To Whom It May Concern:

I am delighted to write this letter in strong support of this proposal to develop training modules to enhance the rigor and reproducibility of biomedical research. The content of this proposal is incredibly timely as I believe there is a tremendous need to develop and advance tools for doing principled and reproducible data analysis in a coherent framework. The recent development of the “tidyverse” series of tools provides a unique opportunity to develop the proposed training modules in a manner that is accessible to laboratory-based scientists.

I have written extensively over the past 10 years about the importance of computational reproducibility in all research. As computing power has grown over time, we have developed the ability to generate large datasets and to apply sophisticated statistical methodology at the “press of a button”. However, the complexity of the resulting analyses dramatically decreases their transparency and there is a need to make sure proper software tools are used in order to guarantee that the results can be reproduced in the future. I believe the proposed modules will go a long way towards providing laboratory-based scientists the training they need to apply the best available tools.

To date, much coursework has been developed (particularly in the online space) instructing people on the use of tools like R, RStudio, and other packages. However, much of this instruction, including my own, is directed at data scientists, statisticians, programmers, or other quantitative analysts. I have seen little material that is specifically directed at laboratory scientists and is able to address their particular needs. I believe the proposed modules fill an important gap.

The proposed training modules make use of the now-popular “tidyverse” framework for data analysis. I have taught R programming for 15 years to public health students from a broad range of backgrounds onsite at Johns Hopkins and now teach the largest course on R programming in the world through the online Coursera platform. The tidyverse framework allows scientists who are new to R to get up and running with R programming very quickly, and to learn tools that are powerful but also very expressive, so that the same simple tools can be joined together to solve many different complex problems. As the tidyverse framework has developed over time, I have quickly transferred much of what I teach into this new framework. I have found it to be intuitive to new users in particular, and all of my new course development uses this framework. Furthermore, many new introductory R books, workshops, and courses (both onsite and online) now start with the tidyverse framework, so there is a rich foundation of resources on which the materials in this proposal could build.

I have worked extensively with Dr. Anderson for many years now and believe she is uniquely suited to lead this work. She and I co-wrote (with Sean Kross) the book Mastering Software Development in R (along with an online course by the same name) which teaches software development principles in R. The book is an online book written with the bookdown software package and has been downloaded over 14,000 times (https://leanpub.com/msdr). The corresponding online course has been taken by over 50,000 people.

In addition to being an excellent writer, R programmer, and researcher, she has extensive experience building reproducible computational workflows and teaching these processes at a variety of levels. I eagerly await the production of these training modules and I would be delighted to disseminate them to my own colleagues at Johns Hopkins and to others around the world.

In summary, I strongly support this proposal.

Sincerely,



Roger D. Peng, PhD

Professor of Biostatistics

Johns Hopkins Bloomberg School of Public Health