The integration of the data scientist into the team: implications and challenges

Abstract

Modern biomedical research is complex and requires a cross section of experts collaborating using multi-, inter-, or transdisciplinary approaches to address scientific questions. Known as team science, such approaches have become so critical it has given rise to a new field – the science of team science. In biomedical research, team-based collaborations have great need for data scientists. Integration of data scientists into research teams has multiple advantages to the clinical and translational investigator as well as to the data scientist. Clinical and translational investigators benefit from having an invested dedicated collaborator who can assume principal responsibility for essential data-related activities, while the data scientist can build a career developing tools that are relevant and data-driven. Participation in team science, however, can pose challenges to the promotion of the data scientist. One particular challenge is the ability to appropriately evaluate the data scientist's scholarly contributions, necessary for promotion. Only a minority of academic health centers have attempted to address this challenge. In order for team science to thrive on academic campuses, leaders of institutions need to hire data science faculty for the purpose of doing team science, with novel systems in place that incentivize the data scientist's engagement in team science and that allow for appropriate evaluation of performance. Until such systems are adopted at the institutional level, the ability to conduct team science to address modern biomedical research with its increasingly complex data needs will be compromised. Fostering team science on campuses by putting supportive systems in place will benefit not only clinical and translational investigators as well as data scientists, but also the larger academic institution.

Team science involves a collaboration of scientists across multiple disciplines to jointly solve scientific questions. While there is still a place for the traditional single investigator-led initiative, there is a trend toward such multi-, inter- and transdisciplinary approaches, particularly in modern biomedical research, as the complexity of research questions and the data to address them has increased.[1] Further, individual labs have become more specialized, as they have consequently been generating and/or handling more specific types of data for which a unique skill set is required. Thus, today, the modern scientist cannot be expected to know all that is necessary about the data generated or the implications of how to address the research question raised. Leveraging expertise outside the scientist's lab and across multiple labs or disciplines can therefore prove efficient and advantageous.

An example of a study where such team-building may be beneficial includes one initiated by a clinical endocrinologist interested in determining whether data from continuous glucose monitoring devices can inform behavior to reduce postprandial hyperglycemic events among diabetics. The investigator may consider a study team that additionally includes a biomedical engineer with knowledge of such devices and the type of data generated, an informatics expert who can extract relevant signal from the continuous measurements, and a statistician with expertise in predictive modeling who can also help design the study. Another example includes a study of the role of microbes in maintaining a healthy skin barrier, which may require a team with expertise in clinical dermatology, statistics, research informatics, systems biology, molecular biology, and genomics, as Dr. Julie Segre from the National Human Genome Research Institute assembled herself when addressing such questions. [2] I am involved as a data scientist in a randomized clinical trial of obese children to evaluate whether a new intervention can reduce obesity. One goal is to assess the role of physical activity – measured using accelerometers -- in reducing obesity. The accelerometers generate over 180 million data points per person, and analyses involving such data are not trivial. Our team includes experts in preventive medicine, sleep medicine, and exercise physiology, as well as data scientists with expertise in software engineering, clinical informatics, distributive computing, and biostatistics. Such teams have proven to be essential to addressing modern scientific problems.

Team science has been so critical to the success of many of today's scientific endeavors that a new field has emerged – the science of team science. [3] The science of team science involves the study and development of methods to understand and improve upon the process and outcomes of conducting team science. Conferences that discuss new tools and strategies for successful cross-disciplinary (including multi-, inter-, and transdisciplinary) engagement have been established. In addition, research articles in scientific journals that describe methods for cross-disciplinary collaboration are appearing (e.g.,See Börner et al., 2010 and Bennett and Galdlin, 2010), [4-5] and field guides on approaches to create successfully productive teams have been developed. [6] As the need for team science persists, the field for underlying tools and processes that guide cross-disciplinary research will continue to grow.

Data scientists are natural and often critically important candidates for membership of a team conducting cross-disciplinary research. For example, as shown in the examples above, most biomedical research has a great need for expertise specifically in data science. While developing skills and training in data science may not necessarily include a biomedical context, in practice, understanding the nuances of the methods applied to a biomedical context is essential and best done in a team environment with content expertise. For example, as part of a group working with investigators to study the comparative effectiveness of HIV anti-retroviral agents on cardiovascular disease, team members with expertise in treating HIV noted the importance of adjusting for potential confounders like CD4 count, viral load, and cholesterol levels. Further, they emphasized that while the literature implicated single agents as having a role in cardiovascular disease, in practice, therapy was prescribed and taken in combinations, making the entire combined therapy a more pertinent focus. In addition, as we observed in the data, patients switched their combinations with great frequency. The clinicians on the team confirmed that this was common because if the treating physician saw poor management of HIV or an unfavorable cardiovascular profile, a different regimen would be prescribed. This information directly informed how we data scientists designed the study and developed our statistical models.[7] Without the context provided by the HIV experts, the data scientists would have developed a less relevant model, and possibly provided misleading findings. Thus, integration of the data scientist into the team environment enabled the data scientists to arrive at an approach that yielded clinically appropriate interpretation. The iterative team-based process also more generally ensures that products from the collaboration will be relevant.

In addition to providing important context to developing the application, there are other benefits to team integration for the data scientist. When data scientists dig deep as collaborators, gaps in methodology can easily be identified. For example, while a data scientist may recommend use of a specific statistically ideal modeling approach it is only in the team setting that the data scientist would be faced with any issues posed by that suggestion. In the study described above, recall that we wanted to incorporate potential confounders like CD4 count and cholesterol and how they changed over time in order to mitigate potential confounding by indication. However, we found that not all patients had the same amount of information, particularly at or just prior to important changes in regimen or cardiovascular events. Different ways of handling the missing data resulted in markedly different inference about the research question. While issues of missing data are not new to data scientists and can commonly be addressed using maximum likelihood or multiple imputation methods, this study posed unique issues. Specifically, the model for the analysis assumed independence across observations even though patients contributed multiple records, and corrected for this by robustly estimating the standard errors of point estimates of interest. It was unclear, however, how to use a tool like multiple imputation to correspond to this model in a way that would be deemed "proper" and that would exploit and account for the correlation in the data. Further, this had not been addressed previously in the missing data literature. The data scientists on the team therefore led efforts to study this particular question as a contribution to the field of multiple imputation as well as a contribution to the specific project. Thus, this deeper-level collaboration can be career-building for the data scientist as well.

Investigators initiating research of biomedical (or other scientific) relevance will gain from including the data scientist on the team in ways beyond simply having the project's data-related needs met. In my experience, having integrated team members generally means that the roles in the team are somewhat fluid. In the example given above, being part of the team means that some of the resulting papers will be led by the data scientists in the group, some will be led by the clinical investigator who initiated the study, and some may be led by a clinical fellow being mentored by the clinical investigator. Having additional papers led by data scientists will increase productivity for the research team. Further, those methodological findings will directly benefit the investigator's scientific program. Finally, opportunities to take the lead on various aspects of the project ensures that all members are invested. These are all qualities that will lead to a successfully productive collaboration.

While there are great benefits to team science, participation in team science is not without issues. One such issue is obtaining appropriate academic credit for one's work. While I have led teams in efforts that have directly impacted the direction of the research, for many manuscripts I am middle author. This placement of authorship does not signal insignificant contribution, but it does pose a difficulty with the classical metrics used in faculty evaluation:

leading manuscripts and serving as principal investigator on grants. Today, only about one quarter of institutions have revised their metrics for promotion to include some aspect of team science contributions. [8] By adhering to metrics that do not appropriately value team science contributions, academic institutions discourage faculty involvement in team science. This is an area that can be addressed by academic leaders invested in fostering team science on campus.

One such suggestion as to how recognition can be assessed was provided by Mazumdar et al. The authors developed an excellent and systematic approach for how evaluation and promotion could better recognize the intellectual leadership of the team scientist. [8] They further described how to evaluate contributions to publications, grants, and research programs, in order to summarize overall scholarship in a way that appropriately weights contributions to team-based science. While the approach is incredibly valuable, one challenge in its implementation is automating the process which also relies on evaluators outside of the members of the promotion committee. To mitigate this, the authors encourage the development of systems to collect such data at the institutional level. Such systems would enable appropriate recognition of significant intellectual contributions by faculty through team science, allowing team-based collaborations to thrive on campus.

I believe more can be done to foster team science that goes beyond academic leaders acknowledging that team science is critical to their mission. It includes hiring faculty in data science for the specific purpose of doing team science, formally adopting novel metrics that can appropriately evaluate their scholarly contributions, [8] and supporting development of research programs for these individuals. Faculty lines that prominently feature team science as part of the programmatic need should be utilized with incentives to engage in team science, retention plans, and no compromise in expectations for excellence in scholarship. Differences in career trajectories for data scientists who are and are not primarily team scientists should be recognized so that evaluation can be appropriately tailored.

Without establishment of such systems by academic leaders, the possibilities of team science will not be fully realized. Many data science faculty will view engaging in team science as a distraction from more promotable activities and will continue to develop tools with perhaps less relevance than would be possible. Investigators will continue to rely on outside consultants (i.e., non-team members) to meet data-related needs, potentially compromising the science. If academic leaders invest in systems that incentivize clinical and translational investigators working with data scientists toward shared goals that are aligned with the mission of the institution, the benefit will not only be for the investigator and data scientist; it will also be for the institution.

References

1. Elfner L, Falk-Krzensinski H, Sullivan K, Velkey A, Illman D, Baker J, Pita-Szezesniewski A. Team Science-Heaving Walls & Melding Silos, A Sigma Xi White Paper. 2011.
2. Grice, E.A., Kong, H.H., Renaud, G., Young, A.C., Bouffard, G.G., Blakesley, R.W., Wolfsberg, T.G., Turner, M.L. and Segre, J.A., 2008. A diversity profile of the human skin microbiota. *Genome research*, *18*(7), pp.1043-1050.
3. Stokols, D., Hall, K.L., Taylor, B.K. and Moser, R.P., 2008. The science of team science: overview of the field and introduction to the supplement. *American journal of preventive medicine*, *35*(2), pp.S77-S89.
4. Börner, K., Contractor, N., Falk-Krzesinski, H.J., Fiore, S.M., Hall, K.L., Keyton, J., Spring, B., Stokols, D., Trochim, W. and Uzzi, B., 2010. A multi-level systems

perspective for the science of team science. *Science Translational Medicine*, *2*(49), pp.49cm24-49cm24.

5. Bennett, L.M. and Gadlin, H., 2012. Collaboration and team science. *Journal of Investigative Medicine*, *60*(5), pp.768-775.
6. Bennett, L.M., Gadlin, H. and Levine-Finley, S., 2010. *Collaboration & team science: a field guide*. NIH Office of the Ombudsman, Center for Cooperative Resolution.
7. Desai, M., Joyce, V., Bendavid, E., Olshen, R.A., Hlatky, M., Chow, A., Holodniy, M., Barnett, P. and Owens, D.K., 2015. Risk of cardiovascular events associated with current exposure to HIV antiretroviral therapies in a US veteran population. *Clinical Infectious Diseases*, p.civ316.
8. Mazumdar, M., Messinger, S., Finkelstein, D.M., Goldberg, J.D., Lindsell, C.J., Morton, S.C., Pollock, B.H., Rahbar, M.H., Welty, L.J. and Parker, R.A., 2015. Evaluating Academic Scientists Collaborating in Team-Based Research: A Proposed Framework. *Academic medicine: journal of the Association of American Medical Colleges*, *90*(10), p.1302.