Developing a Transdisciplinary Team Using Online Data Science Training and a Real-World Case Study on the Pandemic, COVID-19

***Sarah E. Hooper*[1], *Heather Habacker*[2], *Dean Frazier*[3]**

[1]Department of Biomedical Sciences, Ross University School of Veterinary Medicine, Basseterre, St Kitts, WI.

[2]Department of Psychology and Neuroscience, Baylor University, Waco, TX, USA.

[3]Institute for Data Science and Informatics, University of Missouri, Columbia, MO, USA.

Correspondence should be addressed to Sarah E. Hooper; Shooper@rossvet.edu.kn

Received 9 January, 2023; Revised January 30, 2023; Accepted January 30, 2023

Available online January 30, 2023 at www.atlas-tjes.org, doi: 10.22545/2023/00224

Abstract:*Using a real-world, team-led, virtual case study focused on the international pandemic, SARS-CoV-2, more commonly referred to as COVID-19, we assessed if data science could serve as the common discipline for the development of a professional transdisciplinary team. Team success was measured by assessing if all members contributed to the virtual, student-led case study, which occurred, and if the product of the collaboration transcended the discipline, which occurred when the team created this report due to the success of the online data science program preventing the commonly encountered difficulties such as communication among team members and cognitive obstacles of experts being unable to relate to others outside their expertise. We suggest that future pandemic response training programs consider incorporating data science as a fundamental discipline when developing pandemic response teams.*

Keywords:COVID-19, case study, online data science, pandemic response, transdisciplinary, team.

1 Introduction

The ongoing outbreak of coronavirus-associated acute respiratory disease, initially designated COVID-19 by the World Health Organization, emerged from Wuhan, China in December 2019 [1]. Although initial investigations by the Chinese authorities found “no clear evidence of human-to-human transmission” [2], within two weeks COVID-19 had spread to Thailand, and by mid-February COVID-19 was present on every continent, except Antarctica [2]. Within three months the World Health Organization declared the outbreak a pandemic and the International Committee on Taxonomy of Viruses (ICTV) officially named COVID-19 as severe acute respiratory syndrome coronavirus (SARS-CoV-2) based upon the phylogeny, taxonomy, and established naming practices [1, 2]. Concurrently, the rapidly rising death toll and socio-economic costs, led to the United Nations declaring COVID-19 "the greatest challenge our world has faced since World War 2" and is the "defining global health crisis of our time" [3].

Nearly two decades prior to the emergence of SARS-CoV-2, the world briefly experienced another coronavirus epidemic, severe acute respiratory syndrome (SARS-CoV), which emerged from Guangdong, China [4]. Transmission of SARS-CoV occurred primarily in health care facilities where inadequate infection control precautions were employed or in households where human-to-human SARS-CoV transmission was possible due to close person-to-person contact [5, 6]. With implementation of appropriate infectious disease control principles and management strategies, including "social distancing", the SARS-CoV epidemic came to an abrupt halt [4, 5].

A year after the disappearance of SARS-CoV, a multidisciplinary meeting was held to help the world prepare for future public-health emergencies of international concern by synthesizing the knowledge gained from SARS-CoV [7]. Special warnings were drafted for public health officials regarding contingency planning and preparedness of our public health professionals for future outbreaks and pandemics. Of particular note, they warned the next pandemic would likely be caused by a highly contagious virus where asymptomatic individuals would transmit the virus person-to-person prior to becoming symptomatic and some individuals would serve as "super spreaders"—a situation where some asymptomatic individuals would be capable of infecting hundreds of people [7]. Additionally, they predicted world-wide spread of the next viral pandemic pathogen would occur within days with the help of international air travel and that regions such as Africa where health conditions such as HIV-1 infection were highly prevalent, the pandemic pathogen could become endemic [7]. The experts also warned of the expansive social and economic impacts beyond what occurred during previous infectious disease outbreaks [7]. While there was little discussion on how to best train the next generation of healthcare workers and pandemic response teams, transdisciplinary approaches have become thematic in recent emerging infectious disease expert discussions [8]. Further evidence such as the lack of a unified response and mixed messages, particularly in the United States, over such topics as mask wearing, demonstrate the need for transdisciplinary pandemic response team to include public health experts, politicians, community leaders, and other essential personal in order for an effective response to be formulated, implemented, and communicated to the general public.

Despite many calls for transdisciplinary infectious disease training, there remains little development of transdisciplinary training programs at institutions of higher learning [9]. Exhaustive online searches yield most training programs focus on enhancing cooperation of disciplines through the training of students to work with students from other disciplines to contribute to a common goal. It is important to emphasize that this aligns more with interdisciplinary training [10]. True transdisciplinary training for pandemic preparedness and planning involves both academic and non-academic participants, and these individuals must train together in order for their expertise to become integrated into a completely new approach that transcends each of their separate disciplines [10].

With the explosion of digital data from smartphones, social media, genomics, and other healthcare data, the field of data science, a discipline designed around the emergence of "big data" [11], has repeatedly been shown to potentially be an important aspect in managing infectious disease outbreaks [12]. For example, several years after the 2010 cholera outbreak in Haiti, it was shown the initial cholera outbreak as well as the spread of the disease could have been predicted using publicly available Haitian cell phone mobility data [12].

With the recognized need for transdisciplinary approaches to responding and managing emerging infectious disease and the need for transdisciplinary training programs, we explore the possibility of the data science field serving as the common discipline, or link, for the formation of a transdisciplinary team. Specifically, we propose that data science be the common discipline for professionals trained in divergent disciplines (e.g., anthropology and veterinary medicine) and located in different countries. Furthermore, we assess the use of the massive amounts of digital data being generated by the international pandemic, SARS-CoV-2 (COVID-19) to serve as a real-world platform for a virtual, team-led, real-world case study.

2 Methodology

The University of Missouri Data Science and Analytics Master's Program is offered as an online-only master's degree program for working professionals with one week a year spent on campus culminating the skills learned over the previous year and networking with industry professionals. Program participants can emphasize in biohealth analytics, high performance computing, human-centered science design, and data journalism and strategic communication. Part of the core curriculum is an intensive case study focused on a topic selected by a small cohort of 3 to 4 students with an assigned data science advisor. The purpose of this case study is to allow students to synthesize the data science techniques learned in the first year of the program while developing a data story for a stake-holder.

This study focuses on a team of four students who selected the topic SARS-CoV-2 (COVID-19) and who met the criteria outlined by Benesh et al. (2015) [13]. Briefly, each member must be a confident expert in one's own field, have time for additional training activities, and recognize that single discipline training has benefits as well as limitations [13]. The team members represented a diversity of fields with expertise in anthropology, psychology, ecology, education, and veterinary medicine. Team members represented all academic career stages, ranging from a PhD student, a post-doctoral researcher who transitioned to an Assistant Professor position, and two non-academic working career professionals. Data science served as the common, or "linking" discipline for the transdisciplinary team. All members were competent in basic data science techniques with 18 hours of data science courses.

Due to members being located internationally and within the United States, What's App was determined to be an effective communication tool along with video-conferencing services. All data analysis work was completed in JupyterHub, a multi-user notebook, and Google Drive used for all other work with minimal employment of e-mail. Authorship of publications was discussed during the early stages of manuscript drafting. Author order was based upon the contributions to the generation of ideas and synthesizing the papers.

Since potential conflicts between team members are inevitable, all conflicts were addressed directly and respectfully with all team members incorporated into the conflict resolution process via e-mail, What's App, and video-conferencing.

The goal of the summer case study is to coalesce the first year's acquired data science skills in R and/or python data carpentry, exploratory data analysis, statistical analysis, and visualizations by addressing a real-world problem. To accomplish this goal, a series of five milestones (see Figure 1) with deadlines spaced 7-14 days apart were provided. Each week, the team met with the "stakeholder" for the project who was a faculty member. The team's first two milestones focused on developing specific questions or hypothesis, identifying real-world datasets that could be potentially used to answer the proposed questions, and starting relevant data cleaning (data carpentry). The third milestone focused on initial visualizations and descriptive statistics. The fourth milestone focused on refining data carpentry while conducting bivariate and multivariate analysis with supporting visualizations. The fifth milestone focused on developing the data story into a rough draft with detailed methods. The final deliverables were detailed, step-by-step directions on how to reproduce all aspects of the data storying including data ingestion, carpentry, analysis, visualizations along with the completed data story.

All data science work was completed in JupyterHub notebooks within each team member's single-user notebook server which were subsequently compiled into a single group notebook or set of group notebooks. Each milestone was fulfilled by the uploading of these group notebooks to the GitLab site dedicated to the University of Missouri's Data Science and Analytics program. The teams JupyterHub notebooks detailing the data science analysis methods utilized and the complete coding are available on Mendeley Data [14]. To briefly summarize, all data sets were acquired from online, opensource data repositories including Kaggle, National Institutes of Health (NIH), National Center for Health Statistics, Behavioral Risk Factor Surveillance System, and US Census Population Estimates. Since all team members were

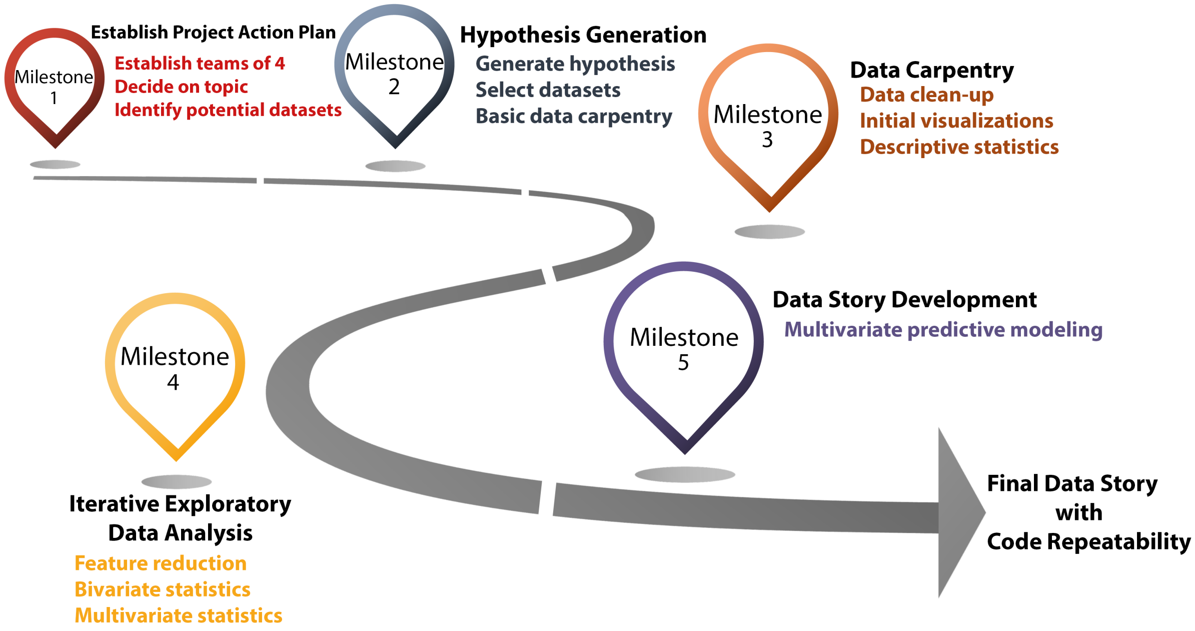


Figure 1: The five milestones of the virtual, student-led case study leading to the production of a final data story with documentation of the code to ensure the project could be reproducible.

most comfortable with the R-programming language, R Jupyter-Hub notebooks were employed to import datasets for basic visualization of the data such as histograms to visualization distributions and data cleaning including joining datasets, elimination of outliers, elimination of repeated variables, and running R-packages containing functions for Monte Carlo algorithms to deal with missing data. Considering the ever-changing amounts of new data generated by COVID-19, a specific cutoff-date was decided upon by the group, with newer data able to be imported later after the case-study for future projects. Once data was cleaned, normalized when possible, and standardized, regression trees were built to help identify potentially relevant variables for answering the group's hypothesis on predicting what variables can be used to predict the number of COVID-19 infections (positive cases) and the death rate. Linear multivariate regression models were built to answer the specific hypothesis surrounding the prediction of COVID-19 infections and deaths.

The outcome measurement of the case study was the development of a data story in the form of a blog post, newspaper article, or in this team's case, the preparation of a manuscript for a peer-review journal with the goal of publishing the findings (e.g. predictors of COVID-19 infections and deaths as determined by machine learning).

Outcomes measurements traditionally used for assessing transdisciplinary teams are primarily limited to the research setting and focus on traditional academic markers of performance including number of grant submissions, number of publications, average number of coauthors per publication, and average journal impact factor per publication [15]. The only alternative metric reported in the literature was a team decision making questionnaire [16] which was designed to test transdisciplinary team decision making for healthcare workers, however, since the study cohort was not a true transdisciplinary team (e.g. only a single discipline of doctors and nurses) and since the questionnaire was designed for use in a clinical case setting it was deemed inappropriate to use in this study [17].

Therefore, we assessed the outcome by several mechanisms. First, we assessed if each team members respective expertise were included in all steps of the process, from hypothesis generation, reproducibility, and storytelling with visualizations. Second, did the products of the collaboration transcend their disciplines by resulting in a completely new approach. Lastly, to assess the viability of the longevity of the team (teamwork), we used the more traditional methods of assessing the number of collaborative projects arising during or immediately after the case study and the number of publications arising or planned from the case study. Grant proposals were not selected as an outcome as half of the team members were non-academic working career professionals.

3 Results and Discussion

3.1 Summary of overall Teamwork/Goals

Overall, three of the team members had great synergy and enjoyed working together to complete each week's goals. However, one team member failed to participate in discussions, contribute adequate quality work, and meet deadlines. Conflict resolution was handled within the group as it became apparent the team member did not meet the criteria of having time for addition training activities. After deliberation via What's App and e-mail correspondence, it was decided that the team member would work with the faculty course coordinator to select an alternative, independent project without specific weekly deadlines.

While the team's machine learning models failed to find any significant predictors of COVID-19 infection rates and deaths, the transdisciplinary team was successful in terms of developing a new approach to the data story by shifting the desired outcome, a peer-reviewed journal article, to focus on the "story" of how data science could serve as a "linking" discipline for training transdisciplinary teams to respond to future pandemics. The subsequent sections describe this process and the results of each step of the case-study with an emphasis on highlighting the effectiveness of 6the use of a real-world case study to developed a transdisciplinary team.

3.2 Identification of Datasets and Hypothesis Generation (Milestones 1-2)

The initial two milestones focused on developing working relationships and establishing the best form and frequency of communication. Often the greatest challenge for interdisciplinary and transdisciplinary teams is the establishment of effective communication and management of the communication with a main stumbling block being "how to communicate" with experts in other disciplines [18-20]. The structure of the online course with clear milestones (project goals) and the project constraints of immovable deadlines necessitated this early establishment of communication.

What's App group texting proved to be the most efficient communication tool due to team members being located both within the continental US as well as in the Caribbean, however e-mail was also utilized along with weekly or bi-weekly video conferences. Earlier studies assessing communication in higher education indicate that e-mail and phone were preferred over texting. This suggests new studies assessing communication preferences in higher education are needed [21, 22]. Additionally, Google Drive was employed to address the need to communicate written information in real-time and this tool allowed the team to work concurrently on a word document during videocalls as well as on the required supportive information for the milestones.

The development of effective communication facilitated the completion of the initial task of gathering high quality datasets. These datasets would serve as a basis for establishing data science research questions on the selected topic of COVID-19. All datasets except one were selected from the Roche Data Science Coalition (RDSC) UNCOVER COVID-19 Challenge on Kaggle, a crowd-sourced data science platform [23]. This specific challenge hosted a curated collection of over 200 publicly available COVID-19 related datasets with original sources ranging from a variety of reputable organizations and government agencies including Johns Hopkins and the World Health Organization [23]. An additional dataset of county-level UV radiation exposure dataset was obtained from the National Institutes of Health National Cancer Institute GIS Portal for Cancer Research [24].

Once the datasets were selected, via a video call, two hypotheses were generated based upon the initial dataset. Hypothesis one focused on COVID-19 cases, "When controlling for population density, can COVID-

19 infection (cases) in US counties be predicted by basic demographic factors, including age, sex/gender and sociobehavioral factors; concurrent health conditions, including diabetes, HIV, and chlamydia; and/or the environmental factor of UV radiation." Whereas the second hypothesis focused on COVID-19 deaths, "When controlling for population density, can COVID-19 deaths in US counties be predicted by basic demographic factors, including age, sex/gender and sociobehavioral factors, concurrent health conditions, including diabetes, HIV, and chlamydia, and/or the environmental factor of UV radiation."

3.3 Data Clean-up (Milestones 2-3)

Many of the milestones and "check-ins" between milestones focused on cleaning and organizing the data. These milestones relied heavily upon the team to synthesize knowledge and utilize the skills developed during prior courses in data carpentry and visualizations. Additionally, these expectations are reflective of the time required to resolve the inherent problems of real-world data sets as data scientists report most of their time is spent on data cleaning [25]. Using real-world data compiled from multiple depositories resulted in the dataset having entries with duplicate information, different data for the same variable (e.g. number of COVID-19 infections reported may be different pending on the cut-off date), erroneous data (e.g. more chlamydia cases reported in a county than the population of the county), and missing data—all issues commonly seen in "big data" [26].

The individual datasets were merged using the common column of the county federal information processing system (FIPS) code. This initial single initial large dataset contained hundreds of potential variables and required team members to work together to reduce the dataset variables which could be potential predictors of COVID-19 cases and deaths. The final experimental dataset contained 53 variables. While the variables selected were broad, it is important to note a measure of the case study's success in developing a transdisciplinary team depended upon each member's respective expertise being represented in the hypothesis generation and data selection process. We can clearly identify the demographic and sociobehavioral factors were contributed by the team members with expertise in the social sciences, the concurrent health conditions contributed by the anthropologist and veterinarian while the environmental variable was reflective of the One Health expertise of the veterinarian. It is clear the majority of the project dedicated to data cleaning supports our hypothesis that data science can be used as the linking discipline for a transdisciplinary team as the expertise of the team members.

Part of the effectiveness of using data science as the "link" is the continued emphasis on effective communication. Communication was essential for the data cleaning stage as the data carpentry was completed directly in the group Jupyter Hub notebooks, because the large dataset was utilized rather than subdividing the dataset so each member could work in their own individual notebooks within Jupyter Hub. This tactic was selected by the team as it avoided having to remerge the data and perform additional clean-up resulting from the merge. The possible disadvantage of this method was that only one team member could work at a time within the notebook. This apparent disadvantage had an unanticipated benefit as it caused team members to reassess how to best facilitate communication after the first hinderance of the production of the team's milestone deliverables caused by lost work due to merging conflicts and work in progress being over-written. At this stage, What's App became the most common form of communication and served as the preferred communication mechanism throughout the rest of the project which is contradictory to a prior report that e-mail and face-to-face are preferred when communication is work related [21, 22], suggesting that texting is better for short, rapid communications as required for this project.

A component of the data cleaning stage, once the smaller dataset of selected variables was established, the team was responsible for determining how to deal with missing data and assess if any of the real-world data was improbable. Basic visualizations were created to help identify the distribution of the data and identify outliers and/or improbable data. For instance, if co-morbidities reported in a county was equal to the reported population of the county, this was recognized and could be corrected or eliminated. However, the more challenging issue was how to deal with the missing completely at random (MCAR) data [27] commonly found in healthcare data [28]. The team was hesitant to accept the most common method for dealing with missing data, removing the entire row or replace empty values with the maximum likelihood or other similar statistical methods [29], because a disproportionate number of counties with small population sizes or in specific geographic regions would be eliminated. An alternative method was selected, a multivariate imputation by chained equations (MICE) algorithm as when employed has been shown to perform best when dealing with healthcare datasets to impute the missing values [28]. Further data pre-processing included identifying outliers using Cook's distance [30] and standardizing select variables (e.g. number of physicians per 100K population). The full data carpentry Jupyter workbook is available at https://data.mendeley.com/datasets/sgngmzxzbr/1 [14].

Reflecting on the team's approach of simultaneously working together on these milestones demonstrates how the data science discipline leads to the development of a transdisciplinary team rather than a multidisciplinary team. In multidisciplinary teams, team members function as independent experts who work independently to develop their goals for the project. Work on the project occurs by each team member working parallel to each other and/or sequentially, and periodically teammates will meet to discuss their progress towards those goals [31, 32]. In contrast, transdisciplinary team members work simultaneously on a common problem, with common goals developed together [31]. In our COVID-19 case study, team members shared data science as foundational common conceptual framework and contributed discipline-specific theories in the form of variables towards the common goal of predicting COVID-19 cases and deaths. Furthermore, thru the use of the Jupyter Hub platform, the team members were able to help the others to develop their skills by using the other team members expertise to resolve coding issues as transdisciplinary teams not only share a common goal(s), but their skills are also shared unlike multidisciplinary teams [31].

3.4 Feature Reduction and Bivariate Analysis (Milestone 4)

With the initial selection of the variables relying upon each group member's expertise, feature reduction methods were employed to further reduce the variables. Feature reduction algorithms seek to improve the performance of machine learning models and generally categorize features as relevant, irrelevant, or redundant [33]. Both the raw values and standardized variables (e.g. physicians per 100K population) were reduced using mutual information methods and regression trees to identify variables that explain the most variation in the outcomes associated with COVID-19.

Khalid et al 2014, defines mutual information methods as "the reduction of uncertainty regarding variable X after the observation of Y" [33]. Therefore, by using mutual information, the team sought the variables with maximal independence [33]. Using mutual information, relatively few variables were relevant, and most were considered redundant. This was also collaborated thru the use of regression trees for feature selection [34]. Few variables were utilized for tree building with minimal overlap of variables when training, testing, and full datasets were employed. Bivariate analysis (Pearson's r) showed few variables with strong correlations. An interactive heat map was developed to help more easily identify positive or negative correlations. Highly correlated, redundant, and irrelevant variables were eliminated from the experimental data set prior to multivariate modelling.

3.5 Multivariate Predictive Modelling (Milestone 5)

The dataset obtained after feature reduction, the final experimental dataset, was utilized for linear regression modelling approaches to determine if the team could identify any predictive variables for COVID-19 cases or deaths. While some models indicated a significant predictor, visualizations did not support this claim. Since each week the teams progress was reported to the "stakeholder" faculty member, feedback was provided to the team. The most valued feedback was about creating a naïve model to assess if the observed significant predictors were truly significant since visualizations did not support this.

The dataset for naïve models was created by taking the average number of cases in a state rather than the reported cases per county and running the models with the potentially significant predictors. The naïve models performed better than the actual data indicating our results were non-significant. By generating these naïve models and the figures, it became clear that Simpson's paradox occurred, a concept explained by the faculty stakeholder. Essentially Simpson's paradox is when one set of data shows a particular trend, but when the full data is employed with all groups represented the trend is reversed. This means that the potential correlations of some variables we observed with COVID-19 cases or deaths may be true in certain groups or regions, however, overall, the variables did not support predictions of either cases or deaths.

The team's findings, Simpson's paradox or negative results, were disappointing as it eliminated the possibility of publishing a manuscript reporting our predictors since our selected methods or early COVID-19 dataset did not enable us to find significant predictors. Ironically, if the traditional assessments of transdisciplinary teamwork were utilized, then our team would have been found to be highly ineffective by not publishing a manuscript on the results of the study [15].

3.6 Effective Integration of Data Science with Hypothesis Generation, Storytelling and Visualizations (Final Product)

Since the COVID-19 modelling failed to find predictors of significance, the team reassessed what "product" would be produced for the case study. There was not much interest in creating a blog post or newspaper article, therefore the team elected to continue with the goal of a peer-reviewed manuscript. The team's manuscript focus altered to the creating of a transdisciplinary team and the use of the massive amounts of digital data generated by SARS-CoV-2 (COVID-19) to serve as a real-world platform for a virtual, team-led, case study where data science could serve as the linking discipline for professionals trained in divergent disciplines (e.g. anthropology and veterinary medicine). This manuscript is the final product of the team's revised hypothesis and goals of the project, showcasing the strength of a transdisciplinary team and suggesting evidence that data science can mitigate many of the challenges faced by transdisciplinary work.

With transdisciplinary approaches thematic in recent emerging infectious disease expert discussions [8], it was important to highlight to the academic community how online professional data science training programs could improve pandemic responses and epidemiologic monitoring by integrating data science as a foundational discipline. This case study provides strong support that data science should be incorporated into infectious disease training to facilitate the development of transdisciplinary teams. Because "big data" is able to provide real-life, important information about the spatial and temporal spread of disease [12], it is likely that data science will be an integral component to managing pandemics.

In particular, the use of data science as a common discipline appeared to help avoid the challenges often described in transdisciplinary research. Methodological conflicts [35] have been cited as a common issue associated with transdisciplinary teams, however our team agreed unanimously on the methodology employed for the case study which may be in part due to drawing upon the previous 18 hours of data science courses. It was also unanimously recognized that our methodology choice of linear regression models could have resulted in the failure to determine specific predictors of new COVID-19 cases and deaths. However, the team could not eliminate the possibility of the limited dataset (March 2020-June 2020) or the initial selection of available variables.

Another commonly cited challenge is the cognitive obstacles of experts being only able to relate to and work with others in their field of expertise [36]. This obstacle has been linked to the style of education within universities with professional and graduate students often working only within a single department and within a single discipline [36]. The graduate students attracted to an online professional degree, have experience working with a variety of professionals within their jobs, and especially those who work daily with clients need to be able to relate to the client to facilitate business transactions and/or services provided. Likely the "soft skills" such as communication and interpersonal skillsets developed prior to engaging in the case study avoided any "cognitive obstacles" despite the diversity of team members backgrounds. Additionally, these skillsets likely contributed to the respecting of the diversity of each team members background as there are reports of other transdisciplinary team members struggling to respect individual and collective diversities [35].

This respect of each members' backgrounds helped to contribute to the integration of each members' expertise into the case study—a key to the success of transdisciplinary teams. Integration easily occurred without specific effort in part likely due to the selected case study topic. COVID-19 has affected each individual member due to the world-wide lockdowns, out-of-stock commodities such as toilet paper, and friends or family contracting the virus. Integration is considered key as it is essential for creating "new knowledge" which "transcends" a single discipline [37]. New conceptual frameworks, theories, and methodological applications are potential products of transdisciplinary teams [37-39] which have to be disseminated and diffused to relevant audiences.

3.7 Co-authorship on Publications and Long-term Collaborations

As discussed earlier, our team's success was measured by assessing if all members contributed to the case study, which occurred, and if the product of the collaboration transcended the discipline, which this publication documents. The last measurement of success is the number of publications arising from the case study or future collaborations. The team members expect two publications to arise from this case study. In addition to the current manuscript, a second one was completed focusing on how implicit and explicit bias affect the mortality rate of COVID-19 in the Black population compared to the White population. Initially, the team had worked on developed a single data story, but it was split into two potential data stories, one focused on COVID-19 predictors and one focused on racial bias due to the limited timeline and feedback from the faculty member stakeholder.

Long-term collaborations are in progress with three additional team members included with expertise in machine learning, communication within medical education, and political psychology. Without time limitations and hard deadlines, the team will reassess the data science methodology as well as include temporal models and the explicit and implicit bias data to further reassess if predictors can be determined for developing COVID-19 and risk of dying from COVID-19.

4 Conclusions

4.1 Summary of Benefits to Using Data Science as Transdisciplinary Discipline

Data science was shown to be effective in developing a transdisciplinary team and the production of novel outputs in part due to the common learning process of all team members who were part of an online professional data science and analytics master's degree program. This online curriculum helped the team members to find a common process that allowed them to learn in common as transdisciplinary learning is a key component of transdisciplinary teamwork [40]. Each member learned from each other through a variety of different online platforms such as Jupyter Hub and Google Drive, and communicated thru a large variety of media formats including video meetings, texting via What's App, and e-mail. By using data science and a real-world case study focused on the pandemic, COVID-19, a team was formed which did not encounter commonly reported difficulties when attempting to form and function as a transdisciplinary team. Therefore, we suggest that future pandemic response training programs consider incorporating data science as a fundamental discipline when developing pandemic response teams.

Authors Contribution: SH, HH, DF all participated in idea generation, discussion, group observations, study design, and data analysis. SH drafted the manuscript. SH, HH, DF edited the manuscript.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflicts of Interest: The authors report no conflict of interest.

Acknowledgments
We would like to thank Dr. Tozammel Hossain for serving as the faculty "stakeholder" and the University of Missouri Data Science and Analytics (DSA) Program.

Copyright © 2023 by the authors. This is an open access article distributed under the Creative Commons Attribution License (CC BY-NC International, https://creativecommons.org/licenses/by/4.0/), which allow others to share, make adaptations, tweak, and build upon your work non-commercially, provided the original work is properly cited. The authors can reuse their work commercially.

References

[1] Gorbalenya, A. E., Baker, S. C., Baric, R. S., de Groot, R. J., Drosten C., Gulyaeva, A. A., Haagmans, B. L., Lauber, C., Leontovich, A. M., Neuman, B. W., Penzar, D., Perlman, S., Poon, L. L. M., Samborskiy, D. V., Sidorov, I. A., Sola, I., Ziebuhr, J. , (2020). The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nature Microbiology,* 5(4), 536-544. doi:10.1038/s41564-020-0695-z.

[2] World Health Organization. *Listings of WHO's Response to COVID-19.* https://www.who.int/news-room/detail/29-06-2020-covidtimeline, accessed September 20, 2020.

[3] United Nations Develoment Programme. *Coronavirus disease COVID-19 pandemic.* https://www.undp.org/coronavirus, accessed October 4, 2020.

[4] Zhong, N.S., Zheng, B. J., Li, Y. M., Poon, L. L. M. Xie, Z. H., Chan, K. H., Li, P. H. Tan, S. Y., Chang, Q., Xie, J. P., (2003). Epidemiology and cause of severe acute respiratory syndrome (SARS) in Guangdong, People's Republic of China, in February 2003. *The Lancet,* 362(9393): 1353-1358. doi:10.1016/S0140-6736(03)14630-2

[5] Bell, D.M. and World Health Organization Working Group on International and Community Transmission of SARS, (2003). Public health interventions and SARS spread. Emerg Infect Dis, 2004. 10(11), 1900-1906. doi:10.3201/eid1011.040729

[6] Low, D.E. (2004). SARS: lessons from Toronto. *in Learning from SARS: Preparing for the Next Disease Outbreak: Workshop Summary.* (pp. 63-83). National Academies Press Washington.

[7] Weiss, R.A. and A.R. McLean, (2004). What have we learned from SARS? *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences,* 359(1447), 1137-1140.

[8] Jacobsen K. H., Aguirre A. A., Bailey C.L., Baranova A.V., Crooks A.T., Croitoru A., Delamater P.L., Gupta J., Kehn-Hall K., Narayanan A., Pierobon M., Rowan K. E., Schwebach J. R., Seshaiyer P., Sklarew D. M., Stefanidis A., Agouris P., (2016). Lessons from the Ebola Outbreak: Action Items for Emerging Infectious Disease Preparedness and Response. *EcoHealth,* 13(1), 200-212. doi: 10.1007/s10393-016-1100-5

[9] Evans, T.L., (2015). Transdisciplinary collaborations for sustainability education: Institutional and intragroup challenges and opportunities. *Policy Futures in Education,* 13(1): 70-96.

[10] Stock, P. and Burton, R. J. F., (2011). Defining Terms for Integrated (Multi-Inter-Trans-Disciplinary) Sustainability Research. *Sustainability,* 3(8), 1090-1113.

[11] Bhadani, A. and D. Jothimani, (2017). Big Data: Challenges, Opportunities and Realities. *ArXiv*, 2017. doi:10.48550/arXiv.1705.04928.

[12] Bengtsson, L., Gaudart, J., Lu, X., Moore, S., Wetter, E., Sallah, K., Rebaudet, S., Piarroux, R., (2015). Using mobile phone data to predict the spatial spread of cholera. *Scientific Reports,* 2015. 5: 8923. doi:10.1038/srep08923.

[13] Benesh, E., Lamb, L. E., Connors, S. K., Farmer, G. W., Fuh, K. C., Hunleth, J., Montgomery, K. L., Ramsey, A. T., Moley, K. H., Colditz, G. A., Gehlert, S. J., (2015). A Case Study Approach to Train Early-Stage Investigators in Transdisciplinary Research. *Transdisciplinary Journal of Engineering & Science,* 2015. 6. 13-22. doi:10.22545/2015/00071.

[14] Hooper, S.E., Heather, H., Frazier, D., (2020). Transdisciplinary Team Building Using a Real-World Case Study on the Pandemic COVID-19. *Mendeley Data.* doi: 10.17632/sgngmzxzbr.1.

[15] Hall, K.L., Stokols, D., Vogel, B. A., Feng, A. L., Masimore, B., Morgan, G., Moser, R. P., Marcus, S. E., Berrigan, D., (2012). Assessing the Value of Team Science: A Study Comparing Center- and Investigator-Initiated Grants. *American Journal of Preventive Medicine,* 42(2), 157-163. doi:10.1016/j.amepre.2011.10.011.

[16] Batorowicz, B., Shepherd, T. A., (2008). Measuring the quality of transdisciplinary teams. *Journal of Interprofessional Care,* 22(6): 612-20.

[17] Anthoine, E., Delmas, C., Coutherut, J., Moret, L., (2014). Development and psychometric testing of a scale assessing the sharing of medical information and interprofessional communication: the CSI scale. *BMC Health Services Research,* 14. 126.

[18] Dale, A., Newman, L., Ling, C., (2010). Facilitating transdisciplinary sustainable development research teams through online collaboration. *International Journal of Sustainability in Higher Education,* 11(1), 36-48.

[19] McNeill, D., (1999). On Interdisciplinary Research: with particular reference to the field of environment and development. *Higher Education Quarterly,* 53(4), 312-332.

[20] Thompson, J.L., (2009). Building Collective Communication Competence in Interdisciplinary Research Teams.*Journal of Applied Communication Research,* 37(3), 278-297.

[21] Swanson, J.A., Renes, S. L., A.T. Strange, A. T., (2020). *The Communication Preferences of Collegiate Students, in Online Teaching and Learning in Higher Education* (pp 65-78). Springer International Publishing.

[22] Robinson, S. and H.A. Stubberud, (2012). Communication preferences among university students. *Academy of Educational Leadership Journal,* 16(2), 105.

[23] Roche Data Science Coalition (RDSC). UNCOVER COVID-19 Challenge, Version 4. https://www.kaggle.com/roche-data-science-coalition/uncover (accessed June 15, 2020).

[24] National Cancer Institute. GIS Portal for Cancer Research NCI Cancer Atlas. https://gis.cancer.gov/canceratlas/ (accessed October 15, 2020).

[25] Press, G., (2016). Cleaning big data: Most time-consuming, least enjoyable data science task, survey says. *Forbes,* March, 2016. 23, 15.

[26] Hernández, M.A. and S.J. Stolfo, (1998). Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem. *Data Mining and Knowledge Discovery,* 2(1): 9-37.

[27] Mack, C., Z. Su, and D. Westreich, (2018). *Managing missing data in patient registries: addendum to registries for evaluating patient outcomes: a user's guide.* Third Edition Rockville, MD: Agency for Healthcare Research and Quality (US).

[28] Chowdhury, M.H., M.K. Islam, and S.I. Khan. (2017). Imputation of missing healthcare data. *In 2017 20th International Conference of Computer and Information Technology (ICCIT).* 2017. IEEE. Paper presented at Dhaka, Bangladesh, 2017, (pp. 1-6). doi:10.1109/ICCITECHN.2017.8281805.

[29] Grzymala-Busse, J.W. and W.J. Grzymala-Busse, (2010). *Handling Missing Attribute Values, in Data Mining and Knowledge Discovery Handbook.* (pp 33-51). Boston, MA: Springer US.

[30] Williams, D.A., (1987). Generalized Linear Model Diagnostics Using the Deviance and Single Case Deletions. *Journal of the Royal Statistical Society: Series C (Applied Statistics),* 36(2): 181-191.

[31] Choi, B.C. and A.W. Pak, (2007). Multidisciplinarity, interdisciplinarity, and transdisciplinarity in health research, services, education and policy: 2. Promotors, barriers, and strategies of enhancement. *Clinical and Investigative Medicine,* 30(6), E224-E232. doi: 10.25011/cim.v30i6.2950

[32] Norris, P. E., O'Rourke, M., Mayer, A. S., Halvorsen, K. E., (2016). Managing the wicked problem of transdisciplinary team formation in socio-ecological systems. *Landscape and Urban Planning,* 154, 115-122.

[33] Khalid, S., Khalil, T., Nasreen, S., (2014). A survey of feature selection and feature extraction techniques in machine learning. Paper presented at 2014 Science and Information Conference. London, UK, August 27-29, 2014 (pp 1-8).

[34] Bittencourt, H. and R. Clarke, (2004). Feature selection by using classification and regression trees (CART). *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences,* https://www.semanticscholar.org/paper/FEATURE-SELECTION-BY-USING-CLASSIFICATION-AND-TREES-Bittencourt-Clarke/e4f02f48769937e2db37d8770df68b1c0b85735f (accessed October 1, 2020).

[35] Pohl, C. and Hadorn, G. H., (2008). Methodological challenges of transdisciplinary research. *Natures Sciences Sociétés,* 16(2), 111-121.

[36] Ramadier, T., (2004). Transdisciplinarity and its challenges: the case of urban studies.*Futures,* 36(4), 423-439. doi: 10.1016/j.futures.2003.10.009

[37] Godemann, J., (2008). Knowledge integration: a key challenge for transdisciplinary cooperation. *Environmental Education Research,* 14(6), 625-641.

[38] Wuchty, S., Jones, B.F., Uzzi, B., (2007). The Increasing Dominance of Teams in Production of Knowledge. *Science,* 316(5827), 1036-1039.

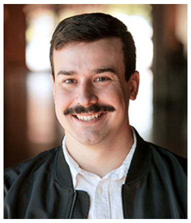
[39] Rosenfield, P.L., (1992). The potential of transdisciplinary research for sustaining and extending linkages between the health and social sciences. *Social Science & Medicine,* 35(11), 1343-1357.

[40] Gibbs, P. (Ed.), (2015). *Transdisciplinary Professional Learning and Practice.* Switzerland: Springer International Publishing.

About the Authors

Dr. Sarah Hooper is an integrative physiologist and serves as an Assistant Professor of Physiology at Ross University School of Veterinary Medicine. After completing her Doctor of Veterinary Medicine (DVM) degree at the University of Georgia, she was selected as a trainee on a National Institutes of Health (NIH) Institutional Research Training Grant (T-32) awarded through the Comparative Medicine Program at the University of Missouri to pursue a residency and PhD. Developing a passion for veterinary education during her residency program, she continued to train veterinary students during her postdoctoral research fellowship with the US Forest Service Northern Research Station. Recognizing the importance of emerging technologies in teaching, learning, and assessment, Dr. Hooper concurrently pursued a MS in data science and analytics to complement her research skills and teaching interests.

Dean Frazier completed his Master's degree in Data Science and Analytics from the University of Missouri in 2021 and joined The Crossing in Columbia, MO, as a Data and Analytics Engineer. Previously he was a teacher at Hickman High School for six years teaching a variety of mathematics, statistics, and AP courses.



Dean Frazier completed his Master's degree in Data Science and Analytics from the University of Missouri in 2021 and joined The Crossing in Columbia, MO, as a Data and Analytics Engineer. Previously he was a teacher at Hickman High School for six years teaching a variety of mathematics, statistics, and AP courses.