A Targeted Simulation-Extrapolation Method for Evaluating Biomarkers Based on Emerging Technologies in Precision Medicine

Dong Wang[1*], Sue-Jane Wang[2], Joshua Xu[1], Samir Lababidi[3]

1. National Center for Toxicological Research, FDA 2. Center for Drug Evaluation and Research, FDA 3. Office of Chief Scientist, FDA

* dong.wang@fda.hhs.gov

Abstract

New technologies for novel biomarkers have transformed the field of precision medicine. On the other hand, in applications like liquid biopsy for early tumor detection, the misclassification rates of next generation sequencing (NGS) and other technologies have become an unavoidable feature of biomarker evaluations. As initial experiments are usually confined to specific technology choices and application setting, a statistical method that can project the performance metrics of other scenarios with different misclassification rates will be very helpful for planning further biomarker development and future trials. In this paper, we describe an approach based on an extended version of simulation extrapolation (SIMEX) to project the performance of biomarkers measured with varying misclassification rates due to technological or application choices when experimental results are only available for one specific setting. Through simulation studies for logistic regression and proportional hazards models, we show that the proposed method can be used to project the performance of the biomarker when switching technology or application setting with good precision. Similar to the original SIMEX, the proposed method can be implemented with existing software in a straightforward manner. A data analysis example is also presented using a lung cancer dataset and performance metrics of two gene panel based biomarkers studied by the Microarray and Sequencing Quality Control Consortia . The results demonstrate the feasibility to infer the potential implications of using a range of technologies or application scenarios for biomarkers with limited human trial data.

Introduction

The advancement of precision medicine has widely impacted both patient care and regulatory considerations with the companion diagnostic tests as a prominent example. Next generation sequencing (NGS) is widely used in this context. However, all sequencing platforms have their inherent error rates especially in applications like liquid biopsy. For example, Cohen et al. 4 described a noninvasive blood test, CancerSEEK, that can detect multiple types of cancer with a sensitivity of 69 to 98% (depending on the cancer type) and a 99% specificity. In the early phases of biomarker development, usually only one technological approach is used, which is not optimized for any one biomarker. It can be extremely beneficial to have the ability to infer performance when using another technological platform (gene panel, sequencing method, or dPCR) with different misclassification rates. This capacity can significantly facilitate the design for the subsequent trials and enhance confidence in clinical decision making based on those findings.

This need can be addressed by adapting the general framework of measurement error models. Among available approaches for correcting bias due to measurement error, SIMEX has been shown to be useful in a wide range of settings. Because it is based on simulation, SIMEX is relatively easy to implement. By adapting the SIMEX method, we propose a targeted SIMEX approach by which performance characteristics for technologies with differing misclassification rates can be inferred based on an existing study that has used a specific technology. This new approach can facilitate the efficient discovery of potential biomarkers and can aid in planning subsequent trials for precision medicine.

Materials and Methods

Error prone versions of biomarkers

Consider two technological or experimental approaches A and B, which could represent two different sequencing approaches, two separate technologies, or two different application scenarios. Suppose neither approach A nor approach B can measure the biomarker X accurately. Instead, the biomarker X_A has a misclassification matrix Π_A regarding X and biomarker X_B has a misclassification Π_B. For example:

$$\Pi_A = \begin{bmatrix} 0.95 & 0.4 \\ 0.05 & 0.6 \end{bmatrix}, \Pi_B = \begin{bmatrix} 0.98 & 0.2 \\ 0.02 & 0.8 \end{bmatrix},$$

where 0.95 and 0.98 are specificity, 0.6 and 0.8 are sensitivity for A and B respectively.

Statistical model using the biomarker

$$h(y_i) = \beta_0 + \beta_1 T_i + \beta_2 x_i + \beta_3 (T_i x_i)$$

where X is the biomarker (available as either X_A or X_B), T is the treatment indicator, y is the response, and h is a link function. The link function h can be the identity function, logistic link, or the Cox proportional hazard function. We are primarily interested in the biomarker-treatment interaction term β_3.

Targeted Simulation

Suppose we have clinical trial data with X_B but want to project the performance for the less accurate biomarker X_A. Let $\Pi_A = \Pi_B \Pi_{BA}$, where $\Pi_{BA} = \Pi_B^{-1}\Pi_A$. We can project the regression coefficient under X_A by generating M pseudo datasets. In each pseudo dataset we generate X_A by randomly misclassify X_B by Π_{BA}. The naïve estimator of the coefficients can be calculated with the pseudo X_A, and the targeted simulation estimator can be obtained by averaging the naïve estimators over the M pseudo datasets. The variance can be estimated with bootstrap

Targeted SIMEX

If we have data obtained with X_A but want to project the performance of the more accurate biomarker X_B, we can use an approach similar to SIMEX. Write the eigen decomposition $\Pi_{BA} = E\Lambda E^{-1}$, let $\Pi_{BA}^{\lambda} = E\Lambda^{\lambda}E^{-1}$. For a given value of λ we can generate pseudo data of x_{λ}^{*} by randomly misclassifying X_A with the matrix Π_{BA}^{λ}, and an estimator at λ can be obtained by averaging over a large number of pseudo datasets. With a suitable grid of λ, we can plot out the function of the estimator regarding λ. The T-SIMEX estimator for X_B is then obtained by extrapolate this function to λ=-1. This this illustrated in Figure 1.

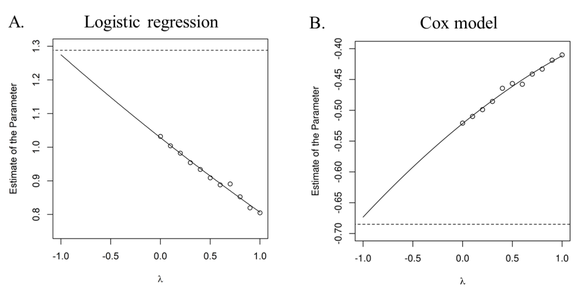

Figure 1. The quadratic approximations of the regression coefficient as a function of λ, extrapolated to -1 for the projection of the targeted approach.

Results and Discussion

Simulation study

We performed simulation studies to evaluate the proposed methods for logistic regression and survival analysis. The simulation design considered a two-arm experiment with 600 patients in which 300 patients were randomly assigned to each arm. The proportion of biomarker positive patients is φ=0.1, 0.3, or 0.5. We are interested in β_3, the coefficient for the biomarker-treatment interaction. For logistic regression, β_3 =1.0 or 1.5. It takes values of -0.4 or -0.8 for Cox proportional hazards model or Weibull regression. Table 1. shows the results for targeted SIMEX. The results for targeted simulation lead to similar conclusions, we omit it due to space constraint.

Table 1. The mean and empirical standard error (ESE) of the naïve estimator; the T-SIMEX estimator for the standard error of the naïve estimator (T-S.SE); and the mean, ESE, bootstrap standard error (BSE) for the T-SIMEX estimator.

		Logistic Model					
		Naïve estimator			Targeted SIMEX estimator		
φ	β	Mean	ESE	T-S.SE	Mean	ESE	BSE
0.3	1.5	1.288	0.457	0.464	1.286	0.633	0.625
	1	0.871	0.462	0.467	0.879	0.642	0.636
0.5	1.5	1.151	0.4	0.397	1.141	0.543	0.551
	1	0.77	0.406	0.403	0.767	0.561	0.565
		Cox proportional hazards model					
		Naïve estimator			Targeted SIMEX estimator		
φ	β	Mean	ESE	T-S.SE	Mean	ESE	BSE
0.1	-0.8	-0.657	0.381	0.385	-0.662	0.576	0.587
	-0.4	-0.335	0.365	0.368	-0.327	0.536	0.55
0.3	-0.8	-0.685	0.256	0.257	-0.679	0.359	0.36
	-0.4	-0.352	0.245	0.246	-0.361	0.339	0.341
0.5	-0.8	-0.637	0.224	0.223	-0.644	0.309	0.308
	-0.4	-0.32	0.216	0.215	-0.312	0.294	0.294
		Weibull regression model					
		Naïve estimator			Targeted SIMEX estimator		
φ	β	Mean	ESE	T-S.SE	Mean	ESE	BSE
0.1	-0.8	-0.645	0.378	0.382	-0.647	0.565	0.566
	-0.4	-0.313	0.363	0.368	-0.325	0.547	0.54
0.3	-0.8	-0.683	0.255	0.256	-0.685	0.363	0.361
	-0.4	-0.351	0.245	0.245	-0.358	0.344	0.348
0.3	-0.8	-0.63	0.223	0.222	-0.622	0.309	0.306
	-0.4	-0.312	0.215	0.214	-0.304	0.307	0.296

For both logistic regression and survival analysis, T-SIMEX estimator mostly coincided with that of the usual naïve estimator when X_B is observed. The T-SIMEX estimator for the standard error for the naïve estimator also performed well. Thus T-SIMEX can provide a detailed picture of how the biomarker might perform when switching methods for obtaining the measurement.

DISCLAIMER: The opinions expressed in this paper are those of the authors, and not necessarily reflect the position of U.S. Food and Drug Administration.

Application to a lung cancer dataset

We applied the proposed methods to a lung cancer data set using gene panels to identify subgroups of patients who might benefit from adjutant cisplatin/vinorelbine (ACT) treatment in early-stage non–small-cell lung cancer (NSCLC). The assumed error free biomarker (X) has an estimated coefficient β_3 regarding the interaction as -1.774 with a standard error of 0.714, which suggests a significant interaction between the treatment and the biomarker. For the misclassification rates, we used those for two gene panels studied in detail by FDA-led Microarray and Sequencing Quality Control (MAQC/SEQC) Consortia. Two gene panels (A and B) both has specificity of 1, while sensitivity was 0.838 for panel A and 0.955 for panel B. We then generated misclassified versions of the biomarker X corresponding to X_A and X_B, respectively.

Table 2. The mean and average standard error (ASE) of the naive estimator for β_3; the TS or the T-SIMEX estimator for the standard error (TSSE for targeted simulation or T-S.SE for T-SIMEX); and the mean, bootstrap standard error (BSE) for targeted simulation or T-SIMEX for analysis based on the lung cancer dataset.

Targeted Simulation				
Naïve estimator			Targeted Sim. Estimator	
Mean	ASE	TSSE	Mean	BSE
-1.163	0.571	0.571	-1.167	0.471
Targeted SIMEX				
Naïve estimator			T-SIMEX estimator	
Mean	ASE	T-S.SE	Mean	BSE
-1.499	0.64	0.629	-1.489	0.825

Conclusion

In this paper, we describe an approach to project biomarker performance both under an increased misclassification rate (targeted simulation) or reduced misclassification (targeted SIMEX). Our approach shares the advantage of SIMEX in that available software can be used to perform computations with straightforward programming for simulation and extrapolation. The misclassification rates are assumed to be known in our examples. However, even if the misclassification rates for different technologies are not precisely known, it is possible to use the approach described here to carry out exploratory studies to obtain a range of possible outcomes when planning future development and trials. As with the original SIMEX method, the performance for the proposed method may decrease as the difference in misclassification rates between technologies becomes very large. The original SIMEX and MC-SIMEX focus on targeting the estimates with error free covariates. Here, we extend it to scenarios with different magnitudes of measurement errors, including misclassification errors. By applying this approach, we can better understand the potential performance characteristics of biomarkers under different technological and application choices with limited clinical data.

Acknowledgements

We would like to thank Dr. Pamela Gallagher from the U.S. FDA Center for Devices and Radiological Health for valuable input regarding the conception and design of the study.