The Journal of Open Source Software

riskclustr: Functions to Study Etiologic Heterogeneity

Emily C. Zabor[1]

1 Memorial Sloan Kettering Cancer Center

DOI: 10.21105/joss.01269

Software

- Review
- Repository
- Archive

Submitted: 18 February 2019
Published: 28 March 2019

Summary

Etiologic heterogeneity refers to the concept of subtypes of disease that are influenced by different risk factors. In cancer epidemiology, it is well known that many types of cancer demonstrate etiologic heterogeneity with respect to subtypes formed by genetic markers and/or other tumor characteristics. Likely the best known example of this is in breast cancer research, where subtypes are often formed based on immunohistochemical staining of estrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor receptor 2 (HER2). Risk factors, including patient characteristics such as age and body mass index as well as hormonal risk factors such as age at menarche, parity, and menopausal status, have been shown to have different relative risks for these disease subtypes (Gaudet et al., 2011; Kwan et al., 2009; Ma et al., 2010; Phipps, Malone, Porter, Daling, & Li, 2008; Yang et al., 2007). With the growing use of genetic testing, the search for disease subtypes will only become more common in cancer research, and across other disease areas. There may be three main goals in any study of etiologic heterogeneity: 1) quantifying the extent of etiologic heterogeneity for a given set of disease subtypes according to all known risk factors, 2) testing the etiologic heterogeneity of individual risk factors with respect to a given set of disease subtypes, and 3) identifying the most etiologically heterogeneous subtype solution from possibly high-dimensional disease marker data. The R (R Core Team, 2018) package `riskclustr` (E. C. Zabor, 2019) provides user-friendly functions to address all three areas of study.

In previous work we introduced a scalar measure that can be used to quantify the extent of etiologic heterogeneity of pre-defined disease subtypes based on known risk factors in the context of a case-control study (Begg et al., 2013). Later work generalized this methodology to the context of case-only studies, as researchers are often working in hospital settings without access to control subjects (Begg et al., 2014). The R package `riskclustr` introduces original functionality to implement these methods to estimate the extent of etiologic heterogeneity in case-control studies using the `d` function and in case-only studies using the `dstar` function.

Separately, we compared available statistical methods for the study of etiologic heterogeneity in case-control studies (Zabor & Begg, 2017). By unifying the notation of the different methods and employing simulations, we showed that the methods are all able to address two key research questions: 1) whether risk factor effects differ across subtypes of disease and 2) whether risk factor effects differ across levels of each individual disease marker of which the disease subtypes are comprised. Our research also showed that adaptation of polytomous logistic regression has statistical properties at least as good as the more sophisticated methods that have been proposed, provided that the number of disease markers is sufficiently small that the analysis is feasible (Zabor & Begg, 2017). In R polytomous logistic regression can be implemented using the `multinom` function from the `nnet` package (Venables & Ripley, 2002) or the `mlogit` function from the `mlogit` package (Croissant, 2018), but the additional calculations needed to perform an analysis

of etiologic heterogeneity are cumbersome. To facilitate use of this method the R package `riskclustr` introduces functions `eh_test_subtype` and `eh_test_marker` that first fit a standard polytomous logistic regression model using the `mlogit` function from the `mlogit` package (Croissant, 2018) and then perform additional calculations to address the two preceding questions regarding etiologic heterogeneity.

Finally, it is increasingly common for statistical or epidemiologic researchers to be confronted with high dimensional disease marker data and when disease subtypes are not pre-defined, this disease marker data can be clustered, and the optimally etiologically heterogeneous subtype solution can be identified. This methodology has been applied to breast cancer (Begg et al., 2015, 2013) and melanoma (Mauguen et al., 2017) and the statistical properties of the approach have been investigated using simulation studies (Zabor et al, under review). In R (R Core Team, 2018) unsupervised k-means clustering can be implemented using the `kmeans` function in the `stats` package (R Core Team, 2018), and the R package `riskclustr` includes a wrapper for the `kmeans` function that calculates the extent of etiologic heterogeneity using the `d` function for each cluster solution resulting from many random starts of the clustering algorithm, and returns the subtype solution that maximizes etiologic heterogeneity.

The R package `riskclustr` was designed for use by researchers in epidemiology and biostatistics, and the open-source software package includes user-friendly tutorials that include examples of how to use the various functions and cover details of all underlying statistical calculations.

Acknowledgements

The research was supported by the National Cancer Institute, awards CA163251 and CA167237. Additional funding support was provided by the Core Grant (P30 CA008748).

References

Begg, C. B., Orlow, I., Zabor, E. C., Arora, A., Sharma, A., Seshan, V. E., & Bernstein, J. L. (2015). Identifying etiologically distinct sub-types of cancer: A demonstration project involving breast cancer. *Cancer Med*, *4*(9), 1432–9. Journal Article. doi:10.1002/cam4.456

Begg, C. B., Seshan, V. E., Zabor, E. C., Furberg, H., Arora, A., Shen, R., Maranchie, J. K., et al. (2014). Genomic investigation of etiologic heterogeneity: Methodologic challenges. *BMC Med Res Methodol*, *14*, 138. Journal Article. doi:10.1186/1471-2288-14-138

Begg, C. B., Zabor, E. C., Bernstein, J. L., Bernstein, L., Press, M. F., & Seshan, V. E. (2013). A conceptual and methodological framework for investigating etiologic heterogeneity. *Stat Med*, *32*(29), 5039–52. Journal Article. doi:10.1002/sim.5902

Croissant, Y. (2018). *mlogit: Multinomial logit models*. Retrieved from https://CRAN.R-project.org/package=mlogit

Gaudet, M. M., Press, M. F., Haile, R. W., Lynch, C. F., Glaser, S. L., Schildkraut, J., Gammon, M. D., et al. (2011). Risk factors by molecular subtypes of breast cancer across a population-based study of women 56 years or younger. *Breast Cancer Res Treat*, *130*(2), 587–97. Journal Article. doi:10.1007/s10549-011-1616-x

Kwan, M. L., Kushi, L. H., Weltzien, E., Maring, B., Kutner, S. E., Fulton, R. S., Lee, M. M., et al. (2009). Epidemiology of breast cancer subtypes in two prospective cohort

studies of breast cancer survivors. *Breast Cancer Res*, *11*(3), R31. Journal Article. doi:10.1186/bcr2261

Ma, H., Wang, Y., Sullivan-Halley, J., Weiss, L., Marchbanks, P. A., Spirtas, R., Ursin, G., et al. (2010). Use of four biomarkers to evaluate the risk of breast cancer subtypes in the women's contraceptive and reproductive experiences study. *Cancer Res*, *70*(2), 575–87. Journal Article. doi:10.1158/0008-5472.can-09-3460

Mauguen, A., Zabor, E. C., Thomas, N. E., Berwick, M., Seshan, V. E., & Begg, C. B. (2017). Defining cancer subtypes with distinctive etiologic profiles: An application to the epidemiology of melanoma. *J Am Stat Assoc*, *112*(517), 54–63. Journal Article. doi:10.1080/01621459.2016.1191499

Phipps, A. I., Malone, K. E., Porter, P. L., Daling, J. R., & Li, C. I. (2008). Body size and risk of luminal, her2-overexpressing, and triple-negative breast cancer in postmenopausal women. *Cancer Epidemiol Biomarkers Prev*, *17*(8), 2078–86. Journal Article. doi:10.1158/1055-9965.epi-08-0206

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (Fourth.). New York: Springer. Retrieved from http://www.stats.ox.ac.uk/pub/MASS4

Yang, X. R., Pfeiffer, R. M., Garcia-Closas, M., Rimm, D. L., Lissowska, J., Brinton, L. A., Peplonska, B., et al. (2007). Hormonal markers in breast cancer: Coexpression, relationship with pathologic characteristics, and risk factor associations in a population-based study. *Cancer Res*, *67*(21), 10608–17. Journal Article. doi:10.1158/0008-5472.can-07-2142

Zabor, E. C. (2019). *Riskclustr: Functions to study etiologic heterogeneity*. https://github.com/zabore/riskclustr.

Zabor, E. C., & Begg, C. B. (2017). A comparison of statistical methods for the study of etiologic heterogeneity. *Stat Med*, *36*(25), 4050–4060. Journal Article. doi:10.1002/sim.7405