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, data scientists often play a critical role in team-based collaborations. 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. 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.

The role of team science in solving modern scientific problems

Team science is a collaborative effort by scientists that join together across multiple disciplines to solve scientific questions (1). While there is still a place for the traditional single investigator-led initiative, there is a trend toward such multi-, inter- and transdisciplinary approaches. This is largely due to a significant increase, particularly in modern biomedical research, in the body of scientific knowledge, the complexity of research questions and the data generated to address them (1). Further, in response to the growing complexity of the scientific landscape, individual labs have become more specialized. Thus, today, addressing scientific problems often requires extensive knowledge across a range of perspectives to fully comprehend the salient issues. Further, leveraging expertise outside the scientist's lab and across multiple labs or disciplines in a deeply collaborative manner can 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 including the annual International Science of Team Science Conference and the Interdisciplinary Network for Group Research Conference. In addition, research articles in scientific journals that describe methods for cross-disciplinary collaboration are appearing (e.g.,See van der Haar., 2017, Börner et al., 2010 and Bennett and Galdlin, 2010 (4-6)), and field guides on approaches to create successfully productive teams have been developed (7). 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 as team scientists

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, I was part of a group of investigators studying the comparative effectiveness of HIV anti-retroviral agents on cardiovascular disease, where 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 (8). 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 meaningful interpretation.

Participation in team science should not be confused with serving as a consultant. Consulting on projects implies providing superficial high-level expertise and does not involve an invested and iterative (or collaborative) effort. Consulting has its place. It can be particularly beneficial when limited high-level advice is needed and can serve as a bridge to a deeper collaboration when appropriate. If relied upon for meeting data-related needs in a project, however, it can potentially compromise the quality of the science. Suppose that for the study above the investigator initiating the question wanted to include preliminary findings in a proposal that demonstrated a particular HIV antiretroviral drug was associated with increased risk of stroke. Using a registry of HIV patients, suppose the investigator assembled data on exposure to various HIV antiretroviral drugs and stroke and then asked a consulting biostatistician to evaluate whether ever being exposed to this antiretroviral drug was associated with stroke in order to justify future experiments that investigate mechanisms for the relationship. A consulting biostatistician could accomplish this simply through a Cox model. However, it would only be through interaction with other investigators, deeper exploration of the data, and perhaps through leading the design and extraction of relevant data herself, that the biostatistician would find that this association varies with time and exists only in the presence of two particular antiretroviral therapies that are no longer administered in combination. Importantly, the association exists only in the earlier periods of the highly active antiretroviral era. Thus, outsourcing this analysis to a consulting setting could have yielded misleading conclusions. Importantly, the iterative team-based process more generally ensures that products resulting from the effort will be relevant.

In addition to providing critical context to developing a method or its application, there are other benefits to team integration particularly 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. This would likely be missed in the consulting context. In the comparative effectiveness 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 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. This work is ongoing and supported by the Patient-Centered Outcomes Research Institute (PCORI). 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 and raise the overall quality of the research. Finally, opportunities to take the lead on various aspects of the project ensures that all members are invested. These are qualities that will lead to a successfully productive research endeavor.

Challenges and solutions to promotion of the team scientist

While there are great benefits, participation in team science is not without issues. One 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 of the resulting manuscripts I am middle author. This placement of authorship may be appropriate and 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. For example, if the data scientist leads in developing the analysis approach, oversees the analysis, provides principal interpretation for the study findings, and writes the methods and results section of the manuscript, a second place in the authorship order may be appropriate. At the same time, when being considered for promotion, such effort should be considered along the same lines as a first author paper, particularly as the effort reflects leadership in the contribution of that discipline. Today, only about one quarter of institutions have revised their metrics for promotion to include some aspect of team science contributions (9). By adhering to metrics that do not appropriately value team science contributions, academic institutions discourage faculty involvement in team science. Consequently, data scientists may opt not to be involved in cutting-edge research where they are not listed as first or senior authors, potentially impacting studies that rely on data science expertise. 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 (9). They described how to evaluate contributions to publications, grants, and research programs, in order to summarize overall scholarship in a way that appropriately weights efforts toward team-based science. More specifically, the authors presented a general framework for evaluating faculty conducting team science among three key domain areas that are typically assessed using traditional metrics: scholarship, teaching, and service. Within each of the three domains, the authors illustrated an approach for assessing team-based contributions and stressed the importance of basing evaluations on similarly well-articulated criteria. For example, for the data scientist, four main activity areas were identified: Design, Implementation, Analysis and Manuscript Reporting. Assessment of an activity or project would rely on evaluators weighing in on whether the contribution was major, moderate or minor with sample comments to guide and justify the choice. The authors recommend that major contributions be considered as on the same level as a first or senior author contribution. While the approach is incredibly valuable, challenges in its implementation are automating the process which relies on evaluators outside of the members of the promotion committee and tailoring the system for the various types of team scientists. To mitigate this, the authors encourage institutions to invest in 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 the scholarly contributions of these faculty and others that participate in team science (9), and supporting development of research programs for data scientists that are team scientists. 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. Expectations and career trajectories for faculty on such lines should be clearly delineated. 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. The latter would be facilitated through adoption of a system such as the one proposed by Mazumdar et al (9).

Conclusions

Without establishment by academic leaders of systems tailored to support team science, 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. van der Haar, S., Koeslag-Kreunen, M., Euwe, E. and Segers, M., 2017. Team Leader Structuring for Team Effectiveness and Team Learning in Command-and-Control Teams. *Small Group Research*, *48*(2), pp.215-248.

5. 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.
6. Bennett, L.M. and Gadlin, H., 2012. Collaboration and team science. *Journal of Investigative Medicine*, *60*(5), pp.768-775.
7. 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.
8. 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.
9. 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.