The expansion of high-throughput laboratory techniques and availability of large public databases has made it clear that the ability to generate data has far outpaced most biomedical scientists’ ability to analyze those data. Although many institutions have core facilities that provide statistical and bioinformatic consulting services, these facilities are overrun with clients and are typically a cost sink for their institution. As a solution, researchers attempt to develop their data analysis skills through workshops and online tutorials. The demand and number of learning resources that have become available through organizations like The Carpentries, Data Camp, and Code Academy are a testament to their popularity. Unfortunately, empirical analysis of outcomes from workshops has shown that although learners universally love the format and content, they have minimal long term retention of the material. **The lack of repeated reinforcement of the content over time is the most likely explanation for the poor outcomes of these workshops and resources.** Consequently, there is a need to create a library of tutorials that present concepts in different contexts that are relevant that allow learners to assess their retention and confidence with employing the concepts. Surprisingly, most online tutorials focus on how to implement individual concepts to answer abstract questions rather than integrating different concepts to answer interesting questions. If these tutorials are designed with the intent of also building local and more distributed communities around data analysis (i.e. communities of practice), then it will be possible to significantly improve the retention of material covered in data analysis workshops. As a solution to these problems, my research group has developed the concept of a weekly **Code Club**, which is analogous to a traditional Journal Club, but focused around programming. These interactive sessions have successfully helped bench scientists develop skills in data analysis that are strong enough to go on to careers as data scientists at leading universities and pharmaceutical companies.

My ***long-term*** goal is to enable bench scientists to analyze biomedical data with robust, rigorous, and reproducible approaches. The ***overall objective*** of this proposal is to develop a a collection of virtual Code Club sessions that researchers can use on their own or with colleagues. These sessions will cover concepts important for performing rigorous and reproducible data science, will be intentionally designed to develop communities of practice, and use robust pedagogical approaches to teaching. This is aligned with the overall goal of this RFA to create “exportable training modules designed to enhance the rigor, reproducibility, and responsible conduct of biomedical and behavioral data science research.” The ***central hypothesis*** is that completing Code Club sessions will improve the retention of concepts covered in prior workshops and allow learners to more quickly develop their skills expand beyond those covered in a workshop. I arrived at this hypothesis based on 20 years of experience helping bench scientists learn to do their own data analysis and the excitement of colleagues who have run their own Code Clubs. Furthermore, during the COVID-19 pandemic I have posted weekly Code Club sessions to help bench scientists developing their R programming skills. The ***rationale*** for developing additional Code Club sessions is that by increasing the diversity and number of videos available, researchers will make quicker and deeper gains in their knowledge of reproducible research practices. I am uniquely poised to achieve my overall objective by executing the following ***Proposed Research Education Program***:

**Produce Code Club sessions that highlight concepts important for performing rigorous and reproducible data science.** Each session will be composed of a brief web tutorial with exercises for participants to complete along with a video version of the tutorial along with solutions to the exercises. The selection of content and datasets will be motivated by questions relevant to diverse areas across biomedical research that cover a range of concepts related to rigor and reproducibility in data science. To help build a community around the videos, we will solicit questions from viewers. I ***hypothesize*** that researchers who take a workshop and then go on to complete the Code Club sessions will have longer retention of the concepts covered in the workshop than researchers who only take the workshop. I further expect that researchers who complete the Code Club tutorials will more quickly pursue concepts beyond the scope of the workshop. To evaluate the success of this Education Program, I will leverage a network of programming workshops to track retention of content with and without the supplemental tutorials.

Successful completion of this Proposed Research Education Program will significantly enhance the availability and efficacy of materials for improving reproducible data science. In addition to being available through the NIH Clearing House, the tutorials developed through this project will be added to the Riffomonas Project website and YouTube channel. The Riffomonas Project was initiated in 2015 to develop a workshop-style instructional series to help microbiome researchers develop skills that foster reproducibility. The material developed through this proposal will complement earlier microbiome content and help to broaden its audience to include microbiology, immunology, metabolomics, biochemistry, cell biology, and genomics.