Task Force on Enhancing Translational Discovery in Biomedical Research
September, 2017

Executive Summary

A group of senior faculty and other school leaders[1] representing different disciplines in basic, translational, and clinical research was constituted to review widely-publicized difficulties in translating research findings from animal and cell models into new diagnostic tests and treatments for humans. The discussion expanded to consideration of the rigor and reproducibility of animal and cell data and the use of electronic laboratory notebooks. The Committee formulated several findings with unanimous support from the group. First, it is important to emphasize that, despite the publicized lack of reproducibility and failures of translation, the Committee noted the essential role that animal and cell research has played in driving the vast majority of all clinical advances over the past several generations. Second, the Committee identified a large number of wide-ranging factors that contribute to the reported lack of reproducibility and translation, including the use of insufficiently optimized animal and cell models, wrong incentives in current grant and journal policies, and deficiencies in many efforts at translation. Third, based on these findings, the Committee identified several steps to improve reproducibility and translation: some of these efforts can be instituted at Mount Sinai, whereas others require national action. Such steps include regular training for faculty—to augment the current training of students and postdoctoral fellows—in the responsible conduct of research and in data rigor and reproducibility, as well as improvements in the use of animal and cell models and the translation of basic findings to humans.

A History of Success

The committee began its work by rejecting the widely-publicized criticism that animal and cell data are poorly reproducible and have largely failed to advance clinical diagnostics and treatment. Committee members documented a large number of celebrated advances in biomedical research in each of their areas of expertise, which reflect highly successful translation and important improvements in patient care. There are enumerable examples over the past generation. A few recent examples among many—from the past few years alone—are the following. 1) The use of agents of several mechanisms of action to prevent restenosis of coronary arteries after stent placements was based on initial discoveries of the ability of these agents to inhibit vascular smooth muscle cell proliferation in animals. 2) Protein tyrosine kinases were discovered as oncogenes in cell and animal models and led to the development of numerous inhibitors of this class of enzyme which are now used widely in the treatment of several cancers as well as non-cancerous proliferative disorders. 3) The receptor for one of these protein tyrosine kinases is HER2 (ERBB2), which was discovered as an oncogene in an initial screen of cultured cells and later found to be amplified in a breast cancer cell line; antibodies to HER2 represent a major advance in the treatment of HER2+ breast cancer. 4) The discovery of so-called check-point inhibitors, one of the most exciting advances in cancer treatment today, was based on identifying a novel cell-surface protein on lymphocytes that controlled their activity in culture, which led to the discovery that antibodies to this protein inhibit tumor growth in mice. 5) The discovery of toll-like receptors (TLRs) and dendritic cells, two major components of the immune response, in *Drosophila* and mice led to their study in humans and their eventual use in improving vaccine design.

The point that the "crisis" in animal data reproducibility is overstated is supported by a study[2], which found that replication in humans was achieved two-thirds of the time when attempted for highly cited animal studies, but that half of animal findings remained without attempts at human

[1] Task Force: Eric Nestler (Chair), Jonathan Cohen, Marta Filizola, Scott Friedman, Bruce Gelb, Annetine Gelijns, Alison Goate, Roger Hajjar, Paul Kenny, Paul Lawrence, Erik Lium, Madhu Mazumdar, Miriam Merad, Reginald Miller, Ramon Parsons, and Andrew Stewart.

[2] *JAMA*, October 11, 2006—Vol 296, No. 14, pg. 1731.

replication. A separate study documented the dramatic contribution of academic research in leading to the most transformational advances in clinical treatment over the past generation.[3]

Nevertheless, there are also many spectacular failures of translation across all fields of medicine. Additionally, there are several reports, which have received widespread attention in the lay press, about the lack of reproducibility of animal and cell findings including those reported in top-ranked journals. The Task Force identified several reasons for these failures as well as tangible steps that the field as a whole, and we at Mount Sinai in particular, can take to advance the success of translational research. The Task Force also noted the related lack of replication of clinical findings, but did not address this challenge in detail. It is recommended that a separate Task Force be considered to focus on reproducibility of human studies.

Factors that Hinder Translation

Clinical Relevance of Animal Models

Animal models have provided many crucial insights into normal organ physiology, which have led to major advances in medicine and surgery. Animal modeling can also refine dosing and duration of novel therapies, establish proof of principle for target engagement, and identify unanticipated effects of the therapies. However, numerous weaknesses are also identified:

- Animal models need to evolve to continue to reflect the latest understanding of human pathophysiology. One example is human tumor xenografts in immuno-deficient animals. Once considered the gold standard of cancer research, this model which notably lacks an adaptive immune system, cannot be used to probe immunotherapy benefits. Too many studies across all fields perseverate in using older models.
- Animal models also need to be tailored to the research question of interest. Too often, the animal studies are not designed to generate data that would then advance a discovery into the clinic. As one example, manipulation of a gene in one cell type might produce a therapeutic-like effect even though manipulation of that gene globally either does not produce a therapeutic-like effect or exerts unacceptable toxicity. (In parallel, too often clinical studies are designed in a way that does not make optimal translation of animal findings, as discussed below.)
- In prior decades, animal models were validated solely by phenocopying a physiological or behavioral feature of a disease. Today's advances permit far more penetrating validation, such as by aligning transcriptomic signatures seen in an animal or cell model with those of diseased human tissue or by replicating a known etiology in an animal.
- Too many animal studies are performed on a single genetic (incred) background of rodent or on a single species. Replicating discoveries across several genetic backgrounds, using outbred rodent lines, and across multiplef species may increase translatability to humans.
- The lack of data rigor and reproducibility likely reflects numerous factors, including underpowered studies, reporting non-replicated findings, cherry-picking experiments that work, inadequate reporting of experimental details, wrong incentives in scientific journals, among many others. Some of these are discussed further below. Interestingly, while all students and postdocs have required training in the responsible conduct of research and in data rigor and reproducibility, faculty do not currently have such required training.

Better Linking Animal and Cell Data with Human Investigations

Too many researchers make insufficient attempts to pair animal data with human data, either within their own labs or via cross-lab collaborations. This is a major gap in animal research and one that can be addressed at the institutional level by: 1) emphasizing to trainees the urgency to

[3] *Health Affairs*, February 2015—Vol 234, No. 2, pg. 286

pair animal data with clinical data as early as possible in a research project, and 2) facilitating access to human samples. Pairing results is bidirectional, as noted above: creating a basic dataset that can be tested in humans, while testing in humans what the basic data tell us, and then taking human findings back to animal models for mechanistic exploration. One key way to bridge animal and human data is to expand the study of human tissue samples from normal and pathological conditions. A great deal could be done to increase the size and access of human tissues to basic researchers; this includes the reducing the regulatory obstacles to access such tissues.

Curating Public Profiling Datasets

The accessibility of sequencing platforms and other types of "big data" has led to an explosion of publically available human data with little knowledge on tissue collection, tissue quality, patient demographics, etc. It is essential for the field to do a better job of reviewing the quality of any dataset examined prior to further analysis.

Problems with Journals and Publication Policies

The competitive nature of the field continues to increase, with faculty, postdocs, and students under ever-increasing pressure to not only publish their research findings but to do so in top journals of the field. The ability of a postdoc to secure a job in academia, or the ability of a faculty member to get promoted or get a grant, requires publications in top journals, with too little consideration given by the hiring institutions or grant agencies to the quality of the research per se. This rush to publish creates the wrong incentives. And the rush to publish in top journals makes matters worse. Top journals force authors to shorten papers by removing any confounding data—these journals want a simple, linear story with no complications; and few journals allow the space needed to report detailed experimental methods (see next section below). Most journals also do not provide the statistical review of data to ensure rigor and reproducibility.

Experimental Details Can Profoundly Influence Experimental Findings

There are enumerable examples where factors thought by some to be "minor" have a dramatic effect on experimental results, with such experimental details rarely being included in publications. Some examples are as follows. 1) Different brands of standard rodent chow can make the difference between a given mutant mouse line—on the same genetic background—being obese or not. 2) Different brands of bedding have a dramatic effect on agonistic encounters between rodents, and make the difference in the development of behavioral abnormalities versus normal behavioral function. 3) The same inbred mouse line (e.g., C57BL6) or same outbred rat strain (e.g., Sprague-Dawley) obtained from different vendors can exhibit widely different physiological functions and responses to pharmacological agents. 4) The infection history and immune status of laboratory rats and mice can have a dramatic effect on physiological functions and responses to pharmacological agents. Such details are rarely if ever reported in most scientific publications. The Task Force believes that accounting for such differences would improve the replication of experimental findings.

Quality of the Research Team

The Task Force believes that some cases of the failure to replicate experimental findings are due to the fact that the replicating team is not as expert in the experimental approach as the original group. Just as the burden of proof is on the discovery team that their data can be replicated, the burden of proof is also on the replicating team to do so in a similarly competent, expert manner.

Underpowered Studies

Too many studies are underpowered. This is the case for both animal and human studies. As alluded to above, some basic experiments are reported without replication. There is selective reporting of only those experiments that "worked," without mention of those that didn't. NIH now

requires that sex be considered as a biological variable in animal and human studies, although most studies in both cases are not sufficiently powered to study sex differences.

Optimizing Clinical Study Design to Test Pre-Clinical Findings

Although this Task Force focused primarily on enhancing the rigor, reproducibility, and translatability of basic research, there was some discussion of the challenges in clinical research that have also contributed to today's difficulties with successful translation. Many, perhaps even most, first in human clinical studies are not designed to specifically translate precisely what the animal studies taught the field. This has prompted researchers to design experimental paradigms that can be carried out across species. There is also an extreme paucity of the right tool compounds. Too often basic research implicates a specific mechanism of action, yet that mechanism is tested in humans with compounds that are non-selective or only weakly engaging the mechanism of interest. One way to frame the challenge of translation is that only a miniscule portion of the best-validated animal findings ever get tested in humans due to the lack of the right compounds and the regulatory burden of testing new mechanisms in humans.

In terms of clinical study design, too many clinical trials are small and insufficiently powered, and do not rigorously adhere to data collection standards. Additionally, there is too high an attrition rate from early and late phase clinical trials. Increasing evidence suggests that later phase clinical studies fail in part due to poorly designed early phase studies (in terms of optimal dose, etc.). The use of multiple sites and CROs around the world, which likely vary in quality—employed to enroll very large numbers of subjects as quickly as possible—also contributes to the failure of late phase clinical trials.

Recommendations

Based on these findings, the Task Force formulated numerous recommendations, some of which we can institute at Mount Sinai, but many others that would require a sea change in the field.

General Steps to Increase Rigor & Reproducibility at Mount Sinai

- Continue to optimize research rigor and experimental design courses for students and postdoctoral fellows. (This is ongoing with the Graduate School.)
- Initiate a new web-based training module for all new faculty focused on the responsible conduct of research and the rigor and reproducibility of data, with a triennial refresher module for all faculty. (This is being developed by a subcommittee of this Task Force.)
- Introduce a recommended campus-wide electronic notebook system. This will not be required, but all PI's laboratory notes will have to meet the same standards as in the electronic system. One consideration for an ELN platform is that few adequately support both chemistry and biology equally. As a result, two general approaches may be needed. Additionally, an ELN which is incorporated into a larger Laboratory Information Management System (LIMS) and supporting ancillary modules (Biobanking) is ideal. One obvious benefit of improved recordkeeping practices, such as with an ELN, is during the patent application process. A separate group on campus is working to evaluate steps to institute this recommendation (report pending).

Optimal Use of Animal Models of Human Disease

- Use optimally validated animal models and, before translation, provide a preclinical package (i.e., intention-to-treat approach).
- Animal models used to test novel therapies should replicate the major molecular and cellular features of the human disease, and the key pathogenic drivers, where known.
- Complement the use of genetically modified animals—which can yield misleading evidence—with other approaches, especially in wild type animals.

- Therapies tested in animals (e.g., antibodies) should be fully cross-reactive in humans, or utilize surrogate therapies with a very similar target or identical profile of activity.
- Dosing and duration of experimental therapies in animal models should seek to achieve similar PK and PD profiles as in human disease, and replicate the anticipated route of human administration.
- Use of more than one strain of animal, more than one species, and/or the use of outbred animals may enhance confidence in the translatability of animal findings to human disease, but this approach runs up against practical considerations of cost and time.
- Use of 'hybrid' models such as human tissue slices, iPS-derived tissues, or organoid cultures may reduce the reliance on animal models and/or refine experimental conditions for testing in vivo.
- Every effort should be made to control for numerous variables including age, sex, housing conditions, temperature, circadian rhythm (lighting), and especially the microbiome's contribution to a phenotype (see next section).
- Investigators assessing animal responses should be blinded to experimental conditions, and results should be objectively measured.

Full Reporting of All Experimental Details Related to Animals

- The following details on animals and husbandry conditions should be reported for all experimental studies: animal species, strain, vendor, sex, age, and weight, health/immune status, size of caging (tanks for fish), bedding material, number of cagemates or singly housed, and husbandry conditions (e.g., breeding program, light/dark cycle, temperature, humidity, quality of water, type of food, access to food and water, environmental enrichment).
- The following variables on experimental design should be reported for all studies: drug formulation, source, and dose, site and route of administration, anesthesia and analgesia, surgical procedure, method of euthanasia, and time of day and location of experiment (e.g., home cage, laboratory, behavioral chamber).
- Welfare-related assessments and clinical interventions that were carried out prior to, during, or after the experiment.
- All genetic mutant rodent lines should be verified with regular, periodic genotyping.

Improving the Use of Cell Lines

- Confirm the quality of cell lines with regular, periodic genotyping.
- All cell studies should be fully blinded.

Improving the Use of Chemicals

To be tested in biological assays, the identity of chemical compounds and their purity (>95%) must be confirmed by well-established analytical methods, including ^{1}H and ^{13}C NMR (nuclear magnetic resonance), HRMS high-resolution mass spectrometry), and HPLC (high performance liquid chromatography).

Improving the Use of Computations

Replicability and reproducibility of computational protocols must be enforced by keeping manual intervention to a minimum and utilizing documented and version-controlled automatic scripts for the execution of all steps in each protocol. Input files, codes, and protocols are encouraged to be made accessible in the Open Science Framework (https://osf.io/) maintained and developed by the Center for Open Science (COS), which is supported, in part, by NIH and NSF to foster openness, integrity, and reproducibility of scientific research. Analysis scripts must be drafted following the logic of literate programming[4] and, where possible, the code for the generation of

[4] Gandrud, C., *Reproducible Research with R and R Studio*. Chapman & Hall/CRC The R Series. 2013: Chapman and Hall/CRC.

tables and figures must be reported. For computational strategies that allow estimation of uncertainties of the predicted values (e.g., free-energies, interaction probabilities, rates, etc.), error bars must be calculated and the statistical significance of the results tested and reported.

Facilitating Human Tissue Access to Basic Scientists

- Streamline the process by which basic scientists can access human tissues.
- Strengthen the biobanking effort through comprehensive banking of curated human samples with extensive clinical annotations and user-friendly query tools.
- Foster the use of the biobank by basic scientists through institution-funded pilot programs.
- Expand the technology platforms that facilitate the mining of human samples.
- Pay greater attention to the quality of publically available human datasets.

Fostering Better Translation

- Answer these three questions:
    1. Do you have the right data from your animal model to help you choose which humans to study (subtypes) or how to study them (biomarkers, dose, etc.)?
    2. Are you performing true translation: are you really studying the same thing in humans as you did in animals?
    3. Are you willing (is it possible) to test your precise hypothesis even if it makes your studies costly, inefficient, and slow to complete? (For example, if you have to screen many to get a few, if you have to encumber trials with burdensome and costly biomarkers, if you have to have fewer sites to reduce the noise in recruitment, etc.).
- Foster better coordination between clinicians and basic scientists through monthly WIPS.
- Consider more internal funding opportunities for basic-clinical partnerships.
- Greater access to human tissues, as noted above.
- Consider increased use of CROs for preclinical studies.
- Consider developing a new type of WIP focused on translational studies that “failed” to better optimize follow up investigations.

Statistical and Other Experimental Considerations

- Specify the total number of animals or cell preps used in each experiment, and the number in each experimental group.
- Explain how the number of animals/cell preps was arrived at, and provide details of sample size calculation used. Likewise, report results of power analysis.
- Avoid underpowered pilot studies when animals are involved.
- Use the appropriate types of statistical tests to analyze data.
- Indicate the number of independent replications of each experiment.
- Specify whether any animals/cell preps were removed from analysis and objective rationale to support such removal.
- Specify whether any experimental replications were not included in final analysis and objective rationale to support such removal.
- Understand the difference between biological vs. technical replicates and the ambiguity of this dichotomy.
- Use methods and experimental designs that include internal controls to assess the validity of findings.
- Acknowledge difference between learning a method and repeating an experiment.

- Work with experts and collaborate in areas outside of expertise, e.g., biostatistics, pathology, etc.

Improving Research and Journal Practices

- Improve peer review including better evaluation of statistical rigor and data reproducibility.
- As above, full reporting of all experimental details
- Deposit all raw data in publically available dababases.
- Validate all key reagents and report their full details including source, antibody batches, etc.
- Use a pre-determined statistical framework for analyzing cell and animal studies akin to current practices with clinical trials.

Improving Clinical Research

- The Task Force recommends that the new Institute for Transformative Clinical Trials undertake a review of clinical study practices to make parallel improvements in this domain. Some considerations are as follows.
- Improving the rigor and reproducibility of clinical trials by carefully attending to and pre-specifying study design. Trial protocols need to specify endpoints, careful definition of adverse events, sample size, patient eligibility criteria, and ways to reduce bias (such as blinding). They also should pre-specify how one plans to deal with missing data.
- Better align early and late clinical studies. Incorporate as early as possible studies to identify optimal drug dose. Adaptive designs and other newer methodological approaches may be helpful to evaluate multiple doses in an efficient manner.
- In terms of trial execution, investigators need to develop a structured database, carefully document screened and eligible patients, and review for completeness the accuracy and consistency of collected data (especially for the primary endpoint).
- Place all human data in the public domain. Following trial completion, the analysis dataset needs to be "locked." This dataset needs to be accompanied by a data dictionary (specify derived variables, e.g., composite endpoints, age, algorithm for missing data analysis), and the statistical code used for primary and secondary analyses.
- The use of electronic health record data to derive treatment outcomes should be accompanied by several approaches to improve reproducibility of the findings, such as specifying the process by which data are downloaded from the EHR (window of download, patient selection criteria, etc.), documenting the "data cleaning" process, and identifying how unstructured data are used in the database (file algorithms used to mine unstructured fields).
- The careful and complete specification of study design and the detailed reporting of trial results on ClinicalTrials.Gov will allow for greater reproducibility.

Promote a Culture of Transparency and Integrity in Research

Mount Sinai and other institutions already do a lot to achieve this goal. Much of what is recommended above can be understood within this crucial theme. We refer the community to a recent report of the Institute of Medicine (now National Academy of Medicine)[5] on how to promote such a culture, and emphasize its primary importance to what we do.

[5] Integrity in Scientific Research, Institute of Medicine, 2002 (https://www.nap.edu/read/10430/chapter/2#2).