frontiers in Artificial Intelligence

PERSPECTIVE
published: 15 May 2020
doi: 10.3389/frai.2020.00031

Bringing Big Data to Bear in Environmental Public Health: Challenges and Recommendations

Saskia Comess[1,2], *Alexia Akbay*[1,3], *Melpomene Vasiliou*[1], *Ronald N. Hines*[4], *Lucas Joppa*[5], *Vasilis Vasiliou*[1]* *and Nicole Kleinstreuer*[1,6]*

[1] Department of Environmental Health Sciences, Yale School of Public Health, New Haven, CT, United States, [2] Department of Statistics and Data Science, Yale University, New Haven, CT, United States, [3] Symbrosia Inc, Kailua-Kona, HI, United States, [4] US Environmental Protection Agency, Center for Public Health and Environmental Assessment, Research Triangle Park, NC, United States, [5] Microsoft Corporation, AI for Earth, Redmond, WA, United States, [6] National Toxicology Program Interagency Center for the Evaluation of Alternative Toxicological Methods, National Institute of Environmental Health Sciences, Research Triangle Park, NC, United States

Understanding the role that the environment plays in influencing public health often involves collecting and studying large, complex data sets. There have been a number of private and public efforts to gather sufficient information and confront significant unknowns in the field of environmental public health, yet there is a persistent and largely unmet need for findable, accessible, interoperable, and reusable (FAIR) data. Even when data are readily available, the ability to create, analyze, and draw conclusions from these data using emerging computational tools, such as augmented and artificial intelligence (AI) and machine learning, requires technical skills not currently implemented on a programmatic level across research hubs and academic institutions. We argue that collaborative efforts in data curation and storage, scientific computing, and training are of paramount importance to empower researchers within environmental sciences and the broader public health community to apply AI approaches and fully realize their potential. Leaders in the field were asked to prioritize challenges in incorporating big data in environmental public health research: inconsistent implementation of FAIR principles in data collection and sharing, a lack of skilled data scientists and appropriate cyber-infrastructures, and limited understanding of possibilities and communication of benefits were among those identified. These issues are discussed, and actionable recommendations are provided.

Keywords: artificial intelligence, public health, machine learning, open data, environmental health sciences, big data

OPEN ACCESS

Edited by:
Frank Emmert-Streib, Tampere University, Finland

Reviewed by:
Kristina Hettne, Center for Digital Scholarship at the Leiden University Library, Netherlands
Thomas Luechtefeld, Toxtrack LLC, United States

***Correspondence:**
Vasilis Vasiliou
vasilis.vasiliou@yale.edu
Nicole Kleinstreuer
nicole.kleinstreuer@nih.gov

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Artificial Intelligence

Received: *14 June 2019*
Accepted: *06 April 2020*
Published: *15 May 2020*

Citation:
Comess S, Akbay A, Vasiliou M, Hines RN, Joppa L, Vasiliou V and Kleinstreuer N (2020) Bringing Big Data to Bear in Environmental Public Health: Challenges and Recommendations. Front. Artif. Intell. 3:31. doi: 10.3389/frai.2020.00031

INTRODUCTION

Out of the tens of thousands of individual chemicals currently in commerce (and many more mixtures, natural products, and metabolites) <10% have been screened for safety. The U.S. EPA's Toxic Substances Control Act (TSCA) Chemical Substances Control Inventory contains roughly 85,000 chemicals (U.S. EPA, 2016), and the European Chemicals Agency (ECHA) Inventory lists over 100,000 unique substances (as of the most recent update in August 2017), of which ∼22,000 are registered substances with some information on structure, usage, or toxicity (ECHA, 2017). Understanding which chemicals in the environment, both with and without safety data,

pose a risk to human health requires that we more effectively leverage the data that we already have, and that we take intelligent approaches to generating new data. While the traditional means of collecting chemical safety data (animal models) are laborious and of variable accuracy and human relevance (Hartung, 2009), such reference data can still be used to train models for prioritizing and predicting toxicity of new chemicals, provided the data are curated in a computationally accessible format and, ideally, integrated with other lines of evidence providing mechanistic information. This requires significant effort, both in collecting, and extracting information as well as annotating it appropriately.

These toxicological problems are mirrored in public and environmental health more generally: huge, complex issues with inadequately curated data, and analytic power. Recent research in toxicology has focused on high-throughput screening to rapidly produce quantitative data on thousands of human biological targets (e.g., Thomas et al., 2019), data-mining to identify relevant end-points building predictive models for adverse toxicological outcomes (e.g., Saili et al., 2019), and application of cutting-edge machine learning (ML) and artificial or augmented intelligence (AI) techniques (e.g., Luechtefeld et al., 2018). Collectively, these technologies facilitate enhanced mechanistic insights and may obviate the need for inefficient testing in animal models, but they are still not considered mainstream approaches nor are they widely accepted by regulatory agencies. Individual research programs generate large data sets, but without centralized coordination, standardized reporting, and common storage structures, the data cannot be effectively combined and used to its full potential. The federal Tox21 research consortium has, to date, tested more than 9,000 chemicals to varying degrees in 1,600 assays and demonstrated environmental chemical interactions with critical human and ecologically-relevant targets (Tice et al., 2013). Translational systems approaches are being employed by this and other programs (e.g., Horizon 2020, EUToxRisk, CEFIC LRI, OpenTox) to produce diverse data streams and predict chemical effects on human health and disease outcomes (e.g., Kleinstreuer et al., 2014). At the same time, there have been substantial efforts to develop and deploy sensors and satellite systems that yield additional large and complex data sets that provide information about chemical exposures (Dons et al., 2017; Ring et al., 2019; Weichenthal et al., 2019). Further, epidemiologists are actively developing ML and AI approaches to enhance understanding of chemical exposures and associated disease risks (Brandon et al., 2018). However, these efforts also are largely disconnected from one another and operate independently, despite the clear potential benefits if such data could be combined and jointly analyzed. Given the need to integrate and analyze large, multifactorial data sets, researchers in public health and the environmental health sciences (EHS) would greatly benefit from the ability to collect, process, analyze, and make inferences on data using ML and AI. However, in these fields, a general lack of relevant knowledge among many researchers, sparse, distributed, or inaccessible data, and an inadequate framework for sharing and disseminating innovations impede efforts to implement these approaches. Here, we discuss three specific areas with room for improvement in the public health/EHS field: data collection and sharing, researcher knowledgebase, and a recognition of the benefits AI/ML can bring to current problems. Recommendations are provided in each of these areas to facilitate bringing big data to bear on public health and EHS challenges.

DATA COLLECTION AND SHARING

Challenge

A major hurdle confronting investigators conducting public health and EHS research is a lack of comprehensive human and environmental exposure and effects data that are annotated using controlled vocabularies. Addressing this problem is a prerequisite to applying AI and ML, as without sufficient, high-quality data and metadata, the analytic methods themselves are irrelevant. Quantifying environmental exposure, such as from air, water, soil, and food, is difficult both at the micro (localized to individuals and small geographic units) and macro (national and international) levels. For instance, air pollution can vary up to eight-fold within a given city block, but most U.S. cities have only one air quality monitor (EDF, 2019). Epidemiologic studies of air pollution health effects often must rely on disparate data that lack both temporal and spatial specificity and cannot account for the movement of people across different areas of pollution. Without continuous and advanced monitoring, and robust computer modeling methods, illnesses related to transient exposures might not be recognized as part of a significant pattern until substantial adverse health effects have occurred. This is one example where the development of AI tools in the EHS space has been hindered not by the AI technology capacity itself but instead by a lack of reliable, interconnected data (NAS, 2019). This is equally true in the medical sector with respect to patient treatment and outcomes. IBM's ambitious partnership with the MD Anderson Cancer Center to develop AI to expedite clinical decision-making has been at a standstill after years of development due to a lack of standardized, accessible data (Jaklevic, 2017).

Even when standardized data are available, finding, accessing, and processing it can be a monumental task. The absence of a uniform framework for openly sharing and storing data means that researchers devote significant time to locating relevant data. Knowledge of where to find data is often highly sector-specific, inhibiting cross-disciplinary research. For example, a climate scientist interested in public health would need knowledge of health-specific data repositories to conduct the search. Rather than waste manual effort and time in locating data, let alone integrating it, coordinated efforts could result in processes that could be automated and simplified. Ethical concerns have been voiced in regard to organizing large repositories of these types (Ienca et al., 2018). Of these, patient data privacy is a major concern, and breaches of patient records databases are a constant challenge. Unique patient identifier numbers and other de-identification/anonymization techniques can protect patient privacy, while allowing for meaningful research and analysis (Emam et al., 2015). New encryption based techniques allow for predictive modeling while maintaining the privacy of sensitive information, such as the application of homomorphic encryption to patient data in predicting cardiovascular disease (Bos et al.,

2014). However, inconsistent regulations and lack of practical protocols around handling sensitive information have resulted in unethical scenarios, where data is being sourced from countries where patients have minimal rights (Mittelstadt and Floridi, 2016). Not only is this problematic from an ethical perspective, it also limits AI innovation to only those who have access to these obscure datasets. Specific tools developed by startups who have the luxury of sourcing data from elsewhere are often acquired by large corporations, making innovation an exclusive pursuit. Thus, the environmental public health field requires a revolution in the collection and organization of environmental exposure and effects data as a first step in democratizing information access and building better models to improve predictions.

Recommendations

Further work is clearly needed in data collection and sharing, but recent attempts in specific sectors are exemplar in the aggregation of data and development of open, accessible repositories that maintain necessary privacy standards. In 2016, over 50 contributing researchers from global institutions proposed the "Findable, Accessible, Interoperable, and Reusable" (FAIR) Guiding Principles for scientific data management and stewardship (Wilkinson et al., 2016). These principles bridge the divide between human-conducted and machine-driven research behaviors. Using FAIR principles, the NIH is creating Data Commons, a platform for data management, and metadata cataloging for terminologies and ontologies (Mahony et al., 2018). This framework has been one of the key drivers behind new repositories and tools such as the National Toxicology Program's Integrated Chemical Environment (ICE) (Bell et al., 2017) portal and the U.S. EPA's CompTox Chemicals Dashboard (Williams et al., 2017), which allow FAIR principles to be applied to non-animal *in vitro* and *in silico* data, along with *in vivo* animal data and human exposure information. A collaboration between the US FDA, the non-profit Clinical Data Interchange Standards Consortium (CDISC), and other stakeholders, resulted in the development of study data standards for non-clinical, clinical, analysis and metadata (https://www.cdisc.org/standards) to create common reporting formats. These concepts are cornerstones of the 2018 U.S. Strategic Roadmap for Modernizing Safety Testing of Chemicals and Medical Products, developed by 16 U.S. federal agencies, which advocates for practices that increase confidence in new data-driven research methods (ICCVAM, 2018). A significant portion of the work done by these data powerhouses is retrospective data curation, often performed manually (e.g., Kleinstreuer et al., 2016). Work is ongoing to automate some aspects of the information extraction pipeline, but additional efforts to standardize reporting formats, and metadata terminologies in emerging research could lighten the curation burden on these institutions and streamline data annotation and storage, allocating greater resources to the development of novel applications.

Many of the recent advances in developing openly accessible databases of environmental exposure information have come from the private sector, often in partnership with non-profit organizations and academic institutions. The Monarch Initiative (https://monarchinitiative.org/) is one such collaboration to apply ontologies, or semantic descriptions, to disease phenotypes and enable intra- and inter-species comparisons and connection to genotypes, pathways, and experimental models. Another example is a pilot project in Oakland, California, between the Environmental Defense Fund (EDF) and Google Earth Outreach which involved attaching air quality sensors to Google Street View cars. This was recently extended to a partnership with an environmental sensor company (Aclima) to equip Street View cars with mobile air quality sensors in cities around the world (Business Wire, 2018). The sensors capture detailed air quality and emissions data at high (street-block level) resolution and temporal frequency. These data will be aggregated and made available on a Google database. Google also recently announced that it will report estimates of city-level greenhouse gas emissions and annual driving, biking, and transit ridership (data gathered via Google Maps and Waze) (Meyer, 2018; Google, 2019). Google's new Dataset Search is an initial attempt to apply distributed search to datasets from the environmental and social sciences, government data, and news organizations (Noy, 2018). By providing access to data from multiple disciplines via a single platform, researchers can conduct interdisciplinary work fundamental to environmental and public health (Vincent, 2018). Applying these powerful methods to better curate and integrate diverse sources of data will promote greater understanding of complex and dynamic systems. However, acceptance and implementation of these improvements are not yet widespread, particularly in the public and environmental health sectors. Following the lead of these innovative pilot studies, a greater emphasis needs to be placed on developing the appropriate infrastructure for effective, standardized data collection approaches, common ontologies, and uniform sharing protocols.

RESEARCHER KNOWLEDGE BASE

Challenge

Public health and EHS researchers are not traditionally trained in scientific computing and data science, and computer scientists do not typically apply their skill sets to EHS problems. This impedes the introduction and utilization of powerful data science techniques to public and environmental health practice and research. The American Association for Public Health, the body responsible for accrediting public health programs in the United States, currently does not include computer skills in the core competencies required of a Master's of Public Health (MPH) program. This omission is a disservice both to public health students and the field of public health research more broadly. Just as being able to read and write are fundamental skills and required baseline competencies for entering a graduate program, computer science will eventually become a similar prerequisite for comprehensive and effective analytical approaches across the biological and toxicological realm. Ultimately, the EHS and public health disciplines need a culture and skill shift. Over the next decade, scientists will need to understand the fundamentals of computer and data science in order to be successful in their field. Having this core computer science competency will be

critical to the future success of transdisciplinary research in the EHS and public health disciplines.

Recommendations

Current public health and EHS students are interested in these issues, which means that public health schools need to integrate computer and data science into their core curriculums. While students interested in big data and public health are not at a total loss for resources, provided they are willing to seek them out, there are major gaps in the academic arena. There are a handful of existing programs that currently provide the skill set needed to apply data science to public health research. The Master of Science in Data Science is an option offered from a number of Universities, but few offer, or even consider, the integration of this degree with applications in the field of public health. Within the status quo, students are presented with the option of either an MPH or an MS in Data Science, with little crossover between the two. Even programs such as Harvard's Health Data Science MS, which is offered through the School of Public Health, only requires one epidemiology course and then places the onus on the student to integrate the public health perspective into their own research.

The lack of computer skills in the core competencies of the MPH is one example of an area with obvious and immediate room for improvement. However, the need for cross-disciplinary training and communication is equally relevant from the perspective of computer scientists, who should be educated and engaged in EHS and public health applications. Students from both academic disciplines should be encouraged, if not required, to engage in coursework in the other. This idea of "cross-departmental partnerships" would arm public health students with the technical skills needed to integrate computer science, AI, and ML into their work while providing insight for potential EHS projects to which computer science students could apply their skills. The thoughtful design of such a program also would foster a culture of transdisciplinary research which will be key to finding solutions to complex environmental and public health problems. The Massachusetts Institute of Technology is laying the necessary groundwork for such a program by creating a new college for AI. With a $1 billion investment, the college is expected to start in the Fall of 2019 with the goal of integrating AI systems across academic fields (Lohr, 2018). Continuing education programs, such as the New Approach Methodology Use in Regulatory Application training series (https://www.pcrm.org/ethical-science/animal-testing-and-alternatives/nura), are being offered to ensure that environmental public health professionals also begin to develop the skills and expertise needed to leverage big data and implement AI and ML based approaches. A highly topical example of the benefit of teaching public health students computational skills is the widely referenced Johns Hopkins University resource for tracking the spread of the novel SARS COV-2 coronavirus (https://coronavirus.jhu.edu/map.html).

While the above focus on cross-discipline training is of paramount importance for future successful applications of AI and ML in the EHS, incentives for cross-disciplinary collaboration would have a more rapid impact and also would inform the integration of computer and data science into EHS curricula. Recognition of this opportunity by current EHS leadership and appropriate investments to achieve this goal would be of substantial benefit.

WHAT CAN AI AND ML DO FOR PUBLIC HEALTH AND EHS?

Challenge

A downstream consequence of the challenges detailed above is that the majority of researchers in the scientific community are still unaware of the benefits that AI and ML could provide when coupled with large, annotated, integrated datasets. A lack of familiarity with AI and ML as tools means that even when presented with examples of effective predictive models, potential end-users may not adopt them due to a lack of understanding, leading to decreased confidence in their utility. Further, without substantial investment in data curation and integration, the ability to apply AI and ML and build such models is severely limited.

Recommendations

While these approaches are not yet mainstream, there are many examples of extremely successful implementations of combining big data with AI and ML to build high-performing predictive models, and such case studies should be widely distributed and serve as the catalysts for increasing support in these research areas. Beyond establishing the cyber-infrastructure to generate and store open, accessible data, select researchers, and government agencies are developing modeling approaches to effectively leverage those data. Aggregation of scholarly data into more structured computational models, such as quantitative structure activity relationship (QSAR)-based chemical predictions, demonstrates the efficacy of pipelines which turn decentralized data inputs into centralized models (Mansouri et al., 2016). Keeping these models in siloed communities is counter-productive, as the fundamental methods of model creation relies on open data provided by researchers. Drawing on standards for collaboration and sharing within the computer sciences, the National Institutes of Health (NIH) and the National Institute for Standards and Technology (NIST) have organized multiple hackathons and public-private partnerships to automate data extraction efforts and to create computational models that map the biological effects of chemical exposures (e.g., Kleinstreuer et al., 2018). Models resulting from such enterprises have surpassed the accuracy and efficiency of traditional, manual-labor driven animal testing (e.g., Browne et al., 2017). Projects such as the NCATS Biomedical Data Translator are designed to establish cross-cutting infrastructures to facilitate these data integration and modeling efforts (Austin et al., 2019).

If the above-mentioned hurdles can be overcome, big data, AI, and ML represent a huge opportunity for the expansion and application of effective environmental public health research, as discussed at a recent 2019 National Academies of Sciences workshop on "Leveraging Artificial Intelligence and Machine Learning to Advance Environmental Health Research and Decisions" (http://nas-sites.org/emergingscience/meetings/ai/). There exist a number of examples of researchers who are

TABLE 1 | Challenges and recommendations for fostering the big data revolution in environmental public health, summarized here and detailed in text.

Challenge	Recommendation
Data are not collected using controlled terminologies or standardized reporting formats.	Study data standards and ontologies should be designed and widely implemented. Such resources should be both specific to individual data types and coordinated across sources to optimize data utility and potential for integration.
Silos of information prevent access, integration, and effective data science analyses and applications.	Implement FAIR principles in data curation and development of scientific cyber-infrastructures. Establishing a culture of openness and data accessibility requires multiple nodes of cross-sector communication. Funding opportunities should emphasize this necessity.
Current environmental public health curricula do not emphasize data science skills.	Public health curricula should be data-minded and with ample resources for learning and improving technical skills. Establish tech-focused learning by offering relevant course, digital resources, and hosting speakers. Acquiring new skills needs to be celebrated and credentialed. Similarly, computer science programs should offer training in environmental health sciences and other biological application-oriented foci.
Lack of technical skills among environmental public health professionals leads to out-sourcing of data-science tasks and lack of adoption.	Training in AI, ML, and data science is not exclusive to universities or educational institutions. Continuing education courses, seminars, and conference sessions should provide environmental public health professionals with specialized resources and hands-on training in data science and new approach methodologies.
Public health culture does not prioritize data expertise and experimentation.	The scientific community must build an agile network of ambassadors and support curiosity and experimentation. Major institutions should appoint a team of experts within the organization to steer institutional culture and develop a wider network of expertise while also encouraging students, faculty, regulators, and professionals to apply data science tools in innovative ways.

already bridging the fields of environmental public health, data science, and AI. The "AI for Earth" initiative partners Microsoft's data science acumen with researchers who have environmental expertise in the areas of agriculture, climate change, biodiversity, and water. Grants from AI for Earth have funded AI projects on population health model projections, image analysis for biodiversity, crop forecasting, climate-related landslide projections, modeling carbon sequestration, understanding global pathogen spread, and much more (Microsoft, 2018). AI and ML have facilitated a more nuanced and accurate understanding of climate patterns, improving the accuracy of forecasting extreme weather events to >90 percent (Cho, 2018). Researchers already use AI to improve air pollution forecasts (Fontes et al., 2014; Bellinger et al., 2017; Bai et al., 2018), disease diagnosis (Xiong et al., 2018), infectious disease monitoring (Milinovich et al., 2014 https://nextstrain.org/), tracking antibiotic resistance (Li et al., 2018), computational chemistry (Goh et al., 2017), exposure and chemical-mixture modeling (Bobb et al., 2014; Park et al., 2017; Vopham et al., 2018) and improving classification of climate regions (Liss et al., 2014). These innovative approaches and their successes are just the tip of the iceberg, and point to the potential benefit of sufficiently resourced investments in the application of AI and ML in environmental public health research.

CONCLUSIONS AND RECOMMENDATIONS

While grassroots innovation is increasing, top-down influencers within academia, government, industry, and funding bodies need to facilitate the conditions for AI to flourish in the fields of public and environmental health sciences. The philosophy of the "Fourth Industrial Revolution" (i.e., rapid technological advancement) is wildly different than the competition involved in typical academic research, as it requires enhanced interdisciplinary collaboration and universal data sharing and organization. As others have noted (Rubens et al., 2014), public health is generally slower than other scientific disciplines to embrace the use of advanced technologies. If we do not collectively aspire to change the framing of public health research and education, the discipline could impede its own progress. Technical talent will gravitate to the open, collective opportunities offered by private enterprise. To prevent this, change should percolate from those currently leading the field: researchers, educators, and practitioners need to understand the ingenious applications of emerging technologies and foster such opportunities within EHS research. Students should be encouraged, and even required, to explore these fields in core curriculums. During this time of unprecedented access to technical domains, cross-disciplinary training in computer science and EHS research will empower students and professionals alike to make meaningful contributions to perceivably the greatest revolution of their lifetime. We therefore propose actionable recommendations for leaders in the public and environmental health fields to implement and create an environment that will foster the data revolution (**Table 1**).

AUTHOR CONTRIBUTIONS

SC, AA, and MV formulated the research question/premise, conducted research and literature reviews, and wrote the first draft of the manuscript. NK and RH contributed writing and research to sections of the manuscript. VV, NK, LJ, and RH assisted with conceptualizing the research question and contributed to manuscript editing. All authors contributed to manuscript revision and read and approved the final manuscript.

REFERENCES

Austin, C. P., Colvis, C. M., and Southall, N. T. (2019). Deconstructing the translational tower of babel. *Clin. Translat. Sci.* 12:85. doi: 10.1111/cts.12595

Bai, L., Wang, J., Ma, X., and Lu, H. (2018). Air pollution forecasts: an overview. *Int. J. Environ. Res. Public Health* 15:780. doi: 10.3390/ijerph15040780

Bell, S. M., Sprankle, C., Morefield, S. Q., Allen, D., Phillips, J., Sedykh, A., et al. (2017). An integrated chemical environment to support 21st-century toxicology. *Environmental Health Perspectives* 125, 1–4. doi: 10.1289/EHP1759

Bellinger, C., Jabbar, M., Zaïane, O., and Osornio-Vargas, A. (2017). A systematic review of data mining and machine learning for air pollution epidemiology. *BMC Public Health* 17:907. doi: 10.1186/s12889-017-4914-3

Bobb, J. F., Valeri, L., Henn, B. C., Christiani, D. C., Wright, R. O., Mazumdar, M., et al. (2014). Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. *Biostatistics* 16, 493–508. doi: 10.1093/biostatistics/kxu058

Bos, J. W., Lauter, K., and Naehrig, M. (2014). Private predictive analysis on encrypted medical data. *J. Biomed. Inform.* 50, 234–243. doi: 10.1016/j.jbi.2014.04.003

Brandon, N., Dionisio, K. L., Isaacs, K., Tornero-Velez, R., Kapraun, D., Setzer, R. W., et al. (2018). Simulating exposure-related behaviors using agent-based models embedded with needs-based artificial intelligence. *J. Expo. Sci. Environ. Epidemiol.* 30, 184–193. doi: 10.1038/s41370-018-0052-y

Browne, P., Judson, R. S., Casey, W., Kleinstreuer, N., and Thomas, R. S. (2017). Correction to: screening chemicals for estrogen receptor bioactivity using a computational model. *Environ. Sci. Technol.* 51:9415. doi: 10.1021/acs.est.7b03317

Business Wire (2018). *Aclima & Google Scale Air Quality Mapping to More Places Around the World*. Available online at: www.businesswire.com/news/home/20180912005440/en/Aclima-Google-Scale-Air-Quality-Mapping-Places

Cho, R. (2018). *Artificial Intelligence—a Game Changer for Climate Change and the Environment*. State of the Planet: Earth Institute, Columbia University. Available online at: https://blogs.ei.columbia.edu/2018/06/05/artificial-intelligence-climate-environment/

Dons, E., Laeremans, M., Orjuela, J. P., Avila-Palencia, I., Carrasco-Turigas, G., Cole-Hunter, T., et al. (2017). Wearable sensors for personal monitoring and estimation of inhaled traffic-related air pollution: evaluation of methods. *Environ. Sci. Technol.* 51, 1859–1867. doi: 10.1021/acs.est.6b05782

ECHA (European Chemicals Agency) (2017). *ECHA Inventory*. Available online at: https://echa.europa.eu/information-on-chemicals/ec-inventory (accessed January 4, 2019).

EDF (Environmental Defense Fund) (2019). *Why New Sensor Technology is Critical for Tackling Air Pollution*. Available online at: www.edf.org/airqualitymaps

Emam, K. E., Rodgers, S., and Malin, B. (2015). Anonymising and sharing individual patient data. *BMJ* 350:h1139. doi: 10.1136/bmj.h1139

Fontes, T., Silva, L. M., Silva, M. P., Barros, N., and Carvalho, A. C. (2014). Can artificial neural networks be used to predict the origin of ozone episodes? *Sci. Total Environ.* 488–489, 197–207. doi: 10.1016/j.scitotenv.2014.04.077

Goh, G. B., Hodas, N. O., and Vishnu, A. (2017). Deep learning for computational chemistry. *J. Comput. Chem.* 38, 1291–1307. doi: 10.1002/jcc.24764

Google (2019). *Environmental Insights Explorer*. Insights Sustainability.Google/

Hartung, T. (2009). Toxicology for the twenty-first century. *Nature* 460, 208–212. doi: 10.1038/460208a

ICCVAM (Interagency Coordinating Committee on the Validation of Alternative Methods) (2018). *A Strategic Roadmap for Establishing New Approaches to Evaluate the Safety of Chemicals and Medical Products in the United States*. National Institute of Environmental Health Sciences. Available online at: https://ntp.niehs.nih.gov/pubhealth/evalatm/natl-strategy/index.html (accessed January 4, 2019).

Ienca, M., Ferretti, A., Hurst, S., Puhan, M., Lovis, C., and Vayena, E. (2018). Considerations for ethics review of big data health research: a scoping review. *PLoS ONE* 13:e0204937. doi: 10.1371/journal.pone.0204937

Jaklevic, M. C. (2017). *MD Anderson Cancer Center's IBM Watson Project Fails, and So Did the Journalism Related to it*. HealthNewsReview. Available online at: www.healthnewsreview.org/2017/02/md-anderson-cancer-centers-ibm-watson-project-fails-journalism-related/

Kleinstreuer, N. C., Ceger, P. C., Allen, D. G., Strickland, J., Chang, X., and Hamm, J. T. (2016). A curated database of rodent uterotrophic bioactivity. *Environ. Health Perspect.* 124, 556–562. doi: 10.1289/ehp.1510183

Kleinstreuer, N. C., Karmaus, A. L., Mansouri, K., Allen, D. G., Fitzpatrick, J. M., and Patlewicz, G. (2018). Predictive models for acute oral systemic toxicity: a workshop to bridge the gap from research to regulation. *Comput. Toxicol.* 8, 21–24. doi: 10.1016/j.comtox.2018.08.002

Kleinstreuer, N. C., Yang, J., Berg, E. L., Knudsen, T. B., Richard, A. M., Martin, M. T., et al. (2014). Phenotypic screening of the toxcast chemical library to classify toxic and therapeutic mechanisms. *Nat. Biotechnol.* 32, 583–591. doi: 10.1038/nbt.2914

Li, L. G., Yin, X., and Zhang, T. (2018). Tracking antibiotic resistance gene pollution from different sources using machine-learning classification. *Microbiome* 6:93. doi: 10.1186/s40168-018-0480-x

Liss, A., Koch, M., and Naumova, E. N. (2014). Redefining climate regions in the United States of America using satellite remote sensing and machine learning for public health applications. *Geospatial Health* 8, S647–659. doi: 10.4081/gh.2014.294

Lohr, S. (2018). *M.I.T. Plans College for Artificial Intelligence, Backed by $1 Billion*. The New York Times. Available online at: https://nyti.ms/2AcLqi9 (accessed January 4, 2019).

Luechtefeld, T., Rowlands, C., and Hartung, T. (2018). Big-data and machine learning to revamp computational toxicology and its use in risk assessment. *Toxicol. Res.* 7, 732–744. doi: 10.1039/C8TX00051D

Mahony, C., Currie, R., Daston, G., Kleinstreuer, N., and van de Water, B. (2018). Highlight report: 'big data in the 3R's: outlook and recommendations', a roundtable summary. *Arch. Toxicol.* 92, 1015–1020. doi: 10.1007/s00204-017-2145-0

Mansouri, K., Grulke, C. M., Richard, A. M., Judson, R. S., and Williams, A. J. (2016). An automated curation procedure for addressing chemical errors and inconsistencies in public datasets used in QSAR modelling. *SAR QSAR Environ. Res.* 27, 939–965. doi: 10.1080/1062936X.2016.1253611

Meyer, R. (2018). *Google's New Tool to Fight Climate Change*. The Atlantic. Available online at: www.theatlantic.com/technology/archive/2018/09/google-climate-change-greenhouse-gas-emissions/571144/

Microsoft (2018). *AI for Earth Current Grant Recipients*. Available online at: https://www.microsoft.com/en-us/ai/ai-for-earth-grant

Milinovich, G. J., Williams, G. M., Clements, A. C. A., and Hu, W. (2014). Internet-based surveillance systems for monitoring emerging infectious diseases. *Lancet Infect. Dis.* 14, 160–168. doi: 10.1016/S1473-3099(13)70244-5

Mittelstadt, B. D., and Floridi, L. (2016). The ethics of big data: current and foreseeable issues in biomedical contexts. *Sci. Eng. Ethics* 22:303. doi: 10.1007/978-3-319-33525-4

NAS (2019). *Leveraging Artificial Intelligence and Machine Learning to Advance Environmental Health Research and Decisions." Leveraging Artificial Intelligence and Machine Learning to Advance Environmental Health Research and* Decisions.

Noy, N., (2018). *Making it Easier to Discover Data Sets*. Google. Available online at: www.blog.google/products/search/making-it-easier-discover-datasets/

Park, S. K., Zhao, Z., and Mukherjee, B. (2017). Construction of environmental risk score beyond standard linear models using machine learning methods: application to metal mixtures, oxidative stress and cardiovascular disease in NHANES. *Environ. Health.* 16:102. doi: 10.1186/s12940-017-0310-9

Ring, C. L., Arnot, J. A., Bennett, D. H., Egeghy, P. P., Fantke, P., Huang, L., et al. (2019). Consensus modeling of median chemical intake for the U.S. population based on predictions of exposure pathways. *Environ. Sci. Technol.* 53, 719–732. doi: 10.1021/acs.est.8b04056

Rubens, M., Ramamoorthy, V., Saxena, A., and Shehadeh, N. (2014). Public health in the twenty-first century: the role of advanced technologies. *Front. Public Health* 2:224. doi: 10.3389/fpubh.2014.00224

Saili, K. S., Franzosa, J. A., Baker, N. C., Ellis-Hutchings, R. G., Settivari, R. S., Carney, E. W., et al. (2019). Systems modeling of developmental vascular toxicity. *Curr. Opin. Toxicol.* 15:55–63. doi: 10.1016/j.cotox.2019.04.004

Thomas, R. S., Bahadori, T., Buckley, T. J., Cowden, J., Deisenroth, C., Dionisio, K. L., et al. (2019). The next generation blueprint of computational toxicology

at the U.S. environmental protection agency. *Toxicol. Sci.* 169, 317–332. doi: 10.1093/toxsci/kfz058

Tice, R. R., Austin, C. P., Kavlock, R. J., and Bucher, J. R. (2013). Improving the human hazard characterization of chemicals: a Tox21 update. *Environ. Health Perspect.* 121, 756–765. doi: 10.1289/ehp.1205784

U.S. EPA (United States Environmental Protection Agency) (2016). *About the TSCA Chemical Substance Inventory*. Available online at: https://www.epa.gov/tsca-inventory/about-tsca-chemical-substance-inventory (accessed January 4, 2019).

Vincent, J. (2018). *Google Launches New Search Engine to Help Scientists Find the Datasets They Need*. The Verge. Available online at: www.theverge.com/2018/9/5/17822562/google-dataset-search-service-scholar-scientific-journal-open-data-access

Vopham, T., Hart, J. E., Laden, F., and Chiang, Y. Y. (2018). Emerging trends in Geospatial Artificial Intelligence (GeoAI): potential applications for environmental epidemiology. *Environ. Health* 17:40. doi: 10.1186/s12940-018-0386-x

Weichenthal, S., Hatzopoulou, M., and Brauer, M. (2019). A picture tells a thousand…exposures: opportunities and challenges of deep learning image analyses in exposure science and environmental epidemiology. *Environ. Int.* 122, 3–10. doi: 10.1016/j.envint.2018.11.042

Wilkinson, M. D., Dumontier, M., Aalbersberg, J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Nat. Sci. Data* 3:160018. doi: 10.1038/sdata.2016.18

Williams, A. J., Grulke, C. M., Edwards, J., McEachran, A. D., Mansouri, K., Baker, N. C., et al. (2017). The CompTox chemistry dashboard: a community data resource for environmental chemistry. *J. Cheminform.* 9:61. doi: 10.1186/s13321-017-0247-6

Xiong, Y., Ba, X., Hou, A., Zhang, K., Chen, L., and Li, T. (2018). Automatic detection of mycobacterium tuberculosis using artificial intelligence. *J. Thoracic Dis.* 10, 1936–1940. doi: 10.21037/jtd.2018.01.91

Conflict of Interest: LJ is employed by Microsoft Corporation.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer TL declared a past co-authorship with one of the authors NK to the handling editor.