METHODS TO ENHANCE THE REPRODUCIBILITY OF PRECISION MEDICINE

ARJUN K MANRAI
Department of Biomedical Informatics
Harvard Medical School, Boston, MA
Email: Manrai@post.harvard.edu

CHIRAG J PATEL
Department of Biomedical Informatics
Harvard Medical School, Boston, MA
Email: Chirag_Patel@hms.harvard.edu

NILS GEHLENBORG
Department of Biomedical Informatics
Harvard Medical School, Boston, MA
Email: Nils@hms.harvard.edu

NICHOLAS P TATONETTI
Department of Biomedical Informatics
Columbia University, New York, NY
Email: Nick.Tatonetti@columbia.edu

JOHN P.A. IOANNIDIS
Department of Medicine
Stanford University School of Medicine, Stanford, CA
Email: Jioannid@stanford.edu

ISAAC S KOHANE
Department of Biomedical Informatics
Harvard Medical School, Boston, MA
Email: Isaac_Kohane@hms.harvard.edu

During January 2015, President Obama announced the Precision Medicine Initiative [1], strengthening communal efforts to integrate patient-centric molecular, environmental, and clinical "big" data. Such efforts have already improved aspects of clinical management for diseases such as non-small cell lung carcinoma [2], breast cancer [3], and hypertrophic cardiomyopathy [4]. To maintain this track record, it is necessary to cultivate practices that ensure reproducibility as large-scale heterogeneous datasets and databases proliferate. For example, the NIH has outlined initiatives to enhance reproducibility in preclinical research [5], both *Science* [6] and *Nature* [7] have featured recent editorials on reproducibility, and several authors have noted the issues of utilizing big data for public health [8], but few methods exist to ensure that big data resources motivated by precision medicine are being used reproducibly. Relevant challenges include: (1) integrative analyses of heterogeneous measurement platforms (e.g. genomic, clinical, quantified self, and exposure data), (2) the tradeoff in making personalized decisions using more targeted (e.g. individual-level) but potentially much noisier subsets

of data, and (3) the unprecedented scale of asynchronous observational and population-level inquiry (i.e. many investigators separately mining shared/publicly-available data).

In this session of the Pacific Symposium on Biocomputing (PSB) 2016, we feature manuscripts that explore and propose solutions to some of the challenges of reproducibility in the era of precision medicine.

Two submissions to the session address challenges to reproducibility in observational (e.g., Electronic Health Record [EHR]) and clinical trial settings. Chen et al. [9] study the stability of predicting clinical practice patterns by varying the duration of EHR data used in training clinical order association rules, finding that larger longitudinal datasets may not improve, and might worsen, some predictions given the importance of secular practice trends. Ma et al. [10] provide a method for finding questionable exclusion criteria commonly used in clinical trials for mental disorders deposited in ClinicalTrials.gov.

Another challenge for the implementation of precision medicine involves novel methods for assessing data quality. Koire et al. [11] study threats to reproducibility when repurposing publicly available genome sequencing data, using data from The Cancer Genome Atlas [12] to study false positive variant calls and systematically evaluate variant call quality.

Software that enables analysts to transparently document analysis protocols can also help ensure reproducibility. Callahan et al. [13] create a reproducible workflow for microbiome studies using the Bioconductor [14] and knitr [15] *R* packages, providing a principled way to share protocols and explore how a multiplicity of analysis choices can sway results [16], [17]. Further, Manrai et al. [18] develop a shareable computational framework for quantifying widely-used pathogenicity assertions that relate genetic variation to disease, enabling users to identify how genetic model parameters influence risk estimates for genetic variants used in clinical practice.

These manuscripts address aspects of maintaining reproducibility as large-scale and heterogeneous datasets become increasingly common in the era of precision medicine. Concerted community-wide efforts will be critical to ensure that our ability to collect diverse types of patient-centric data is tantamount to our ability to distill reproducible findings from these data.

References

[1] F. S. Collins and H. Varmus, “A New Initiative on Precision Medicine.,” *N. Engl. J. Med.*, vol. 372, no. 9, pp. 793–5, Jan. 2015.

[2] W. Pao and N. Girard, “New driver mutations in non-small-cell lung cancer.,” *Lancet. Oncol.*, vol. 12, no. 2, pp. 175–80, Feb. 2011.

[3] S. M. Domchek, T. M. Friebel, C. F. Singer, D. G. Evans, H. T. Lynch, C. Isaacs, J. E. Garber, S. L. Neuhausen, E. Matloff, R. Eeles, G. Pichert, L. Van t’veer, N. Tung, J. N. Weitzel, F. J. Couch, W. S. Rubinstein, P. A. Ganz, M. B. Daly, O. I. Olopade, G. Tomlinson, J. Schildkraut, J. L. Blum, and T. R. Rebbeck, “Association of risk-reducing surgery in BRCA1 or BRCA2 mutation carriers with cancer risk and mortality.,” *JAMA*, vol. 304, no. 9, pp. 967–75, Sep. 2010.

[4] H. L. Rehm, “Disease-targeted sequencing: a cornerstone in the clinic.,” *Nat. Rev. Genet.*, vol. 14, no. 4, pp. 295–300, May 2013.

[5] F. S. Collins and L. A. Tabak, “Policy: NIH plans to enhance reproducibility.,” *Nature*, vol. 505, no. 7485, pp. 612–3, Jan. 2014.
[6] M. McNutt, “Reproducibility.,” *Science*, vol. 343, no. 6168, p. 229, Jan. 2014.
[7] “Journals unite for reproducibility.,” *Nature*, vol. 515, no. 7525, p. 7, Nov. 2014.
[8] M. J. Khoury and J. P. A. Ioannidis, “Medicine. Big data meets public health.,” *Science*, vol. 346, no. 6213, pp. 1054–5, Nov. 2014.
[9] J. H. Chen, M. K. Goldstein, S. M. Asch, and R. B. Altman, “Dynamically evolving clinical practices and implications for predicting medical decisions,” *Pac Symp Biocomput.*, 2016.
[10] H. Ma and C. Weng, “Identification of questionable exclusion criteria in mental disorder clinical trials using a medical encyclopedia,” *Pac Symp Biocomput.*, 2016.
[11] A. Koire, P. Katsonis, and O. Lichtarge, “Repurposing germline exomes of the cancer genome atlas demands a cautious approach and sample-specific variant filtering,” *Pac Symp Biocomput.*, 2016.
[12] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart, “The Cancer Genome Atlas Pan-Cancer analysis project.,” *Nat. Genet.*, vol. 45, no. 10, pp. 1113–20, Oct. 2013.
[13] B. Callahan, D. Proctor, D. Relman, J. Fukuyama, and S. Holmes, “Reproducible research for the fine scale analyses of personalized human microbiome data,” *Pac Symp Biocomput.*, 2016.
[14] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. H. Yang, and J. Zhang, “Bioconductor: open software development for computational biology and bioinformatics.,” *Genome Biol.*, vol. 5, no. 10, p. R80, Jan. 2004.
[15] Y. Xie, *Dynamic Documents with R and knitr*. 2014.
[16] S. S. Young and A. Karr, “Deming, data and observational studies.”
[17] C. J. Patel, B. Burford, and J. P. A. Ioannidis, “Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations,” *J. Clin. Epidemiol.*, Jun. 2015.
[18] A. K. Manrai, B. L. Wang, C. J. Patel, and I. S. Kohane, “Reproducible and shareable quantifications of pathogenicity,” *Pac Symp Biocomput.*, 2016.