3416

Generalizing and Transporting Diagnostic Test Performance to a Target Population

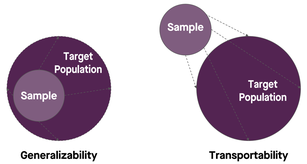
JSM 2023
August 5-10, 2023
Toronto, Canada

Yongqi Zhong, MB, MPH, PhD,[1] James Dai, PhD,[1] Nan Zhang, PhD,[1] Ying Luan, MBA, PhD[1]
[1]GRAIL, LLC, Menlo Park, CA, USA

INTRODUCTION

- Sensitivity, specificity, and positive/negative predictive values (PPV/NPV) are common metrics for evaluating the performance of a diagnostic test.
- Classic calculations of these metrics do not consider the causal relations between the gold standard condition and the diagnostic test. Choi (1997) used causal modeling approaches to show that sensitivity and specificity could be a function of the prevalence of a gold standard condition depending on the causal relations.
- However, diagnostic test performance is usually evaluated in a case-control study. Such performance metrics from case-control studies might or might not be quantitatively generalized to a target population. Further, with advances in causal inference, such performance metrics may be transported (extrapolated) to a new population.
  - For example, the performance of Galleri®, a multi-cancer early detection test, was clinically evaluated in the Circulating Cell-free Genome Atlas (CCGA) Study, a case-control study in the United States. Investigators want to generalize its performance in the target population where the case-control study was sampled from (i.e., general population in the United States), and transport (extrapolate) the performance to a new target population, which is different from the source population of the case-control study (e.g., symptomatic patients in the United Kingdom).
- Generalizability and transportability
  - Generalizability: Sample is a subset of a well-defined target population.
  - Transportability: Sample is external to a well-defined target population (i.e., sample is not a subset but could be partially overlapped in the distribution with the target population).

Figure 1. Illustration of Generalizability and Transportability



OBJECTIVES

- Propose marginally adjusted sensitivity, specificity, PPV, and NPV for a diagnostic test that allow adjusting for characteristics in the sample.
- Illustrate and evaluate generalizability and transportability of the marginally adjusted performance metrics from the sample to a target population using simulations in a hypothetical setting.

DIAGNOSTIC TEST PERFORMANCE CAN BE GENERALIZED AND TRANSPORTED TO A TARGET POPULATION OF INTEREST WITH MARGINAL ADJUSTMENT

KEY METHODS

Estimands: Diagnostic Test Performance With Marginal Adjustment

- Setup: Consider a binary true disease status (X), a binary diagnostic test result of the disease (Y), and a set of characteristics (C) that causes X and Y. Selection of the sample denotes as $S=1$.
- Sensitivity and specificity: $P(Y=1|X=1)$ *and* $P(Y=0|X=0)$.
  - Inspired by Robin's g-formula, the marginal adjusted sensitivity and specificity can be expressed as $P(Y|X)=E_C[P(Y|X,C)P(C)]$ under causal identifiability assumptions described below. These expressions are analogous to causal adjustment for confounders in the exposure-outcome analysis.
  - Accounting for the difference in these characteristics allows generalizability and transportability of the sample to a target population; that is, sensitivity and specificity in the target population $P(Y|X,S=0)$ are estimated with the information in the sample $P(Y|X,S=1)$.
- PPV and NPV: $P(X=1|Y=1)$ *and* $P(X=0|Y=0)$.
  - $P(X|Y,S=1)=P(Y|X,S=1)P(X|S=1)/\Sigma P(Y|X,S=1)P(X|S=1)$, where $P(Y|X,S=1)$ is the marginally adjusted sensitivity and specificity in the sample.
  - To generalize/transport the PPV and NPV to a target population $P(X|Y,S=0)$, causally identified $P(Y|X,S=0)$ and prevalence of disease status $P(X|S=0)$ are required.

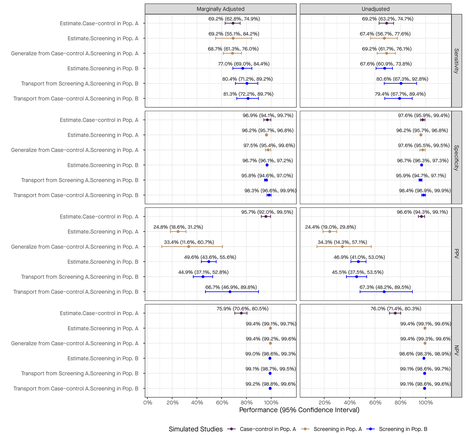
Estimators Used in the Hypothetical Example

- Marginally adjusted performance metrics
  - Inverse probability treatment weighted (IPTW) sensitivity and specificity were estimated with cancer status as the exposure (X), diagnostic test results as outcome (Y), and characteristics (C) as covariates. Empirical prevalence of cancer was used for PPV and NPV estimations.
- Generalizability
  - Inverse probability of sampling weights (IPSW) on the sample selection was estimated with cancer status (X), cancer stage (Z), and covariates (C).
- Transportability
  - Inverse odds of sampling weights (IOSW): Similar to the IPSW.
- Weights were estimated by logistic regressions. Performance metrics were estimated by marginal logistic regressions with IPTW (for marginal adjustment), IPSW (for generalizability), or IOSW (for transportability) as well as the products of IPTW and IPSW, or IPTW and IOSW (for marginal adjustment with generalizability or transportability).

KEY RESULTS

Performance in the Hypothetical Example

Figure 2. Estimated, Generalized, and Transported Performance Metrics in Three Simulated Studies From Two Populations



95% confidence intervals were obtained from 200 nonparametric bootstrapping.

- In this hypothetical setting, we assume both gold standard disease conditions and diagnostic test results were known at the same time, which may not reflect the actual setting in cancer screening. For illustration purposes, we thereafter use test sensitivity as episode sensitivity in actual cancer screening studies.
- Marginally adjusted performance metrics vs unadjusted metrics
  - In Population A, differences in the estimated performance between marginally adjusted and unadjusted metrics are trivial in both case-control (<1%) and screening studies (<1.8%).
  - In Population B, such differences are similar to Population A (differences<2.7%), except for sensitivity (difference=9.4%): adjusted 77.0% (69.0%, 84.4%) vs unadjusted 67.6% (60.9%, 73.8%).
- Generalized and transported performance metrics
  - The generalized and transported performance metrics with marginal adjustment are similar to the unadjusted metrics (differences<1.9%).
  - Generalized performance in Population A (Case-control → Screening)
    - The generalized performance metrics are similar to the estimated performance (differences<1.3%), except for PPV (difference=8.6%): estimated 24.8% (18.6%, 31.2%) vs generalized 33.4% (11.6%, 60.7%).
  - Transported performance in Population B
    - Screening in Population A → Screening in Population B:
      - Transported sensitivity is larger than the estimated sensitivity (difference=3.4%), while the transported PPV is smaller than the estimated PPV (difference=4.7%).
      - Transported specificity and NPV using screening study in population is similar to the estimated (differences<1%).
    - Case-control in Population A → Screening in Population B
      - The differences between transported metrics using case-control study in Population A and the estimated metrics are larger than the transported metrics using screening study in Population B.

Empirical Bias and Mean Squared Errors (MSE)

Table 1. Bias and MSE of the Generalized and Transported Performance

Adjustment	Target Population (Screening Sample)	Generalized/Transported	Bias				MSE			
			Sensitivity	Specificity	PPV	NPV	Sensitivity (10^{-5})	Specificity (10^{-5})	PPV (10^{-5})	NPV (10^{-6})
Marginally Adjusted	Population A	Generalized Performance from Case-Control Study in Population A	-0.18%	-0.02%	1.68%	<0.01%	5.45	13.28	8.05	2.49
	Population B	Transported Performance from Screening Study in Population A	<0.01%	0.01%	0.23%	<0.01%	6.82	3.08	2.54	14.48
		Transported Performance from Case-Control Study in Population A	-0.80%	-0.03%	4.50%	-0.04%	4.81	53.77	28.07	10.17
Unadjusted	Population A	Generalized Performance from Case-Control Study in Population A	0.11%	<0.01%	1.86%	<0.01%	2.99	13.02	7.71	1.35
	Population B	Transported Performance from Screening Study in Population A	0.12%	0.01%	0.10%	0.01%	10.72	2.78	3.01	22.67
		Transported Performance from Case-Control Study in Population A	-0.04%	<0.01%	4.63%	<0.01%	4.52	48.64	27.14	9.44

Bias and MSE were calculated from 1,000 Monte Carlo simulations.

- Empirical bias and MSE of generalized and transported performance metrics
  - $Bias=\frac{1}{N}\Sigma(\hat{P}-\bar{P}_0)$, where N is the number of simulations, $\hat{P}$ is the generalized/transported performance in the target population, and $\bar{P}_0$ is the estimated performance in the target population if test results are available.
  - $MSE=\frac{1}{N}\Sigma(\hat{P}-\bar{P}_0)^2$.
- Generalized performance (Case-control → Screening in Population A):
  - Small bias and MSE were found in both marginally adjusted and unadjusted performance of screening study generalized from case-control study in Population A.
- Transported performance
  - Screening in Population A → Screening in Population B has generally smaller bias and MSE than the performance of Case-control in Population A → Screening in Population B.
  - Transported performance with marginal adjustment has generally larger bias and MSE than the unadjusted performance.

CONCLUSIONS

- The newly proposed marginally adjusted metrics improves over naive methods (adjusting for two or three factors) by allowing all factors related to diagnostic test performance to be adjusted at the same time because diagnostic test performance is rarely dependent on only three factors.
- The marginally adjusted metrics further allow generalizing the performance of diagnostic tests to a larger population and transporting to a different one, empowering the design of new tests.
- Other methods of marginal adjustment, transportability, and generalizability could also be applied, such as outcome regressions and doubly robust approaches.
- Future studies for generalizing and transporting the performance of other tests need to be done (e.g., prognostic tests where test result is a cause of the prognostic condition rather than an outcome of such a condition).

References

1. Choi, B.C., 1997. Causal modeling to estimate sensitivity and specificity of a test when prevalence changes. *Epidemiology*, pp.80-86.
2. Barenboim, E. and Pearl, J., 2016. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27), pp.7345-7352.
3. Cole, S.R. and Stuart, E.A., 2010. Generalizing evidence from randomized clinical trials to target populations: the ACTG 320 trial. *American journal of epidemiology*, 172(1), pp.107-115.
4. Westreich, D., Edwards, J.K., Lesko, C.R., Stuart, E. and Cole, S.R., 2017. Transportability of trial results using inverse odds of sampling weights. *American journal of epidemiology*, 186(8), pp.1010-1014.

Disclosures

Research funded by GRAIL, LLC. YZ, JD, NZ, and YL are current employees of GRAIL, LLC, and hold stock in Illumina, Inc.

Acknowlegements

Funded by GRAIL, LLC. Graphic assistance provided by Prescott Medical Communications Group (Chicago, IL).

SUPPORTING INFORMATION

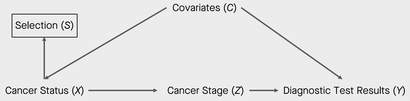
Causal Identifiability Assumptions

- Consistency and no interference: The disease condition is well-defined (gold standard). For each individual, there are not different versions of a disease status leading to different potential outcomes of diagnostic test results. The potential outcomes of diagnostic tests results for a given individual are not influenced by the disease status of other individuals. $P(Y)=P(Y|X=x)=P(Y=|X=x)$.
- (Conditional) Exchangeability
  - Within sample: Conditioning on the C in the sample, the potential outcome of diagnostic test results is independent of the observed disease status. $Y \perp\!\!\!\perp X|C$.
  - Between sample and target population (S-admissibility): Conditioning on the C, which accounts for the difference between sample and target population, the potential outcome of diagnostic test results is independent of the selection of sample. $Y \perp\!\!\!\perp S|C$.
- Positivity:
  - Within sample: In both disease and healthy groups, every stratum of the characteristics has at least one value of the outcome (e.g. diagnostic test results). $P(X=x|C=c)>0$ *for all* c *with* $P(C=c)\neq 0$.
  - Between sample and target population: $P(S=1|C=c)>0$ *for all* c *with* $P(C=c)\neq 0$.

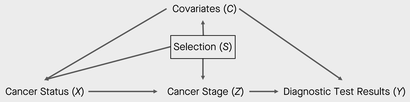
Simulation Setups and Characteristics

Figure S1. Causal Diagrams of Generalizability and Transportability for Hypothetical Example Simulations

A. Generalizing Case-control Study to Screening Study



B. Transporting Population A to Population B



- Variables
  - Covariates (C) are sex, age, smoking with prevalence (marginal probability) in Table S1.
  - Cancer Status $X \sim Binom(N, P(X))$, where $logit(P(X)) = \beta_0 + C\beta'$, β is the log odds ratios of covariates C in Table S1, and the β_0 is the balanced intercept for setting cancer prevalence listed in Table S1.
  - Cancer Stage (Z) follows a multinomial distribution with the prevalence in Table S1 among cancer patients.
  - Diagnostic Test Results (Y)
    - Sensitivity: $P(Y=1|X=1)=0.2+0.1*(cancer\ stage\text{-}1)+0.05*(age\ group\text{-}1)+0.1*ever\ smoked$.
    - Specificity: $P(Y=0|X=0)=0.8+0.02*(age\ group\text{-}1)+0.05*ever\ smoked$.
    - Probability of Diagnostic Test Positive: $P(Y=1)=|(X=1)\cdot sensitivity+|(X=0)\cdot(1\text{-}specificity)$.
- Population size $N=10^6$ was simulated for each population (A and B) with the parameters listed above and in Table S1.
  - Compared with Population A, Population B has more females, older age groups, more ever-smokers, and higher cancer prevalence with more late stages.
- Sampling for hypothetical studies
  - Case-control study in Population A ($N=500$): Stratified random sampling from Population A with 50% cancer (case, $N=250$) and 50% noncancer (control, $N=250$).
  - Screening study in Population A ($N=5,000$): Random sampling.
  - Screening study in Population B ($N=5,000$): Random sampling.
- Constructing analytic datasets
  - In generalizing and transporting performance estimations, test results *not* in the sample ($S=0$, screening studies) were set to unknown.
- Simulations and analyses were performed in the R statistical programming environment (v4.1.2).

Table S1. Population Parameter and Simulated Sample Characteristics in the Hypothetical Example

	Population A ($N=10^6$)				Population B ($N=10^6$)			
	Parameter		Sample		Parameter		Sample	
	Prevalence	Odds Ratio for Cancer Status	Case-control study N (%)	Screening Study N (%)	Prevalence	Odds Ratio for Cancer Status	Screening Study N (%)	
N			500	5,000			5,000	
Sex								
Female	50%	0.8	240 (48.0%)	2,520 (50.4%)	60%	0.8	3,002 (60.0%)	
Male	50%	Reference	260 (52.0%)	2,480 (49.6%)	40%	Reference	1,998 (40.0%)	
Age Group								
<50	20%	0.5	79 (15.8%)	1,028 (20.6%)	10%	0.5	528 (10.6%)	
50-64	50%	Reference	212 (42.4%)	2,504 (50.1%)	20%	Reference	990 (19.8%)	
≥65	30%	2.5	209 (41.8%)	1,468 (29.4%)	70%	2.5	3,482 (69.6%)	
Ever Smoked								
Yes	10%	5	117 (23.4%)	458 (9.2%)	30%	5	1,456 (29.1%)	
No	90%	Reference	383 (76.6%)	4,542 (90.8%)	70%	Reference	3,544 (70.9%)	
Cancer Status								
Presence	1.5%	-	250 (50.0%)	89 (1.8%)	3%	-	204 (4.1%)	
Absence	98.5%	-	250 (50.0%)	4,911 (98.2%)	97%	-	4,796 (95.9%)	
Cancer Stage								
1	30%	-	76 (30.4%)	26 (29.2%)	10%	-	20 (9.8%)	
2	40%	-	102 (40.8%)	31 (34.8%)	10%	-	24 (11.8%)	
3	20%	-	53 (21.2%)	23 (25.8%)	40%	-	79 (38.7%)	
4	10%	-	19 (7.6%)	9 (10.1%)	40%	-	81 (39.7%)	
Diagnostic Test Results								
Positive	-	-	179 (35.8%)	246 (4.9%)	-	-	294 (5.9%)	
Negative	-	-	321 (64.2%)	4,754 (95.1%)	-	-	4,706 (94.1%)	