University of Zurich[UZH]

How Small Is Big Enough?

Tackling Challenges in Fusing Big Numbers of Small Data Sets

Prof. Dr. Reinhard Furrer

Associate Professor at the Department of Mathematics and Affiliate Faculty Member at the Institute for Computational Science, University of Zurich

Application for the Emeritus Foundation for:

Servan L. Grüninger

MSc Biostatistics UZH, MSc Computational Science and Engineering ETH Lausanne

Zürich, April 24 2022

1 Objective: How to make small data big

In recent years, "big data" combined with machine learning has become a major interest in statistical research. In areas where data are plentiful and comparably cheap to acquire, major advances have been made in aggregating and analyzing large data sets. These advances have become part of our everyday lives in the form of automated text translations, targeted advertisements, autonomously driving vehicles, or improved health diagnostics using data from wearable devices such as smartwatches.

The focus on such advances should not hide the fact that in the majority of cases data is neither plentiful nor cheap. This is especially true in biomedical research with humans and animals. In clinical research involving humans, but especially in basic and preclinical research involving animals, the sample sizes of individual experiments are comparably low, ranging from a few dozens subjects to a few hundreds – a far cry from what could be considered "big data". Taken together, however, these numerous small sets of data can rival the volume and variety of more "traditional" big sets of data from genomics or transcriptomics (Ferguson et al., 2014) and are very helpful for answering scientific questions, both old and new, for example in research on neurotraumas (Ferguson et al., 2014), epilepsy (Wagenaar et al., 2015; Lapinlampi et al., 2017) or autism (Hall et al., 2012).

The amount of available preclinical data is growing continuously, both in the form of summary statistics published in journal articles as well as in the form of sets of raw data stored in data repositories. When combined, these sources of information offer the possibility to improve the accuracy of preclinical assessments of potential drug candidates. When it comes to drawing statistical inferences from available preclinical data, the numerous sets of data consisting of small data tables stemming from individual studies are of great relevance because they often contain valuable information from which to draw valid inferences about the safety and efficacy of a novel substance. However, they often consist of only a few data tables containing granular experimental results, for example from pilot studies, small exploratory studies, or larger confirmatory experiments usually collected by individual laboratories during day-to-day research. As such, there is often not much consistency with regard to the format and content of these data tables, sometimes even within the same set of data.

Hence, combining data from various sources is easier said than done and still requires a lot of statistical research to be considered feasible. One of the major challenges concerns the large variability of the data and the heterogeneity of the source from which they stem–something that is of particular relevance in preclinical research, that is, the phase of biomedical research just before first-in-human trials.

In addition, existing methods to integrate data from heterogeneous biomedical sources are scarce and were developed with clinical data in mind. Hence, they have not been adapted to the specific settings of preclinical research—settings which are commonly riddled with more systematic biases and exhibit more heterogeneity than clinical trials.

We therefore propose to **build a hierarchical modeling framework to fuse and analyze data from heterogeneous preclinical sources.** Such a framework will allow to draw more reliable statistical inferences from a large body of preclinical evidence that consists of many sets of data with a high degree of variety and variability due to study heterogeneity. The framework will be based on real and simulated data from preclinical research settings.

2 Why this project matters

Modern biomedical research questions are complex and multifaceted, thereby producing a large number of varied sets of data stemming from heterogeneous sources. To answer the same biomedical question, preclinical researchers may rely on vastly differing experimental models, designs and outcome measures, such as *in vitro* methods that use animal and human tissue and cell cultures, *in silico* approaches created by mathematically modeling biochemical reactions, and *in vivo* studies in different animal species. This can lead to large heterogeneity in the data published by different laboratories—something that is particularly true for preclinical research.

Preclinical research describes the phase after the discovery of a potential therapeutic intervention in basic research and before any first-in-human trials. For scientists, but more importantly for patients and doctors it is crucial to know which substances are safe and effective enough to be tested in humans and which are not. To this end, preclinical research provides the scientific information needed to assess the toxicological safety, the pharmacological efficacy and the economic feasibility of new therapeutic interventions.

Because preclinical research is conducted using a wide range of different methods and frameworks and is much less standardized than research it is much more heterogeneous than clinical research involving humans, it is a challenge to combine data from different preclinical experiments even if these experiments were trying to answer the same scientific questions. For example, efficacy assessments of a medical treatment against at specific type of cancer can be done based on various biomedical outcomes (e.g., tumor size, tumor number, survival rate), experimental setups (e.g., oral or intravenous drug administration), experimental models (e.g., human cell culture vs. animal models), strains of the same species (e.g., "black 6" mice vs. "BALB/c" mice) or different species (e.g., mouse vs. non-human primates).

In fact, when it comes to external validity, that is, the generalisability of scientific findings, such heterogeneity can be desirable if it moves preclinical settings closer to clinical scenarios (Voelkl et al., 2018). Because replication serves to test whether existing models are able to predict outcomes that have not yet been observed (Nosek and Errington, 2020), a replication that is limited to exactly the same experimental settings as in the original experiment does not provide much additional inferential value to the original experiment. As a result, the importance of collecting, merging and analyzing data from heterogeneous sources is increasingly recognized by experimental and theoretical researchers alike (e.g., Searls 2005; Ramirez 2013; Ma'ayan et al. 2014; Bodden et al. 2019). Unfortunately, there are currently no statistical frameworks available that can rise to this challenge within preclinical settings, especially in the face of the high degree of heterogeneity mentioned. In addition, a considerable number of preclinical (and clinical) research findings are only available in the form of summary statistics (Chan et al., 2014), even though there is increasing pressure from funding agencies, publishers, and regulatory bodies to make raw data publicly accessible. Hence, a potential modeling framework would need to be able to incorporate raw data as well as summary statistics from heterogeneous sources in order to achieve high accuracy for parameter estimation and statistical inference. Furthermore, the framework would need to prevent potentially new biases from arising, such as confounding, sampling selection, or cross-population biases, especially regarding causal inferences (Bareinboim and Pearl, 2016).

3 Planned research

To overcome the challenges in fusing data from heterogeneous sources, we will develop a statistical modeling framework that allows the combination of varied data from heterogeneous sources. This framework will hinge on the assumption inherent to all preclinical research, namely that the scientific outcomes with regard to a specific biomedical question assessed in heterogeneous experimental settings are produced by a common underlying biological process. For example, if a treatment shows similar effects against cancer in different experimental settings, we assume that these effects are based on the same underlying biological processes, e.g., by influencing a specific molecular pathway which affects the various outcomes in different experimental conditions.

We start by building the statistical and computational framework of the fusion model based on common scenarios that are encountered in preclinical research. The framework includes all relevant experimental settings as well as available sets of data that are likely to be encountered in the preclinical setting. More precisely, the fusion model will allow us to draw statistical inferences from a large body of evidence that consists of many small and possibly several big data tables with a high degree of variety and variability collected in different sets of data from heterogeneous sources. In the next steps, we test the fusion model using simulated data based on common experimental settings, test its performance using empirical data from preclinical and clinical trials in order to assess how well it can estimate the relevant summary statistics, and improve it based on the test results.

Figure 1 schematically illustrates the conceptual setting of investigating an effect. Different labs or groups perform experiments and may or may not release information to the community. Alternative evidence-generating approaches from labs are meta-analyses or expert predictions.

For a more detailed description of the fusion model, please see the Appendix in Section 6

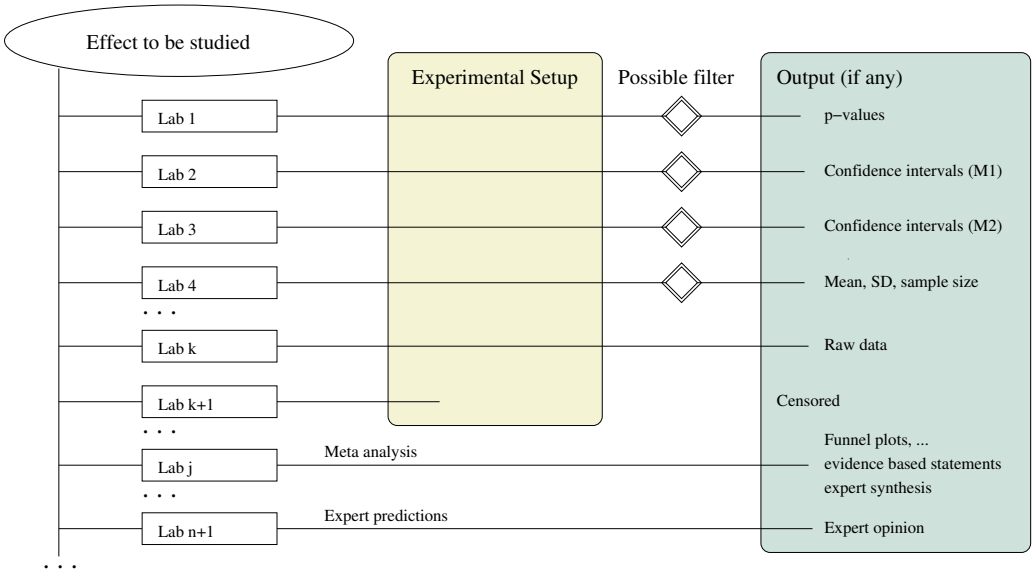

Figure 1: Schematic illustration of different labs investigating a particular effect. The experimental setup determines the covariates and introduces potential biases. The output of a lab may be very different, ranging from single p-values to fully open data.

4 Budget

Overall, the proposed project is expected to last 18 months of which 6 months are already financed by the Forschungskredit *Candoc* of the University of Zurich. The sole costs involved are the wages for the designated PhD student Servan L. Grüninger (started in 2020). In total, this amounts to total expenditures of CHF 85'456 of which CHF 27'911 are covered by *Candoc*. Hence, we request the amount of CHF 57'545 for the realisation of this research project.

5 Significance of the expected results of the project

Academic researchers, pharmaceutical companies and regulatory bodies are all in great need of reliable and powerful statistical tools to fuse varied preclinical data from heterogeneous sources, combine them with summary statistics and draw valid inferences. Reproducible, valid data and the appropriate statistical models to analyze them not only save time and money, but also reduce the necessary number of experimental animals and improve the safety and efficacy of therapeutic candidates tested in humans.

For successful use in preclinical settings, statistical frameworks must be general enough to allow for versatile application across settings, but specific enough that they can yield meaningful and reliable inferences in these settings. In this project, we can make a major step in providing tools to properly analyze varied sets of data containing preclinical data tables of different sizes and affected by different methodological and statistical choices. Nowadays, Bayesian hierarchical models are widespread, yet the proposed fusion approach pushes the methodology and software implementation to a new level, beyond a mere academic exercise to one with a significant impact in preclinical research and subsequent scientific domains. Hence, I expect the results of this project to be welcomed by a range of neighboring scientific disciplines focused on drug development, disease research, biomedical basic research and applied mathematics, as well as in other experimental fields such as psychology, in which the integration of heterogeneous data from different sources also pose a methodological challenge.

6 Appendix: Conceptual description of the fusion model

Every lab produces a set of data, which is denoted as $\{\boldsymbol{y}_{ik}, \boldsymbol{x}_{ik}\}$, for lab k, containing the outcome $\boldsymbol{y}_{ij}$ and covariate (i.e., explanatory variables) $\boldsymbol{x}_{ik}$, $i = 1, \ldots, n_k$. In a publication or repository, the results of lab k are typically filtered and reported as summary statistics (including for example p-values, confidence intervals, means and standard errors) but in the ideal case contain actual raw data as well. Filtered results are denoted as $T_k\big(\{\boldsymbol{y}_{ik}, \boldsymbol{x}_{ik}\}\big)$. The examples of Figure 1 are expressed, as

$$
\begin{aligned}
&T_1\big(\{\boldsymbol{y}_i, \boldsymbol{x}_i\}\big) = \Pr\big(|\bar{\boldsymbol{y}}_i| > c_{\text{crit}} \mid H_0\big) && \text{p-value}\\
&T_2\big(\{\boldsymbol{y}_i, \boldsymbol{x}_i\}\big) = \big\{b_{\text{lower}}, b_{\text{upper}}\big\}, && \text{confidence interval}\\
&T_4\big(\{\boldsymbol{y}_i, \boldsymbol{x}_i\}\big) = \Big\{n, \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{y}_i, \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{y}_i - \bar{\boldsymbol{y}}_i)^2\Big\}, && \text{sample size, mean and variance}\\
&T_k\big(\{\boldsymbol{y}_i, \boldsymbol{x}_i\}\big) = \{\boldsymbol{y}_i, \boldsymbol{x}_i\} && \text{set of data, including relevant predictors}
\end{aligned}
$$

In some situations, nothing at all is observed due to, e.g., publication bias. Hence, $T_{k+1}\big(\{\boldsymbol{y}_i, \boldsymbol{x}_i\}\big) = \emptyset$ based on the censoring mechanism $\Pr\big(|\bar{\boldsymbol{y}}_i| > c_{\text{crit}} \mid H_0\big) > \alpha$. Of course, this specific type of censoring not only hides the results of the lab but most likely also the fact that the lab has studied the effect.

To combine the output of all labs in a hierarchical model, the effect to be studied is denoted as a generic parameter vector $\boldsymbol{\theta}$, representing, e.g., a single value, a parameterized density, spline curve, etc. Figuratively said, $\boldsymbol{\theta}$ can be seen as the parameterized largest common component of all studies involved. For example, if all lab studies are for female mice with a single treatment dose, then $\boldsymbol{\theta}$ is scalar representing the effect for female mice. If both sexes are included in the studies with various doses, $\boldsymbol{\theta}$ contains the information for both sexes including a parametrization of the dose effect.

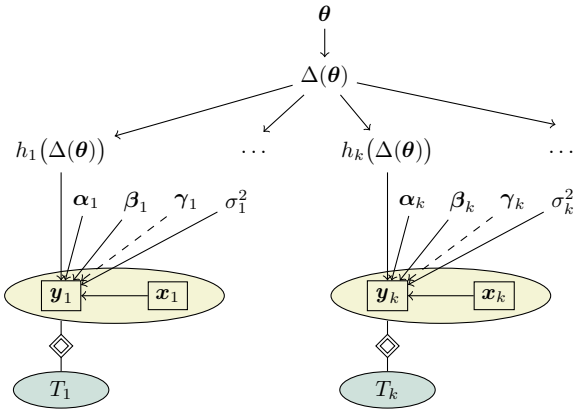

Figure 2: A graphical model representation of a hierarchical fusion model. Boxes represent data tables; ellipses represent sets of data at the lab level or available to us. $\boldsymbol{\alpha}_k$, $\boldsymbol{\beta}_k$ and $\boldsymbol{\gamma}_k$ represent biases and covariate parameters. Dashed arrows represent optional components that depend on modeling assumptions of the outcome variable. σ_k^2 represent generically measurement errors.

Figure 2 provides a schematic, graphical model representation of a proposed fusion model framework where, simply put, the goal is to infer the effect Δ parameterized by $\boldsymbol{\theta}$ given the different summary statistics $\{T_1, \ldots, T_K\}$. Different labs observe partial information of Δ through a transfer function h_k. The problem would be fairly standard if the entire sets of data were available from the labs. We could provide individual models that are combined over the different labs through the common parameter $\boldsymbol{\theta}$.

As a simple illustration, let us assume that two labs report the success probability of a treatment under different doses. Further, we assume a logistic relation between the success probability of a treatment and its dose. At the lab level, we would propose a logistic regression model with link function g, parameters $\boldsymbol{\beta}_k$ and lab specific biases $\boldsymbol{\alpha}_k$:

$$\begin{aligned} g\big(\mathrm{E}(Y_{i1} \mid \boldsymbol{x}_{i1})\big) &= \boldsymbol{x}_{i1}^{(1)\mathsf{T}}\boldsymbol{\theta} + \boldsymbol{x}_{i1}^{(2)\mathsf{T}}\boldsymbol{\beta}_1 + \boldsymbol{\alpha}_1, \\ g\big(\mathrm{E}(Y_{i2} \mid \boldsymbol{x}_{i2})\big) &= \boldsymbol{x}_{i2}^{(1)\mathsf{T}}\boldsymbol{\theta} + \boldsymbol{x}_{i2}^{(2)\mathsf{T}}\boldsymbol{\beta}_2 + \boldsymbol{\alpha}_2. \end{aligned} \tag{1}$$

We then employ a Bayesian setting and are therefore able to reuse the knowledge and software components established in Wang et al. (2017, 2018); Wang and Furrer (2019b,a). The framework will be implemented in Stan (Gelman et al., 2015) via the R interface `rstan` (Stan Development Team, 2020). The Bayesian embedding allows a straightforward extension to all summary statistics. For example, the prior for θ_i will yield lower and upper bounds for the effect such that $\Pr(b_{\text{lower}} < \Delta(\theta_i) < b_{\text{upper}}) = 95\%$. we consider the published confidence intervals as an observed instance of such credible intervals.

Model (1) is over-parameterized, and more complex models as illustrated by Figure 2 are even more so. It is only possible to determine a single *lab offset* which would include all lab specific biases. Even if several labs work with the same organism, but with slightly different experimental setups, the parameters are hardly identifiable. We propose to address the identifiability issues and over-parametrization with the following approaches. First, we include dependencies between the individual labs that report similar summary statistics. The rationale is that for the same class of summary statistics, similar biases exist (because often, several studies from a particular lab are published). Such a model is closely related to what has been developed in Wang et al. (2018) and subsequent work. Although these dependencies introduce additional parameters, we claim that these will stabilize the model. A second class of approaches are based on classical penalization methods which we embed in the Bayesian framework (Park and Casella, 2008). We will mainly work with the SLOPE (Sorted L-One Penalized Estimation) selection procedure (Bogdan and Frommlet, 2020), which has substantially smaller prediction errors compared to the optimal version of LASSO (Tibshirani, 1996).

7 Bibliography

Bareinboim, E. and Pearl, J. (2016). Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113, 7345–7352.

Bodden, C., von Kortzfleisch, V. T., Karwinkel, F., Kaiser, S., Sachser, N., and Richter, S. H. (2019). Heterogenising study samples across testing time improves reproducibility of behavioural data. *Scientific Reports*, 9, 1–9.

Bogdan, M. and Frommlet, F. (2020). Identifying important predictors in large data bases - multiple testing and model selection. In Cui, X., Dickhaus, T., Ding, Y., and Hsu, J. C., editors, *Handbook of Multiple Comparisons*, in preparation.

Chan, A.-W., Song, F., Vickers, A., Jefferson, T., Dickersin, K., Gøtzsche, P. C., Krumholz, H. M., Ghersi, D., and van der Worp, H. B. (2014). Increasing value and reducing waste: addressing inaccessible research. *The Lancet*, 383, 257–266.

Ferguson, A. R., Nielson, J. L., Cragin, M. H., Bandrowski, A. E., and Martone, M. E. (2014). Big data from small data: data-sharing in the 'long tail' of neuroscience. *Nature Neuroscience*, 17, 1442–1447.

Gelman, A., Lee, D., and Guo, J. (2015). Stan: A Probabilistic Programming Language for Bayesian Inference and Optimization. *Journal of Educational and Behavioral Statistics*, 40, 530–543.

Hall, D., Huerta, M. F., McAuliffe, M. J., and Farber, G. K. (2012). Sharing Heterogeneous Data: The National Database for Autism Research. *Neuroinformatics*, 10, 331–339.

Lapinlampi, N., Melin, E., Aronica, E., Bankstahl, J. P., Becker, A., Bernard, C., Gorter, J. A., Gröhn, O., Lipsanen, A., Lukasiuk, K., Löscher, W., Paananen, J., Ravizza, T., Roncon, P., Simonato, M., Vezzani, A., Kokaia, M., and Pitkänen, A. (2017). Common data elements and data management: Remedy to cure underpowered preclinical studies. *Epilepsy Research*, 129, 87–90.

Ma'ayan, A., Rouillard, A. D., Clark, N. R., Wang, Z., Duan, Q., and Kou, Y. (2014). Lean Big Data integration in systems biology and systems pharmacology. *Trends in Pharmacological Sciences*, 35, 450–460.

Nosek, B. A. and Errington, T. M. (2020). What is replication? *PLoS Biology*, 18, 1–8.

Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103, 681–686.

Ramirez, T. (2013). Metabolomics in toxicology and preclinical research. *ALTEX*, 30, 209–225.

Searls, D. B. (2005). Data integration: challenges for drug discovery. *Nature Reviews Drug Discovery*, 4, 45–58.

Stan Development Team (2020). *RStan: the R interface to Stan.* R package version 2.19.3, https://CRAN.R-project.org/package=rstan.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, 58, 267–288.

Voelkl, B., Vogt, L., Sena, E. S., and Würbel, H. (2018). Reproducibility of preclinical animal research improves with heterogeneity of study samples. *PLoS Biology*, 16, 1–13.

Wagenaar, J. B., Worrell, G. A., Ives, Z., Matthias, D., Litt, B., and Schulze-Bonhage, A. (2015). Collaborating and Sharing Data in Epilepsy Research:. *Journal of Clinical Neurophysiology*, 32, 235–239.

Wang, C. and Furrer, R. (2019a). Combining Heterogeneous Spatial Datasets with Process-based Spatial Fusion Models: A Unifying Framework. *arXiv:1906.00364 [stat]*.

Wang, C. and Furrer, R. (2019b). Efficient inference of generalized spatial fusion models with flexible specification. *Stat*, 8, 1–9.

Wang, C., Puhan, M. A., and Furrer, R. (2018). Generalized spatial fusion model framework for joint analysis of point and areal data. *Spatial Statistics*, 23, 72–90.

Wang, C., Torgerson, P. R., Höglund, J., and Furrer, R. (2017). Zero-inflated hierarchical models for faecal egg counts to assess anthelmintic efficacy. *Veterinary Parasitology*, 235, 20–28.