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

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

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

Research Category (please highlight or circle which category best describes your research)

Methodological research

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 if there are (i) no interactions between causal exposures and (ii) all other effects of exposures are removed. (2–4) ARs support the inference of whether a given outcome was *caused* by a particular exposure. (5) Such an assessment can be made at the population level (*global inference*) or the individual level (*local inference*). (6,7) Ideally, estimation of AR would be based on knowledge of the relevant causal graph, in which relationships between exposures, outcomes, and confounders are made explicit. But in the setting of many potential exposures – the *high-dimensional setting* – causal graph construction may be infeasible. As an alternative, confounders to AR estimation may be controlled for through propensity-score modeling. However, this approach may be inefficient when estimating AR for many exposures, and does not lend itself to making local inferences about a particular patient. To estimate the ARs of many exposures simultaneously without knowledge of the causal graph, the primary question is one of model specification. In this work, we explore a particular model specification for estimating attributable risks in the context of unstructured binary exposures and outcomes. In this setting, typical methods of AR estimation include the (i) calculation of excess risk by the Levin, 1953 definition (8); (ii) approximation using disproportionality methods for signal detection such as RRs or the Gamma Poisson Shrinker (GPS) (1,9–11); (iii) and regression- based methods such as Penalized Logistic Regression (PLR) (25–29). The Levin Definition and approximation by disproportionality methods, make interpretable global estimates of ARs, but lack inferences for individuals. Furthermore, these methods are univariate and cannot account for confounding. Regression-based methods are powerful tools for the prediction of individual-level outcomes, but may often result in global estimates that are unstable or lack interpretability as AR estimates.

This research proposes the Noisy-Or Risk Allocation (NORA) model. NORA is a multivariate latent variable model with a likelihood that captures the notion of causal independence. Unlike comparator methods, NORA is able to estimate global ARs, predict outcomes, and estimate ARs of exposures at the individual-level.

THE MODEL. NORA is a Bayesian, probabilistic model that supports AR estimation of an uncertain causal system in which many potential risk factors exist. Let N be the number of patients and K be the number of unique exposures. $X_{n,k}$ is a binary indicator of exposure k for patient n; $Z_{n,k}$ is a binary indicator of activation of exposure k for patient n; R_k is the AR of exposure k; and Y_n is a binary indicator of outcome for patient n. Activation of an exposure is defined as a binary variable representing whether that exposure is a cause of the outcome for patient n. In other words, for an exposure $X_{n,k}$ to contribute to the outcome Y, it must be present ($X_{n,k} = 1$) and activated ($Z_{n,k} = 1$). The activation of the k^{th} risk for the n^{th} person is given by $Z_{n,k}$, which is dependent on both the presence ($X_{n,k}$) and the risk (R_k) of the exposure, k.

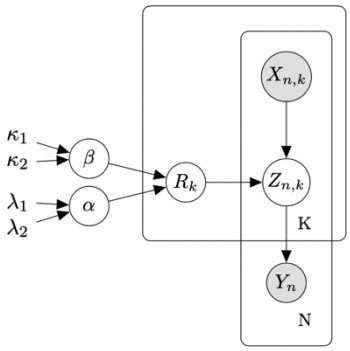

Figure 1:
Noisy-Or Risk Allocation Model

The model is predicated upon the Noisy-Or Gate, a model which expresses the conditional probabilities of one or more binary exposures on a single binary outcome. (30,32–34) The modeling assumptions of NORA are intuitive and many causal problems of interest can be distilled into a binary representation of the data. The exposures are assumed to affect the outcome independently, a property known as causal independence. (35,36). This assumption means that the probability of surviving an outcome given an exposure, is independent of the probability of surviving, given other exposures.

EXPERIMENTATION. To support our understanding of NORA's ability to support AR estimation, we conducted two studies. A simulation in which we are able to assess correctness of implementation and the extent to which NORA is resilient to confounding; and an investigation into the ability of NORA to recover known, clinically-meaningful causal relationships from noisy observational data, using electronic health record (EHR) data from NewYork-Presbyterian Hospital. For both studies, the model was learned by Gibbs Sampling with an additional Metropolis-Hastings sequence to estimate the priors on the risks (R). Simulation. We simulated data according to the toy causal system represented in Figure 2, wherein High Cholesterol is the exposure with the greatest risk (*true cause*) for the outcome, myocardial infarction (MI). We hypothesized that we would recover

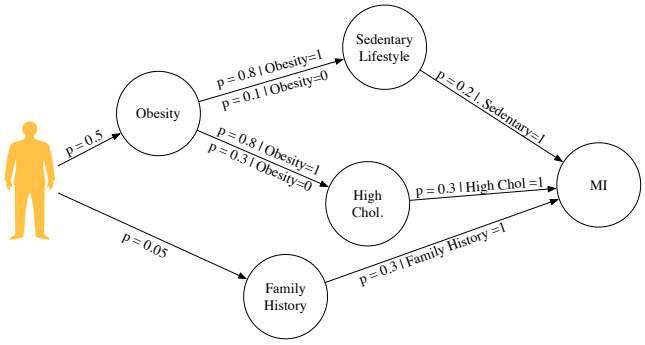

Figure 2: *NORA Simulation Schema*

the risk of MI associated with High Cholesterol within a small margin of error when all variables are observed despite disregarding knowledge of the causal graph. To evaluate robustness to confounding, we will compare the risk estimates of exposures on the outcome, as determined by the NORA model and L1 logistic regression (LR). Both NORA and LR will be applied to the simulated data twice; (i) the Main Effect Model, in which the confounding variables (obesity and sedentary lifestyle) are not observed; and (ii) the Adjusted Model, in which the confounding variables are observed. Application to Clinical Data. To determine NORA's utility in real-world practice, we applied the model to ten different observational cohorts. Outcomes of our causal system include -- disseminated intravascular coagulation (DIC), glaucoma, hearing loss, heart failure, Kaposi sarcoma, mucositis, renal impairment, disorder of the spleen, and hypothyroidism. ARs were estimated for the causal systems within the EHR using NORA; L1- regularized Logistic Regression (L1); the Levin-AR calculation; AR estimation using DPAs, including the RR and GPS. We present a three-part evaluation. (i) *The local inference of the outcome*, which is assessed through the predictive performance of a held-out dataset (NORA and L1 only). (ii) *The global inference of the exposures*, which is assessed by comparing the gold-standard, real-world AR estimates with the model-based AR estimates. We present an analysis of the exposure of HIV for the outcome of Kaposi sarcoma. (iii) *The local inference of the exposures*, which is assessed through an inspection of high-AR exposures for an individual with an outcome of interest. We present an analysis of one individual with the outcome of heart failure.

RESULTS. Simulation. NORA, modeling main-effects only, resulted in 1000-trial average for the risk of MI from High Cholesterol of 0.177; with full adjustment for confounding, NORA found the 1000-trial average for attributable risk of MI from High Cholesterol to be 0.298 (Truth = 0.30); LR, when modeling main effect only found the attributable risk of MI from High Cholesterol to be 0.459; and when modeled with all confounders, resulted in a probability of 0.821. When the backdoor-path/confounders were unobserved, the risk estimate of High Cholesterol from the NORA model changed 37.1%, versus 78.9% when estimated from LR. Application to Clinical Data. (i) *The local inference of the outcome.* The results of this evaluation are summarized by the area under the receiver operating curve (AUROC). Across all outcomes, NORA had a higher average AUROC of 0.6817 as compared to L1 with an AUROC of 0.6669 (data not shown for brevity). (ii) *The global inference of the exposures.* A 2018 article by Liu, et al reported the AR of HIV for Kaposi sarcoma to be 0.0048. (39) For each method, the estimated AR for HIV is presented in Table 1. Estimates from L1, the Levin-AR calculation, and GPS are extremely high, only the estimate from NORA is the correct order of magnitude. (iii) *The local inference of the exposures.* Results are shown in Figure 3. This heart failure patient had twenty unique exposures. Of these, cardiomyopathy (0.367), preinfarction syndrome (0.084), coronary arteriosclerosis in native artery (0.045) were the highest AR exposures for this patient. These exposures are known risk-factors of heart failure and are biologically sensible. Though this patient had other exposures, such as abdominal pain and sprains and strains of joints and adjacent muscles, these do not have a causal relationship with the outcome and have an estimated AR of near zero.

	Liu, et al	NORA	L1	Levin AR	RR	GPS
HIV	0.0048	0.0070	0.7872	0.2022	0.9566	0.9603

Table 1: *AR estimates of HIV for Kaposi sarcoma vs gold-standard estimate*

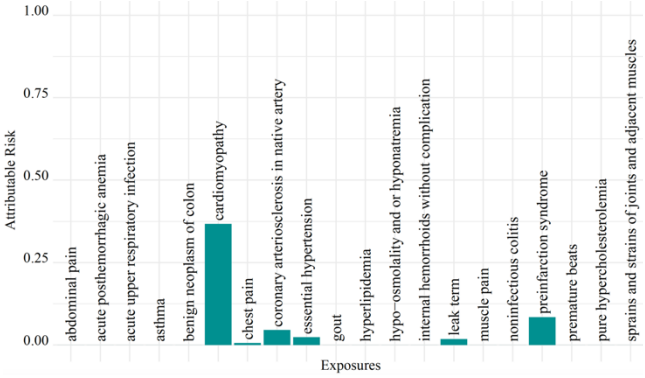

Figure 3: *The average AR of each exposure for a single heart failure patient.*

CONCLUSIONS. These results demonstrate that NORA is able to recover known, clinically-meaningful causal relationships with similar or better performance than the state-of-the-art. Furthermore, simulations suggest that NORA may be more robust to confounding than comparator methods. To our knowledge, NORA is the only method that is able to support local and global inferences. This may support causal reasoning at the global-level informing public-health with local estimates of risks across the entire population.

References

1. ROTHMAN KJ. CAUSES. Am J Epidemiol [Internet]. 1976 Dec [cited 2020 Jan 31];104(6):587–92. Available from: http://www.ncbi.nlm.nih.gov/pubmed/998606
2. LEVITON A. DEFINITIONS OF ATTRIBUTABLE RISK. Am J Epidemiol [Internet]. 1973 Sep 1 [cited 2020 Jan 17];98(3):231–231. Available from: https://academic.oup.com/aje/article/167532/DEFINITIONS
3. Miettinen OS. PROPORTION OF DISEASE CAUSED OR PREVENTED BY A GIVEN EXPOSURE, TRAIT OR INTERVENTION. Am J or Epidemiol [Internet]. 1974 [cited 2020 Mar 26];99(5). Available from: http://www.med.mcgill.ca/epidemiology/hanley/c609/material/Miettinen1974.pdf
4. MARKUSH RE. LEVIN'S ATTRIBUTABLE RISK STATISTIC FOR ANALYTIC STUDIES AND VITAL STATISTICS. Am J Epidemiol [Internet]. 1977 May [cited 2020 Mar 26];105(5):401–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/860704
5. Rosen DA. In Defense of a Probabilistic Theory of Causality [Internet]. Vol. 45, Philosophy of Science. The University of Chicago PressPhilosophy of Science Association; 1978 [cited 2017 Dec 12]. p. 604. Available from: https://www.jstor.org/stable/186972
6. Rockhill B, Newman B, Weinberg C. Use and misuse of population attributable fractions. Am J Public Health [Internet]. 1998 Jan [cited 2020 Jan 31];88(1):15–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9584027
7. Greenland S, Robins JM. Reviews and Commentary CONCEPTUAL PROBLEMS IN THE DEFINITION AND INTERPRETATION OF ATTRIBUTABLE FRACTIONS [Internet]. AMERICAN Journal of Epidemiology Formed. 1988 [cited 2020 Jan 31]. Available from: https://cdn1.sph.harvard.edu/wp-content/uploads/sites/343/2013/03/conceptual-problems-in-the-definition-and-interpretation-of-attributable-fractions.pdf
8. LEVIN ML. The occurrence of lung cancer in man. Acta Unio Int Contra Cancrum [Internet]. 1953 [cited 2020 Jan 31];9(3):531–41. Available from: http://www.ncbi.nlm.nih.gov/pubmed/13124110
9. Crowson CS, Therneau TM, Michael O'fallon W. Attributable Risk Estimation in Cohort Studies [Internet]. 2009 [cited 2020 Feb 3]. Available from: https://www.mayo.edu/research/documents/biostat-82pdf/doc-10027843
10. DuMouchel W, Pregibon D. Empirical bayes screening for multi-item associations. In: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '01 [Internet]. New York, New York, USA: ACM Press; 2001 [cited 2020 Feb 11]. p. 67–76. Available from: http://portal.acm.org/citation.cfm?doid=502512.502526
11. Dumouchel W. Bayesian Data Mining in Large Frequency Tables, with an Application to the FDA Spontaneous Reporting System. Am Stat [Internet]. 1999 Aug [cited 2020 Feb 11];53(3):177–90. Available from: http://www.tandfonline.com/doi/abs/10.1080/00031305.1999.10474456