Making the case for data sharing: ImmPort

Atul Butte, MD, PhD
Chief, Division of Systems Medicine,
Departments of Pediatrics and Genetics, and by courtesy,
Departments of Medicine, Pathology, and Computer Science
Center for Pediatric Bioinformatics, LPCH
Stanford University

abutte@stanford.edu
@atulbutte
@ImmPortDB
FLOCK (FLOw Clustering without K means) uses a density-based clustering approach to algorithmically identify biologically relevant cell populations. Provides statistical analysis of populations.

Two dimensional and three dimensional visualization of cell populations.

Compare population statistics from multiple samples.

Flow Cytometry Analysis (FLOCK)

Flow cytometry analysis component includes:
- Automated cell population identification
- Result visualization in 2D and 3D
- Statistical visualization of population characteristics
- Automated mapping of populations across multiple samples

New Data Release
August 16, 2013 - The National Institute of Allergy and Infectious Diseases (NIAID) released to the ImmPort user community new data from 6 clinical studies or trials and updates to 7 additional studies available here. Research areas include predictive influenza biomarkers, antibody responses to pH1N1 and oral immunotherapy for childhood allergies. This release brings the total number of shared studies to 60.

MHC Validation and Analysis
MHC Sequence Feature Variant Type (SFVT) Analysis enables genetic association analysis of classical HLA protein sub-regions defined with structural (e.g. helix) and functional (e.g. binding site) information.

MHC Alleles
Complete DNA and protein sequences, sequence features, and population frequencies of MHC, MIC and TAP alleles. Align MHC sequences horizontally to visualize extent of polymorphisms across all alleles in a locus.

Data Summary
- Studies: 60
- Subjects: 13859
- Experiments: 569
- Total Results: 226998
- ELISA Results: 126976
Making the case for data sharing

• Reproducibility
• Transparency
• Enable learning
• Return data to the community
• Enable new ventures
• New science
Reproducibility

Repeatability of published microarray gene expression analyses

John P A Ioannidis¹–³, David B Allison⁴, Catherin Mario Falchi⁸, Cesare Furlanello¹⁰, Laurence Gai Michael Nitzberg⁵, Grier P Page⁴,¹², Enrico Petretti

Given the complexity of microarray-based gene expression studies, guidelines encourage transparent design and public data availability. Several journals require public data deposition and several public databases exist. However, not all data is publicly available, and even when available, it is unknown whether the published results are reproducible by independent scientists. Here we evaluated the replication of data analysis in 18 articles on microarray-based gene expression profiling published in Nature Genetics in 2005–2006. One table or figure from each article was independently evaluated by teams of analysts. We reproduced two analyses in principle and six partially or with some discrepancies; ten could not be reproduced. The data is available from the authors.

Figure 1 Summary of the efforts to replicate the published analyses.
Disclose all data in publications

After thousands of hours of investigation, three clinical trials at Duke University in Durham, North Carolina, were suspended in late 2009 because of the irreproducibility of the genomic ‘signatures’ used to select cancer therapies for patients. Journals have a duty to help the community by maintaining reproducibility as a cornerstone of the scientific process.

The independent reanalysis of these signatures took so long because the information accompanying the associated publications was incomplete. Unfortunately, this is common: for example, a survey of 18 productive biomedical journals found that only 15% of articles provide comprehensive data access statements that are clear, complete, and unambiguous.”

Why the inaction on biodiversity?

We are launching an initiative to assess whether or not decision-makers are serious about wanting to halt the biodiversity crisis (S. H. M. Butchart et al. Science 328, 1164–1168; 2010).
Recommendations

Discovery Phase

Phenotypic phase are intended for discovery (should be satisfied and fully optimized (e.g., from a high-throughput application) to enable
translation).

The candidate omics-based test is designed using an independent computational model and,
through phenotypic data until after the phenotype is validated down and the candidate omics-based
proteins; The candidate omics-based test is analyzed and managed database (e.g., using the
X-LiP [MtbMAP]) in standard format; the computational procedures used to develop the
candidate omics-based test should be made
available, described precisely, including the computational procedures, and the intended
uses of the omics-based test.

Validation Phase

3. For investigators conducting a clinical trial to assess the clinical utility and use of an
omics-based test that has been confirmed and validated as described above (Recommendations 1-2), the committee recommends that:
   a. Investigators should communicate early with FDA regarding the investiga-
tional device exemption (IDE) process and validation requirements.
   b. Omics-based tests should not be changed during a clinical trial without a
protocol amendment and discussion with FDA. A substantive change to the
omics-based test may require that the study be restarted.

---

^For publicly funded research, code and fully specified computational procedures should
be made publicly available either at the time of publication or at the end of funding. For
commercially developed tests, this information would be submitted for FDA review if seeking
approval or clearance, or would be described in a publication in the case of a laboratory-
developed test.
Making the case for data sharing

- Reproducibility
- Transparency
- Enable learning
- Return data to the community
- Enable new ventures
- New science
Access to Patient-Level Data from GlaxoSmithKline Clinical Trials

Perry Nisen, M.D., Ph.D., and Frank Rockhold, Ph.D.

Efforts are under way to increase the transparency of clinical trial data. Among the efforts are those being undertaken at GlaxoSmithKline, where we have recently made a commitment to provide access to deidentified patient-level data. We are taking this step because it is the right thing to do, both scientifically and for society, and it is in line with our company’s commitment to transparency in clinical trial reporting. As of May 2013, investigators are able to request access to deidentified patient-level data from a subset of GlaxoSmithKline-sponsored clinical trials. We expect that research teams requesting data will have sufficient statistical and data-management expertise to evaluate these data sets, and we call for standards in analysis to be defined by the scientific community. This article describes the approach we are adopting and answers some of the most common questions we have received since our announcement of this policy.

Which GlaxoSmithKline Studies Will Be Available?

The GlaxoSmithKline Clinical Study Requests website (https://clinicalstudydata.gsk.com) lists studies available as of May 7, 2013, when the website was launched. These studies are also listed in Section 1 in the Supplementary Appendix, available with the full text of this article at NEJM.org. The list includes clinical studies of medicines that have been approved for specific indications (anywhere in the world) and medicines that have been terminated from development. Manuscripts describing the results of a clinical study are made available in the Supplementary Appendix after publication.

Initially, we are including globally conducted clinical studies (phases 1 through 4) started since January 2000. The GlaxoSmithKline database contains data collected during studies conducted anywhere in the world and will be updated periodically. In the future, we will consider making available additional studies with other characteristics.

Roche offers researchers access to all Tamiflu trials

Deborah Cohen

BMJ

More than three years after the Cochrane Collaboration first asked Roche for the full clinical study reports for its influenza drug oseltamivir (Tamiflu), the Swiss company has offered the collaboration access to “all 74 Roche sponsored trials.”

Don MacLean, life cycle leader for Tamiflu at Roche, emailed the Cochrane researchers on 2 April to propose providing data in a staggered approach over the next few months.

“In line with European Union law, each CSR [clinical study reports] will be edited by Roche to ensure patient confidentiality and to protect legitimate commercial interests,” he wrote.

A full phase III clinical study report typically consists of 2000 to 3000 pages, and redaction would be a “large undertaking,” he added.

The Cochrane group has cautiously welcomed the move, pointing out that Roche has previously promised access to data. MacLean’s email follows GlaxoSmithKline’s decision to release 30 clinical study reports on its influenza drug zanamivir (Relenza) to the same Cochrane group. The group is concerned, however, that data redaction and other problems may make analysis and interpretation impossible.

Follows that regulators were in the same situation as we were: lacking data to come to firm conclusions.”

Then in March, Roche said that the company had appointed various third parties to review the data on Tamiflu, identify any gaps, and decide on an analysis plan. It had set up a group for this purpose and invited the Cochrane Collaboration to participate.

The so-called Multiparty Group for Advice on Science (MUGAS) will be organised by the European Scientific Working Group on Influenza, funded by an unrestricted grant from Roche. This group is due to meet on 18 June.

However, the Cochrane Collaboration has asked for further clarification before deciding whether it will accept the invitation.

The researchers are concerned that three of the four key MUGAS partners are scientific advisers or consultants to Roche. A previous BMJ investigation has also highlighted that the European Scientific Working Group on Influenza has a remit to lobby politicians, to “develop a policy for antiviral stockpiling” and to impress on government representatives that “use of antivirals is beneficial and safe.”
John Holdren, Director of the Office of Science and Technology Policy, “has directed Federal agencies with more than $100M in R&D expenditures to develop plans to make the published results of federally funded research freely available to the public within one year of publication and requiring researchers to better account for and manage the digital data resulting from federally funded scientific research.”
Making the case for data sharing

• Reproducibility
• Transparency
• Enable learning
• Return data to the community
• Enable new ventures
• New science
Enabling Learning

Would the training of pathologists (or other professionals) change if hundreds/thousands of trial-labeled images were publicly available?
Making the case for data sharing

• Reproducibility
• Transparency
• Enable learning
• Return data to the community
• Enable new ventures
• New science
• Research subjects and their support groups could conceivably access data related to their diseases
Making the case for data sharing

• Reproducibility
• Transparency
• Enable learning
• Return data to the community
• Enable new ventures
• New science
Validation and Reproducibility of a Microarray-Based Gene Expression Test for Tumor Identification in Formalin-Fixed, Paraffin-Embedded Specimens

Tumors whose primary site is challenging to determine represent a considerable proportion of all cases. We present validation study results of a novel expression-based diagnostic test (the International Genomics Consortium’s Expression Project for Oncology [expO], deposited in the GEO database) under the GSE number 2109, and from commercial data sources. The goal was to include files that spanned an adequate range of laboratories and operators to ensure that the genes used for standardization were highly robust.

Standardized expression data are evaluated in the test’s classification algorithm, which was developed using a database of 2032 frozen and 104 FFPE specimens, divided into independent training and test data sets. The training set data are a combination of public data sets, commercial data sources, and private correspondence. The publicly available component has been derived from GEO (GSE number 2109). All training specimens had been assigned one of 15 tissue of origin diagnoses according to standard clinical and pathological practices.
All the world’s genomic data at your fingertips?

Modern medicine is quickly becoming an information-driven field. Developing personalized therapies depends on intelligently leveraging complex molecular and public sources.

NextBio's big data technology enables users to access proprietary molecular data and clinical information from model organisms, thus applying genomic data to human health in ways never before possible.
The Immunological Genome Project is a collaborative group of Immunologists and Computational Biologists who are generating, under carefully standardized conditions, a complete microarray dissection of gene expression and its regulation in the immune system of...
Making the case for data sharing

• Reproducibility
• Transparency
• Enable learning
• Return data to the community
• Enable new ventures
• New science
Take home points

• ImmPort isn’t the end of the line, it’s a beginning
• Many successful uploads into ImmPort
• Looking at data is good; reanalyzing it is great
  – Clinical trial results can be reproduced and expanded
  – For external investigators to do this, they need to find the data
• While individual studies are great, pooling studies may be more powerful
  – Pooling enables new questions that coexist with investigators
  – Let’s learn why trials work and fail
  – Need better annotations (ontologies) to make this work
• Linking molecular and clinical trial data will enable even more virtual studies and findings
  – Need catalogs and annotations to make this work
Collaborators

- Northrop Grumman: Jeff Wiser, Patrick Dunn, Mike Atassi, Patty Berger, John Campbell, Vincent Desborough, Keith Ferguson, Jason Lucas, Prabhu Byrappagowda, Thomas Smith, Liz Thomson, Bryan Walters
- NIAID: Ashley Xia and Quan Chen
- Stanford: Yannick Pouliont, Li Li, Mazen Nasrallah, Sanchita Bhattacharya, Rachel Finck, Gary Nolan, Petter Brodin, Mark Davis, Susan Aptekar, Rhonda Pisk
- Technion: Shai Shen-Orr
- Buffalo: Barry Smith, Alan Ruttenberg, Alex Diehl, Susan Smith
- ESAC Inc: Henry Schaefer, Yezhou Sun
- Glenys Thomson and Richard Single