SUMMARY STATEMENT
(Privileged Communication)

PROGRAM CONTACT:
Veerasamy Ravichandran
veerasamy.ravichandra@nih.gov

Release Date: **12/10/2020**
Revised Date:

Application Number: **1 R25 GM141481-01**

Principal Investigator

SCHLOSS, PATRICK DAVID

Applicant Organization: UNIVERSITY OF MICHIGAN AT ANN ARBOR

Review Group: **ZGM1 TWD-8 (RR)**
National Institute of General Medical Sciences Special Emphasis Panel
Review of R25 Training Modules to Enhance the Rigor, Reproducibility and Responsible Conduct of Biomedical Data Science Research applications.

Meeting Date: **11/12/2020**
Council: **JAN 2021**
Requested Start: **04/01/2021**

RFA/PA: **GM20-001**
PCC: **B120VR**
Dual PCC: **X86**
Dual IC(s): **AI, AT, DE, EB, LM**

Project Title: **Code Clubs: Repeated practice opportunities to develop reproducible data analysis skills**
SRG Action: **Impact Score:40**
Next Steps: **Visit https://grants.nih.gov/grants/next_steps.htm**
Human Subjects: **10-No human subjects involved**
Animal Subjects: **10-No live vertebrate animals involved for competing appl.**

Project Year	Direct Costs Requested	Estimated Total Cost
1	76,075	82,161
2	76,075	82,161
3	77,575	83,781
TOTAL	229,725	248,103

ADMINISTRATIVE BUDGET NOTE: The budget shown is the requested budget and has not been adjusted to reflect any recommendations made by reviewers. If an award is planned, the costs will be calculated by Institute grants management staff based on the recommendations outlined below in the COMMITTEE BUDGET RECOMMENDATIONS section.

1R25GM141481-01 Schloss, Patrick

RESUME AND SUMMARY OF DISCUSSION: The PI of this R25 Education project from the University of Michigan proposes to develop a series of 100 short, online training modules called Code Clubs about data analysis and reproducibility, and then compare the effectiveness of this teaching with traditional boot camp sessions. The Code Clubs cover highly relevant topics and are an innovative way to reinforce learning. The PI has extensive experience developing Code Clubs and leading 3-day workshops to teach data analysis and is a major strength of the proposal. Enthusiasm was somewhat diminished by some weaknesses in the application. The specific aims are not clearly presented, and the expected audience is not well defined. The evaluation plan lacks detail; e.g. there is no description of power analyses or measurable outcome variables to test the stated hypothesis. Details about the available facilities and resources for this project are not provided. In general, the application seemed disconnected from data science concepts and may benefit by adding additional expertise in that field. The strengths of the proposal are the PI's experience and the innovative approach and reviewers agreed that, if completed, the proposed Code Clubs are likely to have a positive impact on improving reproducible data analysis skills.

DESCRIPTION (provided by applicant):

The development of high throughput data generation tools used across the biomedical sciences has led to a situation where researchers with excellent bench skills struggle to appropriately and reproducibly analyze their data. With the increased size of the datasets, the complexity of the analyses has also grown. Although many institutions provide bioinformatic and statistical consulting services, the reality is that these services are overburdened and ultimately require the researcher who generated the data to also analyze the data. Researchers that once used paper notebooks to record data and spreadsheet-based tools to analyze their data now struggle to use command line tools. The long-term goal of this work is to enable bench scientists to analyze their biomedical data with robust, rigorous, and reproducible approaches. Traditional training programs have not been able to meet the needs of these researchers. Although very popular and well rated, workshops and bootcamps have proven ineffective at establishing lasting competency. The lack of repeated reinforcement of the content over time is the most likely explanation for the poor outcomes of these workshops and resources. Code Clubs have proven critical for providing this repeated reinforcement. Code Clubs are weekly activities that are analogous to a traditional Journal Club, but that focus on developing data analysis skills. The overall objective of this proposal is to develop a collection of virtual Code Club sessions that researchers can use on their own or with colleagues at their own institutions to improve reproducible data analysis skills. These sessions will cover concepts important for performing rigorous and reproducible data science, will be intentionally designed to develop local communities of practice, and will implement robust pedagogical approaches to teaching. These efforts are 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 and expand beyond those covered in a workshop. This hypothesis is 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. 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. This project will yield a significant vertical step in the field because it will put tools into the hands of researchers performing the analyses, empowering them to perform sophisticated and reproducible analyses. The approach taken in the proposed research is innovative because it represents the first concentrated effort to develop materials that use repeated engagement of the same content in different contexts to help researchers develop data analysis skills.

PUBLIC HEALTH RELEVANCE

The proposed research is relevant to public health because it supports researchers within the domain of biomedical research who need to develop and strengthen their data science skills. Thus, the research is relevant to the part of NIH’s mission that pertains to the development, maintenance, and renewal of scientific resources that will assure our ability to perform robust and reproducible research in order to prevent disease.

CRITIQUE 1

Significance: 3
Investigator(s): 1
Innovation: 3
Approach: 4
Environment: 2

Overall Impact

PI Schloss is a highly regarded scientist in microbiology, and he has spent significant time and resources teaching data analysis to groups of students for many years. He proposes to develop online modules called Code Club sessions, which people can watch individually to learn about data analysis and data reproducibility. He will also evaluate the extent to which these modules effectively teach this material, as compared with more traditional bootcamp sessions.

1.Significance

Strengths

- This proposal will develop 'code clubs', which are short online tutorials for bench scientists increasingly overwhelmed with such data coming out of their labs. By repeating the information and analysis every week, the idea is to provide an approach complementary with one-time bootcamp sessions which according to the PI tend to have poor outcomes. The PI proposes to develop approximately 100 modules, among them 60 on scripting, and a smaller number on topics such as automation, organization, and version control. They consist of a short web content site, plus a video of the same session. Dissemination is done in part through the courses that PI Schloss teaches every year, approximately 6 3-day courses each. Evaluation will be done by comparing students who have or have not taken Course Code sessions with students who have or have not taken more traditional 3-day courses on this topic.
- I find that this course is a novel take on a very important problem to teach reproducible data analysis to scientists. It follows up on existing courses that the PI already teaches, and which are very popular.

Weaknesses

- None noted

2. Investigator(s)

Strengths

- Investigator is very highly cited with an h-factor of 51 and over 29,000 citations. He works in the area of microbiome research which is highly relevant in the context of data science research. He is part of the Michigan Institute for Data Science. He has multiple publications specifically on the topic of reproducible research. He has taught between 4 and 8 sessions per year on reproducible data analysis skills. Overall, this is a highly qualified PI for this proposal

Weaknesses

- The budget consists of 2.4 calendar months per year for the PI and 3 months per year for a postdoc, plus some travel and publication cost funds. I have somewhat of a concern that someone with such a high publication record and active scientific productivity will really dedicate so much time to this specific topic of course creation. On the other hand, the PI is already very actively teaching around 6 3-day courses per year on this topic.

3. Innovation

Strengths

- The material for these courses has already been discussed in a different format by the PI, but the Course Code format is intriguing and has the potential for high impact

Weaknesses

- None noted

4. Approach

Strengths

- The approach of developing individual sessions is in line with what has been suggested for this program announcement. The Course Code sessions are a format that has the potential for having a large impact in the community.

Weaknesses

- Sample Code Club sessions included in the Research Education Program Plan on Table 2 seemed very specialized and geared towards technical aspects of using one program vs another. I would suggest the creation of more basic sessions to make them more useable among bench scientists looking to practice basic quantitative analysis tools. I find it encouraging that the PI will request suggested topics from the audience in the Riffomonas website.

5. Environment

Strengths

- The University of Michigan has a large community of scientists, and the PI has demonstrated to have high quality resources including computing clusters and an established set of courses on this topic. Overall, the environment is very strong.

Weaknesses

- None noted

Protections for Human Subjects

- Not applicable

Data and Safety Monitoring Plan (Applicable for Clinical Trials Only):

  - Not applicable

Vertebrate Animals

Not applicable

Biohazards

Not applicable

Resubmission Not applicable

Renewal Not applicable

Revision Not applicable

Select Agents

Not applicable

Resource Sharing Plans

Acceptable

Budget and Period of Support

Recommended as requested

CRITIQUE 2

Significance: 4
Investigator(s): 4
Innovation: 3
Approach: 3
Environment: 4

Overall Impact:

If successful, the investigators will support individuals within the domain of biomedical research who need to develop and strengthen their data science skills. The investigators aimed at organizing weekly Code Club, which is similar to the traditional Journal Club but focused on programming. The long-term goal of the investigating team is to enable bench scientists to analyze biomedical data with robust, rigorous and reproducible approaches. The investigators claimed that the Code Club sessions will help scientists to develop their R programming skills that will go a long way in improving their knowledge of reproducible research practices. The strength in this proposal is in the innovation and approach. The investigating team is weak without a data scientist, the significance is moderate. The PI spent a lot of time on definition rather than outlining specific approaches to improve rigor, reproducibility and responsible conduct of data science research. The specific aims are not well arranged. The aims are not well defined, it is difficult to say if the proposed modules will be on data analysis or data management. It is very difficult to understand the different populations that will benefit from this proposal.

1.Significance:

Strengths

- If successful, the investigating team will concentrate on developing code modules with emphasis on data management, data curation, data visualization.

Weaknesses

- The investigators spent a lot of time on background and discussed a lot about the issues on rigor, reproducibility and responsible conduct in other projects but failed to outline straightforward approaches in addressing these concerns that the PI outlined.

2. Investigator(s):

Strengths

- The PI Dr. Schloss is an environmental microbiologist with broad research interest.
- The PI Dr. Schloss is a good role model; faculty at University of Michigan in Ann Arbor for more than a decade with 106 publications.

Weaknesses

- The PI Dr. Schloss lacks peer reviewed publication in any of the data science journals.
- The proposal will benefit from the service of an expert in data science.

3. Innovation:

Strengths

- The rationale for developing additional Code Club sessions is to increase diversity and number of videos available, is very innovative since researchers will make quicker and deeper gains in their knowledge of reproducible research practices. This approach is innovative because it represents the first concentrated effort to develop materials that use repeated engagement of the same content in different contexts to help researchers develop data analysis skills.
- The PI has also made a good case for this program to effectively reach an audience. Dr. Schloss has taught more than 1200 students and have a platform to build upon for this project.

Weaknesses

- None

4. Approach:

Strengths

- The PI clearly states the objective of the proposal is to develop a collection of virtual Code Club sessions that researchers can use on their own or with colleagues to strengthen the reproducibility data analysis skills.
- The goal of the proposal is to target biomedical scientists at any career stage.
- The PI Dr. Schloss plans to have modules available as webpages on the Riffomonas project webpage.
- The PI also plan to disseminate materials through advertising, publications and the Riffomonas project webpage.
- There PI has also presented an evaluation plan through surveying participants.

Weaknesses

- Despite the objectives of the study being clearly stated deep within the proposal, it would have been easier stating the objectives within the specific aims page.
- Despite missing data is one aspect of data science that negatively impacts reproducibility, the PI failed to address it.

5. Environment:

Strengths

- The proposal takes advantage of the extensive, excellent, research facilities such as great research support staff, great computing system and great communication system within The University of Michigan Ann Arbor.

Weaknesses

- The PI failed to take advantage of the well-respected data science team at The University of Michigan at Ann Arbor.

Protections for Human Subjects

Not applicable

Vertebrate Animals

Not applicable

Biohazards

Not applicable

Resubmission Not applicable

Renewal Not applicable

Revision Not applicable

Select Agents

Not applicable

Resource Sharing Plans

Acceptable

Budget and Period of Support

Recommended as requested.

CRITIQUE 3

Overall Impact

This application is from one investigator who intends to develop 100 Code Clubs. The investigator has experience in developing Code Clubs and offering workshops on similar topics. He mentions that the current Code Clubs have received a positive response but does not indicate how many people have participated in a Code Club. He has a considerable number of followers on Twitter and has strong attendance at his workshops. The Code Clubs are an innovative way to reinforce one's learning, such as in a workshop. The topics are relevant and fill a niche that will improve reproducibility. The Code

Clubs will use real questions, which will improve the learning, but it is curious why he would include questions from epidemiology or psychology as they are in a completely different field from what he is in as well as the audience he is focused on (i.e. basic scientists). The Facilities and other Resources provides very limited information as to what he can access to help develop the videos, etc. He does have a letter of support from the Center for Research on Learning and Teaching but does not speak to any resources that they have. Also, it is curious why he isn't connecting with the CTSA at Michigan. This could be an excellent way to disseminate the Code Clubs. He does, however, have a strong dissemination plan. It appears that there is only one aim for the proposal and that is to develop Code Clubs. He does propose a hypothesis, but the evaluation plan lacks details as to how this hypothesis will be tested. For example, there is no power analysis and he never specifies the outcome variables and how they will be measured. He does refer to this as a study in the evaluation plan, so I would expect more details. Curiously, he says it will be important to study differences by gender and race but does not qualify why this is important.

THE FOLLOWING SECTIONS WERE PREPARED BY THE SCIENTIFIC REVIEW OFFICER TO SUMMARIZE THE OUTCOME OF DISCUSSIONS OF THE REVIEW COMMITTEE, OR REVIEWERS' WRITTEN CRITIQUES, ON THE FOLLOWING ISSUES:

COMMITTEE BUDGET RECOMMENDATIONS: The budget was recommended as requested.

rhj

Footnotes for 1 R25 GM141481-01; PI Name: Schloss, Patrick David

NIH has modified its policy regarding the receipt of resubmissions (amended applications).See Guide Notice NOT-OD-18-197 at https://grants.nih.gov/grants/guide/notice-files/NOT-OD-18-197.html. The impact/priority score is calculated after discussion of an application by averaging the overall scores (1-9) given by all voting reviewers on the committee and multiplying by 10. The criterion scores are submitted prior to the meeting by the individual reviewers assigned to an application, and are not discussed specifically at the review meeting or calculated into the overall impact score. Some applications also receive a percentile ranking. For details on the review process, see
http://grants.nih.gov/grants/peer_review_process.htm#scoring.

MEETING ROSTER

National Institute of General Medical Sciences Special Emphasis Panel
NATIONAL INSTITUTE OF GENERAL MEDICAL SCIENCES
Review of R25 Training Modules to Enhance the Rigor, Reproducibility and Responsible Conduct of Biomedical Data Science Research applications.
ZGM1 TWD-8 (RR)
11/12/2020

Notice of NIH Policy to All Applicants: Meeting rosters are provided for information purposes only. Applicant investigators and institutional officials must not communicate directly with study section members about an application before or after the review. Failure to observe this policy will create a serious breach of integrity in the peer review process, and may lead to actions outlined in NOT-OD-14-073 at https://grants.nih.gov/grants/guide/notice-files/NOT-OD-14-073.html and NOT-OD-15-106 at https://grants.nih.gov/grants/guide/notice-files/NOT-OD-15-106.html, including removal of the application from immediate review.

CHAIRPERSON(S)

RUBIO, DORIS M., PHD
ASSOCIATE VICE PROVOST FOR FACULTY, PROFESSOR OF MEDICINE - DIRECTOR
INSTITUTE FOR CLINICAL RESEARCH EDUCATION
DEPARTMENT OF MEDICINE
SCHOOL OF MEDICINE
UNIVERSITY OF PITTSBURGH
PITTSBURGH, PA 15213

MEMBERS

ATEM, FOLEFAC D, PHD
ASSISTANT PROFESSOR
DEPARTMENT OF BIOSTATISTICS
HEALTH SCIENCE CENTER AT HOUSTON
UNIVERSITY OF TEXAS
DALLAS, TX 75390

ENCISO, GERMAN ANDRES, PHD
ASSOCIATE PROFESSOR
DEPARTMENT OF MATHEMATICS, DEVEL AND CELL BIOLOGY
UNVERSITY OF CALIFORNIA, IRVINE
IRVINE, CA 92617

GADD, CYNTHIA S, PHD
PROFESSOR
DEPARTMENT OF BIOMEDICAL INFORMATICS
VANDERBILT UNIVERSITY
NASHVILLE, TN 37232

HAJIRASOULIHA, IMAN, PHD
ASSISTANT PROFESSOR
DEPARTMENT OF PHYSIOLOGY AND BIOPHYSICS
INSTITUTE FOR COMPUTATIONAL BIOMEDICINE
INSTITUTE FOR PRECISION MEDICINE
WEILL CORNELL MEDICINE
NEW YORK, NY 10065

KORF, IAN F, PHD
PROFESSOR
UNIVERSITY OF CALIFORNIA GENOME CENTER
UNIVERSITY OF CALIFORNIA DAVIS
DAVIS, CA 95616

LAMBERT, CHARLA, PHD
DIVERSITY, EQUITY, & INCLUSION OFFICER
COLD SPRING HARBOR LABORATORY
COLD SPRING HARBOR, NY 11724

MACEK, MARK D, DDS, DRPH
ASSOCIATE PROFESSOR
SCHOOL OF DENTISTRY
UNIVERSITY OF MARYLAND- BALTIMORE
BALTIMORE, MD 21201

MURRAY, DEBRA DIANNE, PHD
PROFESSOR
HUMAN GENOME SEQUENCING CENTER
BAYLOR COLLEGE OF MEDICINE
HOUSTON, TX 77030

MUSE, SPENCER V., PHD
PROFESSOR
DEPARTMENT OF STATISTICS
BIOINFORMATICS RESEARCH CENTER
NORTH CAROLINA STATE UNIVERSITY
RALEIGH, NC 27695

WELTY, LEAH J, PHD
PROFESSOR
DEPARTMENT OF PREVENTIVE MEDICINE - BIOSTATISTICS
DEPARTMENT OF PSYCHIATRY AND BEHAVIORAL SCIENCES
FEINBERG SCHOOL OF MEDICINE
NORTHWESTERN UNIVERSITY
CHICAGO, IL 60611

SCIENTIFIC REVIEW OFFICER

JOHNSON, REBECCA H., PHD
SCIENTIFIC REVIEW OFFICER
OFFICE OF SCIENTIFIC REVIEW
NATIONAL INSTITUTE OF
GENERAL MEDICAL SCIENCES
NATIONAL INSTITUTES OF HEALTH
BETHESDA, MD 20892

EXTRAMURAL SUPPORT ASSISTANT

MENDOZA, DOUG
EXTRAMURAL SUPPORT ASSISTANT
OFFICE OF SCIENTIFIC REVIEW
NATIONAL INSTITUTE OF GENERAL MEDICAL SCIENCES
NATIONAL INSTITUTES OF HEALTH
BETHESDA, MD 20892

Consultants are required to absent themselves from the room during the review of any application if their presence would constitute or appear to constitute a conflict of interest.