Integrating Biomedical Informatics Training into Existing High School Curricula

Avantika R. Diwadkar, MS[1], Susan Yoon, PhD[2], Jooeun Shim, MSEd[2], Michael Gonzalez, PhD[1], Ryan Urbanowicz, PhD[1], Blanca E. Himes, PhD[1]

[1]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA, US

[2]Graduate School of Education, University of Pennsylvania, Philadelphia, PA, US

Abstract

Growing demand for biomedical informaticists and expertise in areas related to this discipline has accentuated the need to integrate biomedical informatics training into high school curricula. The K-12 Bioinformatics professional development project educates high school teachers about data analysis, biomedical informatics and mobile learning, and partners with them to expose high school students to health and environment-related issues using biomedical informatics knowledge and current technologies. We designed low-cost pollution sensors and created interactive web applications that teachers from six Philadelphia public high schools used during the 2019-2020 school year to successfully implement a problem-based mobile learning unit that included collecting and interpreting air pollution data, as well as relating this data to asthma. Through this project, we sought to improve data and health literacy among the students and teachers, while inspiring student engagement by demonstrating how biomedical informatics can help address problems relevant to communities where students live.

Introduction

Biomedical informatics—an interdisciplinary field that studies and pursues the effective uses of biomedical data, information, and knowledge for scientific inquiry, problem solving, and decision making, motivated by efforts to improve human health[1]—is a specialty that arose in response to the need for computing in biology and medicine. As data-intensive computing has become essential to research and practice in various health domains, the demand for professionals trained in data science and biomedical informatics has steadily risen[2–4]. Further, as data-driven health decisions are increasingly necessary in daily life, health literacy (i.e., the capacity to obtain, communicate, process, and understand basic health information and services to make appropriate health decisions) and data literacy (i.e., the ability to read, work with, analyze, and argue with data) are critical life skills for all citizens[5,6]. Long-standing training programs in biomedical informatics exist at the post-graduate and undergraduate levels, and elements of data science and biomedical informatics are increasingly becoming embedded in the curricula to train health professionals and undergraduates[7–9]. A training gap exists, however, at the level of high school, whereby many high school graduates are under-prepared for, and unaware of, careers in biomedical informatics. Of further concern, levels of data and health literacy may be low in communities relying on out-of-date high school science curricula[10,11]. In response to these issues, several summer programs have been designed by universities to expose high school students to bioinformatics and biomedical informatics research[12–14], and the American Medical Informatics Association (AMIA) began the High School Scholars Program to provide a dissemination venue for these students to report their research findings and network more broadly with trainees and professionals across the U.S.[15]. Although these programs have been very successful, they reach a limited number of students and can be biased toward those who are already prepared and high achieving. To prepare a wide range of students for biomedical informatics careers and equip them with fundamental skills in health and data literacy, teaching in the high school setting is necessary.

Early exposure to science, technology, engineering, and mathematics (STEM) careers and academic preparedness in K-12 are critical factors for success in STEM at the undergraduate level and beyond[16]. Programs seeking to broaden participation in STEM have thus attempted to enhance teaching of, and provide exposure to, STEM-related areas to students of all ages[17,18]. For example, introduction of bioinformatics curricula has shown significant gains in cognitive traits among high school students, as well as an increased interest in STEM careers[19]. Structured summer and afterschool research programs can also be highly effective STEM-strengthening interventions for high school students and may help increase the participation and diversity within the biomedical workforce[20,21]. The viability of incorporating training of STEM, and biomedical informatics specifically, into high school curricula is challenged by the lack of curricular resources and teacher knowledge[10]. Therefore, professional development activities initiated through collaborations between educators and scientists are necessary to effectively improve student proficiency in these areas[22]. Moreover, due to the complexity and continuous evolution of biomedical informatics approaches, there

is a general agreement that new curricula should be inquiry-based, tied to STEM content already taught, and support real-world problem solving[10,23]. Any new material must also impart basic biomedical informatics skills through simple, nonburdensome integration into the existing high school curricula. To imitate real-world scientific practice, mobile learning and technologies (e.g., portable sensors and mobile apps) offer an inexpensive way to collect large-scale data and embed activities into the student's context that enable authentic experimentation and participation[24,25].

Over half of Philadelphia, PA consists of Environmental Justice areas—census tracts where 20 percent or more of individuals live in poverty and/or 30 percent or more of the population is minority[26]. Residents of these areas have historically suffered a disproportionate burden of pollutant exposures from sources that include oil refineries, trash incinerator plants, and vehicular exhaust from major roadways, putting them at increased risk for various diseases, including asthma. Exposure to environmental pollutants such as fine particulate matter ($PM_{2.5}$) has been associated with increased risk of asthma exacerbations[27,28], and Philadelphia is among the most polluted cities in the U.S[29]. Over 20 percent of children in Philadelphia have asthma[30,31], with hospitalizations occurring at a rate of 59.1 per 10,000[32]. Asthma prevalence and hospitalizations disproportionately affect Black and Hispanic residents and those who are socioeconomically disadvantaged[32,33]. Because many Philadelphia children have asthma or know of people in their communities who do, asthma serves as a relatable case study for students to understand how the environment (air pollution), genetics, and lifestyle factors (smoking) contribute to disease. Further, biomedical informatics approaches can readily be applied to studies of asthma and its various risk factors which facilitates teaching of biomedical informatics in different high school classes (environmental science, biology).

We initiated a professional development project entitled "K-12 Bioinformatics" in 2019 to educate Philadelphia area high school teachers about data analysis and mobile learning, and partner with them to expose high school students to health and environment-related issues using biomedical informatics knowledge and current technologies. Through this project, we sought to impart skills that will aid students to investigate real-world health problems and inspire them to take action in their local communities, while also generating interest in future biomedical informatics and STEM careers. Here, we describe some of the materials we created to facilitate teaching of biomedical informatics to high school teachers, and report how teachers used these materials during the academic year.

Methods

Summer Institute for Teacher Professional Development. During the summer of 2019, three environmental science and three biology teachers from the School District of Philadelphia attended a three-week professional development course to bring current science research into the secondary classroom through a problem-based learning curriculum in biomedical informatics. Three teachers identified as African American and three identified as White. The schools they taught in ranged in race and ethnic diversity, each qualified for U.S. federal Title I Funds (i.e., low income families made up at least 40 percent of the enrollment), and the percent of students per school deemed *proficient or advanced* in the Keystone Exam ranged from 2 to 100 percent. In partnership with researchers from the University of Pennsylvania Graduate School of Education and Perelman School of Medicine, teachers learned to build and integrate mobile technologies into classroom activities and designed a problem-based learning unit for classroom implementation during the 2019-2020 school year. The course work consisted of in-person lectures that introduced concepts such as bioinformatics and air pollution, hands-on trainings with a "teacher-as-student" pedagogy such as investigating and interpreting air quality data using tools, modifying problem-based learning units for their existing curricula, and pilot testing their units with high school students invited to participate in the summer program.

Low-Cost Pollution Sensor Assembly. We designed and assembled low-cost $PM_{2.5}$ and carbon monoxide (CO) portable sensors using commercially available Android components. The prototype design was improved by the Penn Electronic Design shop who ordered custom printed circuit boards to facilitate assembly of 100 sensors. Each pollution sensor kit included a sensor, Android smartphone and 9V batteries (Figure 1A). The sensors paired via Bluetooth with an Android smartphone to record pollution measures with an app created with App Inventor, a block-based programming language for building Android apps (Figure 1B). Sensor measurements, phone location, time, de-identified student IDs, and other general information captured by the app was saved to a Google Sheet (Figure 1C). Details regarding the sensor components and assembly, as well as the related code to capture measures are available at https://github.com/HimesGroup/k12bioinformatics.

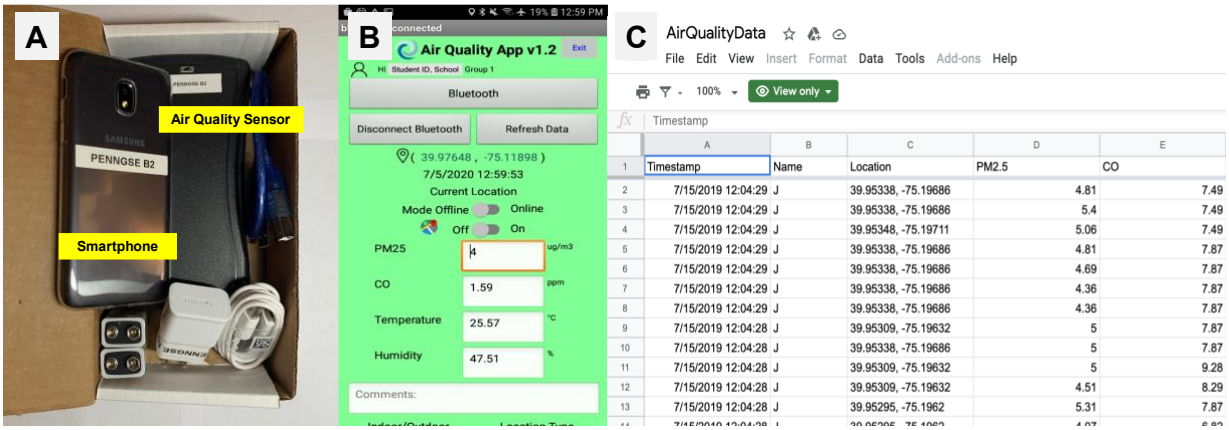

Figure 1. Low-cost pollution sensor measurements. (A) Air quality sensor kit included sensor, smartphone, cables, and 9V batteries. (B) Air quality app created with App Inventor collects measures of $PM_{2.5}$, CO, temperature, humidity, and geographic data. (C) A Google sheet was used to store data from all sensors distributed to a school.

Regulatory Monitor Air Pollution Data. To provide context for low-cost sensor pollution measures, data from Environmental Protection Agency (EPA) regulatory monitors was obtained for eight diverse U.S. cities (Philadelphia, PA, New York, NY, Los Angeles, CA, Miami, FL, Pierre, SD, Billings, MO, Standing Rock, NM and Portland, OR). Daily average $PM_{2.5}$ measures for September 2017 were computed from AirData[34] hourly estimates for all available monitors within each city. Monthly averages of $PM_{2.5}$ and CO for each city based on measures at a single geocoordinate during the years 2007-2017 were obtained with the R package *pargasite*[35].

Gene Expression Microarray Data. We searched for a gene expression microarray study related to the effects of cigarette smoking in the Gene Expression Omnibus (GEO) and selected one that measured differences in gene expression in macrophages from 13 cigarette smokers versus 11 non-smokers (GEO accession number GSE8823)[36]. The RAVED pipeline was used to analyze this dataset (https://github.com/HimesGroup/raved)[37], which included obtaining quality control metrics using outlier scoring methods from the *arrayQualityMetrics* R package[38], transforming the raw data using the robust multi-array average (RMA)[39] method, and performing differential expression analysis with the *limma* R package[40]. The Benjamini-Hochberg approach was used to correct for multiple comparisons, and an adjusted p-value <0.05 was considered significant.

Web Application Development. Prior to the summer institute for high school teacher professional development, we designed prototypes of three web-based tools that would provide an interactive user-interface to teach biomedical informatics concepts applied to topics that are highly relevant to students in the Philadelphia area, namely asthma and environmental exposures (e.g., air pollution and smoking). During the summer institute, as teachers designed problem-based mobile learning modules for use during the school year, the apps were updated based on teacher feedback to meet their needs. The apps, available at http://www.k12bioinformatics.org/, assisted with teaching of biomedical informatics concepts: 1) an exploratory data analysis tool for basic data visualization, 2) an app to visualize air pollution measures from low-cost sensors and regulatory monitors, and 3) an app demonstrating steps of gene expression microarray analysis. The apps were created with the R *Shiny* package[41] and deployed on a DigitalOcean droplet containing RStudio Connect and RStudio Server Pro. Various R packages, including *leaflet*[42] and *ggiraph*[43] were used to create the apps that displayed interactive plots and maps. Full code for the apps is available at https://github.com/HimesGroup/k12bioinformatics.

Results

Exploratory Data Analysis.

Description of App. The goal of this app was to help convey concepts related to descriptive statistical analysis. Users can upload either their own tabular data file in comma-separated-value (csv) format or use an example file available in the app. Tabular data is parsed by the app to identify columns that contain categorical and continuous variables, and subsequently, visualize univariate (e.g., bar plot for categorical variable, histogram for continuous variable) and bivariate (e.g., split bar plots, split box plots) distributions of user-selected variables. Students can learn to interpret

these visualizations using their own spreadsheets without the need to code. This app was intended for use with, for example, the phenotype file associated with the gene expression microarray study.

Use During the School Year. Students initially tested the app by uploading the example data file to understand types of variables and learn how to interpret bar plots, histograms and box plots. Subsequently, they utilized their own data to generate summary statistics and visualizations. For example, in one lesson, teachers provided students with large air quality datasets from two regulatory monitor sites (i.e., Torresdale Station and Car-Barn Montgomery I -76), and students uploaded the dataset and made plots (e.g., bar plot of mean values and box plot of all values) to compare differences between the two sites and examine air quality measures over time. Teachers guided students to identify what was familiar and what was confusing. Additionally, teachers asked students to respond to questions such as "What can we learn from a box plot that we can't learn from a bar plot and vice versa?" to ensure students learned proper usage of each visualization. For final student project reports, teachers instructed students to use the air quality data they collected along with Philadelphia asthma rate data to generate plots with this app.

Air Pollution Analysis and Visualization.

Description of App. The goal of this app was to help describe characteristics and visualize the geospatial distribution of low-cost air pollution sensor measures recorded by the students at all participating schools. Additionally, this app provides regulatory monitor pollution data that students can use as gold-standard pollution measures to contrast with those of low-cost sensors. The "EPA Measures Map" tab shows average $PM_{2.5}$ estimates in September 2017 for each of the eight cities, with the highest value for Portland, OR (10.5 $\mu g/m^3$) and lowest for Standing Rock, NM (5.6 $\mu g/m^3$) (Figure 2A). Interactive scatter plots of monthly average $PM_{2.5}$ and CO levels for a user-selected location and time period that provide information regarding pollution trends in the eight cities across a 10-year period are provided in the next tab. Seasonal variation in $PM_{2.5}$ and CO monthly average estimates can be observed, with differing trends based on location. For example, monthly $PM_{2.5}$ estimates in Philadelphia across the years 2007 to 2017 were consistently higher in the period from May to August while such trends were generally absent in Portland for the same time period (Figure 2B). The "Crowdsourced Map" and "Crowdsourced Summary Plots" tabs enable users to access, process, and analyze crowdsourced sensor data. Specifically, students can compute daily averages of sensor measures and compare them with the available regulatory monitor data and visualize in an interactive map of Philadelphia the estimates recorded for a user-selected pollutant, date range, school and sensor (Figure 2C). The overall characteristics of the selected data can be studied through univariate (box plot and histogram) plots for each pollutant as well as bivariate (scatter) plots comparing $PM_{2.5}$ and CO distributions.

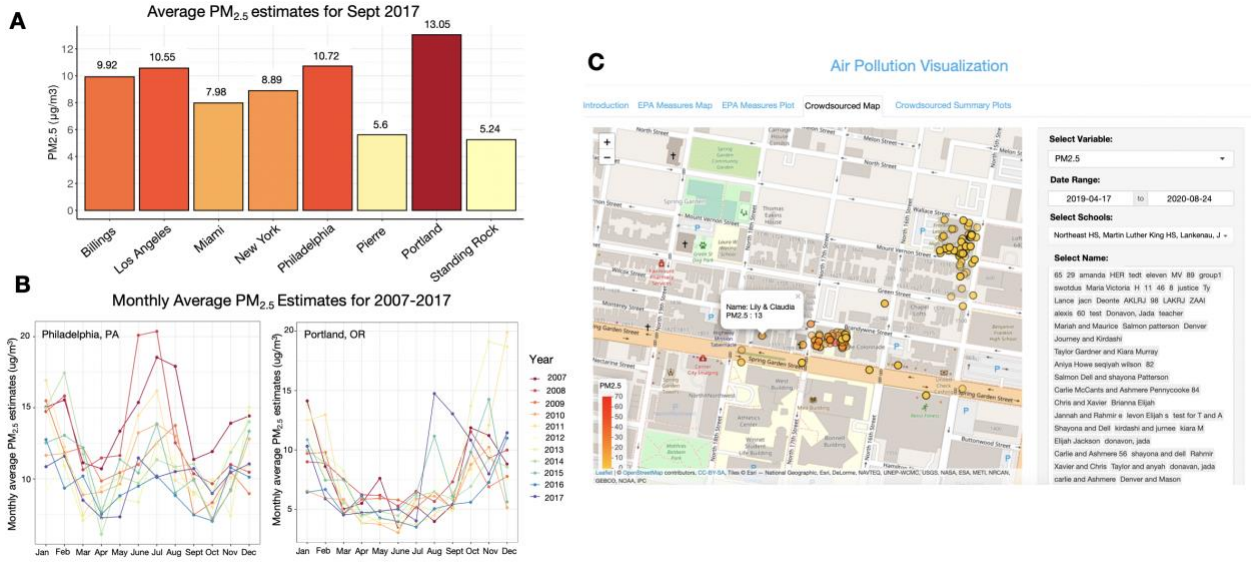

Figure 2. Features of air pollution visualization app. (A) Bar plot of average EPA AirData $PM_{2.5}$ estimates for September 2017 in eight U.S. cities. (B) Monthly $PM_{2.5}$ estimates for Philadelphia, PA and Portland, OR from 2007-2017. (C) Interactive map of crowdsourced $PM_{2.5}$ and CO sensor data acquired by students and teachers of six Philadelphia high schools.

Use During the School Year. During the school year, 153 students and teachers from six School District of Philadelphia high schools participated in the air pollution problem-based learning units. On average, 15 sensors were distributed to each of the six schools and managed using technology tracker sheets. In classrooms, one sensor package was assigned to a group of 2-4 students with distributed roles (e.g., sensor carrier, phone carrier, map navigator, observer and cartographer). Students were introduced to concepts related to air pollution, types of pollutants, EPA regulatory monitors in the U.S., and how the low-cost sensors worked. For the data collection activities, teachers first set classroom expectations (e.g., carry yourself as though you are a scientist) and then students gathered outdoor measures for 30-40 minutes. The research team provided on-site support when teachers introduced the sensors to the students and 1-2 backup sensor packages were available in case of technical difficulties. During the school year, 2001 measures of $PM_{2.5}$ and CO were successfully recorded. Using the app, students visualized data for their own schools and compared the pollutant levels across dates and geolocations to identify the most polluted areas among the surveyed neighborhoods in Philadelphia, recognizing that low-cost sensors are limited compared to research grade or regulatory monitors. Additionally, students downloaded the AirData $PM_{2.5}$ pollution estimates for the eight U.S. cities provided in the app and calculated the daily average estimate in each city for the month of September 2017. They compared these estimates with the *pargasite*-acquired estimates displayed in the app and explored seasonal trends across the years 2007-2017 via interactive scatter plots.

Gene Expression Microarray Analysis.

Description of App. The goal of this app was to allow students to explore the workflow of a gene expression microarray analysis without having to write code to do so. The "Sample Characteristics" tab explores the phenotype data information (smoking status, age, ethnicity and sex of the sample donors) associated with the study through univariate and bivariate plots of user-selected variables. The "Quality Control" analysis tab includes normalization of gene expression raw data using the RMA method, outlier identification through both log_2-transformed/normalized intensity distribution box plots (Figure 3A) and intensity curves of all samples in the dataset, and dimensionality reduction with principal component analysis (PCA) plots to visualize clustering patterns and batch effects for user-selected variables (Figure 3B). Under the "Differential Expression Results" tab, the top 50 significant differentially expressed gene probes (adjusted p-value < 0.05) are displayed in tabular form along with a volcano plot, a box plot comparing normalized read counts of smokers versus non-smokers for a user-selected gene (Figure 3C), and a heatmap of a user-selected number of top ranking probes (Figure 3D).

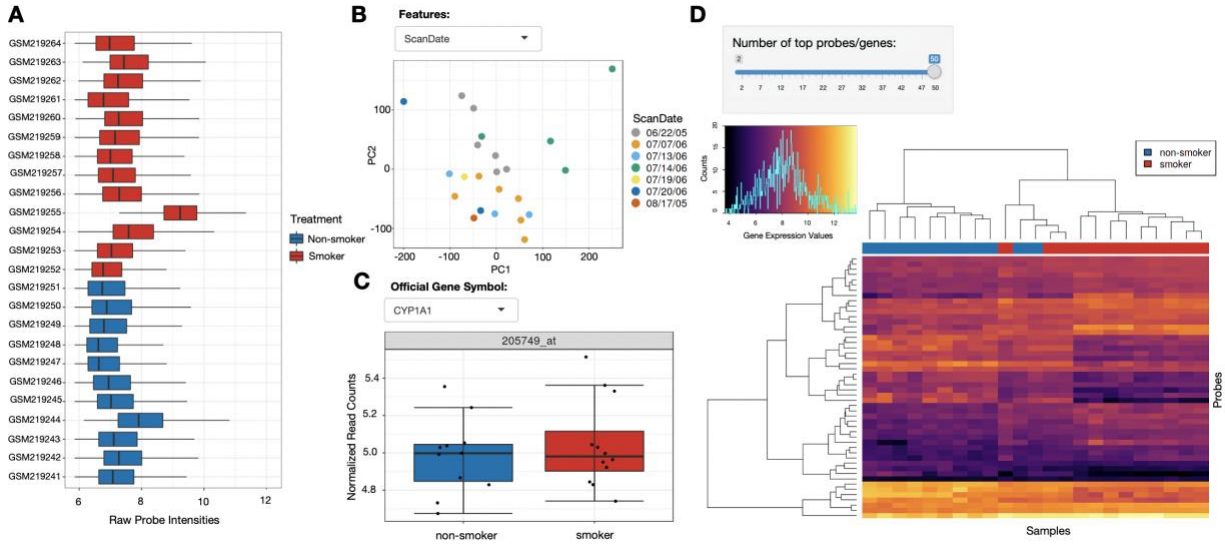

Figure 3. Features of gene expression microarray analysis app. (A) Log_2-transformed/normalized intensity distribution box plots of each sample. (B) Principal Component Analysis (PCA) plot colored according to a user-selected variable. (C) Box plots comparing normalized read count distributions of a user-selected gene *CYP1A1* in smokers versus non-smokers. (D) Heatmap based on a user-selected top number of probes/genes.

Use During the School Year. This app was not used by teachers during the school year, as they thought too many concepts would have to be introduced that were beyond the scope of their current curriculum. However, one teacher had students use the "Sample Characteristics" tab as a homework assignment in which students were asked to make an inference by writing a claim-evidence-reasoning essay on smoker/non-smoker data. The teacher asked students to

identify what they noticed and think of possible research questions based on this data set. Teacher prompts to support students making a claim and inference included, "Can you hypothesize answers to those questions?" and "What evidence do you have to support these inferences?".

Discussion

Despite its increasing scientific and practical importance, biomedical informatics remains a complex field and activities imparting related theoretical knowledge coupled with informatics skills and scientific thinking are atypical in current school tasks. Due to its relatively recent emergence and the continuous evolution of its scope, optimally integrating biomedical informatics in the high school scientific curricula is not straightforward. A previous study that evaluated existing U.S. secondary school science standards for bioinformatics found low and vague representation of bioinformatics across various topics, the lowest of which was "Human Genome Project/genomics and computer use" (implemented in only 4 of 49 U.S. states and the District of Columbia)[44]. There are major challenges to incorporating bioinformatics, and biomedical informatics more broadly, into high school lessons: 1) no guidance is provided on how to teach bioinformatics topics among existing frameworks[44], 2) teachers often lack any prior experience in bioinformatics research or its instruction[45], 3) schools lack sufficient computational infrastructure (e.g., computers, internet access, qualified IT personnel)[45], and 4) the majority of students do not have basic programming skills. To help overcome these challenges, biomedical informatics should be introduced as a practical education elective compatible with the time frame, cognitive level and available resources of the target population, while requiring minimal technical support and providing long-term training to the teachers[45]. Several external educational outreach programs offering bioinformatics training and research opportunities to high school students exist across the U.S.[13,14], but most engage directly with the students on an individual basis, and very few provide professional development and training strategies for the teachers[19]. Furthermore, public-facing education programs that engage with a mobile-learning curriculum designed around community issues of low-income minority neighborhoods to boost youth participation in local science campaigns and policymaking are still few in number[24].

The K-12 Bioinformatics program is the first of its kind in the Philadelphia area to provide advanced scientific seminars and professional training to educators, while equipping them with open-source, freely available web-based interactive tools that aid in imparting biomedical informatics training to high school students via effective visualizations and concise analysis workflows. We chose web-based resources that are akin to the facilities available at the participating public schools with majority low-income enrollments. Additionally, the resources such as databases and analysis tools that help build upon the data literacy of the teachers were procurable and nonintimidating for usage. Our initiative included problem-based mobile learning activities focused on locally contingent health hazards that aimed to empower students to address real-world problems through technological and data-driven investigations and propel evidence-based arguments with community stakeholders.

While planning our initial summer training session, we experienced minor setbacks, especially in designing the classroom activity using the air pollution sensors. Assembling the sensors in-house was more cumbersome than expected: initial sensor prototypes took 75 minutes to build, including soldering wires, gluing buttons using epoxy, and testing devices. We had considered including sensor assembly with the school activity but opted to work with a local electronics design shop to speed up production and improve the fabrication process. A major issue was experienced by the biology teachers who deemed the gene expression app, designed specifically to incorporate bioinformatics, as too complicated and elaborate for usage. Moreover, due to an existing heavy academic course load, the biology teachers found that the integration of bioinformatics-related modules was too cumbersome and time-consuming. In contrast, the environmental science teachers found it easier to incorporate additional materials related to air pollution measures into their existing curricula. As the summer institute progressed, the biology teachers decided to use the air pollution visualization app for their classroom activity sessions too. Lastly, during the school year, we faced an unanticipated challenge: the school district's firewall blocked the server where the shiny apps were hosted because it was integrated into the primary k12bioinformatics.org website without a named domain (i.e., with an IP address only). Communicating with the appropriate high school's IT team and the school district to identify and resolve this issue delayed the implementation of the new unit with the first teacher's class.

Our first summer training session in 2019 was successful, and teachers provided helpful feedback that was used to improve the program. To minimize teacher cognitive load and ensure long-term effectiveness, we introduced redesigned activities into our professional development program for the second cohort of teachers that included a fully annotated teacher guide armed with student-facing instructional presentations elucidating relevant scientific information translated for high school populations and familiar readymade curricular resources. We also extended the

summer program duration to a longer period with fewer daily hours. Due to COVID-19-related public health measures, the 2020 summer institute was held virtually. Publicly available video lectures introducing concepts such as bioinformatics, public health informatics, air pollution, and genetics were recorded in advance for asynchronous delivery, and the program was delivered via the edX platform. The new group of eight teachers from seven School District of Philadelphia high schools along with five teachers from the first session will implement newly developed units into their schools over the 2020-2021 school year, which includes collecting indoor air quality estimates at home and performing data analysis. Additionally, teachers are designing novel teaching strategies to implement mobile learning units that are appropriate to the circumstances faced at their schools.

In ongoing work, we are improving the interface of our apps to add more background content for coherence and help ease individual differences in contextualizing and processing biomedical informatics-specific knowledge and skills. We are also re-evaluating multiple aspects of the existing apps to build more on the foundations of data analysis as a concept, rather than fast-track through advanced topics. Increasing the compatibility of the material by matching the information base with the cognitive level of the participating teachers and students will help enhance the usability of the tools, making them more impactful in fulfilling the overall goals of the initiative. To ensure that our materials can be disseminated more widely and given that sensors may be unavailable or their use cost prohibitive in some settings, we are creating modules that do not rely on sensors. As computer programming becomes a requirement in more high schools across the country, expanding teaching modules to include more coding will enable the design of higher-level training materials in biomedical informatics.

Conclusion

We launched the K-12 Bioinformatics professional development project to educate high school teachers about biomedical informatics and partner with them to expose high school students to health and environment-related issues using biomedical informatics knowledge and current technologies. We successfully implemented a problem-based mobile learning unit focused on measuring neighborhood air pollution with low-costs sensors and relating these measures to asthma, thereby providing a hands-on biomedical informatics research experience that can potentially set in motion increased youth engagement in health and environmental hazards at a community level and generate interest in pursuing higher education in biomedical informatics.

Acknowledgements

We would like to thank Phil Bowsher and RStudio for providing licenses to RStudio Server Pro and RStudio Connect, which facilitated the web application development and deployment. We thank Miguel E. Hernandez and Alexander Santos from the Penn Electronic Design Shop for helping with low-cost air pollution sensor design and assembly. This work was supported by National Science Foundation award DRK12 1812738.

References

1. Kulikowski CA, Shortliffe EH, Currie LM, Elkin PL, Hunter LE, Johnson TR, et al. AMIA Board white paper: definition of biomedical informatics and specification of core competencies for graduate education in the discipline. J Am Med Inform Assoc JAMIA. 2012 Dec;19(6):931–8.

2. Hersh W. The health information technology workforce: Estimations of demands and a framework for requirements. Appl Clin Inform. 2010;1(2):197–212.

3. Investing in America's data science and analytics talent. Business Higher Education Forum and PWC; [cited 2020 Aug 25]. Available from: https://www.bhef.com/sites/default/files/bhef_2017_investing_in_dsa.pdf

4. Missed opportunities? The labor market in health informatics, 2014. Burning Glass; [cited 2020 Aug 25]. Available from: https://www.burning-glass.com/wp-content/uploads/BG-Health_Informatics_2014.pdf

5. Wolff A, Gooch D, Montaner JJC, Rashid U, Kortuem G. Creating an understanding of data literacy for a data-driven society. J Community Inform. [cited 2020 Aug 25]; Available from: http://www.ci-journal.net/index.php/ciej/article/view/1286

6. Berkman ND, Sheridan SL, Donahue KE, Halpern DJ, Viera A, Crotty K, et al. Health literacy interventions and outcomes: An updated systematic review. Evid ReportTechnology Assess. 2011 Mar;(199):1–941.

7. Florance V. Training for informatics research careers: History of extramural informatics training at the national library of medicine. In: Berner ES, editor. Informatics Education in Healthcare. London: Springer London; 2014. p. 27–42. (Health Informatics). Available from: http://link.springer.com/10.1007/978-1-4471-4078-8_3

8. Zhan YA, Wray CG, Namburi S, Glantz ST, Laubenbacher R, Chuang JH. Fostering bioinformatics education through skill development of professors: Big genomic data skills training for professors. Ouellette F, editor. PLOS Comput Biol. 2019 Jun 13;15(6):e1007026.

9. Moore JH, Boland MR, Camara PG, Chervitz H, Gonzalez G, Himes BE, et al. Preparing next-generation scientists for biomedical big data: Artificial intelligence approaches. Pers Med. 2019 May;16(3):247–57.

10. Machluf Y, Gelbart H, Ben-Dor S, Yarden A. Making authentic science accessible—the benefits and challenges of integrating bioinformatics into a high-school science curriculum. Brief Bioinform. 2017 Jan;18(1):145–59.

11. Smith PS. What does a national survey tell us about progress toward the vision of the NGSS? J Sci Teach Educ. 2020 Aug 17;31(6):601–9.

12. Teen Research and Education in Environmental Science (TREES) by Center of Excellence in Environmental Toxicology (CEET) at University of Pennsylvania. [cited 2020 Aug 25]. Available from: http://ceet.upenn.edu/training-career-development/summer-programs/teen-research-and-education-in-environmental-science/

13. Dutta-Moscato J, Gopalakrishnan V, Lotze M, Becich M. Creating a pipeline of talent for informatics: STEM initiative for high school students in computer science, biology, and biomedical informatics. J Pathol Inform. 2014;5(1):12.

14. Stanford Institute of Medicine summer Research program (SIMR). [cited 2020 Aug 25]. Available from: https://simr.stanford.edu

15. Unertl KM, Finnell JT, Sarkar IN. Developing new pathways into the biomedical informatics field: the AMIA high school scholars program. J Am Med Inform Assoc. 2016 Jul;23(4):819–23.

16. Honey M, Pearson G, Schweingruber HA, National Academy of Engineering, National Research Council (U.S.), editors. STEM integration in K-12 education: Status, prospects, and an agenda for research. Washington, D.C: The National Academies Press; 2014. 165 p.

17. Museus SD, Palmer RT, Davis RJ, Maramba DC, Ward K, Wolf-Wendel L. Racial and ethnic minority students' success in STEM education. San Francisco, Calif.: Hoboken, N.J: Jossey-Bass Inc.; Wiley Periodicals; 2011. 140 p. (ASHE higher education report).

18. Fisher AJ, Mendoza-Denton R, Patt C, Young I, Eppig A, Garrell RL, et al. Structure and belonging: Pathways to success for underrepresented minority and women PhD students in STEM fields. PloS One. 2019;14(1):e0209279.

19. Kovarik DN, Patterson DG, Cohen C, Sanders EA, Peterson KA, Porter SG, et al. Bioinformatics education in high school: implications for promoting science, technology, engineering, and mathematics careers. CBE Life Sci Educ. 2013;12(3):441–59.

20. Salto LM, Riggs ML, Delgado De Leon D, Casiano CA, De Leon M. Underrepresented minority high school and college students report STEM-pipeline sustaining gains after participating in the Loma Linda University Summer Health Disparities Research Program. PloS One. 2014;9(9):e108497.

21. Chittum JR, Jones BD, Akalin S, Schram ÁB. The effects of an afterschool STEM program on students' motivation and engagement. Int J STEM Educ. 2017;4(1):11.

22. Pierret C, Sonju JD, Leicester JE, Hoody M, LaBounty TJ, Frimannsdottir KR, et al. Improvement in student science proficiency through InSciEd out. Zebrafish. 2012 Dec;9(4):155–68.

23. Shuster M, Claussen K, Locke M, Glazewski K. Bioinformatics in the K-8 classroom: Designing innovative activities for teacher implementation. Int J Des Learn. 2016 Feb 3 [cited 2020 Aug 25];7(1). Available from: https://scholarworks.iu.edu/journals/index.php/ijdl/article/view/19406

24. Taylor KH, Silvis D, Kalir R, Negron A, Cramer C, Bell A, et al. Supporting public-facing education for youth: Spreading (not scaling) ways to learn data science with mobile and geospatial technologies. Contemp Issues Technol Teach Educ 193. [cited 2020 Aug 25]; Available from: https://citejournal.org/volume-19/issue-3-19/current-practice/supporting-public-facing-education-for-youth-spreading-not-scaling-ways-to-learn-data-science-with-mobile-and-geospatial-technologies

25. Herodotou C, Villasclaras-Fernández E, Sharples M. The design and evaluation of a sensor-based mobile application for citizen inquiry science investigations. In: Rensing C, de Freitas S, Ley T, Muñoz-Merino PJ, editors. Open Learning and Teaching in Educational Communities. Cham: Springer International Publishing; 2014 [cited 2020 Aug 25]. p. 434–9. (Lecture Notes in Computer Science; vol. 8719). Available from: http://link.springer.com/10.1007/978-3-319-11200-8_38

26. Department of Environmental Protection, State of Pennsylvania. PA Environmental Justice Areas. Available from: https://www.dep.pa.gov/PublicParticipation/OfficeofEnvironmentalJustice/Pages/PA-Environmental-Justice-Areas.aspx

27. Orellano P, Quaranta N, Reynoso J, Balbi B, Vasquez J. Effect of outdoor air pollution on asthma exacerbations in children and adults: Systematic review and multilevel meta-analysis. PloS One. 2017;12(3):e0174050.

28. Mirabelli MC, Vaidyanathan A, Flanders WD, Qin X, Garbe P. Outdoor PM2.5, ambient air temperature, and asthma symptoms in the past 14 days among adults with active asthma. Environ Health Perspect. 2016;124(12):1882–90.

29. American Lung Association. State of the air 2019. [cited 2020 Aug 25]. Available from: http://www.stateoftheair.org/assets/sota-2019-full.pdf

30. Bryant-Stephens T, West C, Dirl C, Banks T, Briggs V, Rosenthal M. Asthma prevalence in Philadelphia: Description of two community-based methodologies to assess asthma prevalence in an inner-city population. J Asthma. 2012 Aug;49(6):581–5.

31. Mangione S, Yuen EJ, Balsley C. Asthma prevalence in children: A survey of 57 Philadelphia middle schools. (Asthma: guidelines, delivery system and public health). Chest; (No 4). Report No.: Vol 122.

32. Department of Public Health, City of Philadelphia. Philadelphia Community Health Assessment: Health of the City 2019. [cited 2020 Aug 25]. Available from: https://www.phila.gov/media/20191219114641/Health_of_City_2019-FINAL.pdf

33. Bryant-Stephens T. Asthma disparities in urban environments. J Allergy Clin Immunol. 2009 Jun;123(6):1199–206.

34. Pre-generated data files. United States Environmental Protection Agency. [cited 2020 Aug 25]. Available from: https://aqs.epa.gov/aqsweb/airdata/download_files.html

35. Greenblatt RE, Himes BE. Facilitating inclusion of geocoded pollution data into health studies. AMIA Jt Summits Transl Sci Proc AMIA Jt Summits Transl Sci. 2019;2019:553–61.

36. Kazeros A, Harvey B-G, Carolan BJ, Vanni H, Krause A, Crystal RG. Overexpression of apoptotic cell removal receptor MERTK in alveolar macrophages of cigarette smokers. Am J Respir Cell Mol Biol. 2008 Dec;39(6):747–57.

37. Kan M, Shumyatcher M, Diwadkar A, Soliman G, Himes BE. Integration of transcriptomic data identifies global and cell-specific asthma-related gene expression signatures. AMIA Annu Symp Proc. 2018;2018:1338–47.

38. Kauffmann A, Gentleman R, Huber W. arrayQualityMetrics--a bioconductor package for quality assessment of microarray data. Bioinforma Oxf Engl. 2009 Feb 1;25(3):415–6.

39. Carvalho BS, Irizarry RA. A framework for oligonucleotide microarray preprocessing. Bioinformatics. 2010 Oct 1;26(19):2363–7.

40. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015 Apr 20;43(7):e47.

41. Chang W, Cheng J, Allaire JJ, Xie Y, McPherson J, RStudio, et al. Shiny: Web application framework for R. 2019 [cited 2020 Aug 25]. Available from: https://CRAN.R-project.org/package=shiny

42. Cheng J, Karambelkar B, Xie Y, Wickham H, Russell K, Johnson K, et al. Leaflet: Create interactive web maps with the JavaScript "leaflet" library. 2019 [cited 2020 Aug 25]. Available from: https://CRAN.R-project.org/package=leaflet

43. Gohel D, Panagiotis S, Bostock M, Kokenes S, Schull E. Ggiraph makes 'ggplot' graphics interactive. [cited 2020 Aug 25]. Available from: https://davidgohel.github.io/ggiraph/

44. Wefer SH, Sheppard K. Bioinformatics in high school biology curricula: A study of state science standards. Schulz B, editor. CBE—Life Sci Educ. 2008 Mar;7(1):155–62.

45. Machluf Y, Yarden A. Integrating bioinformatics into senior high school: design principles and implications. Brief Bioinform. 2013 Sep 1;14(5):648–60.