Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Environ Res. Author manuscript; available in PMC 2022 June 01.

Published in final edited form as:

Environ Res. 2021 June ; 197: 111019. doi:10.1016/j.envres.2021.111019.

Interdisciplinary Data Science to Advance Environmental Health Research and Improve Birth Outcomes

Jeanette A. Stingone[a,*], **Sofia Triantafillou**[b], **Alexandra Larsen**[c,1], **Jay P. Kitt**[d], **Gary M. Shaw**[e], **Judit Marsillach**[f]

[a]Department of Epidemiology, Columbia University's Mailman School of Public Health, 722 West 168th St, Room 1608, New York, NY 10032, USA

[b]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA

[c]Department of Biostatistics and Bioinformatics, Duke University, Durham, NC, USA

[d]Departments of Chemistry and Biomedical Informatics, University of Utah, Salt Lake City, UT, USA

[e]Department of Pediatrics, Stanford University School of Medicine, Stanford, CA, USA

[f]Department of Environmental and Occupational Health Sciences, University of Washington, Seattle, WA, USA

Abstract

Rates of preterm birth and low birthweight continue to rise in the United States and pose a significant public health problem. Although a variety of environmental exposures are known to contribute to these and other adverse birth outcomes, there has been a limited success in developing policies to prevent these outcomes. A better characterization of the complexities between multiple exposures and their biological responses can provide the evidence needed to inform public health policy and strengthen preventative population-level interventions. In order to achieve this, we encourage the establishment of an interdisciplinary data science framework that integrates epidemiology, toxicology and bioinformatics with biomarker-based research to better define how population-level exposures contribute to these adverse birth outcomes. The proposed interdisciplinary research framework would 1) facilitate data-driven analyses using existing data

[*]Corresponding author at: Department of Epidemiology, Columbia University's Mailman School of Public Health, 722 West 168th St, Room 1608, New York, NY 10032, USA. j.stingone@columbia.edu (J. A. Stingone).

[1]Present address: US Environmental Protection Agency, Office of Research and Development, Center for Public Health and Environmental Assessment, Research Triangle Park, NC

CRediT Author statement

Jeanette A Stingone: Conceptualization, Data Curation, Writing-Original Draft, Sofia Triantafillou: Conceptualization, Formal Analysis, Writing-Original Draft, Jay P. Kitt: Conceptualization, Writing-Review & Editing, Alexandra Larsen: Conceptualization, Writing-Review & Editing, Gary M Shaw: Conceptualization, Writing-Review & Editing, Judit Marsillach: Conceptualization, Writing-Original Draft, Project Administration

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

from health registries and environmental monitoring programs; 2) develop novel algorithms with the ability to predict which exposures are driving, in this case, adverse birth outcomes in the context of simultaneous exposures; and 3) refine biomarker-based research, ultimately leading to new policies and interventions to reduce the incidence of adverse birth outcomes.

Keywords

Preterm Birth; Environmental Mixtures; Multiple Exposures; Public Health Data Science

1. Introduction

In the United States, rates of preterm birth (deliveries before 37 weeks gestation) have risen for the past four years, to 10.02% in 2018[1]; and indeed the frequency of preterm birth has increased in the US for four decades.[2] Incidence of low birth weight has also increased 3% since 2014.[1] There are well-documented disparities in these rates, with Black women having almost double the risk of having preterm births than White women.[1,3] Preterm birth and low birthweight remain a substantial public health problem as they have a significant impact on an infant's survival, development and long-term health.[4–6] Decades of research have provided evidence that environmental exposures contribute to both preterm birth and low birthweight.[7–14] These exposures include air pollution, pesticides, extreme temperatures and aspects of the built environment.

Despite the breadth of the exposures investigated for associations with preterm birth and low birthweight, there has been limited translation to public health policies and interventions that reduce the frequencies of these outcomes, particularly in vulnerable communities. Evidence for prevention would be strengthened by simultaneously investigating the manifold exposures experienced by women during pregnancy and then employing analytic techniques that have the ablity to disentangle effects and identify targeted points for intervention. Characterizing the complex exposure-response relationships that contribute to preterm birth and low birthweight requires both epidemiologic and toxicologic studies that include the examination of multiple exposures simultaneously, account for temporal variability in exposure throughout pregnancy and explore the potential for interaction between environmental exposures, socioeconomic contexts and genetics. Current study designs and analytic methods often fail to capture this complexity.[15] In addition, there has been limited research that combines epidemiology of ambient environmental pollutants with biomarker-based research of biological responses to fully characterize the pathways that define how population-level exposures can contribute to these adverse birth outcomes.[16,17] This knowledge is critical for informing successful public health policy and population-level interventions aimed at prevention. These approaches require collaborations between computational scientists skilled at the analysis of large, complex data with biomarker-based laboratory researchers with the knowledge and techniques to determine the mechanistic paths that connect external exposures to adverse birth outcomes. An interdisciplinary data science framework enables these collaborations and allows for the holistic analysis of complex environmental data to extract primary risk drivers and to guide biological,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

mechanistic and biomarker-based research, enabling reduction and prevention of preterm birth, low birthweight and other adverse birth outcomes.

Much has been written about the promise of data science in advancing the understanding of public health.[18] There have been numerous commentaries arguing for the integration of data science methods into environmental health research.[19,20] Many of these focus on examining the complexity of the exposome, a paradigm describing the totality of endogenous and exogenous exposures that occur throughout a lifetime.[21–26] We are beginning to see examples of research that have successfully adapted data science methods to address the challenges of characterizing the exposome and documenting its effects on adverse birth outcomes.[27,28] While this is a clear sign of progress, applying more complex analytics to ever larger datasets can lead to more questions than answers. This is because data-driven analyses, including those that have marked this initial phase of environmental health data science, are limited by the assumptions required of purely statistical approaches and often fail to include the existing knowledge generated within other fields such as known biophysical pathways underlying disease pathology or toxicological knowledge of environmental pollutants.[27] The causal inference needed to inform public health policy and interventions requires an interdisciplinary data science, one that enables the integration of knowledge across research fields including epidemiologic and toxicologic knowledge with environmental health data on large populations.

The objective of this commentary is to propose an interdisciplinary research framework that integrates data science across the multiple disciplines within environmental health. The proposed framework utilizes multiple lines of scientific inquiry, such as epidemiology, data-driven analytics and biomarker-based investigations, in order to connect the external exposures, modifiable through policy and interventions, to the internal measures of biologic response that may serve as more proximal causes of preterm birth and low-birthweight. In addition, we provide recommendations aimed at the environmental health community to support the interdisciplinary collaborations needed to advance this work.

2. Rationale for an Interdisciplinary Data Science Framework

There are two primary reasons to support an interdisciplinary data science framework for environmental health: the first is to address the limitations of evidence produced by existing approaches and the second is to advance the field of data-driven analytic approaches to navigate the space of hypotheses more efficiently.

As discussed above, epidemiologic studies of environmental contributors to adverse birth outcomes have traditionally focused on a single exposure. This fails to account for the complexities of exposure, pregnancy and fetal development.[29] These complexities include epidemiologic analysis of simultaneous exposure to multiple chemical and non-chemical stressors, temporal variability in both exposure and vulnerability throughout pregnancy and the potential for interaction between environmental exposures, socioeconomic contexts and genetics. Recent studies have attempted to tackle many of these complexities. Previous work in California took a comprehensive approach to assess both spatial and temporal variation in pesticides to explore associations with pregnancy outcomes.[30] A number of studies have

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

applied complex analytic techniques, including distributed lag models and others, to identify critical windows of exposure during pregnancy.[30–33] Smaller cohort studies have begun to utilize biological specimens, such as maternal urine collected during pregnancy and cord blood collected at delivery, to assess more proximal exposures to the fetus.[34–36] While making considerable advances in our knowledge, these studies have tackled the individual limitations of previous studies. In many instances, they remain subject to other challenges, that could be addressed by a more interdisciplinary data-science approach. For example, a study to identify critical exposure windows of a single exposure using statistical approaches remains prone to the confounding influences of unmeasured co-exposures and often only considers biological knowledge about fetal development when interpreting results, not in the analysis itself. Thus, each individual study may be limited in its ability to generate causal inference at the level needed for policy and intervention planning.

Additionally, data science methods have traditionally focused on building robust predictive statistical models. State-of-the-art feature selection algorithms have the ability to select, among large sets of covariates, a set of variables that are maximally predictive for the target variable, and discard the rest as non-significant contributors. This omission is particularly important in environmental health data sets, where exposures can often co-occur or have a spurious association with the outcome due to their correlation with other causal exposures, and therefore carry similar information for the outcome.[37] For example, in Figure 1 we illustrate the challenge of outcome-equivalent co-exposure sets and how to detect them on a county-level data set on preterm birth from California. In this example, the chemicals chloroform and ethylene oxide were found to be interchangeable in a linear model that included 8 other air toxics as covariates. Figure 1 shows the corresponding residuals for preterm birth rates, using chloroform and ethylene oxide. The residuals are almost identical, suggesting that the variables are interchangeable in the model. In regression analysis methods used for feature selection, such as LASSO (least absolute shrinkage and selection operator), chloroform would be discarded as a non-significant contributor, without any consideration of toxicologic knowledge about which exposure may be more relevant to preterm birth. This example illustrates a common issue in environmental health research: the presence of variables that carry similar outcome information can result in some of the variables being overlooked as potential risk factors for the outcome of interest.

Solutions to these limitations require not just expertise in complex analytics, but the ability to integrate data across a variety of fields within the broader domain of environmental health research including epidemiology, toxicology and the omics technologies (e.g. genomics, proteomics, metabolomics, etc). Approaches that incorporate previous knowledge on toxicity of exposures while adapting study designs and analytic methods to address multiple limitations simultaneously can advance our knowledge of environmental contributors to adverse birth outcomes and accelerate efforts to translate research into prevention.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

3. An interdisciplinary data science framework to address challenges within environmental health research

We propose an interdisciplinary data science framework that integrates epidemiologic data with toxicological knowledge of exposures when applying complex analytic methods and then using generated results to inform targeted biomarker-based studies of biological mechanisms.

The proposed framework includes multiple techniques from data science that span the disciplines of epidemiology, bioinformatics and laboratory science: i) the re-use of existing public health data, environmental monitoring and biospecimens to efficiently access information on large, representative populations; ii) the use of bioinformatics toxicological knowledge within quantitative and statistical approaches; and iii) the translation of results to inform biomarker-based investigations of biological mechanisms. Our proposed framework is not meant to serve as a "how-to" for conducting complex analytics related to environmental contributors to adverse birth outcomes. Rather, we aim to encourage a more holistic interpretation of data science, marked by the integration of interdisciplinary approaches and tailored to the research question of interest. Figure 2 illustrates an example of how this framework could be applied to investigations of adverse birth outcomes, such as preterm birth, using publicly-available public health data. We will refer to this example throughout the rest of this commentary. However, we encourage readers to consider the study populations, exposure data sources, analytic techniques and biomarker-based investigations that best fit the scientific complexities of their research question and context.

3.1 Conduct efficient data-driven analyses using big public health data

As stated earlier, much work has been done to incorporate data science into the fields of medicine and healthcare to better utilize electronic health record data to promote precision medicine.[38,39] A parallel challenge is to utilize the vast amounts of data in health registries, environmental monitoring programs and administrative records to catalyze improvements in public health.[40] The need is echoed in the NIEHS Strategic Plan, which specifically calls for work that effectively uses data to generate and translate knowledge into actionable policies to improve public health.[41] Environmental epidemiology has a long history of linking administrative birth records with place-based exposure metrics to investigate environmental contributors to a variety of health outcomes including preterm birth and low birthweight. Data quality, however, remains a limitation as birth records often lack high-quality information on maternal conditions during pregnancy, pre-pregnancy health and other important factors. The use of data integration, to combine features across multiple administrative systems, has the potential to yield a more accurate and holistic picture of women and infants within populations.[42] Research from Northern European countries with integrated health systems routinely use public health data to investigate novel questions related to perinatal environmental health that are not possible relying on birth records alone.[43–45] While the US does not have automatically integrated systems, many municipalities now routinely link birth and delivery hospitalization data to provide higher quality information, enabling more comprehensive investigations into birth outcomes.[46]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

The breadth and depth of data resources for environmental exposures that can be linked to these richer health data records will vary based on availability and the spatial and temporal context of an individual study. However, the ability to capture the complexities of exposure during pregnancy will depend upon using a broad definition of environment, and including resources that capture exposures across all domains of the external exposome.[47] This could include both traditional resources, such as air monitoring databases and pesticide registries maintained by municipal governments for regulatory purposes, as well as data resources built from newer technologies such as crowd-sourced traffic data and citizen-science initiatives.

As shown in Figure 2, the selection and integration of multiple data sources is the first step in our proposed framework. In our example examining preterm birth using publicly available data, the use of birth registries linked with hospital discharge data would provide higher quality data on maternal conditions before and during pregnancy, as well as allow directed investigations of the different phenotypes of preterm birth. One would use the environmental contributors identified in previous research as a starting point, but expanding to include all domains of the external exposome including measures related to built environment. For example, previous research linking historical red-lining[48,49] and housing insecurity[50] to preterm birth would support the inclusion of data resources containing metrics of residential segregation, evictions and gentrification. The resulting big public health data facilitates use of data-driven discovery approaches to characterize complex exposure patterns experienced in pregnancy that span the domains of the exposome.

Another component that can be linked to birth records, and therefore to this big public health data are newborn dried blood spots (NBDS).[51] In the US and many other countries, heel-stick blood samples are collected from newborns at birth to determine inborn errors of metabolism that could be detrimental for the infant's postnatal development if not treated immediately. This screening utilizes only a few of the NDBS collected using a filter paper Guthrie card. The remaining NDBS are typically stored by a state's Newborn Screening program for a particular number of years and under safe storage conditions, defined by state policies, until they are discarded.[52] A number of these programs make these stored residual NBDS available, with appropriate human subjects protection, for research purposes. NDBS offer a unique opportunity to assess external exposures from samples representative of *in utero* conditions and investigate their effects on adverse birth outcomes, childhood developmental disorders and susceptibility to certain diseases during aging.[53–60] Given the limited availability of these valuable stored residual NDBS, secondary research proposals using residual NDBS should prioritize scientific rigour to ensure the outcomes will have a significant impact in improving human health. This imperative supports the need for analytic methods that can identify the exposures most likely to be causal contributors to adverse birth outcomes.

Linkages between administrative data, birth registries and NDBS repositories are not the only source of big public health data that would fit within our proposed interdisciplinary data science framework. There are a growing number of consortia and "big science" initiatives that seek to pool and harmonize existing data from individual scientific studies, as well as implement shared protocols moving forward. Examples include the HELIX project

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

in Europe[28] and the ECHO program in the United States[61]. These initiatives are actively implementing analytic pipelines to investigate the pregnancy exposome and also provide a complementary resource to replicate findings observed in studies that use publicly available data.

3.2. Disentangle correlated exposures

Data science often focuses on prediction: popular machine learning methods mine the data for the most informative features and learn a model that predicts a target outcome. Such models can exploit complex correlation patterns in the data, and have greatly improved the precision of target prediction, compared to traditional statistical models. However, this improved capacity does not necessarily translate to increased knowledge. For example, in epidemiology, there is great interest in understanding the treatment-outcome mechanism, especially modifiable factors that can influence outcomes, rather than simply discovering the features which allow statistical predictions of those outcomes. To tackle this problem, algorithms that are customized for domain-specific problems, and which incorporate expert knowledge in guiding feature selection while mining information from the data are needed.

Again, if we consider the purely data-driven approach, the reason becomes clear. When investigating adverse health outcomes due to environmental factors, simultaneous exposures to multiple pollutants are common; while some of these will lead to poor outcomes with a clear underlying pathophysiology, others are associated only through co-occurrence with the biophysically important compound. Even state-of-the-art statistical algorithms, as part of a statistical software package with no modification, may return either the important or unimportant variable as the most significant predictor, and discard the other. This is because highly-correlated co-exposures, no matter the basis of that correlation (e.g. pollutant source, location, or time period), carry the same statistical information and are thus mathematically indistinguishable to the algorithm. Thus, the information underlying the true source of the adverse outcome, in pathophysiological terms, may be carried in any one of the discarded variables. In the simplest terms, most current algorithms fail to account for the mechanistic nature of a variable and lack a means for discriminating them.

The challenges of analyzing such highly correlated mixtures of exposures are known, and addressing them is an active area of research.[62,63] Many of these approaches include some form of data reduction before statistical analysis, with the goal of building the best predictive model for the outcome of interest. Although epidemiologists often aim to identify potential drivers of the outcome, it is known that such data-reduction techniques hinder the biological interpretation of the results.[64] Some statistical methods are specifically designed for mixtures, but also aim to estimate interaction effects between agents[65] or to asses the overall effect of combined exposures.[66] When it comes to variable selection, state-of-the art methods are known to often be sensitive to even small pertubations of the data when the features are highly correlated.[67] We propose combining such methods with accumulated prior knowledge stored in biochemical databases to improve the data-driven generation of scientific hypotheses.

Our interdisciplinary framework proposes the development of novel algorithms specifically designed for discriminating mechanistic variables based on *a priori* knowledge from broader

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

scientific study. As illustrated in Figure 2, a possible implementation would be the development of an algorithm that returns sets of statistically equivalent features (as opposed to discarding one in favour of another) and that carries out a ranking analysis based on the biochemical and toxicological properties for each feature allowing stratification by likelihood of biological impact. Some methods for identifying multiple molecular "signatures" for a target outcome have been developed in the area of molecular biology, but focus on predictive performance and do not incorporate domain knowledge. For environmental exposures, for example, indications for the likelihood of toxicity can be assessed using a variety of chemical, biological, and toxicological databases.[68,69] As an example, consider again the data presented in Fig. 1. A state-of-the-art feature selection algorithm would return either chloroform and ethylene oxide as risk factors for preterm birth, based on noise in the measurement or even chance. Prior knowledge on the toxicity and action of these compounds, mined from toxicology databases or from the literature could rank them in terms of their likelihood for participating in the mechanism of preterm birth. This type of "ranking" is manually done by researchers to formulate plausible testable hypotheses. Using data science to automatically include this infomration can help the methods navigate the space of hypotheses more efficiently.

3.3 Inform biomarker-based research at the population-level

Biomarker-based research complements large-scale epidemiologic investigations by allowing targeted and a potentially more biologically proximal understanding of the influence of environmental factors. This can potentially lead to better prevention, diagnosis and treatment of adverse birth outcomes. It is practically infeasible, fiscally prohibitive and inefficient to simultaneously measure biomarkers of every single exposure from conception to death when attempting to study the exposome in relation to a particular adverse birth outcome. Application of interdisciplinary data science, as proposed within this commentary, is fundamental in addressing this challenge and will be essential in informing and guiding targeted biomarker-based research of the exposome. As illustrated in Figure 2, hypotheses generated from integrated analysis of epidemiologic research can inform the selection of exposures to be investigated in mechanistic studies using available biospecimens.

Advances in laboratory research are allowing the simultaneous characterization and accurate quantification of a large number of individual biological components such as proteins, metabolites, lipids, etc. These omics technologies offer an ideal strategy at characterizing the internal components of the exposome. However, without information about external and lifestyle components, the omics approaches will yield limited knowledge of the exposome and population-based prevention strategies.[22] Data science can provide an ideal platform to connect all three components of the exposome and refine measurements that together can lead to improved prevention of adverse birth outcomes and disease. For example, in a cohort with well-characterized external exposures, identification of the exposures that are driving a specific biological effect/disease of interest by a data science approach can help biomarker-based research stratify subjects for those exposures of interest and apply omics technologies. This type of study has the potential to provide mechanistic input linking certain exposures and their effects on adverse birth outcomes or diseases, providing evidence in support of prevention strategies aimed at specific external exposures..

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Direct characterization of the *in utero* exposome is unattainable, and to date, it has been mostly estimated from exposures measured in the mother's blood. However, recent improvements in omics technologies are allowing the use of archived residual NDBS for assessing the etiology of diseases and certain environmental exposures that occur during the fetal stage.[55,57–59,70–74] In our example study of environmental contributors to preterm birth (Figure 2), we propose to integrate the data science-provided hypotheses on the specific environmental features most relevant to preterm birth s with the use of proteomics to analyze archived residual NDBS and ascertain the biological signatures of *in utero* exposures. By allowing for much more targeted omics analysis, the interdisciplinary framework described herein can limit the required breadth of patient samples required for laboratory processing and potentially allow more rapid discovery of biomarkers of exposure, furthering the omics field and advancing NDBS-based omics analyses.

4. Recommendations

We present three recommendations to support and advance the use of the proposed interdisciplinary framework. First, we encourage researchers across the multiple disciplines within environmental health to establish interdisciplinary research teams. The challenges facing the field of environmental health are complex and require a broad set of skills to maximize the knowledge generated by the diverse types of environmental health data. The expansion of research teams is not limited to the inclusion of data scientists with expertise in analytic approaches, but also chemists knowledgeable in the use of bioinformatics resources, clinicians and biologists with training in physiological pathways and public health professionals to advise on optimal translation strategies. The formation and function of these interdisciplinary research teams will need to be actively supported by funding agencies, professional organizations and academic centers, through grants and programs that support collaboration and consortia development. An excellent example is the Data Science Innovation Labs currently supported by NIH that led to the creation of the authors' interdisciplinary group.[75] Initially supported by the NIH Big Data to Knowledge Initiative (BD2K), the Data Science Innovation Labs are an annual intensive workshop, created specifically to foster the development of interdisciplinary teams focused on specific biomedical challenges that could benefit from increased use of data science techniques. Guided by professional facilitators and experienced mentors, investigators from diverse disciplines form teams to solve a data science challenge related to a broad biomedical theme.

Second, to encourage efficient and timely interdisciplinary research designs, we recommend greater use of existing biorepositories for research into adverse birth outcomes. In our proposed framework, we discussed the use of archived residual NDBS as a potential source for biomarker-based research. New technologies have enhanced our ability to re-use these public health resources for multiple omics research. Many states have demonstrated the ability to share these resources with researchers, while still protecting individuals' privacy and ensuring data security. A greater commitment to sharing data across birth cohorts could also provide access to larger resources of biospecimens linked with extensive epidemiologic data. The HELIX study in Europe provides a blueprint for constructing novel study designs aimed at integrating interdisciplinary data science within existing birth cohort studies.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

As a third recommendation, we encourage the cross-training of both environmental health and data science investigators to improve communication and integration of skills within interdisciplinary research teams. NIEHS has acknowledged this need through an approved concept to advance workforce development for environmental health data science. Their proposed concept included initiatives to support environmental health training of data scientists and educational resources for skill development to enhance data-intensive environmental health research.

7. Conclusion

Incorporating interdisciplinary data science techniques and greater use of big public health data into environmental health research can address limitations of prior research. We call for the creation and support of interdisciplinary research teams to implement our proposed framework, connecting population-level environmental exposures to biomarker-based targeted interventions of biological mechanisms. By leveraging expertise across the domains of environmental health research and utilizing diverse techniques of data science, we will enable integrative research to generate translatable knowledge on environmental contributors to the etiologies and prevention of adverse birth outcomes.

Acknowledgements

The authors acknowledge Dr. Lynda R. Hardy for her contributions to early discussions of this paper, and the 2019 Data Science Innovation Lab led by Dr. John Van Horn.

Funding sources

JAS' work was supported in part by NIH/NIEHS (ES027022). JPK's work was supported in part by the Utah Center for Clinical and Translational Science funded by NCATS award (1ULTR002538) and the NIH/NLM Training grant (LM007124). AL's work was supported in part by NIH/NCI (CA220693). JM's work was supported in part by NIH/NIEHS (ES04696).

Abbreviations

NDBS	newborn dried blood spots

References

1. Hamilton B, Martin J, Osterman M, Rossen L. Births: Provisional Data for 2018. National Center for Health Statistics; 2019.
2. Goldenberg RL, Culhane JF, Iams JD, Romero R. Epidemiology and causes of preterm birth. Lancet Lond Engl. 2008;371(9606):75–84. doi:10.1016/S0140-6736(08)60074-4
3. Burris HH, Hacker MR Birth outcome racial disparities: A result of intersecting social and environmental factors. Semin Perinatol. 2017;41(6):360–366. doi:10.1053/j.semperi.2017.07.002 [PubMed: 28818300]
4. Twilhaar ES, Wade RM, de Kieviet JF, van Goudoever JB, van Elburg RM, Oosterlaan J. Cognitive Outcomes of Children Born Extremely or Very Preterm Since the 1990s and Associated Risk Factors: A Meta-analysis and Meta-regression. JAMA Pediatr. 2018;172(4):361–367. doi:10.1001/jamapediatrics.2017.5323 [PubMed: 29459939]
5. Frey HA, Klebanoff MA. The epidemiology, etiology, and costs of preterm birth. Semin Fetal Neonatal Med. 2016;21(2):68–73. doi:10.1016/j.siny.2015.12.011 [PubMed: 26794420]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

6. Petrou S, Sach T, Davidson L. The long-term costs of preterm birth and low birth weight: results of a systematic review. Child Care Health Dev. 2001;27(2):97–115. doi:10.1046/j.1365-2214.2001.00203.x [PubMed: 11251610]
7. Nieuwenhuijsen MJ, Dadvand P, Grellier J, Martinez D, Vrijheid M. Environmental risk factors of pregnancy outcomes: a summary of recent meta-analyses of epidemiological studies. Environ Health Glob Access Sci Source. 2013;12:6. doi:10.1186/1476-069X-12-6
8. Stieb DM, Chen L, Eshoul M, Judek S. Ambient air pollution, birth weight and preterm birth: a systematic review and meta-analysis. Environ Res. 2012;117:100–111. doi:10.1016/j.envres.2012.05.007 [PubMed: 22726801]
9. Kloog I Air pollution, ambient temperature, green space and preterm birth. Curr Opin Pediatr. 2019;31(2):237–243. doi:10.1097/MOP.0000000000000736 [PubMed: 30640892]
10. Klepac P, Locatelli I, Korošec S, Künzli N, Kukec A. Ambient air pollution and pregnancy outcomes: A comprehensive review and identification of environmental public health challenges. Environ Res. 2018;167:144–159. doi:10.1016/j.envres.2018.07.008 [PubMed: 30014896]
11. Nieuwenhuijsen MJ, Ristovska G, Dadvand P. WHO Environmental Noise Guidelines for the European Region: A Systematic Review on Environmental Noise and Adverse Birth Outcomes. Int J Environ Res Public Health. 2017;14(10). doi:10.3390/ijerph14101252
12. Stillerman KP, Mattison DR, Giudice LC, Woodruff TJ. Environmental exposures and adverse pregnancy outcomes: a review of the science. Reprod Sci Thousand Oaks Calif. 2008;15(7):631–650. doi:10.1177/1933719108322436
13. Kuehn L, McCormick S. Heat Exposure and Maternal Health in the Face of Climate Change. Int J Environ Res Public Health. 2017;14(8). doi:10.3390/ijerph14080853
14. Shirangi A, Nieuwenhuijsen M, Vienneau D, Holman CDJ. Living near agricultural pesticide applications and the risk of adverse reproductive outcomes: a review of the literature. Paediatr Perinat Epidemiol. 2011;25(2):172–191. doi:10.1111/j.1365-3016.2010.01165.x [PubMed: 21281330]
15. Patel CJ. Analytic Complexity and Challenges in Identifying Mixtures of Exposures Associated with Phenotypes in the Exposome Era. Curr Epidemiol Rep. 2017;4(1):22–30. doi:10.1007/s40471-017-0100-5 [PubMed: 28251040]
16. Vadillo-Ortega F, Osornio-Vargas A, Buxton MA, et al. Air pollution, inflammation and preterm birth: a potential mechanistic link. Med Hypotheses. 2014;82(2):219–224. doi:10.1016/j.mehy.2013.11.042 [PubMed: 24382337]
17. Zhang Y, Wang J, Gong X, et al.. Ambient PM2.5 exposures and systemic biomarkers of lipid peroxidation and total antioxidant capacity in early pregnancy. Environ Pollut Barking Essex 1987. 2020;266(Pt 2):115301. doi:10.1016/j.envpol.2020.115301
18. Jones KH, Ford DV. Population data science: advancing the safe use of population data for public benefit. Epidemiol Health. 2018;40:e2018061. doi:10.4178/epih.e2018061
19. Choirat C, Braun D, Kioumourtzoglou M-A. Data Science in Environmental Health Research. Curr Epidemiol Rep. 2019;6(3):291–299. doi:10.1007/s40471-019-00205-5 [PubMed: 31723546]
20. Stieb DM, Boot CR, Turner MC. Promise and pitfalls in the application of big data to occupational and environmental health. BMC Public Health. 2017;17(1):372. doi:10.1186/s12889-017-4286-8 [PubMed: 28482822]
21. Wild CP. Complementing the genome with an “exposome”: the outstanding challenge of environmental exposure measurement in molecular epidemiology. Cancer Epidemiol Biomark Prev Publ Am Assoc Cancer Res Cosponsored Am Soc Prev Oncol. 2005;14(8):1847–1850. doi:10.1158/1055-9965.EPI-05-0456
22. Wild CP. The exposome: from concept to utility. Int J Epidemiol. 2012;41(1):24–32. doi:10.1093/ije/dyr236 [PubMed: 22296988]
23. Rappaport SM, Smith MT. Epidemiology. Environment and disease risks. Science. 2010;330(6003):460–461. doi:10.1126/science.1192603 [PubMed: 20966241]
24. Miller GW, Jones DP. The nature of nurture: refining the definition of the exposome. Toxicol Sci Off J Soc Toxicol. 2014;137(1):1–2. doi:10.1093/toxsci/kft251

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

25. Siroux V, Agier L, Slama R. The exposome concept: a challenge and a potential driver for environmental health research. Eur Respir Rev Off J Eur Respir Soc. 2016;25(140):124–129. doi:10.1183/16000617.0034-2016
26. Stingone JA, Buck Louis GM, Nakayama SF, et al. Toward Greater Implementation of the Exposome Research Paradigm within Environmental Epidemiology. Annu Rev Public Health. 2017;38:315–327. doi:10.1146/annurev-publhealth-082516-012750 [PubMed: 28125387]
27. Oskar S, Stingone JA. Machine Learning Within Studies of Early-Life Environmental Exposures and Child Health: Review of the Current Literature and Discussion of Next Steps. Curr Environ Health Rep. 2020;7(3):170–184. doi:10.1007/s40572-020-00282-5 [PubMed: 32578067]
28. Maitre L, de Bont J, Casas M, et al. Human Early Life Exposome (HELIX) study: a European population-based exposome cohort. BMJ Open. 2018;8(9):e021311. doi:10.1136/bmjopen-2017-021311
29. Robinson O, Vrijheid M. The Pregnancy Exposome. Curr Environ Health Rep. 2015;2(2):204–213. doi:10.1007/s40572-015-0043-2 [PubMed: 26231368]
30. Shaw GM, Yang W, Roberts EM, et al. Residential Agricultural Pesticide Exposures and Risks of Spontaneous Preterm Birth. Epidemiol Camb Mass. 2018;29(1):8–21. doi:10.1097/EDE.0000000000000757
31. Darrow LA, Klein M, Strickland MJ, Mulholland JA, Tolbert PE. Ambient air pollution and birth weight in full-term infants in Atlanta, 1994–2004. Environ Health Perspect. 2011;119(5):731–737. doi:10.1289/ehp.1002785 [PubMed: 21156397]
32. Liang Z, Lin Y, Ma Y, et al.. The association between ambient temperature and preterm birth in Shenzhen, China: a distributed lag non-linear time series analysis. Environ Health Glob Access Sci Source. 2016;15(1):84. doi:10.1186/s12940-016-0166-4
33. Sheridan P, Ilango S, Bruckner TA, Wang Q, Basu R, Benmarhnia T. Ambient Fine Particulate Matter and Preterm Birth in California: Identification of Critical Exposure Windows. Am J Epidemiol. 2019;188(9):1608–1615. doi:10.1093/aje/kwz120 [PubMed: 31107509]
34. Ashley-Martin J, Lavigne E, Arbuckle TE, et al. Air Pollution During Pregnancy and Cord Blood Immune System Biomarkers. J Occup Environ Med. 2016;58(10):979–986. doi:10.1097/JOM.0000000000000841 [PubMed: 27483336]
35. Minatoya M, Itoh S, Miyashita C, et al. Association of prenatal exposure to perfluoroalkyl substances with cord blood adipokines and birth size: The Hokkaido Study on environment and children's health. Environ Res. 2017;156:175–182. doi:10.1016/j.envres.2017.03.033 [PubMed: 28349882]
36. Haraux E, Tourneux P, Kouakam C, et al. Isolated hypospadias: The impact of prenatal exposure to pesticides, as determined by meconium analysis. Environ Int. 2018;119:20–25. doi:10.1016/j.envint.2018.06.002 [PubMed: 29929047]
37. ZIDEK JV WONG H, LE ND, BURNETT R%JE. Causality, measurement error and multicollinearity in epidemiology. 1996;7(4):441–451.
38. Lopes P, Silva LB, Oliveira JL. Challenges and Opportunities for Exploring Patient-Level Data. BioMed Res Int. 2015;2015:150435. doi:10.1155/2015/150435
39. Alyass A, Turcotte M, Meyre D. From big data analysis to personalized medicine for all: challenges and opportunities. BMC Med Genomics. 2015;8:33. doi:10.1186/s12920-015-0108-y [PubMed: 26112054]
40. Gamache R, Kharrazi H, Weiner JP. Public and Population Health Informatics: The Bridging of Big Data to Benefit Communities. Yearb Med Inform. 2018;27(1):199–206. doi:10.1055/s-0038-1667081 [PubMed: 30157524]
41. NIEHS. 2018–2023 Strategic Plan. Advancing Environmental Health Sciences. Improving Health. National Institutes of Health US Department of Health and Human Services
42. Jutte DP, Roos LL, Brownell MD. Administrative record linkage as a tool for public health research. Annu Rev Public Health. 2011;32:91–108. doi:10.1146/annurev-publhealth031210-100700 [PubMed: 21219160]
43. Li X, Sundquist J, Sundquist K. Parental occupation and risk of small-for-gestational-age births: a nationwide epidemiological study in Sweden. Hum Reprod Oxf Engl. 2010;25(4):1044–1050. doi:10.1093/humrep/deq004

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

44. Lunde A, Melve KK, Gjessing HK, Skjaerven R, Irgens LM Genetic and environmental influences on birth weight, birth length, head circumference, and gestational age by use of population-based parent-offspring data. Am J Epidemiol. 2007;165(7):734–741. doi:10.1093/aje/kwk107 [PubMed: 17311798]
45. Gong T, Dalman C, Wicks S, et al. Perinatal Exposure to Traffic-Related Air Pollution and Autism Spectrum Disorders. Environ Health Perspect. 2017;125(1):119–126. doi:10.1289/EHP118 [PubMed: 27494442]
46. Kim SY, Ahuja S, Stampfel C, Williamson D. Are Birth Certificate and Hospital Discharge Linkages Performed in 52 Jurisdictions in the United States? Matern Child Health J. 2015;19(12):2615–2620. doi:10.1007/s10995-015-1780-4 [PubMed: 26140836]
47. Turner MC, Nieuwenhuijsen M, Anderson K, et al. Assessing the Exposome with External Measures: Commentary on the State of the Science and Research Recommendations. Annu Rev Public Health. 2017;38:215–239. doi:10.1146/annurev-publhealth-082516-012802 [PubMed: 28384083]
48. Krieger N, Van Wye G, Huynh M, et al. Structural Racism, Historical Redlining, and Risk of Preterm Birth in New York City, 2013–2017. Am J Public Health. 2020;110(7):1046–1053. doi:10.2105/AJPH.2020.305656 [PubMed: 32437270]
49. Mendez DD, Hogan VK, Culhane JF. Institutional racism, neighborhood factors, stress, and preterm birth. Ethn Health. 2014;19(5):479–499. doi:10.1080/13557858.2013.846300 [PubMed: 24134165]
50. Leifheit KM, Schwartz GL, Pollack CE, et al. Severe Housing Insecurity during Pregnancy: Association with Adverse Birth and Infant Outcomes. Int J Environ Res Public Health. 2020;17(22). doi:10.3390/ijerph17228659
51. DePasquale JM, Freeman K, Amin MM, et al. Efficient Linking of Birth Certificate and Newborn Screening Databases for Laboratory Investigation of Congenital Cytomegalovirus Infection and Preterm Birth: Florida, 2008. Matern Child Health J. 2012;16(2):486–494. doi:10.1007/s10995-010-0740-2 [PubMed: 21203810]
52. Rothwell E, Johnson E, Riches N, Botkin JR. Secondary research uses of residual newborn screening dried bloodspots: a scoping review. Genet Med. 2019;21(7):1469–1475. doi:10.1038/s41436-018-0387-8 [PubMed: 30531811]
53. Funk WE, Waidyanatha S, Chaing SH, Rappaport SM. Hemoglobin adducts of benzene oxide in neonatal and adult dried blood spots. Cancer Epidemiol Biomark Prev Publ Am Assoc Cancer Res Cosponsored Am Soc Prev Oncol. 2008;17(8):1896–1901. doi:10.1158/1055-9965.EPI-08-0356
54. Funk WE, McGee JK, Olshan AF, Ghio AJ. Quantification of arsenic, lead, mercury and cadmium in newborn dried blood spots. Biomark Biochem Indic Expo Response Susceptibility Chem. 2013;18(2):174–177. doi:10.3109/1354750X.2012.750379
55. Funk WE. Use of Dried Blood Spots for Estimating Children?s Exposures to Heavy Metals in Epidemiological Research. J Environ Anal Toxicol. 2015;s7. doi:10.4172/2161-0525.S7-002
56. Asrani K, Shaw GM, Rine J, Marini NJ. DNA Methylome Profiling on the Infinium HumanMethylation450 Array from Limiting Quantities of Genomic DNA from a Single, Small Archived Bloodspot. Genet Test Mol Biomark. 2017;21(8):516–519. doi:10.1089/gtmb.2017.0019
57. Petrick L, Edmands W, Schiffman C, et al. An untargeted metabolomics method for archived newborn dried blood spots in epidemiologic studies. Metabolomics Off J Metabolomic Soc. 2017;13(3). doi:10.1007/s11306-016-1153-z
58. Yano Y, Grigoryan H, Schiffman C, et al. Untargeted adductomics of Cys34 modifications to human serum albumin in newborn dried blood spots. Anal Bioanal Chem. 2019;411(11):2351–2362. doi:10.1007/s00216-019-01675-8 [PubMed: 30783713]
59. Gonseth S, Shaw GM, Roy R, et al.. Epigenomic profiling of newborns with isolated orofacial clefts reveals widespread DNA methylation changes and implicates metastable epiallele regions in disease risk. Epigenetics. 2019;14(2):198–213. doi:10.1080/15592294.2019.1581591 [PubMed: 30870065]
60. Ma W-L, Gao C, Bell EM, et al. Analysis of polychlorinated biphenyls and organochlorine pesticides in archived dried blood spots and its application to track temporal trends of

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

environmental chemicals in newborns. Environ Res. 2014;133:204–210. doi:10.1016/j.envres.2014.05.029 [PubMed: 24968082]

61. Buckley JP, Barrett ES, Beamer PI, et al. Opportunities for evaluating chemical exposures and child health in the United States: the Environmental influences on Child Health Outcomes (ECHO) Program. J Expo Sci Environ Epidemiol. 2020;30(3):397–419. doi:10.1038/s41370-020-0211-9 [PubMed: 32066883]
62. Taylor KW, Joubert BR, Braun JM, et al. Statistical Approaches for Assessing Health Effects of Environmental Chemical Mixtures in Epidemiology: Lessons from an Innovative Workshop. Environ Health Perspect. 2016;124(12):A227–A229. doi:10.1289/EHP547 [PubMed: 27905274]
63. Gibson EA, Goldsmith J, Kioumourtzoglou M-A. Complex Mixtures, Complex Analyses: an Emphasis on Interpretable Results. Curr Environ Health Rep. 2019;6(2):53–61. doi:10.1007/s40572-019-00229-5 [PubMed: 31069725]
64. Patel CJ, Kerr J, Thomas DC, et al. Opportunities and Challenges for Environmental Exposure Assessment in Population-Based Studies. Cancer Epidemiol Biomark Prev Publ Am Assoc Cancer Res Cosponsored Am Soc Prev Oncol. 2017;26(9):1370–1380. doi:10.1158/1055-9965.EPI-17-0459
65. Bobb JF, Valeri L, Claus Henn B, et al. Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. Biostat Oxf Engl. 2015;16(3):493–508. doi:10.1093/biostatistics/kxu058
66. Carrico C, Gennings C, Wheeler DC, Factor-Litvak P. Characterization of Weighted Quantile Sum Regression for Highly Correlated Data in a Risk Analysis Setting. J Agric Biol Environ Stat. 2015;20(1):100–120. doi:10.1007/s13253-014-0180-3 [PubMed: 30505142]
67. Dougherty ER, Brun M. On the number of close-to-optimal feature sets. Cancer Inform. 2007;2:189–196. [PubMed: 19458767]
68. Judson R Public databases supporting computational toxicology. J Toxicol Environ Health B Crit Rev. 2010;13(2–4):218–231. doi:10.1080/10937404.2010.483937 [PubMed: 20574898]
69. Pawar G, Madden JC, Ebbrell D, Firman JW, Cronin MTD In Silico Toxicology Data Resources to Support Read-Across and (Q)SAR. Front Pharmacol. 2019;10:561. doi:10.3389/fphar.2019.00561 [PubMed: 31244651]
70. Yu M, Dolios G, Yong-Gonzalez V, et al. Untargeted metabolomics profiling and hemoglobin normalization for archived newborn dried blood spots from a refrigerated biorepository. J Pharm Biomed Anal. 2020;191:113574. doi:10.1016/j.jpba.2020.113574
71. Bell EM, Yeung EH, Ma W, et al. Concentrations of endocrine disrupting chemicals in newborn blood spots and infant outcomes in the upstate KIDS study. Environ Int. 2018;121(Pt 1):232–239. doi:10.1016/j.envint.2018.09.005 [PubMed: 30219610]
72. Yano Y, Schiffman C, Grigoryan H, et al. Untargeted adductomics of newborn dried blood spots identifies modifications to human serum albumin associated with childhood leukemia. Leuk Res. 2020;88:106268. doi:10.1016/j.leukres.2019.106268
73. Yeung EH, Bell EM, Sundaram R, et al. Examining Endocrine Disruptors Measured in Newborn Dried Blood Spots and Early Childhood Growth in a Prospective Cohort. Obes Silver Spring Md. 2019;27(1):145–151. doi:10.1002/oby.22332
74. Ernst M, Rogers S, Lausten-Thomsen U, et al. Gestational age-dependent development of the neonatal metabolome. Pediatr Res. Published online September 17, 2020. doi:10.1038/s41390-020-01149-z
75. Van Horn JD. Biomedical data science innovation labs: an intensive research project development program. Accessed October 15, 2020. https://projectreporter.nih.gov/project_info_description.cfm?aid=10049064&icde=52210839

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Highlights

- Rates of preterm birth and low birthweight continue to rise in the United States.
- Interdisciplinary data science can address challenges of previous research.
- Integrative analysis of complex environmental data can improve translation efforts.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

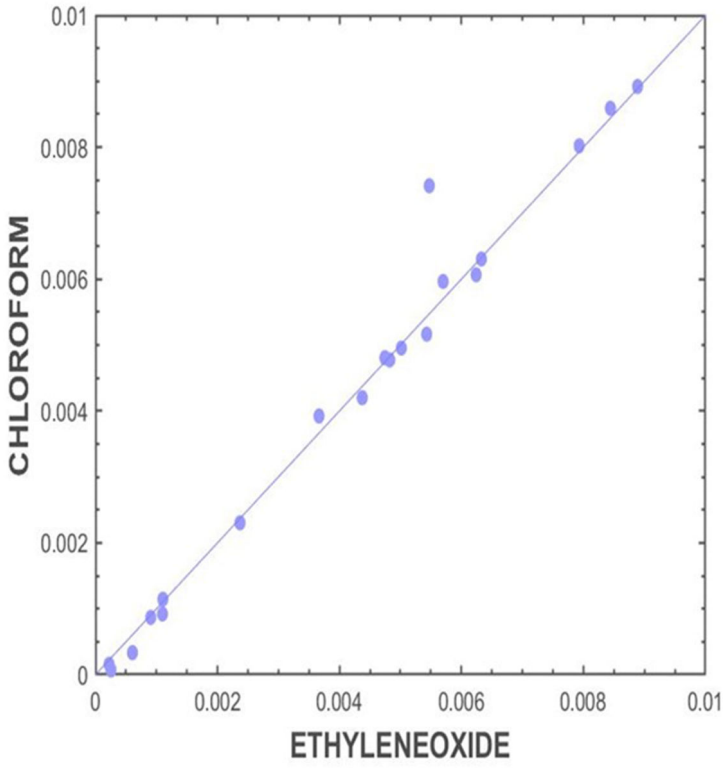

Figure 1.

Example of the challenge of interchangeable co-exposures.

To illustrate the challenge of statistically interchangeable co-exposures in environmental health, we constructed a data set including preterm birth rates by county, along with the estimated ambient concentrations of 175 air toxics from the 2014's EPA national air toxics assessment. To show that equivalent co-exposures can be interchangeable in a predictive model, we ran a 10-fold, cross-validated LASSO regression on the data set and reported the selected co-exposures. We then removed each selected co-exposure from the predictor set and re-ran the algorithm to obtain a new model. We tested the equivalence of the two models with a paired t-test of the two models' residuals. If the difference of the residuals were not statistically significant (at the 0.1 level), the co-exposure sets were considered equivalent. In general, the initial signature included 9 variables, and 8 of them were found to be replaceable.

For example, chloroform and ethylene oxide have equivalent statistical information for preterm birth: They are both predictive of preterm birth and are interchangeable in a model including 8 additional covariates. The corresponding residuals of the two models, plotted above are almost collinear (Pearson's Γ: 0.923).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

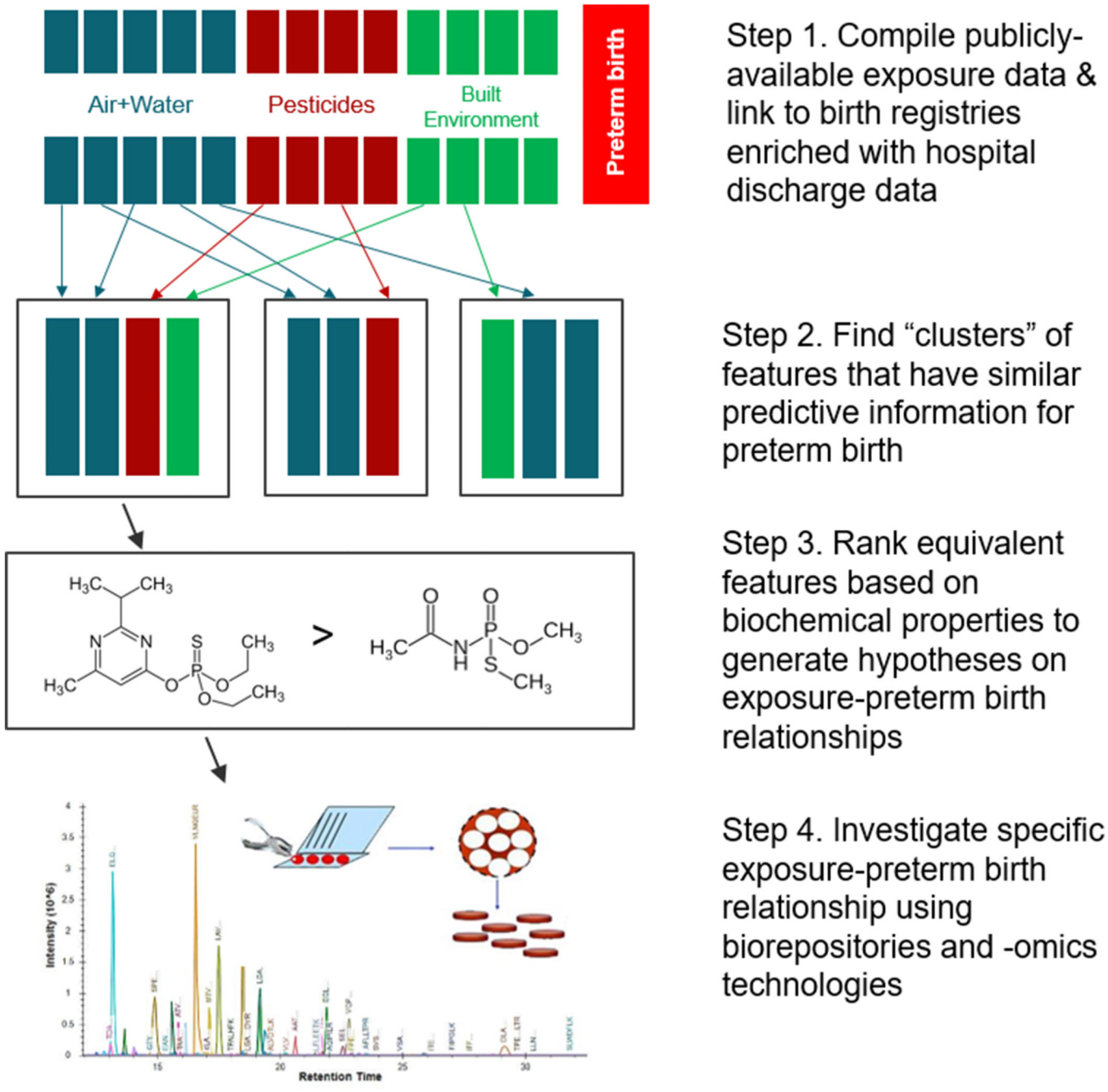

Figure 2.
Example implementation of the proposed framework to investigate environmental contributors to preterm birth