Published in final edited form as:

J Am Stat Assoc. 2016 ; 111(513): 127–129. doi:10.1080/01621459.2016.1149406.

Comment: Addressing the Need for Portability in Big Data Model Building and Calibration

Chirag J. Patel and

Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115

Francesca Dominici

Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115

Chirag J. Patel: chirag_patel@hms.harvard.edu; Francesca Dominici: fdominic@hsph.harvard.edu

Due to rapid advances in information and genomic-related technology, environmental and social epidemiology, cancer research, and many other biomedical related fields, the production and storage of massive observational datasets (both with large number of observations and variables), known as “big data” (National Institutes of Health Big Data to Knowledge 2015), poses new computational and logistic challenges to data analysts. First, one of the primary logistic challenges for the analyst in many big data fields, such as biomedicine (Weber, Mandl, and Kohane 2014), now includes “data integration,” the resource-intensive process required to put disparate datasets together for modeling and analyses. For example, datasets may exist in separate computer file systems, or “silos,” and physically moving them to a central location for modeling is not feasible, and relatedly, contain partially overlapping sets of covariates. Or, datasets may contain sensitive information, such as identifiable information of people whose identity must be concealed or obfuscated from modelers of the data. Second, scientific findings and data analyses must be reproducible. To maximize reproducibility, datasets must be accessible to analysts. How can an analyst integrate large datasets that may exist in different physical locations under diverse data use agreements and security permissions to maximize use, estimate parameters of interests with highest possible efficiency, and assure reproducibility? While our ability to create and store datasets is now immense, our attention now must turn to computational methods to analyze these massive data streams. Further, there is a need to increase the portability of either datasets, or barring that, the computational methods themselves. Infrastructural advances, such as cloud computing (such as Amazon elastic computing), “open” communities (e.g., efforts such as the *Open Science Data Cloud*; Open Science Data Cloud 2015), and digital “notebooks” to enable model and data sharing (and as covered in Leek and Peng 2015) all can make the former possible; however, analytic methods need to catch up.

The recent report by Chatterjee and colleagues proposes a method that allows analysts to combine information from two different datasets. Their proposed approach is aimed at overcoming theoretical and logistic challenges of siloed and sensitive data for model building and calibration. Typically, as readers will know, model building usually requires individual-level data for all observations and variables being modeled, a challenge to modeling datasets that may exist in a number of physical silos or containing sensitive information. The authors address these challenges by developing a new method,

“constrained maximum likelihood estimation for model calibration using summary-level information” (or CML for short).

For example, suppose there are two datasets consisting of two independent populations: A (*the internal dataset*) and B (*the external dataset*). These two datasets (A and B) might have been collected by two institutions or different groups from large consortia. Or, very simply, A may have individual level data on subset of the study population that is included in B .

Now further suppose that dataset B is both very large, might have crude information on the covariates, and it generally not available to the analyst. Let us also suppose dataset A is readily accessible, has a larger number of covariates than B but also is markedly smaller in sample size than B . With CML and as shown in the case studies presented in their article, the authors are interested in predicting Y (e.g., breast cancer risk measured in both A and B) as a function of variables X (e.g., age, height, weight, hormone levels, measured in both A and B) and Z (e.g., mammographic density or the number of biopsies, measured in only in A and for a much smaller number of individuals than B). How can we make the most accurate prediction of breast cancer risk that relies on all the available information (from both A and B), even though individual level data from B is not accessible to us? What is more portable than the entire dataset B includes *summary statistics*, such as effect sizes and standard errors or more generally estimates of model parameters.

The advantages of such a modeling technique seem numerous. This is because CML circumvents the need for having individual level data from dataset B . Instead, CML only relies on summary statistics from dataset B for model building and calibration on the dataset A . In their words, the method uses the summary statistics to a “set of constraints, which [they] use in analysis of internal data (dataset A) to improve efficiency of parameter estimates and generalizability of models.” Or, in our words, they use information from a potential misspecified reduced model fitted to dataset B —from which they obtain an estimated to obtain a constrained nonparametric maximum-likelihood estimate of β , where β is the vector of unknown parameters, which belongs to the full model for dataset A .

The first apparent advantage includes model calibration in scenarios where one dataset is a subset of another larger and perhaps physically inaccessible dataset. For example, take the Medicare population dataset and its subset focused on cancer-related research, the Surveillance, Epidemiology and End Results (SEER)-Medicare dataset (National Cancer Institute 2015). The SEER-Medicare data (dataset A) contain detailed information on treatment (X) and individual level covariates (Z) for patients diagnosed with cancer from the SEER cancer registry. The SEER-Medicare dataset is a nonrandom subset of the entire Medicare population dataset (dataset B). With CML, summary estimates from the whole Medicare population (Θ) may be used to obtain β and calibrate the parameters to improve prediction in dataset A . With the advent of new and perhaps costly measurement technologies, a likely scenario includes one where we collect additional individual data (e.g., covariates Z) from sensors (e.g., physical activity), omic data (e.g., *metabolome*, *microbiome*, *proteome*), and/or biomarkers for a *subset* of individuals enrolled in a large cohort (e.g., the entire Nurses’ Health Study; Colditz, Manson, and Hankinson 1997, or National Health and Nutrition Examination Survey; Centers for Disease Control and

Prevention (CDC) 2013 are examples of datasets B), and we want to predict health risks in the subset (dataset A) while borrowing information from the larger and complete cohort data. CML allows for this scenario. Furthermore, there are many publicly available external databases with summarized information (e.g., summary information for datasets B), such as the *Genome-Wide Association Study (GWAS)* catalog (Welter et al. 2014), which has cataloged effect sizes, and standard errors for over 2000 phenotypic traits (Y) in over 15,000 genetic variants. The emergence of databases of summarized information like the GWAS catalog may spur developments of new summarized datasets in other domains, such as nutritional or environmental epidemiological studies.

Perhaps the greatest yield in integrating of diverse large datasets for model calibration and building will be reduction of false positive results. For example, big data analytics can lead to “Big Error” (Khoury and Ioannidis 2014), whereby the chances for spurious findings increase as a function of number of associations mined. Model calibration constrained to risk factor distributions from a different dataset or population, as proposed by Chatterjee et al., may mitigate chances for finding spurious results. Sensitivity analyses that consider different modeling scenarios given estimates computed in external datasets (B) are also possible (Young and Karr 2011).

But what assumptions must hold to make this a pragmatic solution for data integration?

The proposed approach makes at least two very strong assumptions: (1) the model parameters (β) in the full model (f) are assumed to be the same in the two populations (internal and external dataset, A and B, respectively) and (2) the full model (f) must be correct.

CML has the nice feature that allows the reduced model (g) to be misspecified. Considering that the reduced model is fitted to an enormous external dataset B, one would expect substantial robustness of the parameter estimates (θ) to misspecification of the reduced model g . However, we wished the authors had investigated the sensitivity of the results to the misspecification of the full model (f), in the context where the internal dataset A has limited sample size compared to the dimension of the covariates.

We do applaud the authors for conducting a simulation study and sensitivity analyses; however, it is not clear what would be the impact on the performance of this method in the context where any of the aforementioned two assumptions are violated. In fact, in the data analysis presented in the article, it is indeed the case where the model coefficients are quite different in the two study populations. Furthermore, it seems that, in the era of large datasets, that difference in populations and datasets (e.g., datasets A and B) will be abundant. Therefore, a critical next step for CML would be to characterize the scale of error induced due to violation of these assumptions of the method.

As always is the case in the context of combining information across data sources, the performance of the method will rely on a bias-variance tradeoff. For example, if the parameters (β) are very different between the two populations, it might be better to rely only on the data from the internal sample (A). On the other extreme, if the two populations are homogenous (same parameters, β) then there will be maximum gain from the external data

source (B) when making prediction in the internal data (A). Unfortunately, the authors did not study nor provide diagnostics to inform as whether the borrowing information from the external data source will provide significant gain in efficiency (except for the few simple scenarios in the simulation study). As a speculative side note, this reminds us of Bayesian estimation, where estimation of the parameters β in the internal sample can be improved by borrowing information from the external sample via a constraint (on which we can also call it an informative prior coming from model fitting to the external dataset B).

Another potential tradeoff with the approach is regarding the flexibility of the nonparametric estimate of $R(X,Z)$. Here, flexibility may come at the expense of potential identifiability and computational scalability. For example, let us consider the more realistic case where $R(X,Z)$ is different between the internal and the external study (a case that is indeed considered in the article). What would happen if we have a (1) very large number of covariates (X and Z , where Z is a million of genetic variants only available in A), (2) the sample of A is close to the number of variables in X and Z ? Then the most likely scenario is that $R(X,Z)$ will not be identified in the internal dataset or its estimate will be highly uncertain. Here, authors could have considered the possibility of using a more parametric approach for $R(X,Z)$, as model assumptions would have increased efficiency in this situation—but admittedly also the bias. Relatedly, if the sample size is not much larger than dimensionality of the smaller dataset, then, asymptotic theory will not apply for the standard error calculation. While two covariates (one X and one Z) are considered in the simulations, we remain cautious regarding the computational efficiency and scalability of the proposed approach, especially when the distribution of the risk factors ($R(X,Z)$) is different between the internal and the external datasets and/or when the number of covariates approaches the sample size.

Further, the authors show that in simulations, the CML is substantially better than generalized regression calibration (GRC) in a few situations, that is, in the context of measurement error and the efficiency gain only occurs for the parameters corresponding to the covariates available in both datasets. We note that there was not an efficiency gain in the context of missing covariate (a very common situation and also the situation in the dataset where there are missing data in the external dataset). Finally, although the proposed method is *developed* in the context of risk prediction, it is not *evaluated* in the context of prediction, especially in the context where A is a random sample of B. In the simulations and data analysis, the proposed method is evaluated and compared in the context of mean standard errors of estimated regression coefficients. Further ongoing work should undoubtedly focus on prediction.

In conclusion, there is an urgent need to increase the availability and portability of datasets for model building and calibration; Chatterjee and colleagues propose a new technique extending survey-based methodologies that may serve as a way to save infrastructural resources to calibrate, it is hoped, more generalizable linear models. This is a step in the right direction but as always, more work needs to be done before claiming that this can be broadly applicable.

References

- Centers for Disease Control and Prevention (CDC). National Health and Nutrition Examination Survey. 2013. available at <http://www.cdc.gov/nchs/nhanes/>
- Colditz GA, Manson JE, Hankinson SE. The Nurses' Health Study: 20-Year Contribution to the Understanding of Health Among Women. *Journal of Women's Health*. 1997; 6:49–62.
- Khoury MJ, Ioannidis JP. *Medicine*. Big Data Meets Public Health. *Science*. 2014; 346:1054–1055. [PubMed: 25430753]
- Leek JT, Peng RD. Opinion: Reproducible Research Can Still be Wrong: Adopting a Prevention Approach. *Proceedings of the National Academy of Sciences of the United States of America*. 2015; 112:1645–1646. [PubMed: 25670866]
- National Cancer Institute. SEER-Medicare Linked Database. 2015. available at <http://healthcaredelivery.cancer.gov/seermedicare/>
- National Institutes of Health Big Data to Knowledge. What is Big Data?. 2015. Available at <https://datascience.nih.gov/bd2k/about/what>
- Open Science Data Cloud. Open Science Data Cloud. 2015. available at <https://www.opensciencedatacloud.org>
- Weber GM, Mandl KD, Kohane IS. Finding the Missing Link for Big Biomedical Data. *Journal of the American Medical Association*. 2014; 311:2479–2480. [PubMed: 24854141]
- Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorf L, Parkinson H. The NHGRI GWAS Catalog, a Curated Resource of SNP-trait Associations. *Nucleic Acids Research*. 2014; 42:D1001–D1006. [PubMed: 24316577]
- Young SS, Karr A. Deming, Data and Observational Studies. *Significance*. 2011; 8:116–120.