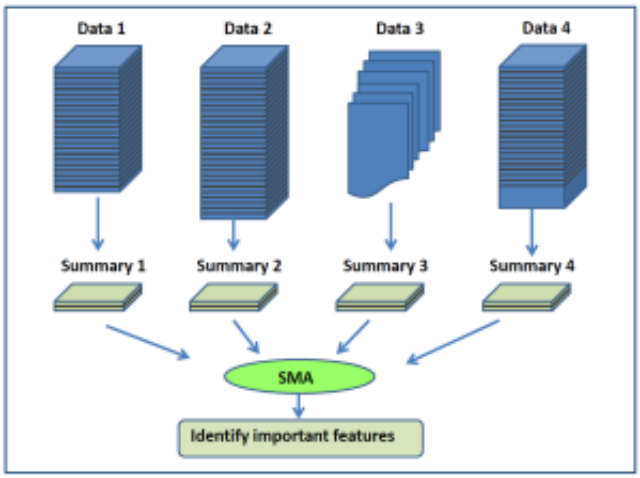
Meta-analysis in the Big-data era

Meta-analysis plays an important role in summarizing and synthesizing scientific evidence derived from multiple studies. (In Greek, ‘meta’ pertains to ‘with, across, or after’, referring to a level above or beyond.) By combining multiple data sources, one can achieve higher statistical power, more accurate estimation, and greater reproducibility. To date, there are more than 97000 publications containing ‘meta-analysis’ in the PubMed database. The role of meta-analysis in biomedical research will only become bigger as we enter the big-data era.



Traditionally, meta-analysis can handle only a handful of covariates. Modern high-throughput genome technology, however, has generated an enormous amount of data with a large number of genomic features. (In statistics, the large number of features is often referred to as high-dimensions.) Under such a scenario, it is imperial to develop variable-selection methods for meta-analysis. (Here, variable-selection refers to the selection of the most important/predictive features out of the vast majority of noise features.) Incorporation of variable-selection into meta-analysis will improve model interpretation, reduce prediction errors, and provide better prioritization of genomic features for follow-up studies.

Existing variable selection methods require direct access to the raw data (i.e., patient-level data). Unfortunately, raw data are often unavailable because of high cost, logistical difficulties, time constraints, IRB restrictions, and other study policies. Summary statistics, instead, reduce the raw data to a much compressed level, and are much easier to be accessed and managed. Taking GWAS as an example, virtually all meta-analyses to date have been conducted at the summary-statistics level rather than the raw-data level. The emergence of big data, such as next-generation sequencing data, makes the collation of raw data even more challenging. A question naturally arises as to whether it is possible to conduct effective variable selection using only summary statistics.

In this article, we propose a new approach, Sparse Meta-Analysis (SMA), in which variable selection for meta-analysis is based solely on summary statistics. (Here, ‘sparse’ refers to the model post variable-selection.) Remarkably, we find that SMA is as efficient as using raw data if the correlation information of the summary statistics from each study is available. Hence, SMA skips the traditional bottleneck of data-aggregation, and provides an innovative tool for conducting high-dimensional meta-analysis. In addition, SMA can harness information shared by different studies while allowing heterogeneity among studies. In the era of big data, SMA will be extremely useful when it is impractical to collect or store all of the raw data (because SMA only needs summary statistics which are much more manageable). Thus, by utilizing summary statistics for variable selection, SMA can help people to conduct research that would be deemed impossible by other meta-analysis approaches.

Publication

Sparse meta-analysis with high-dimensional data.
He Q, Zhang HH, Avery CL, Lin DY
Biostatistics. 2015 Sep 21