Approach (Cont’d.)

Cons (Cont’d.):

An oft-heard criticism: **“But I don’t know what the causal model is!”**

Response: **Of course not!** It is simply a hypothetical paradigm from which estimators can be derived. The same as for the models used in all branches of statistics: designed-experiment models, sample survey models, regression models, analysis of variance models, general linear statistical models, generalized linear statistical linear models, multivariate analysis models and time-series analysis models.

A Pearl structured causal model is a complete, compact, visual representation about what is believed about the causal relationships among the variables of a system under test.

Recall George Box’s aphorism: **“All models are wrong, but some are useful.”**

13. Summary of the Advantages of the Pearl Approach over the NRCM Approach

Pearl’s AI-based Structured Causal Model approach to causal analysis of observational data has **substantial advantages** over the more widely used Neyman-Rubin Causal Model approach.

Easier to assess model validity and causal-effect estimability. More transparent. Less restrictive assumptions. Easier to justify and defend results.

More general. Makes use of modern artificial-intelligence methodology of causal inference, not just traditional statistical methodology. In addition to estimation of causal effects, the Pearl methodology can be used to assist test design and automated scenario development.

Based on a complete causal model, causal-effect estimates can be constructed for populations of interest additional to the specific one used in a field test.

…and several disadvantages:

Not widely accepted by the established statistical community

Not widely known or practiced by statisticians

Not widely documented in statistics reference texts; **limited software support** (SAS, R)

14. If the Pearl AI-Based Approach to Causal Inference Is So Useful, Then Why Has the Statistical Establishment Not Embraced It?

Much of the field of statistics is concerned with description and analysis of associational (probabilistic) relationships and **associational inference, not with causal inference**. The exception to this is the field of experimental design, which has the goal of estimating causal effects. Over the past century, the word "causal" rarely appeared in statistics texts.

Causal inference and statistical inference are intimately related. The fact that, by and large, the field of statistics generally restricts causal inference to the methodology of randomized experiments, and utilizes the NRCM approach rather than Pearl's approach, deserves comment.

The general attitude of the field is reflected in the following statement from Chapter I of the book, *Causal Inference for Statistics, Social and Biomedical Sciences: An Introduction,* by Guido Imbens and Donald Rubin (Cambridge University Press, 2015): **"Pearl's work is interesting, and many researchers find his arguments that path diagrams are a natural and convenient way to express assumptions about causal structures appealing. In our work, perhaps influenced by the type of examples arising in social and medical sciences, we have not found this approach to aid drawing of causal inferences, and we do not discuss it further in this text."**

Here follow a number of speculations on why the field of statistics has not embraced Pearl's approach to causal inference, when it is so useful.

14b. If the Pearl AI-Based Approach to Causal Inference Is So Useful, Then Why Has the Statistical Establishment Not Embraced It? (Cont'd.)

1. **The traditional (non-AI-based) approach is less complex and requires less effort.**

    The Pearl approach requires more work in both the design and analysis phases than the ED or NRCM approaches. A complete causal model must be developed to guide the design, and if, in the analysis, the model is later seen to be incorrect, it must be revised and the causal effects re-estimated in accordance with the revised model.

    The ED approach is conceptually simpler: the causal model is extremely simple (randomized assignment to treatment eliminates all causal links to the treatment variable). If the design assumptions are violated, this is noted and the possibility of bias is acknowledged, but little or no effort is expending in salvaging the situation (e.g., by analyzing the data using NRCM methodology). Analysis proceeds generally in accordance with the original design (suitably modified to accommodate design-structure changes, such as by using a general linear statistical model in place of an analysis-of-variance approach), even if randomization assumptions are violated.

14c. If the Pearl AI-Based Approach to Causal Inference Is So Useful, Then Why Has the Statistical Establishment Not Embraced It? (Cont'd.)

2. **The traditional approach entails less professional risk.** ("Nobody ever got fired for buying IBM.")

The **ED approach**, if it works, **is easy to defend**.

If a randomized experiment fails, e.g., because of treatment noncompliance, **blame is usually not focused on the statistician**.

In the NRCM approach, **the strong ignorability assumption is difficult to explain and to comprehend, and impossible to validate**. It is usually uncritically accepted by reviewers and clients (since it is essential to the NRCM approach, which is generally accepted), even though it represents a point of strong weakness in the methodology in practical settings.

The graphic causal model is central to the Pearl approach: it is the basis for assessing the validity of model assumptions and the estimability of causal effects.

With the Pearl approach**, the analyst must accept responsibility for constructing and defending a complex causal model,** a graphical representation of which is available to the client.

14d. If the Pearl AI-Based Approach to Causal Inference Is So Useful, Then Why Has the Statistical Establishment Not Embraced It? (Cont'd.)

3. **In general, acceptance of new and more complex methodologies may be slow.** Pearl's work is much more recent than the NRCM methodology (1920s for ED, 1970s for observational data; 1990s for Pearl). The situation is similar to the case of Bayesian analysis, which was slow to achieve widespread acceptance and use.

4. **Many researchers prefer to specialize and to work in a single field**, such as statistics, instead of straddling two fields (statistics and artificial intelligence).

5. **Access to Pearl's methodology is limited.** The methodology is discussed in few statistics textbooks, and commercial software for implementing the methodology is limited.

6. **The US government requires the use of randomized controlled trials in biomedical research** (despite difficulties in implementation, such as noncompliance, the ethics of denying treatment, and reluctance of people to participate in randomized trials).

7. **Some federal agencies are encouraging the use of propensity-score-based methods** (a major feature of the NRCM) as a substitute for randomized experimental designs.

15. Summary Comparison and Assessment of Pearl and NRCM Approaches

1. The NRCM approach is based on “strong ignorability” assumptions about an unobservable quantity: the joint distribution of treatment and response at the level of the individual experimental unit. These fundamental assumptions, based on unobservable quantities, are untestable. They are difficult to comprehend and impossible to validate or justify.
2. In the NRCM approach, causal knowledge about a system is represented in a set of distinct estimability assumptions for each causal-effect estimate of interest.
3. In the Pearl approach, causal knowledge is represented in a complete, integrated, graphically represented causal model that describes causal relationships among system variables.
4. The NRCM approach does not provide a methodology for identifying estimable causal effects.
5. The Pearl approach provides two simple criteria (the Back-Door and Front-Door Criteria) for assessing the estimability of any causal effects of interest, and provides estimation formulas for them, from the causal diagram.
6. The Pearl methodology, based on a complete causal model, can be used to assist the design of experimental and observational tests and the construction of test scenarios.
7. The Pearl methodology provides an easy-to-use mechanism for generalizing the results of a field test to other situations of interest: By making modifications to the causal model, causal estimates can be obtained for alternative situations of interest.

Assessment: For OT&E applications, the Pearl approach offers substantial advantages over the NRCM approach. It has greater face validity, involves testable assumptions, represents causal knowledge in a more efficient form, and is generally much more useful than the NRCM approach for OT&E applications.

16. How Pearl's AI-Based Structured Causal Model Methodology Can Be Used in OT&E (Both Experimental and Observational Tests)

In test design:

Construction of test designs

- Identification of design variables based on a **complete causal model**
- Display of causal model (directed acyclic graph (DAG))
- Assessment of estimability of causal effects and specification of estimation formulas
- Matching (of comparison units to treatment units)
- (Other tasks:
  - Specification of levels of design variables
  - Determination of sample size and allocation (statistical precision and power analysis)
  - Allocation of sample to design "cells"
  - Sample selection (variable selection probabilities) and assignment to treatment)

Automated scenario development (in support of test design)

Using the Pearl approach, the test design is specified in terms of the variables of a complete causal model and their values.

These can be used to specify system configurations and environmental conditions for field tests (i.e., to define scenarios).

16b. How Pearl’s AI-Based Structured Causal Model Methodology Can Be Used in OT&E (Cont’d.)

In test-data analysis:

Analysis of test data in accordance with a complete causal model:

Assessment of estimability of desired causal-effect estimates

Specification of estimation formulas for, and calculation of, causal-effect estimates

Post-test updating of the causal model (from the one used in design)

Generalization: Extension of ED and ODA results to populations (PDFs) of interest (alternative scenarios) additional to those represented in field tests:

The causal-effect estimates produced by an ED or ODA are conditional on the design and the population (causal system, joint probability distribution function (PDF) of all model variables) from which the experimental units were selected (randomly or otherwise), and do not represent causal effects relative to other populations of interest (scenarios).

By **modifying the causal model**, causal effects can be estimated for **other populations (PDFs) of interest**. This is very useful for extension of inferential scope, sensitivity analysis, and input to simulation models and scenarios.

17. Recommendations for Future Development and Use of AI-Based Causal Inference in OT&E: General Recommendation

Use the Pearl Structured-Causal-Model AI-based approach for design and analysis of tests in the field of OT&E **to complement or replace alternative methodologies** (experimental design, Neyman-Rubin Causal Model).

AI methodology can **overcome the implementation difficulties** associated with the use of experimental designs in OT&E field tests.

Pearl’s Structured-Causal-Model approach **has significant advantages over other methodologies** used to analyze observational data (NRCM (Rosenbaum-Rubin / Heckman)).

AI methodology can be used **to construct test designs** and to help **automatically generate test scenarios**.

AI methodology can be used **to analyze test data** and **generalize field-test results**.

AI methodology can be used to **combine test data** with (experimental or observational) data **from other sources**.

AI methodology is a powerful tool that can greatly improve the capability, effectiveness and efficiency of OT&E beyond the limits possible using traditional statistical methodologies.

17b. Specific Recommendations

1. **Expand awareness of and access to software that implements the basic Pearl methodology,** which enables the user to interactively specify and display structured causal model graphs (DAGs), assess estimability of causal effects from them, specify estimation formulas, and calculate causal-effect estimates from test data.

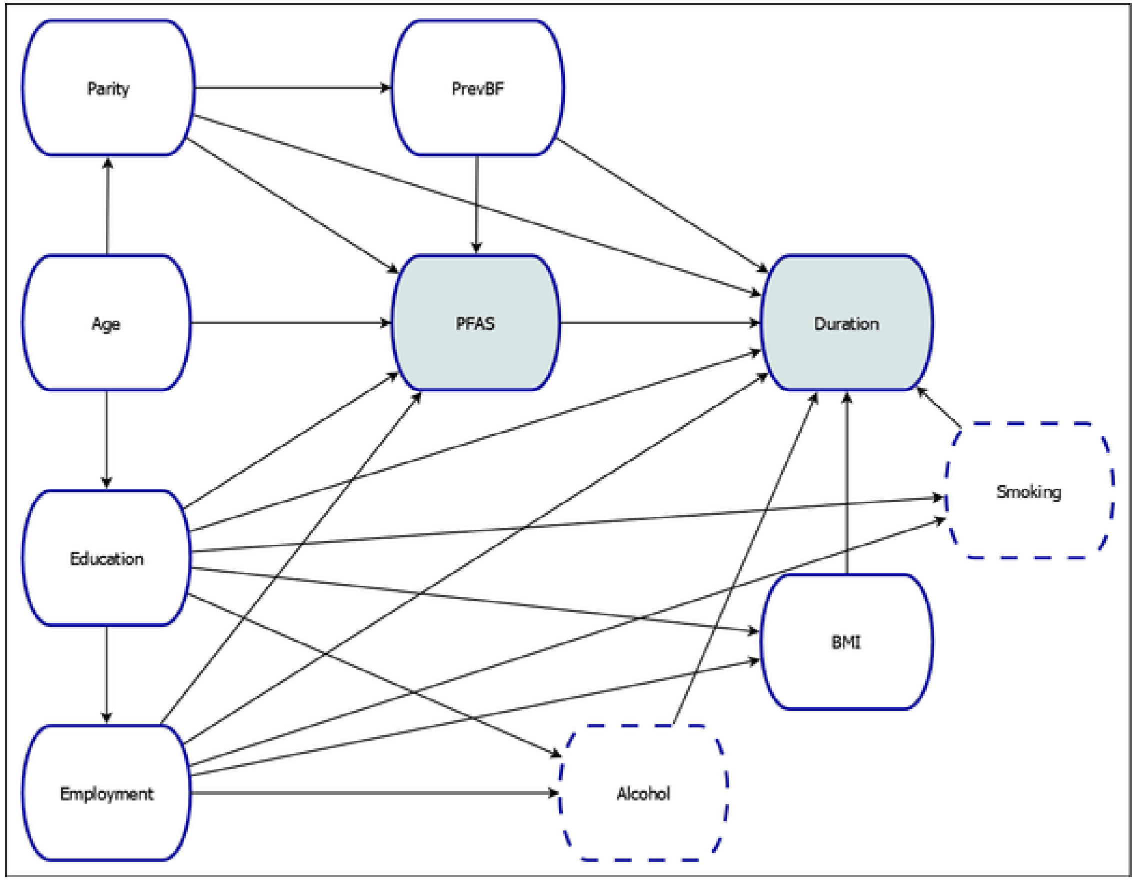
Sample resources:

> *Causal Inference In Statistics: A Companion for R Users* by Johannes Textor, Andrew Forney, and Judea Pearl. Posted at Internet website http://dagitty.net/primer/. This document provides programmatic solutions in the R package for statistical computing for many of the exercises in *Causal Inference in Statistics: A Primer* by Pearl, Glymour, and Jewell.
>
> SAS CAUSALGRAPH procedure, posted at http://documentation.sas.com/doc/en/pgmsascdc/9.4_3.4/statug/statug_causalgraph_gettingstarted.htm

17c. Specific Recommendations (Cont'd.)

The CAUSALGRAPH Procedure (from http://documentation.sas.com/doc/en/pgmsascdc/9.4_3.4/statug/statug_causalgraph_gettingstarted.htm)

Figure 34.1: Causal Model of the Effect of Persistent Perfluoroalkyl Substances on Breastfeeding Duration



17d. Specific Recommendations (Cont'd.)

2. **Develop (expert-system) software to assist a user in constructing test designs** corresponding to specified causal models and test environments.

> The software will guide the user in selection of design variables (or factors), matching (for forming matched pairs or comparison groups), specification of variation (spread, balance, orthogonality) in design variables, and selection of a sample of system/equipment locations and configurations taking into account precision and power requirements, geographic data, military doctrine and system-under-test requirements.
>
> This methodology will be similar to that used to develop analytical sample survey designs, since it involves sampling from finite populations.

17e. Specific Recommendations (Cont’d.)

3. (“Blue Sky”) **Develop software to enable a user to modify a causal model to represent populations (PDFs) other than the one sampled from in field tests, and to construct causal-effect estimates relative to those populations.**

This capability will provide estimates of causal effects (system performance and effectiveness) for a wide range of operational environments, not just for the one represented in field tests.

This feature is perhaps the most useful one that will result from combining AI methodology with traditional statistical methodology. Since each specified population of interest represents a scenario, this extension will represent a significant capability to evaluate systems under alternative scenarios additional to the specific ones used in field tests.

This approach is a **Bayesian approach**, where the causal model represents prior information and the test provides sample data. Alternative scenarios are simply alternative prior distributions.

The causal-effect estimates for alternative scenarios are simply Bayesian estimates corresponding to alternative priors.

This development would do much to address the problem of small sample sizes for some OT&E tests, by identifying conditional estimates of high precision.

17f. Specific Recommendations (Cont'd.)

4. "(Blue Sky") **Develop automated procedures for generating alternative causal models.**

Alternative causal models represent alternative scenarios.

The resulting system is an ***automated scenario-generation system***.

The tremendous advantage of this approach to automated scenario generation is that the scenario is specified in terms of the variables of the causal model, and causal-effect estimates corresponding to the scenario (and the test data) may be determined directly from the causal model.

(Note: To facilitate this development, assure, in Recommendation 2, that the test design is constructed in a way that accommodates generation of scenarios of interest and estimation of causal effects for them (by including variables of interest in these scenarios and assuring adequate variation in those variables). That is, **the field test is used to collect data not only for a specific evaluation situation**, but also for use in generating alternative scenarios and estimating causal effects for them.)

(For an example of an AI-based automated scenario generation system, see documentation for the **Scenarist** development project is posted at http://www.foundationwebsite.org/index16-artificial-intelligence.htm.)

18. Additional Information about Causal Inference

Additional detail on causal inference using the Neyman-Rubin Causal Model and Pearl Structured Causal Model approaches is provided in the following, posted at http://www.foundationwebsite.org/index12-design-of-analytical-sample-surveys.htm:

> *Briefing*: Microsoft PowerPoint file design-and-analysis-of-analytical-sample-surveys-briefing-short-version.pptx, .pdf
> *Briefing*: Microsoft PowerPoint file design-and-analysis-of-analytical-sample-surveys-briefing.pptx, .pdf
> *Briefing Notes*: Microsoft Word file design-and-analysis-of-analytical-sample-surveys-briefing-notes.docx, .htm, .pdf

The preceding pieces discuss application of causal inference to the problem of designing analytical sample surveys, which is similar to the problem of designing tests for operational test and evaluation. References for analytical survey design using causal inference are listed on the following slide.

19. A Methodology for Designing Analytical Sample Surveys

Most of the published material on causal inference is concerned **with *analysis*, not with *design***.

There is no standard reference text that presents a detailed or comprehensive description of procedures or general methodology **for constructing analytical survey designs**.

This author presents a general methodology in the paper:

> *Sample Survey Design for Evaluation (The Design of Analytical Surveys)* posted at Internet website http://www.foundationwebsite.org/SampleSurveyDesignForEvaluation.htm.
>
> Additional material is presented in lecture notes for the courses:
> *Causal Inference and Matching*, at
> http://www.foundationwebsite.org/StatCourse4and5CausalInferenceAndMatching.htm; and
>
> *Statistical Design and Analysis for Evaluation*, at
> http://www.foundationwebsite.org/StatCourse6and7StatisticalDesignAndAnalysisForEvaluation2DayCourse.htm.

The methodology includes elements of all major approaches to causal inference, experimental design, and sample survey design.