

Innovation of Causal Inferential Frameworks for Biostatistics in Real-World Data and Evidence

Anwesha Roy (Senior Statistical Programmer)

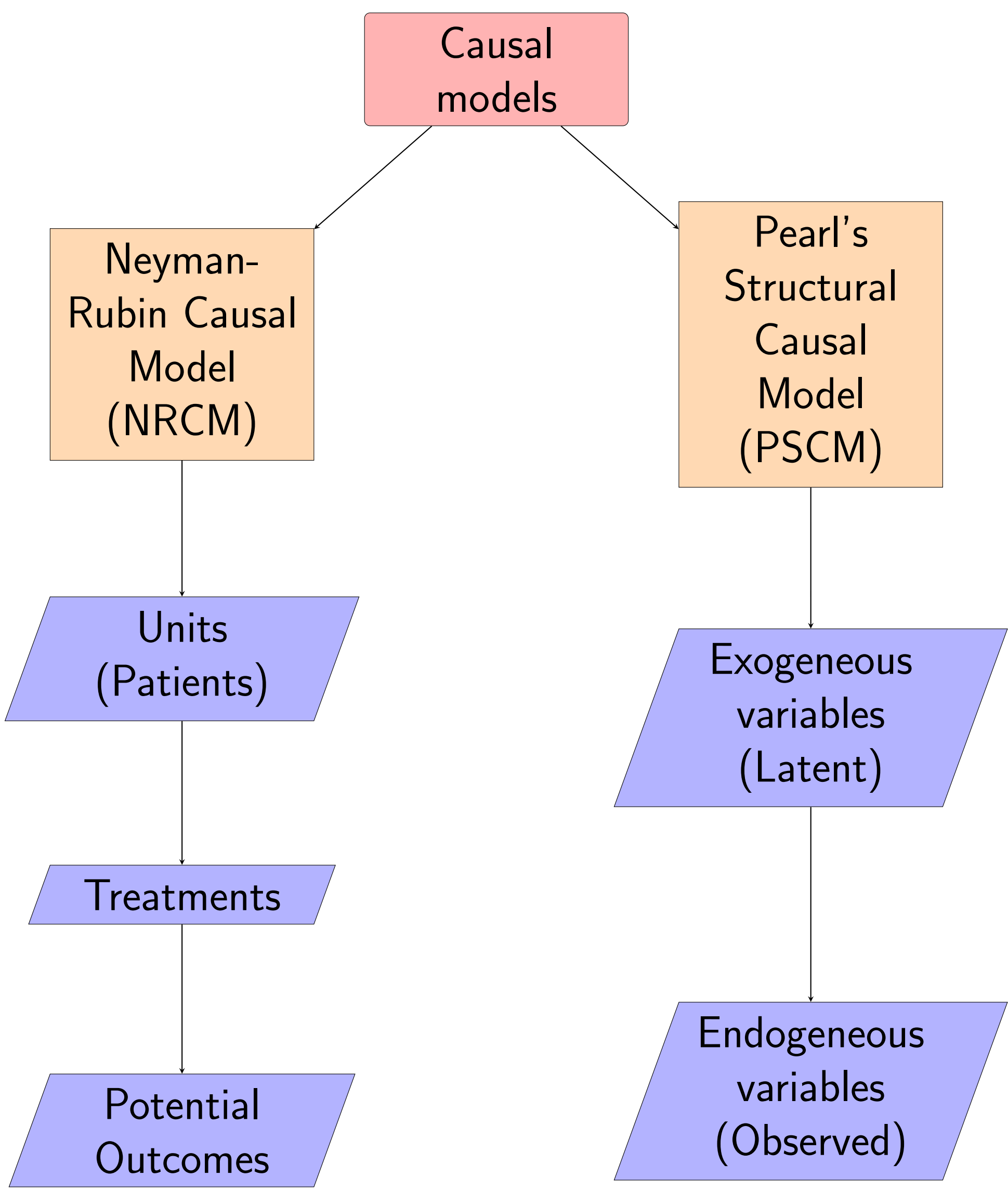
(Contact: anwesha.roy@novartis.com)

Novartis Healthcare Pvt Ltd, Hyderabad, India

Motivation

- To promote the appropriate use of RWE to support medical product development. [Chen \(2020\)](#)
- Advanced modeling and inference for rare diseases using Causal Networks.
- Road map to guide the design and analysis of the clinical studies that use RWD to generate RWE. [Ho \(2021\)](#)

Causal Models



- Data generating process in a clinical study:

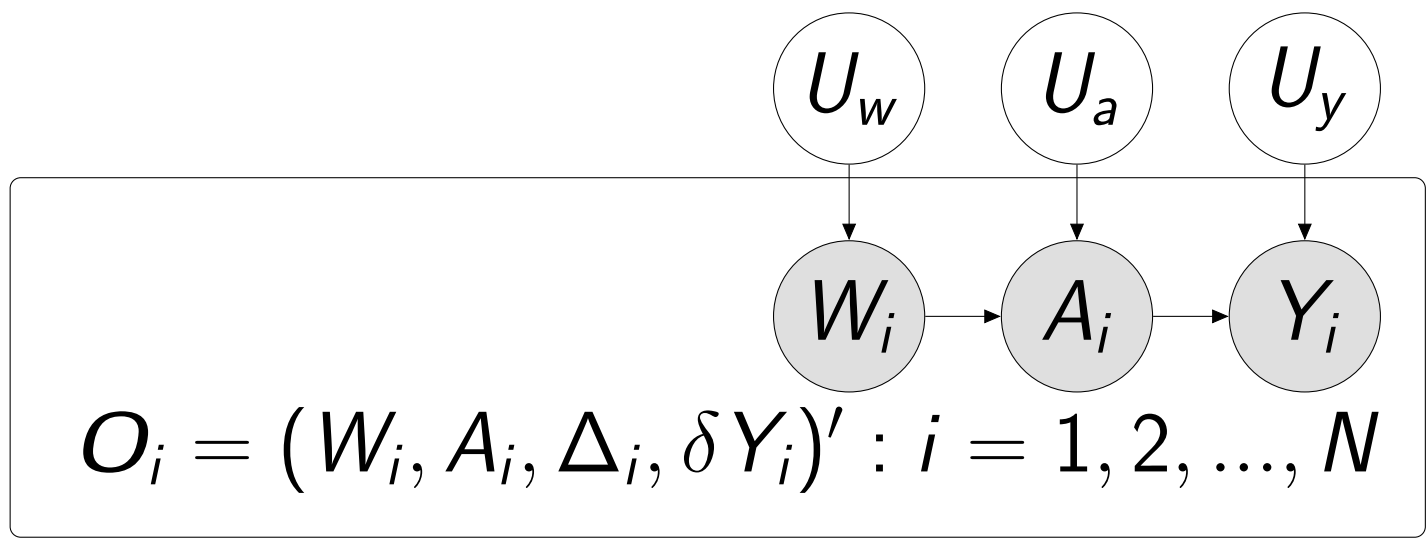


Figure: Directed Graphical Diagram for PSCM

- W : Baseline Covariates (Endogenous)
- A : Assigned Treatment (Endogenous)
- Y : Patient's Outcome (Endogenous)
- $U = (U_w, U_a, U_y)$: Unobserved (Exogenous) variables.
- Δ : An indicator of the patient's final outcome.

- Structural Equations:

$$\begin{aligned} W &= \phi_W(U_w) \\ A &= \phi_a(W, U_a) \\ Y &= \phi_Y(W, A, U_y) \\ \delta Y &= \phi_Y(W, A = 1, U_y) - \phi_Y(W, A = 0, U_y) \end{aligned}$$

- Patient's Causal Estimand:

Average Treatment Effect (ATE): $\mathbb{E}[Y_1] - \mathbb{E}[Y_0]$

Equivalence: The assumption of an unconfounded mechanism in the NRCM $\iff$ PSCM's assumption of the unmeasured factor U_A being independent of latent U_y conditioning on W .

Roadmap

- Describe the observed data and the data-generating experiment.
- Specify a probability model for the observed data.
- Specify an estimator of the target estimand, respecting the statistical model.
- Specify the target estimand of the observed data distribution.
- Specify an estimator of the sampling distribution of the estimator of target estimand for statistical inference.
- Interpret results, sensitivity analysis.

Statistical Learning Algorithm

Step-1 : Probabilistic Model on RWD

- Suppose the study draws n i.i.d copies of $O \sim \mathcal{P}_0$.
 $\mathcal{P}_0$: The unknown "True" probability distribution.
 $\mathcal{M}^{obs}$: Sets of all probability distributions of O . ($\mathcal{P}_0 \in \mathcal{M}^{obs}$)
- Augment $\mathcal{M}^{obs}$ with knowledge about data generating process.
- Find a function $\Psi : \mathcal{M}^{obs} \rightarrow \mathbb{R}^d$

Step-2: Causal World (Counterfactuals)

- Define the [Research Question](#).
- Counterfactual Y_{0i} and Y_{1i} : i th patient's outcome of receiving active and control drug respectively.
- $\mathcal{M}^C$ denotes the set of all possible probability distributions of joint random vector (O, U)
- Translate question into a causal quantity Ψ^C in terms of counterfactual or potential outcomes.
For example, select ATE as the causal quantity, then $\Psi^C = \mathbb{E}[Y_1] - \mathbb{E}[Y_0]$

Step-3: Connect the Causal World and Real World

- Assess the identifiability:** Express the causal quantity Ψ^C in terms of a "target estimand" Ψ^{obs} (a function of true distribution $\mathcal{P}_0$).
- Search for a "target estimand" $\Psi^{obs} : \Psi^C = \Psi^{obs}$ under appropriate causal assumptions.
- Find $\Psi^{obs} = \mathbb{E}_{W,0} [\mathbb{E}_0[Y|A = 1, \Delta = 1, W] - \mathbb{E}_0[Y|A = 0, \Delta = 1, W]]$
- Identification of Ψ^C requires assumption of a) [Randomization](#), b) [Missing at Random](#).
"Are this non-testable assumptions plausible?"
- Commit to the selected target estimand** if assumptions are reasonable to Ψ^C .

Step-4: Inference - Return to the Real World

- Find statistical estimators $\hat{\Psi}$ for $\Psi^{obs}(\mathcal{P}_0)$.
- Candidates of $\hat{\Psi}$:** Parametric g-computation, Matching estimators, Doubly robust estimators, Inverse Probability Weighting, TMLE.
- Specify methods to estimate standard error and confidence interval (by exact or MCMC methods).
- Use a hierarchy of result interpretation with increasing strength of assumptions to augment the output with a causal interpretation.

Conclusion: Current challenges and opportunities

Challenges of data ethics and regulations

- Clinical research incorporating external controls or borrowing strength from external data has rarely been leveraged to substantiate efficacy claims.
- The road map and innovations still lack uptake in generating evidence for regulatory considerations.
- The two-stage propensity score paradigm, which separates the design and analysis stages while conserving outcome masking until the final analysis [Yue et al. \(2014\)](#), appeals to regulators since it closely resembles traditional RCTs.
- This approach necessitates sponsors specifying a major portion of the study protocol and statistical analysis strategy in advance.
- Prospective specifications and result masking are widely acknowledged ideas among regulators as strategies to control biases and the inflating of Type I error rates.

Statistical challenges and open problems

- Methodologies for enhancing inference output are still being developed.
- Feature selection and testing procedures are critically challenging problems when the dimension of covariates is polynomial or exponential in order of sample size ($p \gg n$).
- Employing the nonparametric Bayesian approach for choosing the class of probability distributions ($\mathcal{P}_0$) and functional estimation in structural equations in the presence of latent (exogenous) variables could be an open research challenge.

References

Chen, J. (2020), 'The current landscape in biostatistics of real-world data and evidence: Clinical study design and analysis', *Statistics in Biopharmaceutical Research* pp. 29–42.
URL: <https://doi.org/10.1080/19466315.2021.1883474>

Ho, M. (2021), 'The current landscape in biostatistics of real-world data and evidence: Causal inference frameworks for study design and analysis', *Statistics in Biopharmaceutical Research* pp. 43–56.
URL: <https://doi.org/10.1080/19466315.2021.1883475>

Yue, L. Q., Lu, N. & Xu, Y. (2014), 'Designing premarket observational comparative studies using existing data as controls: Challenges and opportunities', *Journal of Biopharmaceutical Statistics* pp. 994–1010.
URL: <https://doi.org/10.1080/10543406.2014.926367>