ETH zürich

BIOMEDICAL INFORMATICS

Estimation of Area Under Receiver Operating Characteristics Curve (AUCROC) in Matched Case-Control Studies

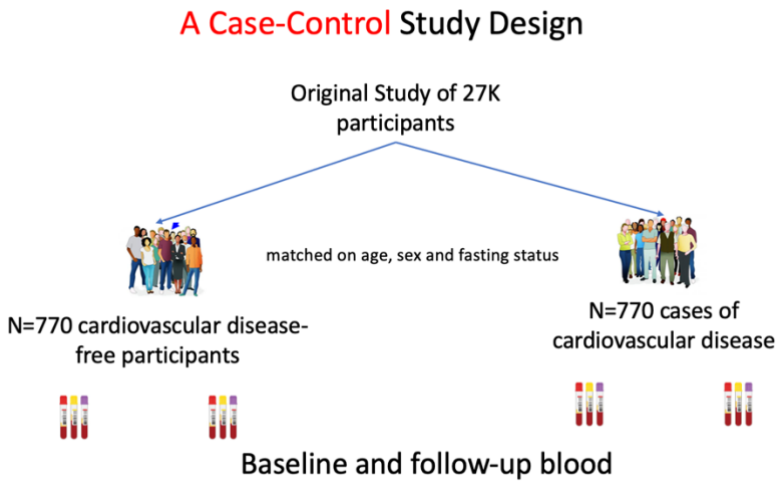
Master's Project Proposal

Matched case-control study design is ubiquitous in biomarker discovery studies, when cost of the assay is often substantial, and outcome of interest is rare. This design provides consistent estimate of parameter of interest while greatly increase cost-efficiency of a study. However, estimation of absolute risk in case-control studies is problematic. For this reason, AUCROC (a commonly reported measure of discriminatory performance of added biomarker) cannot be directly estimated from a case-control study. Subsequently, improvement in discrimination cannot be estimated directly when evaluating impact of a novel biomarker on risk of the outcome of interest.

Our search on google scholar for "we conducted a case-control study" returned more than 12K citations, while "AUCROC" OR "c-statistic" OR "c-index" returned 20K citations.

Two studies proposed to address this problem. Xu et.al. and Pepe et. al. stack parameter estimation in a parent study and a case-control study and use weighted AUC as a final step. However, the two approaches do not preserve a non-decreasing property of AUC for nested models and inverse probability weighting adds considerable variability of AUC estimate.

We developed a novel method of estimation of AUC in matched case-control studies and we need to publish it. We already worked on several practical examples demonstrating that our method produces valid estimate. In this project we need to expand to other practical and simulated examples and show how to incorporate testing-validation setting into developed method.



Project goals:
The goal of this project is to validate our novel method of AUCROC estimation in matched case control studies and extend to training-testing settings. Results of this project will be submitted as a manuscript to a peer-reviewed journal and code will be submitted as an R-package to CRAN.

Work Plan:

1. Literature review of AUCROC estimation in cohort and case-control studies.
2. Develop a novel method for AUCROC estimation in training only settings using explicit solution for AUCROC. Confirm results in simulations and in real data. Compare to Xu et.al. and Pepe et.al. methods.
3. Extend method developed in 2. for AUCROC estimation in training-validation settings. Confirm results in simulations and in real data. Compare to Xu et.al. and Pepe et.al. methods.
4. Investigate robustness of developed methods to violation of assumptions and ways to address this issue.

There will be the possibility to write a scientific publication in place of the thesis.

Prerequisites:

1. An interest in the subject matter

2. Experience in Python and/or R programming
3. Interest in developing novel methodology independently and implement ideas in algorithms

Grading:

- To receive a 6 the student must: meet the general goals and timeline of the thesis; work independently, demonstrate curiosity in the thesis topic and contribute her/his own ideas; communicate intermediate results clearly; write detailed and clear intermediate and final reports; document well project code; give a clear final presentation; work in exceptional and at a similar level expected for acceptance at leading international conferences or peer-reviewed journals.
- To receive a 5 the student must: meet the general goals of the thesis; work independently; write a detailed and clear final report; and give a clear final presentation; document well project code.
- To receive a 4 the student must: partially meet the general goals of the thesis; work somewhat independently; write a satisfactory final report; and give an understandable final presentation; document well project code

Advisors:	Dr. Olga Demler demlero@ethz.ch	ETH D-INFK/ BWH, Harvard Medical School
Supervisor:	Prof. Dr. Gunnar Rätsch	D-INFK

References:

Brentnall, A.R. and Cuzick, J., (2018). Use of the concordance index for predictors of censored survival data. Statistical methods in medical research, 27(8), pp.2359-2373.

Goldstein, L. and Langholz, B. (1992). Asymptotic theory for nested case-control sampling in the Cox regression model. The Annals of Statistics 20, 1903(1928). ISSN 0090-5364.

Langholz, B. and Borgan, Y. (1997). Estimation of absolute risk from nested case-control data. Biometrics 53, 767(774).

Pepe, M.S., Fan, J. and Seymour, C.W., (2013). Estimating the receiver operating characteristic curve in studies that match controls to cases on covariates. Academic radiology, 20(7), pp.863-873.

Pepe, M.S., Fan, J., Seymour, C.W., Li, C., Huang, Y. and Feng, Z., (2012). Biases introduced by choosing controls to match risk factors of cases in biomarker research. Clinical chemistry, 58(8), pp.1242-1251.

Prentice, R. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika 73, 1. ISSN 0006-3444.

Xu, H., Qian, J., Paynter, N.P., Zhang, X., Whitcomb, B.W., Tworoger, S.S., Rexrode, K.M., Hankinson, S.E. and Balasubramanian, R., (2019). Estimating the receiver operating characteristic curve in matched case control studies. Statistics in medicine, 38(3), pp.437-451.