Noisy-Or Risk Allocation: A Probabilistic Model for Attributable Risk Estimation

Amelia J. Averitt, MPH MA PhD[1,2]; Adler Perotte, MD MA[1]

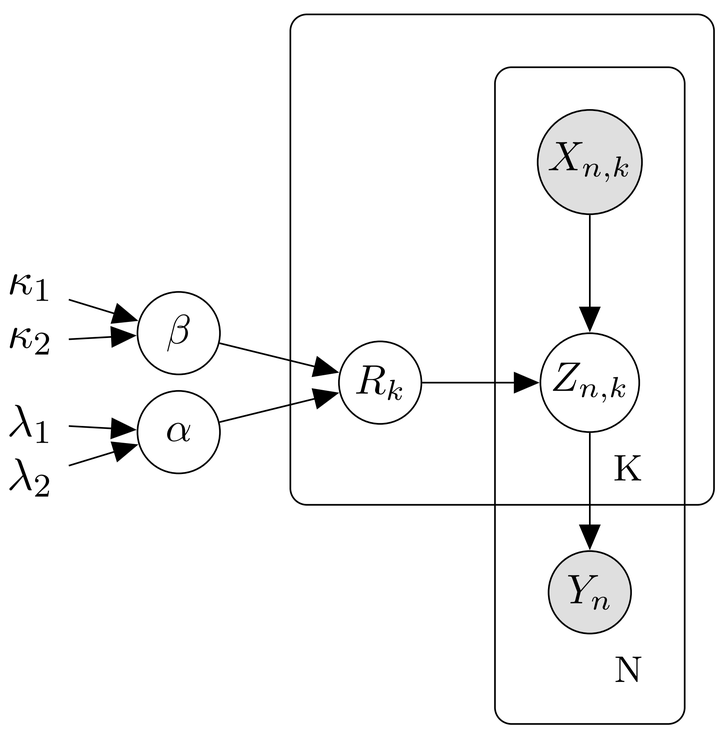
[1] Columbia University, Department of Biomedical Informatics. New York, NY. [2] Regeneron Genetics Center. Tarrytown, NY.

Introduction

Attributable risk (AR) is the proportion of an outcome in a population that could be prevented by elimination of a causal exposure from the population.[1] In the high-dimensional setting, typical methods of AR estimation include the (i) calculation of excess risk [2]; (ii) approximation using disproportionality methods, such as risk ratios (RR) and Gamma Poisson Shrinker (GPS); and (iii) regression-based methods. However, none of these methods are able to estimate global ARs, predict outcomes, and estimate ARs of exposures at the individual-level.

This research proposes the Noisy-Or Risk Allocation (NORA) model for high-dimensional AR estimation from observational data. NORA is a multivariate, latent-variable model with a likelihood that captures the notion of causal independence. Unlike comparator methods, NORA supports both global inferences of risks and local inferences of risks and outcomes.

The Model



$$K = \text{The number of exposures} \tag{1}$$

$$N = \text{The number of subjects} \tag{2}$$

$$Z_{n,k}|X_{n,k}, R_k \sim Bernoulli(X_{n,k}R_k) \tag{3}$$

$$R_k|\alpha, \beta \sim Beta(\alpha, \beta) \tag{4}$$

$$p(Y_n|Z_{1:K}) \sim Bernoulli(1 - \prod_{k=1}^{K}(1 - Z_{n,k}) \tag{5}$$

Key Takeaways

- NORA infers ARs of causal exposures for a single binary outcome.
- Simulations suggest that NORA may be more robust to confounding than logistic regression.
- NORA may support causal reasoning at the patient-level with outcome predictions and causal estimation and at the population-level informing public-health with estimates of risks across the entire population.

[1] Leviton, A. *Definitions of Attributable Risk. Am J Epidemiol Sep 1;98(3):231–231. (1973).*

[2] Levin, M.L. *The Occurrence of Lung Cancer In Man. AUICC, 9, 531-541. (1953).*

[3] Liu Z, et al. *The world-wide incidence of ... HIV Med, 19(5):355–64. (2018).*

Simulation

To assess correctness the extent to which NORA is robust to bias, we simulated a confounded causal system for Myocardial Infarction (see file in MS Teams) in which the true risk (highest) is High Cholesterol (HC). NORA and L1 logistic regression (L1) were applied to learn ARs in the presence of observed and unobserved confounders.

Table 1: Estimated Risk of High Cholesterol by Model (True=0.30)

	Unobserved Confounders	Observed Confounders	% Bias
NORA	0.18	0.30	37.1%
L1	0.46	0.82	78.9%

Experimentation

We additionally applied NORA to 10 outcome-exposure cohorts from the NewYork-Presbyterian EHR (see Table 2 for outcomes). ARs were estimated for these causal systems using (i) NORA; (ii) L1; (iii) excess risk by Levin 1953; (iv) RR; and (v) GPS. We conducted a 3-part evaluation.

(1) Local inference of the outcome was evaluated through the predictive performance of a held-out dataset (NORA and L1 only.) The results (Table 2) indicate that NORA has competitive predictive ability to L1.

Table 2: Area Under the Receiver Operating Curve for NORA and L1

	DIC	Glaucoma	Hearing Loss	Heart Failure	Kaposi sarc.	Mucositis (v drugs)	Renal Imp.	D/O Spleen	Hypothyroidism	Mucositis (v proc.)
NORA	**0.89**	**0.70**	0.51	**0.80**	**0.80**	0.53	**0.82**	**0.65**	0.56	0.59
L1	0.78	0.70	**0.63**	0.80	0.56	**0.66**	0.80	0.50	**0.64**	**0.61**

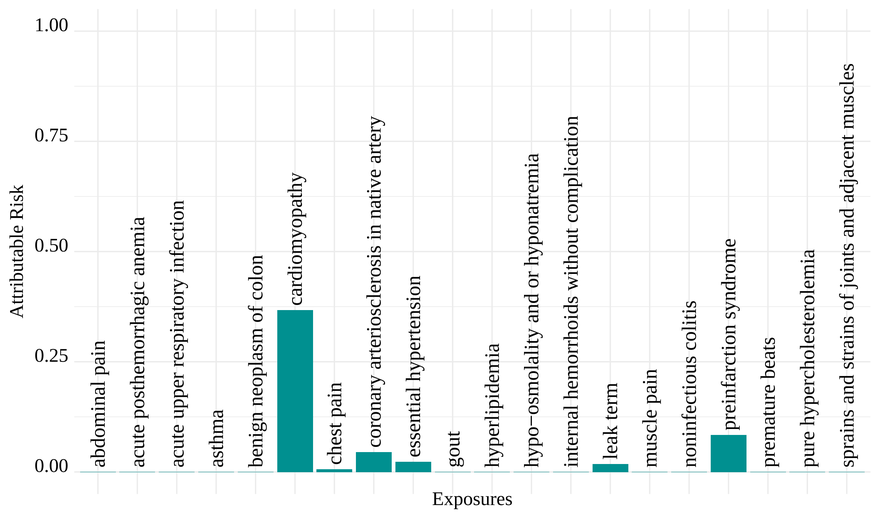
(2) Global inference of the exposures was evaluated by comparing the gold-standard, real-world AR of HIV for the outcome of Kaposi sarcoma [3] with the model-based AR estimates from NORA and comparators. Only NORA yielded an AR estimate in the correct order of magnitude (Table 3).

Table 3: AR Estimates of HIV for Kaposi sarcoma

	Gold-Standard	NORA	L1	Levin 1953	RR	GPS
HIV	0.0048	0.0070	0.7872	0.2022	0.9566	0.9603

(3) Local inference of the exposures was evaluated with an inspection of ARs & exposures for one individual with heart failure (HF). The posterior distribution over Z_n determines the probability that an exposure is a cause of the outcome given the remaining latent variables. The NORA-estimated high-AR exposures are known risk-factors of HF and are biologically sensible.

Figure 1: The Average AR of Exposures for a Single HF Patient



Funding acknowledgement: R01LM009886-10 and T15LM007079