Center for Biomedical Informatics Research
Department of Medicine, Stanford University
1265 Welch Rd.
Stanford, CA 94305

To the Editor,

We are pleased to submit the attached manuscript to *The R Journal* as a short introduction to the stratamatch package. We expect this manuscript to be of immediate interest to applied researchers in fields such as epidemiology, health outcomes, and clinical informatics, while appealing to a broader audiences in the R and statistics community for its methodological contributions – in particular to causal inference from observational studies.

The task of inferring a causal effect from observational data is one of the more difficult and pressing statistical problems of our time – especially because of the increasing availability of large, passively collected "Big" data which has long promised to revolutionize a variety of disciplines. One of the most ubiquitous approaches for making causal inference from observational data is optimal propensity score matching; However, outstanding critiques of the statistical properties of propensity score matching have cast doubt on the statistical efficiency of this technique (see, for example King and Neilson, "Why propensity scores should not be used for matching"), and the poor scalability of optimal matching to large data sets makes this approach inconvenient if not infeasible for sample sizes that are increasingly commonplace in modern observational data.

The 'stratified matching design' implemented by the stratamatch package addresses both of these issues with optimal propensity score matching: increasing the precision and robustness of inference while improving the scalability of optimal matching designs. Prior to the introduction of stratamatch, little support existed for stratified matching designs, and this approach was relatively uncommon in spite of the potential benefits. To address this need, our manuscript briefly introduces stratamatch and the methodological context of the implementation, specifically: (1) the stratified matching design and its statistical and computational benefits (2) a suite of new diagnostic tools for stratified data sets in causal inference studies (3) the implementation of a novel 'pilot design' approach in a 'prognostic score' stratification scheme. In particular, this manuscript represents the first introduction a prognostic score stratification scheme, an approach which compromises the statistical benefits of the prognostic score (see more theoretical discussion by Aikens et al (2020), Antonelli et al (2018), Leacy and Stuart (2014), Hansen (2008), and Rosenbaum (2005)) with the practical computational gains from stratification.

We anticipate that this manuscript – which contains material of substantial methodological merit anchored in application and implementation – will meet a pressing need in epidemiology and clinical informatics, while being of interest and utility to the broader R and statistics community.

For your convenience, we suggest the following potential reviewers:

- Sam Pimentel (spi@berkeley.edu)
- Joseph Antonelli (jantonelli@ufl.edu)
- Brain K Lee (bklee@drexel.edu)
- Adam C. Sales (asales@utexas.edu)
- Trang Quynh Nguyen (tnguye28@jhu.edu)
- Mark Fredrickson (mfredric@umich.edu)
- Noah Greifer (noah.greifer@gmail.com)

Thank you for your time and consideration.

We look forward to your response,

Jonathan H. Chen