UNIVERSITY OF CALIFORNIA

Los Angeles

Maternal Smoking and Metabolic Outcomes among Newborns in California

A dissertation submitted in partial satisfaction of the

requirements for the degree Doctor of Philosophy

in Epidemiology

by

Di He

2022

ABSTRACT OF THE DISSERTATION

Maternal Smoking and Metabolic Outcomes among Newborns in California

by

Di He

Doctor of Philosophy in Epidemiology

University of California, Los Angeles, 2022

Professor Julia E. Heck, Co-Chair

Professor Beate R. Ritz, Co-Chair

Tobacco smoke contains multiple toxic compounds, and maternal tobacco smoking during pregnancy has been related to negative infant and child outcomes. Newborn blood spots are collected shortly after birth to test the baby for a set of metabolic disorders, after which the State of California stores them for research use. High-resolution metabolomics (HRM) is an analytical approach utilizing ultra-high resolution mass spectrometry and data science methods to characterize and quantify small molecules (metabolites) in biological samples. In this dissertation, we employed HRM as a tool to demonstrate the usefulness of neonatal dried blood spots (DBS) and nicotine metabolites in epidemiological studies, and to better understand the biological pathways through which maternal tobacco use may have long-term impacts on child metabolism.

We first examined the utility of archived newborn blood spots which have been stored for 29 years. We used 899 neonatal DBS of children without cancer before age 6. High-resolution

metabolomics with liquid chromatography mass spectrometry (LC-MS) was performed and the relative ion intensities of common metabolites and selected xenobiotic metabolites of nicotine (cotinine and hydroxycotinine) were evaluated. In total, we detected 26,235 mass spectral features across two separate chromatography methods (C18 and HILIC). For most of the 39 metabolites related to nutrition and health status, we found no statistically significant annual trends across the years of storage. Nicotine metabolites were captured in the DBS with relatively stable intensities.

We then assessed the usefulness of nicotine biomarkers in the same population and built a prediction model for maternal tobacco smoking in pregnancy based on birth certificate information using a combination of self- or provider-reported smoking and biomarkers (smoking metabolites cotinine and hydroxycotinine) in neonatal blood spots as the alloyed gold standard. Potential predictors of smoking were selected from the California birth certificate. Logistic regression with stepwise backward selection was used for prediction model building. Five predictors were selected by the stepwise procedure, including maternal race/ethnicity, maternal education, child's birth year, parity, and child's birth weight. We calculated an overall discrimination accuracy of 0.724 and an AUC of 0.805 (0.770-0.839) in the training set. Similar accuracies were achieved in the internal and external validation sets.

Lastly, we performed a HRM analysis in 899 newborns, following an untargeted metabolome-wide association study (MWAS) workflow. A total of 26,183 features (15,562 in HILIC column and 10621 in C18 column) were detected with HRM of which 1,003 were found to be associated with maternal smoking. Smoking affected metabolites and metabolic pathways in neonatal blood included vitamin A (retinol) metabolism, the kynurenine pathway, and tryptophan and arachidonic acid metabolism. The metabolites and pathway perturbations associated with

cigarette smoking that we identified suggested inflammatory responses and have also been implicated in chronic diseases of the central nervous system and the lung.

In summary, our studies support the usefulness of DBS stored long-term for epidemiological studies of the metabolome, build a prediction model that may benefit future birth registry-based studies in California when there is missing maternal smoking information, and suggest that infant metabolism in the early postnatal period reflects smoking specific physiologic responses to maternal smoking with strong biologic plausibility.

The dissertation of Di He is approved.

Roch Arnaud Kibsa Nianogo

Xia Yang

Julia E. Heck, Committee Co-Chair

Beate R. Ritz, Committee Co-Chair

University of California, Los Angeles

2022

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

AUC	Area Under the Receiver Operating Characteristic Curve
C18	C18 Hydrophobic Reversed-Phase Chromatography
C18neg	C18 Column Coupled with Negative Ionization Mode
CI	Confidence Interval
CV	Coefficients Of Variation
DBS	Dried Blood Spots
ESI	Electrospray Ionization
EVP	Events Per Variable
FDR	False Discovery Rate
GDSP	Genetic Disease Screening Program
HILIC	Hydrophilic Interaction Liquid Chromatography
HILICpos	HILIC Column Coupled with Positive Ionization Mode
HMDB	Human Metabolome Database
HRM	High-Resolution Metabolomics
KEGG	Kyoto Encyclopedia of Genes and Genomes
LC-MS	Liquid Chromatography-Mass Spectrometry
m/z	Mass-To-Charge Ratio
MWAS	Metabolome-Wide Association Study
NMR	Nuclear Magnetic Resonance
NPV	Negative Predicted Value
PLS-DA	Partial Least Squares Discriminant Analysis
PPV	Positive Predicted Value

QC	Quality Control
ROC	Receiver Operating Characteristic
SES	Socioeconomic Status
SVM	Support Vector Machine
VIP	Variable Importance in Projection

ACKNOWLEDGMENTS

First and foremost, I would like to express my sincere gratitude to my mentor, the principal investigator of my dissertation projects, and the co-chair of my dissertation committee, Dr. Julia Heck, for her guidance, encouragement, knowledge, and tremendous support throughout my studies at UCLA. I am continuously grateful for her wisdom and insights. Moreover, she has been such a caring and supportive mentor who values her student as a person. It has been a great fortune to work with and learn from her in my education. I also want to extend my greatest appreciation to my academic advisor and the co-chair of my dissertation committee, Dr. Beate Ritz, for her patience, motivation, enthusiasm, and guidance. Dr. Ritz has provided many constructive comments on the conception, analysis, and writing of my dissertation projects. This work could not have been completed without her. It was a great honor to conduct research and teach graduate students with her, which has prepared me for the next steps in my career as an epidemiologist.

I am also grateful for the members of my doctoral committee, Drs. Roch Nianogo and Xia Yang for their incredibly timely and thorough guidance. A special thanks to Dr. Janet S. Sinsheimer for her kindness and support. I would give special thanks to Drs. Zuo-Feng Zhang, Akihiro Nishi, Lisa V. Smith, and Jeffrey Klausner for their encouragement to pursue doctoral training in this program. I would also like to express my gratitude to all the collaborators and colleagues, Dr. Dean Jones, Dr. Douglas Walker, Dr. Karan Uppal, Dr. Onyebuchi A. Arah, Dr. Qi Yan, and Xiwen Huang for their guidance and support of my work.

In addition, I would like to acknowledge all the staff in the Department of Epidemiology at UCLA for their patience and support. Special thanks to Ms. Joy Miller and Mr. Ramces Jimenez. Sincere thanks to the UCLA Department of Epidemiology, the UCLA Graduate Division, and the University of California Tobacco-Related Disease Research Program for their generous financial

support. A special thanks to my Program Officer at the funding agency, Dr. Ginny Delaney, for her thoughtfulness and support during the pandemic.

Last but not least, I would like to thank my family, friends, and my partner for their unconditional love and support. I could not have made it without you.

Chapter 2 is a version of He, D., Yan, Q., Uppal, K., Walker, D.I., Jones, D.P., Ritz, B., Heck, J.E. (2022). Metabolite stability in archived neonatal dried blood spots used for epidemiological research. *American Journal of Epidemiology*. In press.

VITA

EDUCATION

2016	Bachelor of Science in Biotechnology Sun Yat-sen University, Guangzhou, China

RESEARCH EXPERIENCE

2017 – Present	Graduate Student Researcher Department of Epidemiology University of California, Los Angeles
2019 – 2021	Tobacco-Related Disease Research Program Predoctoral Fellow University of California
2017	Research Assistant Department of Medicine -- Division of Infectious Diseases UCLA David Geffen School of Medicine

PUBLICATIONS

He, D., Yan, Q., Uppal, K., Walker, D.I., Jones, D.P., Ritz, B., Heck, J.E. (2022). Metabolite stability in archived neonatal dried blood spots used for epidemiological research. *American Journal of Epidemiology*. In press.

Yeh, K. W., **He, D**., Hansen, J., Carpenter, C. L., Ritz, B., Olsen, J., & Heck, J. E. (2022). The risk of childhood brain tumors associated with delivery interventions: A Danish matched case-control study. *Cancer Epidemiology*, *76*, 102077.

Ritz, B., Yan, Q., **He, D**., Wu, J., Walker, D. I., Uppal, K., ... & Heck, J. E. (2022). Child serum metabolome and traffic-related air pollution exposure in pregnancy. *Environmental Research*, *203*, 111907.

Hall, C., Hansen, J., Olsen, J., **He, D**., von Ehrenstein, O. S., Ritz, B., & Heck, J. E. (2021). Parental occupation and childhood germ cell tumors: a case–control study in Denmark, 1968–2016. *Cancer Causes & Control*, *32*(8), 827-836.

Heck, J. E., Lee, P. C., Wu, C. K., Li, C. Y., **He, D**., Federman, N., ... & Hansen, J. (2020). Spina bifida and pediatric cancers. *Pediatric hematology and oncology*, *37*(7), 630-636.

Heck, J. E., **He, D**., Contreras, Z. A., Ritz, B., Olsen, J., & Hansen, J. (2019). Parental occupational exposure to benzene and the risk of childhood and adolescent acute lymphoblastic leukaemia: a population-based study. *Occupational and environmental medicine*, *76*(8), 527-529.

Heck, J. E., **He, D**., Janzen, C., Federman, N., Olsen, J., Ritz, B., & Hansen, J. (2019). Fetal programming and Wilms tumor. *Pediatric blood & cancer*, *66*(1), e27461.

Kimble, M. M., Javanbakht, M., Chew, K. W., Stafylis, C., **He, D**., Ramirez, S., ... & Klausner, J. D. (2019). Sociodemographic and clinical characteristics of persons who experienced spontaneous hepatitis C viral clearance. *BMC infectious diseases*, *19*(1), 1-7.

CONFERENCE PRESENTATIONS

He D, Huang X, Ritz B, Arah OA, Heck JE. Prediction model for classifying maternal tobacco smoking using newborn dried blood spots among children born from 1983 to 2011 in California. *International Conference on Prenatal Programming and Toxicity,* Virtual, January 2022. (Poster presentation)

He D, Yan Q, Uppal K, Walker DI, Ritz B, Jones DP, Heck JE. Maternal tobacco smoking and childhood metabolic outcomes. *American Public Health Association Annual Meeting,* Virtual, October 2020. (Oral presentation)

He D, Yan Q, Uppal K, Walker DI, Ritz B, Jones DP, Heck JE. Maternal tobacco smoking and childhood metabolic outcomes. *Joining Forces 2020: Ending The Tobacco Epidemic For All,* Postponed, CA 2020. (Poster presentation)

He D, Hansen J, Federman N, Olsen J, Ritz B, Heck JE. Hernia and Ewing sarcoma. *American Association for Cancer Research Annual Meeting*, Virtual, April & June 2020. (Poster presentation)

Chapter 1. Introduction and Background

1.1 Maternal perinatal tobacco use

The tobacco epidemic is one of the biggest public health threats worldwide, leading to more than 8 million death per year around the world. More than 7 million of deaths are results of direct tobacco use exposure and 1.2 million are the result of non-smokers being exposed to secondhand smoke.[1] Tobacco use is responsible for approximately one-third of all cancer deaths [2] and high rates of morbidity and mortality from oral cancer, lung cancer, gastric cancer, cardiovascular disorders, and headache.[3-10] In the United States, tobacco use remains the leading cause of preventable diseases, disability, and death, resulting in more than 480,000 deaths each year.[11] In 2020, nearly 12.5% of U.S. adults aged 18 years or older currently smoked cigarettes.[11] Some demographic groups have a higher percentage of tobacco use and related health problems. According to a report of tobacco product use among U.S. adults in 2020,[12] the current smoking prevalence is higher among men than women (14.1% vs. 11.0%), highest among people aged 25-44 years (14.1%) and 45-64 years (14.9%), and highest among non-Hispanic American Indian/Alaska Native adults (27.1%) followed by non-Hispanic Black adults (14.4%), non-Hispanic White adults (13.3%), Hispanic adults (8.0%) and non-Hispanic Asian adults (8.0%). Moreover, current smoking is generally more prevalent among disadvantaged groups such as people with lower education levels, with lower household income, without health insurance coverage, being sexual minority, or with mental health issues.[12]

Over the past decades, the prevalence of current cigarette smoking among U.S. adults declined from 51% in 1965 to 31% in 2012 among males and from 34% to 23% among females.[11] Among California women, interview-based studies revealed that maternal pregnancy smoking prevalence declined from approximately 50% in the late 1970s [13] to 15% from 1995-2002 [14] and

5% from 2008-2018.[15] The prevalence of maternal smoking estimated using our data was in line with the trend described above , with approximately 40% in the 1980s down to less than 5% in the 2000s (Figure 1-1). Despite the rate declining by more than half since 1964, cigarette smoking remains the most preventable cause of disease and death in the U.S. with disparities persisting across population groups.[16] In recent decades, smoking rates in California pregnant women have been approximately 5% lower than nationwide,[17] which is partly due to demographics as immigrant Mexican women make up about 25% of all pregnant women and most are non-smokers.[18] Components in cigarette smoke including polycyclic aromatic hydrocarbons (PAHs), N-nitrosamines, aromatic amines, aldehydes, volatile organic hydrocarbons, and metals have long been known or suspected carcinogens. Nicotine is also a modifier of cancer progression,[19] and thus may be contributing to childhood cancers due to prenatal and early childhood exposure from smoking mothers. Smoking in pregnancy has also been associated with adverse infant and child outcomes, including low birth weight, preterm birth, and congenital anomalies.[20,21]

1.2 Accuracy of interview-based or birth certificate reported maternal pregnancy smoking

As it is a risk factor for multiple pediatric diseases, maternal pregnancy smoking is an important confounding variable to be adjusted for in studies of maternal exposures and child outcomes, such as childhood cancers.[22] Record-linkage studies are advantageous due to their (often) population-based nature, larger sample size, and lower cost, but frequently have limited or no information on lifestyle factors such as smoking. Systematically recorded smoking data may not be available in vital statistics records for some periods or in some regions possibly resulting in uncontrolled confounding.

Even when these records collect smoking status, the quality of the data might be a concern as self-reported data on pregnancy smoking may be unreliable due to social stigma, thus, the missing data is not missing at random. Two studies in the United States in the 2000s and 2010s showed that approximately 14% of smokers deny smoking in interviews.[23,24] Compared with smoking data from other sources including CDC surveys and medical records, it has been reported that birth certificates only capture 70.6-82.0% of prenatal smoking.[25-27] Maternal smoking may be underreported due to several reasons: women failed to disclose their smoking to their provider; providers failed to accurately record the information in the medical record; the medical record was not received by the hospital; or inaccurate input when clerks fill out the birth certificate.[26]

1.3 Accuracy of blood spots to capture smoking

Given concerns about the accuracy of recorded smoking on the birth certificate, measurement of tobacco smoking using a biomarker of exposure is preferable in research studies. For pregnant women close to the time of delivery, cotinine levels in newborns' dried blood spots can be used as a biomarker of maternal smoking, with high sensitivity (92.3%) and specificity (99.7%).[28,29] Cotinine is the best marker of recent smoking including active and secondhand smoke.[30,31] Although cotinine has a longer half-life than its parent compound, nicotine, it is only detectable for about 17 hours.[32] Hydroxycotinine, metabolized by the hepatic enzyme cytochrome P450 2A6, is the major metabolite of cotinine in most individuals,[33,34] and is detectable for a longer period with an approximated half-life of 18 hours.[35]

1.4 Research into the metabolomics of smoking

The usefulness of neonatal dried blood spots

Approximately 98% of babies born in the United States participate in newborn screening, a public health program that aims to identify early metabolic and genetic defects to prevent disease and disability.[36] Several US states and other nations store neonatal dried blood spots (DBS) for research purposes.[36,37] Research using DBS continues to increase,[38-41] particularly studies employing metabolomics tools including studies of pediatric leukemia,[42,43] estimating gestational age,[44] and those attempting to assess the early-life exposome.[45] Research targeting rare diseases must often rely on biospecimens that are collected routinely and have been stored for long time periods in order to accumulate a sufficient number of samples to address hypotheses with adequate statistical power. Thus, a key question is whether metabolites are sufficiently stable in DBS over time to produce robust estimates of endogenous or exogenous metabolites (such as those from dietary sources or tobacco use), or whether they degrade, limiting their use for metabolomics research.

In previous studies, the stability of metabolites in DBS was shown to be time-and-temperature dependent with good stability in the short term (a week under all temperature conditions); long-term stability of metabolites in DBS varied with the storage environment, depending on temperature and humidity.[46,47] For example, it has been reported that folate, essential for embryonic development, and hemoglobin concentrations in DBS decreased with storage at warmer temperatures (4°C vs. −80°C), and greater humidity (humid vs. ambient).[47] Others found major alterations in metabolites occurring only at room temperature (21°C), with no attenuation observed in frozen samples (-20°C or −80°C) over a 2-year period.[48] Improved understanding of the stability of metabolites in blood spots that were stored in -20°C for several decades will enhance the value of this important resource for environmental epidemiology research.

High-resolution metabolomics

High-resolution metabolomics (HRM) is an analytical approach utilizing ultra-high resolution mass spectrometry and data science methods to characterize and quantify small molecules (metabolites) in biological samples.[49,50] Untargeted metabolomics is a hypothesis-free approach that profiles endogenous and exogenous metabolites in biological samples to gain insights into exposures and pathways underlying disease. Importantly, the development of hybrid approaches to obtain information on identified metabolites without loss of information on un-identified metabolites, considerably enhances translational potential for epidemiologic and model systems research.[51] Untargeted profiling of neonatal DBS has recently been shown to provide a valuable measure of metabolic alterations associated with pediatric disease, including inborn errors of metabolism and the growth in infants associated with breast-feeding.[52-55]

HRM studies on tobacco smoking

Tobacco smoking is related to metabolomic changes as recently documented for adult smokers, children exposed to passive smoke, and pregnant women.[56-58] A study of military personnel observed alterations in lipid and xenobiotic metabolism, and diverse effects on amino acid, sialic acid and purine and pyrimidine metabolism in tobacco smokers compared to non-smokers.[56] Another study among preschool children reported secondhand smoke-induced urinary metabolites including kynurenine, tyrosyl-tryptophan, and 1-(3-pyridinyl)-1,4-butanediol, peptides, and pyridines.[57] Utilizing second-trimester amniotic fluid, a third study found low-level maternal nicotine exposure from light smoking or secondhand smoke to be associated with dysregulated metabolic pathways in the fetus such as aspartate and asparagine metabolism,

pyrimidine metabolism, and metabolism of other amino acids, and also saw decreases in acetylated polyamines.[58]

1.5 Objective

The linkages between maternal smoking during pregnancy and childhood metabolic outcomes remain largely unexplored.[59] Here, we first assessed the feasibility of applying untargeted HRM to archived DBS which were selected for a population-based record-linkage study of childhood cancers in California.[60] We comprehensively assessed the stability of metabolites across all study years (1983-2011) and metabolites with confirmed identifications were characterized for stability over increasing storage times.

Then, we generated a clinical prediction model [61] for maternal late pregnancy smoking based on variables routinely collected on birth certificates and a combination of self- or provider-reported smoking information and metabolomics-derived smoking biomarkers together used as the alloyed gold standard for maternal smoking. Clinical prediction models have been used to predict tobacco-related disease,[62] calibrate self-reported maternal smoking using urinary cotinine,[63] and to characterize smoking patterns in older adults.[64]

Lastly, to better understand the biological pathways through which maternal tobacco use may be increasing the risk of disease, we performed a high-resolution metabolomics (HRM) analysis in 899 children born from 1983 to 2011 in California following an untargeted metabolome-wide association study (MWAS) workflow.

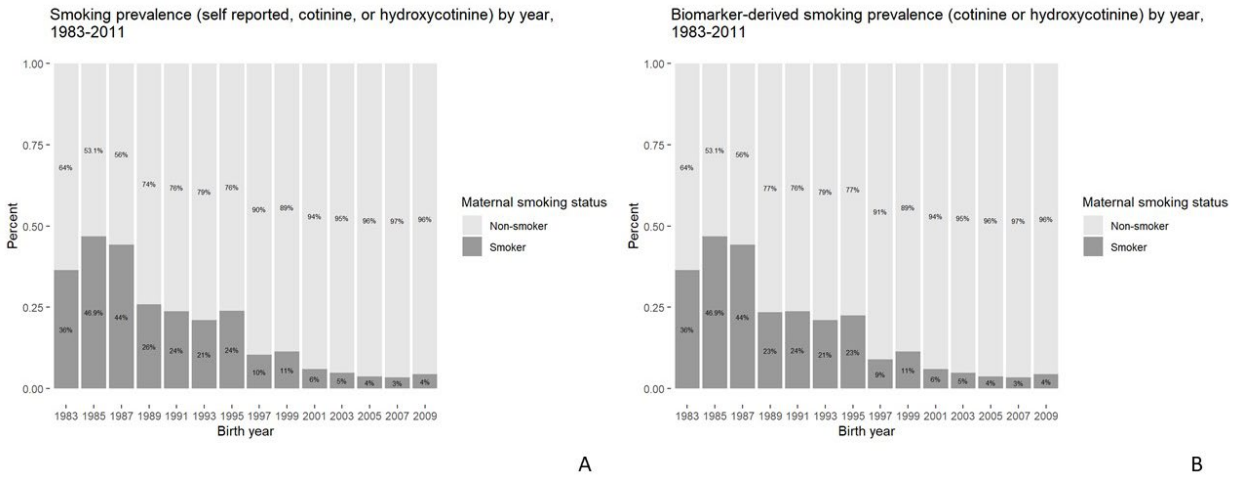


Figure 1-1. Pregnancy smoking rate by year, 1983-2011

Chapter 2. Metabolite stability in archived neonatal dried blood spots used for epidemiological research

2.1 Abstract

Background: Epidemiologic studies of low-frequency exposures or outcomes using metabolomics analyses of neonatal dried blood spots (DBS) often require assembly of samples with substantial differences in duration of storage. Independent assessment of stability of metabolites in archived DBS will enable improved design and interpretation of epidemiologic research utilizing DBS.

Methods: Neonatal DBS routinely collected and stored as part of the California Genetic Disease Screening Program between 1983 and 2011 were used. The study population included 899 children without cancer before age 6 born in California. High-resolution metabolomics with liquid chromatography mass spectrometry (LC-MS) was performed and the relative ion intensities of common metabolites and selected xenobiotic metabolites of nicotine (cotinine and hydroxycotinine) were evaluated.

Results: In total, we detected 26,235 mass spectral features across two separate chromatography methods (C18 and HILIC). For most of the 39 metabolites related to nutrition and health status, we found no statistically significant annual trends across the years of storage. Nicotine metabolites were captured in the DBS with relatively stable intensities.

Conclusions: This study supports the usefulness of DBS stored long-term for epidemiological studies of the metabolome. Omics-based information gained from DBS may also provide a valuable tool for assessing prenatal environmental exposures in child health research.

2.2 Introduction

Approximately 98% of babies born in the United States participate in newborn screening, a public health program that aims to identify early metabolic and genetic defects to prevent disease and disability.[36] Several US states and other nations store neonatal dried blood spots (DBS) for research purposes.[36,37] Research using DBS continues to increase,[38-41] particularly studies employing metabolomics tools including studies of pediatric leukemia,[42,43] estimating gestational age,[44] and those attempting to assess the early-life exposome.[45] Research targeting rare diseases must often rely on biospecimens that are collected routinely and have been stored for long time periods in order to accumulate a sufficient number of samples to address hypotheses with adequate statistical power. Thus, a key question is whether metabolites are sufficiently stable in DBS over time to produce robust estimates of endogenous or exogenous metabolites (such as those from dietary sources or tobacco use), or whether they degrade, limiting their use for metabolomics research.

In previous studies, the stability of metabolites in DBS was shown to be time-and-temperature dependent with good stability in the short term (a week under all temperature conditions); long-term stability of metabolites in DBS varied with the storage environment, depending on temperature and humidity.[46,47] For example, it has been reported that folate, essential for embryonic development, and hemoglobin concentrations in DBS decreased with storage at warmer temperatures (4°C vs. −80°C), and greater humidity (humid vs. ambient).[47] Others found major alterations in metabolites occurring only at room temperature (21°C), with no attenuation observed in frozen samples (-20°C or −80°C) over a 2-year period.[48] Improved understanding of the stability of metabolites in blood spots that were stored in -20°C for several decades will enhance the value of this important resource for environmental epidemiology research.

High-resolution metabolomics (HRM) is an analytical approach utilizing ultra-high resolution mass spectrometry and data science methods to characterize and quantify small molecules (metabolites) in biological samples.[49,50] Untargeted metabolomics is a hypothesis-free approach that profiles endogenous and exogenous metabolites in biological samples to gain insights into exposures and pathways underlying disease. Importantly, the development of hybrid approaches to obtain information on identified metabolites without loss of information on un-identified metabolites, considerably enhances translational potential for epidemiologic and model systems research.[51] Untargeted profiling of neonatal DBS has recently been shown to provide a valuable measure of metabolic alterations associated with pediatric disease, including inborn errors of metabolism and the growth in infants associated with breast-feeding.[52-55]

Here, we assess the feasibility of applying untargeted HRM to archived DBS which were selected for a population-based record-linkage study of childhood cancers in California.[60] We comprehensively assessed the stability of metabolites across all study years (1983-2011) and metabolites with confirmed identifications were characterized for stability over increasing storage times.

2.3 Methods

Study population

We utilized data from the Smoking and Embryonal Tumor Study, a large population-based case-control study of childhood cancers, which ascertained cancer cases from the California Cancer Registry and included controls from among births in California between 1983 and 2011.[65] Population controls ascertained from California birth rolls were frequency matched (20:1) to cancer cases by year of birth. From this population, we randomly selected 1400 children (501

retinoblastoma cases and 899 controls) for HRM analysis. The demographic, socioeconomic, and gestational characteristics of cases and controls were previously reported.[60,66] Starting 2007, smoking status before and during pregnancy was reported on the birth certificate as responses to the questions about the number of cigarettes smoked per day in the 3 months before pregnancy and during each trimester.

California's newborn screening program, the Genetic Disease Screening Program (GDSP), began in 1966.[37] The GDSP obtained blood samples from babies' heel-sticks between 12 and 48 hours after birth.[67] Blood samples were put on six small circles on a specialized filter paper, dried at room temperature for at least 3 hours, and shipped to the Neonatal and Prenatal Screening Laboratory within 24 hours of collection.[67] Since 1982, after the routine screening, left-over specimens were packed and stored at -20°C. Additional details concerning dried blood specimen collection and storage are described elsewhere.[68,69] In preparation for shipment, the GDSP anonymized all samples (i.e. none contained information on year sampled) and placed the samples into plastic bags. All bags were loosely put into a box and shipped to the laboratory for HRM analysis. Thus, the samples were 'naturally randomized' as they were tossed together at random in the box for shipment.

High-resolution metabolomics

HRM profiling was completed according to established methods.[70,71] Samples were punched using a 5 mm hole puncher, extracted with 2:1 acetonitrile in water (containing a mixture of stable isotopic internal standards), and then mixed on an orbital shaker in the dark at low speed for 12 hours at 0-4 °C. Samples were then centrifuged to remove any particulate matter and were analyzed in triplicate using liquid chromatography interfaced to an ultra high-resolution mass spectrometer (Thermo Scientific Q-Exactive HF).[72] DBS samples along with the NIST 1950 and

QSTD (internal quality control) samples were analyzed in batches of 40 study samples using an acetonitrile gradient and two technical columns that include hydrophilic interaction liquid chromatography (HILIC) with positive electrospray ionization (ESI) and C18 hydrophobic reversed-phase chromatography with negative ESI, to enhance the coverage of metabolic feature detection.[73] The NIST 1950 samples were analyzed at the beginning and the end of the entire run and the QSTD samples were analyzed at the beginning, middle, and the end of each batch. Raw data was extracted using apLCMS[74] with modifications by xMSanalyzer,[75] and batch corrected using ComBat.[76] For data analysis, we only included metabolomic features with median coefficients of variation among technical replicates <30% and Pearson correlation coefficients >0.7. Ten samples were considered as outliers and excluded from the analyses, among which 6 were removed from the HILIC positive column and 6 were removed from the C18 negative column (2 overlapping samples). Feature tables of detected signals (referred to as metabolite features) were then generated with mass-to-charge ratio (*m/z*), retention time, and ion intensity as the unique identifier for each signal.[73]

Statistical analysis

Intensities for each feature were generated based on the median of three replicate measures. If more than 2 out of 3 replicates were missing, the summarized value was represented as a missing value. Metabolite features were then filtered to keep only those metabolites detected in greater than 50% of all samples, with the exception of cotinine and hydroxycotinine. For cotinine and hydroxycotinine, signals were retained even though they were present in less than 50% of samples. Missing values were then imputed by using one-half of the lowest signal intensity in the complete dataset and intensities were log2-transformed before analyses. To control for potential confounders including Hispanic ethnicity and maternal age, we calculated residuals of intensities derived from

linear regression against Hispanic ethnicity and maternal age prior to downstream analyses. A total of 894 control subject samples were included in the analysis.

Nutrition and Health Biomarker Assessment

For evaluation of targeted metabolites, we selected a small set of metabolites as examples for the overall quality of the blood spot over 3 decades of storage. We focused on 39 health and nutrition-related metabolites that had previously been assessed in adult serum samples by our team; [77] i.e. a type of sample commonly used to examine metabolites in blood. By focusing on the same set of metabolites that can be found in blood serum in analyses obtained on the same metabolomics platform, we can better understand what we newly see in children's blood samples archived as dried blood spots. We plotted the intensities by year of births using the R packages "ggplot2" and "cowplot" with LOESS smoothing. To examine the feasibility of using cotinine in long-term stored neonatal DBS as a biomarker to assess maternal smoking status, we also extracted cotinine and hydroxycotinine detected using HILIC with positive ESI.[78] A linear trend test was performed for each selected metabolite to assess whether there were intensity changes across samples collected over the 29 years such as increasing/decreasing trends.

Endogenous and Exogenous Features

For analysis of untargeted metabolomics data, the detected mass spectral features were matched to the Human Metabolome Database (HMDB), Kyoto Encyclopedia of Genes and Genomes (KEGG), and LipidMaps with a mass error threshold of 10 ppm. Using this approach, metabolite identification is consistent with Level 4 using Schymanski criteria, meaning we have assigned unequivocal molecular formula using spectral information such as *m/z* and adduct, but have insufficient evidence to propose possible structures.[79] We classified the features as endogenous metabolites or having exogenous origin (e.g. food, plant, microbial, drug, cosmetic,

or toxin/pollutant) using HMDB. For the endogenous metabolites, we plotted the intensities by year of births following the same procedure as described above. For the exogenous features, we followed the statistical strategy of Bunning et al.[80] and fitted LOESS regression models using the R package "stats" to assess feature intensity according to the year of birth with a default span of 0.75. Then we used the fitted model to obtain the predicted values for the year of birth of each feature and created a matrix of birth years (N=29, 1983-2011) and feature intensities. This matrix was the input for a Fuzzy C-means Clustering analysis (R package "Mfuzz"). We calculated the minimum centroid distance for a range of clusters and selected 10 clusters as the optimal number for the HILIC column and the C18 column.

2.4 Results

The demographics of subjects were shown in Table 2-1. Mothers were mostly white and more than half identified themselves as non-Hispanic. The average maternal and paternal ages at birth were 27.2 and 30.2, respectively. More than half (54.2%) of all mothers were born in the US and the remaining were Mexican-born (26.4%) or other foreign-born (19.4%). Forty-one percent were firstborn children. Only 2 (1.6%) mothers of children born in 2007 and later reported having actively smoked during pregnancy.

In total, we detected 26,235 mass spectral features (15,596 in HILIC and 10,639 in C18), and after filtering out missing values and pre-processing, 21,759 features (12,998 in HILIC column and 8,761 in C18 column) were retained for downstream analyses.

The 39 nutritional and health indicator metabolites detected in DBS (Table 2-2) included17 amino acids and their metabolites, 2 health indicators (creatine and cholesterol), 4 vitamin coenzymes, 10 fatty acid or lipid metabolites, 4 nucleotide metabolites, and 2 exogenous chemicals (benzoic acid and caffeine). The MS/MS details of the metabolites with annotation level 1 were

previously published.[56,70] Their log2-transformed, confounder adjusted residuals of relative intensities over years of birth are shown in Figures 2-1 and 2-2. Most of the selected metabolites appeared to remain stable across the years 1983-2011, with the exception of methionine, sphingosine, sphinganine, choline, arachidic acid, and caffeine. Stability over the period was seen for 17 out of the 39 selected metabolites with slopes close to zero and no linear trends were detected for the remaining metabolites.

From the untargeted metabolomics data, we selected 1,137 mass spectral features having accurate mass match to endogenous features (547 in HILIC column and 590 in C18 column) and 1,944 with accurate mass match to exogenous features (1,020 in HILIC column and 924 in C18 column). Note that none of these had confirmed identity; the selection strategy was intended to support examination of groups of signals enriched in endogenous metabolites and exogenous chemicals. There was some fluctuation in levels of the endogenous features over time, but the majority (~80%) of these remained stable across samples collected over the years 1983-2011 (Supplemental Tables 2-S1, Figures 2-S1 & 2-S2). All detectable exogenous features were used as input into clustering analyses and the clustering results are shown in Supplemental Figures 2-S3 & 2-S4. Each feature only contributed to one cluster. No distinct pattern was detected for either of the columns. There were some increasing and some decreasing trends as well as some general fluctuations of features in both columns. The exogenous features with cluster IDs are listed in Supplemental Table 2-S2 & 2-S3.

Residuals of intensities for metabolites of nicotine are shown in Figure 2-3. Among the 899 subjects, we detected cotinine in 79 (8.8%) and hydroxycotinine in 284 (31.8%) DBS and their intensities were highly correlated (Spearman's rho correlation coefficient = 0.77, Supplemental Figure 2-S5). Overall, we did not observe samples from children born in earlier years to have lower

intensity for cotinine or hydroxycotinine compared to those born in the more recent years. A small negative linear trend was seen for hydroxycotinine (beta = -0.04, $P < 0.01$) while a small positive linear trend was seen for cotinine (beta = 0.05, $P = 0.04$).

2.5 Discussion

In this study of archived DBS from a large population of children born in California, we found that almost three decades of storage in -20 °C freezers did not affect the deterioration of metabolites that reflect the general human physiology or exogenous metabolites that are markers of common lifestyle-related exposures such as cotinine and caffeine. This was true for both the nutritional and health indicator metabolites and some of the exogenous chemicals (Table 2-2; Figures 2-1 & 2-2). Our study was conducted in a multi-ethnic California population and our results support the feasibility of future studies that investigate maternal behavior and exposures and rare childhood diseases based on metabolic profiles, such as studies of maternal smoking and caffeine intake, in DBS stored for decades.

There is no gold standard to assess metabolite stability in stored DBS samples. Different studies assessed the stability of various metabolites in DBS with cross-sectionally collected specimens and repeated analysis prospectively conducted over a certain time course. Phosphatidylethanol, a direct ethanol metabolite, remained stable at 4°C or -80°C for at least 9 months.[81] In another study, amino acid reference materials degraded after storage for 28 days at ambient temperature in a dry environment and the degradation rate ranged from 3% - 7% per year for several amino acids including alanine, arginine, leucine, methionine, and phenylalanine. An even higher rate of degradation was seen for acylcarnitines such as free carnitine, acetylcarnitine, and propionylcarnitine.[82] However, degradation occurred when DBS were collected between 3-7 days after birth and stored at ambient temperatures and in a dry environment.[82] Degradation

appears to be minimal for amino acids in our samples, likely because California's neonatal blood spots are collected within 12-48 hours after birth, mailed within 24 hours of sampling, and stored at -20°C. Highly unstable metabolites can be expected to be largely lost by the time of storage, so detected metabolites can be expected to be biased towards those that are relatively more stable. There is a possibility that trapping of metabolites within the proteinaceous matrix may stabilize and inhibit degradation, and this could contribute to differences seen compared to amino acid reference materials.

Another targeted metabolomics study found that most of 404 detected metabolites in DBS remained stable for the first month of storage, declined rapidly within 1-3 months of storage, and remained stable for at least one year regardless of the storage conditions (4°C vs. -20°C vs. −80 °C; with/without desiccant; with/without an O_2 scavenger).[83] Phospholipids, sphingolipids, acylcarnitines, amino acids, and steroids were the main chemical groups that degraded by over 30%.[84] However, there were no significant changes in the total number of metabolites detected over a one-year period. Our study assessed the stability of metabolites by examining the patterns of ion intensities in relation to DBS age. We did not measure changes of metabolite intensities over time; i.e. we did not repeatedly test the same sample over time for metabolite levels. However, since we observed a similar number of metabolites in long-term stored DBS in every year this may suggest some general stability in the number of features that can be identified over a very long time in DBS stored at -20 ℃. Palmer employed untargeted metabolomics and examined the stability of metabolites for dried blood spots and dried urine spots over a 12-month period under different storage temperatures (-20, +4, and +21 °C).[85] They found greater instability in polar compounds measured by HILIC in dried blood spots stored at -20 °C. Although the platform they used was not completely comparable to ours, we did observe slightly higher variation with

confirmed metabolites identified in the HILICpos column (Figure 2-1) than for those in the C18neg column (Figure 2-2).

Amino acids have been a focus of newborn screening and the most up-to-date quantification platforms can now differentiate isomers.[86] We have found methionine to be the least stable amino acids in our samples, which corresponds to previous findings.[77,82,87] During long-term storage, methionine tends to be slowly oxidized into methionine sulfoxide and methionine sulfone.[77] As an essential amino acid, methionine's carbon skeletons cannot be synthesized by the body and needs to be provided in the diet to meet requirements.[88] Dietary intake of methionine is suspected to affect the fetal genome and pregnancy outcomes and thus has been a target in studies of pregnancy outcomes.[88] Our results suggest that a slight degradation of methionine is common in DBS stored long-term.

We also observed varying intensities across sample years for lipid metabolites such as sphingosine, sphinganine, and choline. Sphingolipid metabolites are lipid mediators that regulate cellular functions such as cell growth, immune cell trafficking, inflammation, and cancer.[77,89] A measurable amount of sphingosine and sphinganine in DBS suggests the possibility of measuring these physiologic parameters and use them to evaluate maternal and child health. Choline was classified as an essential nutrient by the Institute of Medicine in 1998 and both too low and too high levels may cause health concerns.[77,90] Maternal choline supplementation during pregnancy has been suggested to benefit several physiologic systems in the offspring.[91] Therefore, the varying intensities of choline over the past decades in California newborns we observed in the DBS may indicate novel use of choline supplements by pregnant women over time or a major change in diets and/or changes in the underlying distribution of chronic diseases among mothers possibly due to a shift in the age or ethnic/race composition of the population of pregnant women.

The intensities of metabolites measured in neonatal DBS are influenced by many factors including the mother's metabolic status, dietary intake, the maturity of the infant, and maternal behaviors.[82] Pregnancy is frequently a motivator for behavior change, with some women choosing to stop smoking, reducing or eliminating caffeine from the diet, and eating healthier foods.[92] Although the prevalence of caffeine consumption in pregnancy varies by countries and cultures, worldwide 60% to 75% of pregnant women drink caffeine-containing beverages.[93] Caffeine was detected in 737 (82%) of our samples with no significant difference in relative intensities over time in the DBS collected in earlier vs. recent years. The fact that caffeine was detected in a majority of samples possibly indicates breastfeeding-related exposures. The mean half-life of caffeine is 5 hours in adults but 82 hours in newborns.[94] In the US, the rate of breastfeeding initiation was about 60% to 70% in the 1980s to 90s.[95] In 2018, 93.8% of California women reported having initiated breastfeeding.[96] Our DBS represent the child's metabolism affected by maternal behavior in late pregnancy and the newborn's postnatal diet.

Although maternal smoking is of great interest in studies of pregnancy and offspring health, there are concerns about the accuracy of self-reported smoking.[97-99] Thus, cotinine and hydroxycotinine measures in DBS may be more accurate measures compared to self-reported smoking data. Neonatal DBS cotinine is a good biomarker of maternal smoking close to the time of delivery and can predict umbilical cord blood cotinine well ($R^2 = 0.80$) both at room temperature and after long-term freezer storage.[68] As shown in several studies, cotinine of ≥10 ng/ml in neonatal blood spots represents maternal active smoking.[68,100] It has been suggested that the optimal threshold for DBS cotinine to distinguish smokers from nonsmokers is lower than the threshold used in adult blood (3.13-6 ng/mL vs. 10 ng/mL).[101] Cotinine is the primary metabolite of nicotine and half of the cotinine is subsequently metabolized into hydroxycotinine.[102] A study

of 13 newborns at San Francisco General Hospital reported an elimination half-life for cotinine and hydroxycotinine of 16.3 hours and 18.8 hours in umbilical cord blood, respectively.[35] Thus, given that the DBS were collected 12-48 hours after birth, it is plausible that we would observe hydroxycotinine (32% in our sample) more frequently than cotinine. However, hydroxycotinine also has other sources including dietary sources such as tomatoes and eggplant.[103] Our study suggests that cotinine and hydroxycotinine remain detectable in DBS after very long-term storage (up to 29 years), which supports the usefulness of cotinine measures to assess maternal smoking status in health studies of pregnancy exposures based on newborn blood.[63,104]

Interpreting our endogenous features plots and the exogenous clustering results will require further investigations. As mentioned above, while the majority of endogenous features remained stable throughout decades of storage some, such as lipid metabolites, appeared to be less stable. Lipids and lipid metabolites are known to be prone to oxidation and hydrolysis.[105,106] Even though lipid profiles derived from DBS are no less reproducible than those from plasma or whole blood, the process of spotting and air drying blood spots might still lead to some level of lipid oxidation and hydrolysis and it is recommended to interpret those results carefully.[106] Clustering analysis allowed us to group exogenous features with similar patterns across time of storage. The patterns may suggest differences in maternal lifestyle factors especially diets that may contribute to metabolite changes in newborns over time or reflect environmental exposures that wax and wane. We examined exogenous chemicals such as benzoic acid, caffeine, and cotinine in this study, and our results for mass spectral signals matching exogenous chemicals indicate that further exploration utilizing MS/MS targeted metabolomics platform together with a clustering approach will be useful for environmental epidemiology.

Despite having high specificity, annotation using the combination of *m/z* and retention time is challenging due to the presence of isomers, similar molecular-weight interferences, and in-source degradation products.[104] Metabolites can degrade or interconvert during the process of extraction and the degradation varies from one metabolite to another. During the extraction process, high-energy, high-abundant compounds are more likely to be lost compared to less-abundant, lower-energy ones.[104] This may explain why we did not identify all nutritional and health-related metabolites that were assessed for stability by Accardi et al.[77] Thus, to improve the accuracy of metabolite annotation, it is recommended to conduct an untargeted analysis in parallel with a targeted MS/MS confirmation.[104]

Other than the characteristics of the metabolites, there are other possible reasons why some metabolites that should be detectable were not detected in all samples including instability due to initial handling and storage of the blood spots that were prepared in clinics and labs all over California before being mailed to the California Neonatal Screening program. Specifically, the blood spots were not collected by one research group but rather for neonatal screening purposes i.e. they represent a public health effort at the community level to collect, transport, extract, and analyze or store these samples. There can be sample-handling issues at every phase of the process, which may cause potential variations that affect sample quality. Thus, it is especially encouraging to find that mostly these samples are of high quality. Nevertheless, limitations include the inability to detect all metabolites that should possibly be detectable in all samples.

In Table 2-3, we summarize the uncontrollable and controllable factors contributing to the quality and consistency of routinely collected neonatal blood spots for epidemiological research. Overall, it is advisable to 1) have enough samples to minimize contributions from factors that may corrupt the sample data and interpretations; 2) design studies with discovery and validation subsets

whenever feasible; 3) work towards a minimum standards guiding the use of blood spots in targeted and untargeted metabolomics in epidemiologic research; and 4) design a quantification strategy and improve comparability between studies.

This study supports the usefulness of archived DBS in epidemiologic studies of rare diseases in later childhood or life that have to rely on archived routine biospecimens and using an untargeted metabolomics approach. This reflects a strength of California's newborn screening program's use of specific protocols for the collection and storage of DBS, which allowed us to assess the stability of DBS derived metabolites in samples collected in different years.[41] Results from this study support the feasibility of conducting metabolomics studies using archived DBS stored for decades. Utilizing existing biospecimens from statewide newborn screening programs will provide a unique and important resource for future epidemiologic health and environmental exposure studies.

2.6 Tables and figures

Table 2-1. Demographic characteristics of the study population (1983-2011)

Demographic characteristics	Subjects (N = 899) N (%)
Infant sex	
Male	440 (48.9)
Female	459 (51.1)
Ethnicity	
Hispanic	424 (47.2)
Non-Hispanic	475 (52.8)
Race	
Native American	5 (0.6)
Asian	84 (9.4)
Pacific Islander	31 (3.5)
Black	57 (6.4)
White	717 (80.2)
Missing	5
Maternal age (years), Mean (SD) = 27.2 (6.2)	
19 or less	102 (11.3)
20-24	224 (24.9)
25-29	240 (26.7)
30-34	214 (23.8)
35 and older	119 (13.2)
Paternal age (years), Mean (SD) = 30.2 (7.1)	
19 or less	33 (3.9)
20-24	175 (20.7)
25-29	215 (25.4)
30-34	198 (23.4)
35-39	137 (16.2)
40+	89 (10.5)
Missing	52
Maternal education (years)	
8 or less years	93 (11.9)
9-11 years	162 (20.7)
12 years	235 (30.1)
13 to 15 years	155 (19.8)
16 more years	136 (17.4)
Missing	118
Paternal education (years)	
8 or less years	109 (14.9)
9-11 years	107 (14.6)
12 years	229 (31.2)
13 to 15 years	146 (19.9)
16 more years	143 (19.5)
Missing	165
Mother's place of birth	
Mexico	237 (26.4)
US	487 (54.2)
Other foreign	174 (19.4)
Missing	1
Parity	
0	374 (41.6)

1	266 (29.6)
2	147 (16.4)
3	72 (8.0)
4 or more	39 (4.3)
Missing	1
Maternal smoking	
Any smoking during pregnancy (2007+)	
Yes	2 (1.6)
No	124 (98.4)
Missing	2
Any smoking in 3 months before pregnancy (2007+)	
Yes	3 (2.4)
No	123 (97.6)
Missing	2

Table 2-2. Selected nutritional and health indicator metabolites and xenobiotic metabolites of nicotine with identification confidence level (Schymanski EL; Jeon J; Gulde R; Fenner K; Ruff M; Singer HP; Hollender (2014) Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence. J Environ. Sci. Technol 48: 2097–8)

m/z	time	Metabolite	Classification	Adduct form	Mode	Confidence
90.0550	52.4	Alanine	Amino Acid	M+H	HILIC+	4
116.0707	64.5	Proline	Amino Acid	M+H	HILIC+	1
120.0656	60.6	Threonine	Amino Acid	M+H	HILIC+	4
126.0220	56.1	Taurine	Amino Acid	M+H	HILIC+	1
132.1020	40.8	Leucine/Isoleucine	Amino Acid	M+H	HILIC+	1
145.0982	42.0	Lysine	Amino Acid	M-H	C18-	4
146.0459	19.0	Glutamate	Amino Acid	M-H	C18-	1
154.0623	23.2	Histidine	Amino Acid	M-H	C18-	1
166.0848	45.6	Phenylalanine	Amino Acid	M+H	HILIC+	4
173.1040	124.4	Arginine	Amino Acid	M-H	C18-	4
182.0811	46.0	Tyrosine	Amino Acid	M+H	HILIC+	1
150.0584	48.5	Methionine	Amino Acid	M+H	HILIC+	1
203.0827	21.9	Tryptophan	Amino Acid	M-H	C18-	1
147.0766	77.1	Glutamine	Amino Acid	M+H	HILIC+	1
104.0707	58.8	2-Aminobutyrate	Amino Acid Metabolites	M+H	HILIC+	1
180.0653	44.2	Hippurate	Amino Acid Metabolites	M+H	HILIC+	1
209.0922	51.0	Kynurenine	Amino Acid Metabolites	M+H	HILIC+	1
132.0766	53.7	Creatine	Health Indicators	M+H	HILIC+	1
369.3513	25.3	Cholesterol	Health Indicators	M-H2O+H	HILIC+	1
124.0586	273.2	Niacin	Vitamins Coenzymes	M+H	HILIC+	4
137.0716	48.3	Methylnicotinamide	Vitamins Coenzymes	M+H	HILIC+	4
169.0965	68.4	Pyridoxamine	Vitamins Coenzymes	M+H	HILIC+	1
170.0809	220.7	Pyridoxine	Vitamins Coenzymes	M+H	HILIC+	4
162.1125	41.5	Carnitine	Fatty Acid Metabolism	M+H	HILIC+	4
204.1230	34.3	Acetyl-carnitine	Fatty Acid Metabolism	M+H	HILIC+	4
277.2176	225.5	a-Linolenic acid	Fatty Acid Metabolism	M-H	C18-	1
284.2575	271.0	Oleic acid	Fatty Acid Metabolism	M-H	C18-	4
311.2960	291.3	Arachidic acid	Fatty Acid Metabolism	M-H	C18-	1
258.1091	113.1	Glycero-3-Phosphocholine	Lipid Metabolism	M+H	HILIC+	1
300.2896	21.8	Sphingosine	Lipid Metabolism	M+H	HILIC+	1
302.3053	20.9	Sphinganine	Lipid Metabolism	M+H	HILIC+	1
104.1071	107.9	Choline	Lipid Metabolism	M+H	HILIC+	4
524.3710	27.5	LysoPC(18:0)	Lipid Metabolism	M+H	HILIC+	1
115.0497	264.4	5,6-Dihydrouracil	Nucleotide metabolism	M+H	HILIC+	4
137.0458	40.7	Hypoxanthine	Nucleotide metabolism	M+H	HILIC+	1
167.0208	19.1	Urate	Nucleotide metabolism	M-H	C18-	1
153.0407	246.5	Xanthine	Nucleotide metabolism	M+H	HILIC+	4
121.0294	284.6	Benzoic acid	Exogenous Chemical	M-H	C18-	4
195.0876	31.8	Caffeine	Exogenous Chemical	M+H	HILIC+	1
177.1023	31	Cotinine	Xenobiotic metabolites of nicotine	M+H	HILIC+	1
193.0973	31	Hydroxycotinine	Xenobiotic metabolites of nicotine	M+H	HILIC+	1

Table 2-3. Factors contributing to sample quality
Uncontrollable
Time and location dependent supplies
Site-specific standard operating procedures for collection, processing and delivery
Duration of storage
Sample-specific differences in lipid content and hematocrit impacting stability of metabolites during blood spot drying and delivery
Sample-specific resolubilization of metabolites/loss of metabolites due to free-radical polymerization and protein trapping
Controllable
LC-MS-specific variation in operation and analytical drifts
Data extraction software-dependent variations in feature identification and intensity determination

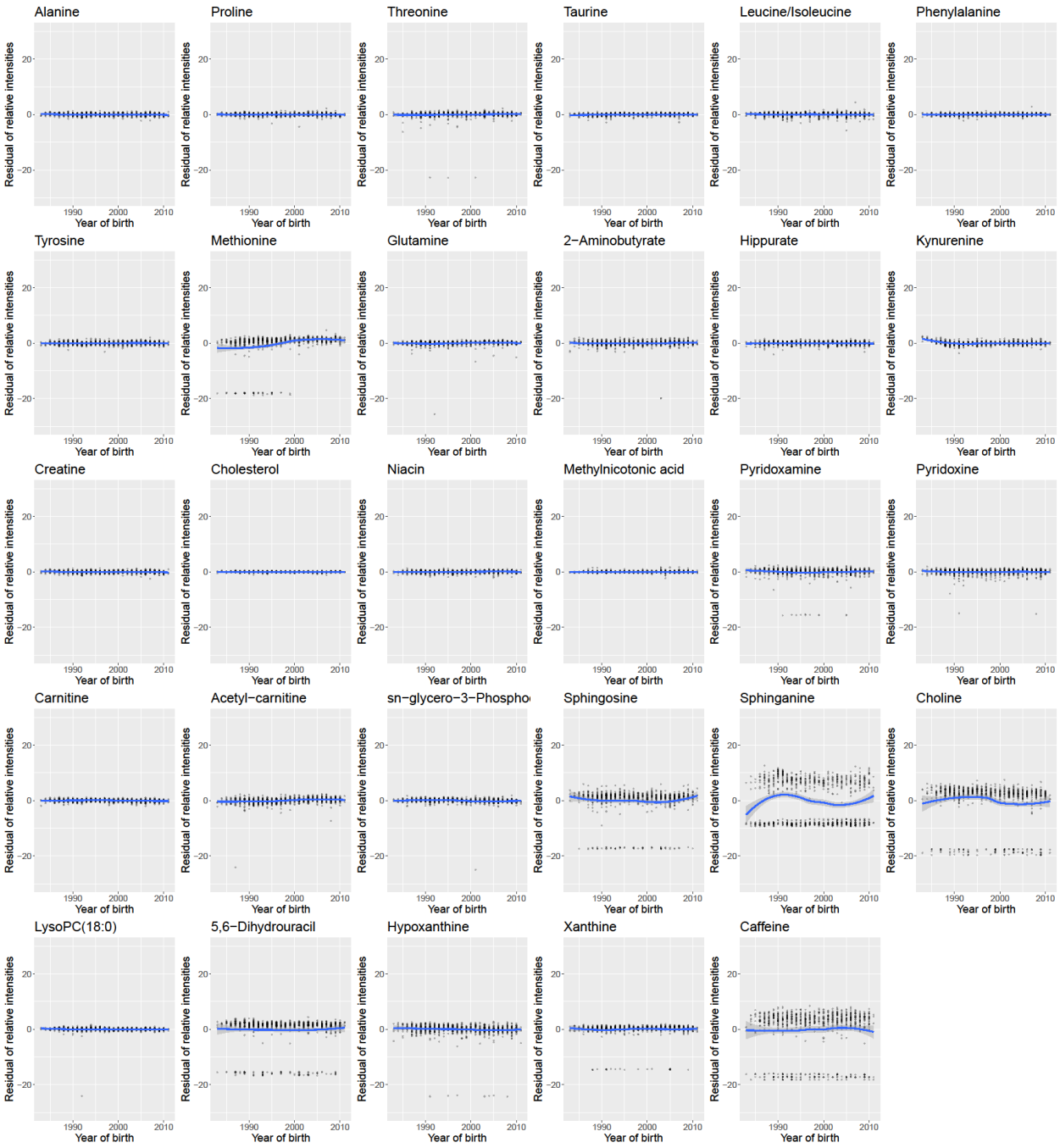


Figure 2-1. Log2-transformed residual of relative ion intensities of selected metabolites detected in HILIC column, adjusted for Hispanic ethnicity and maternal age

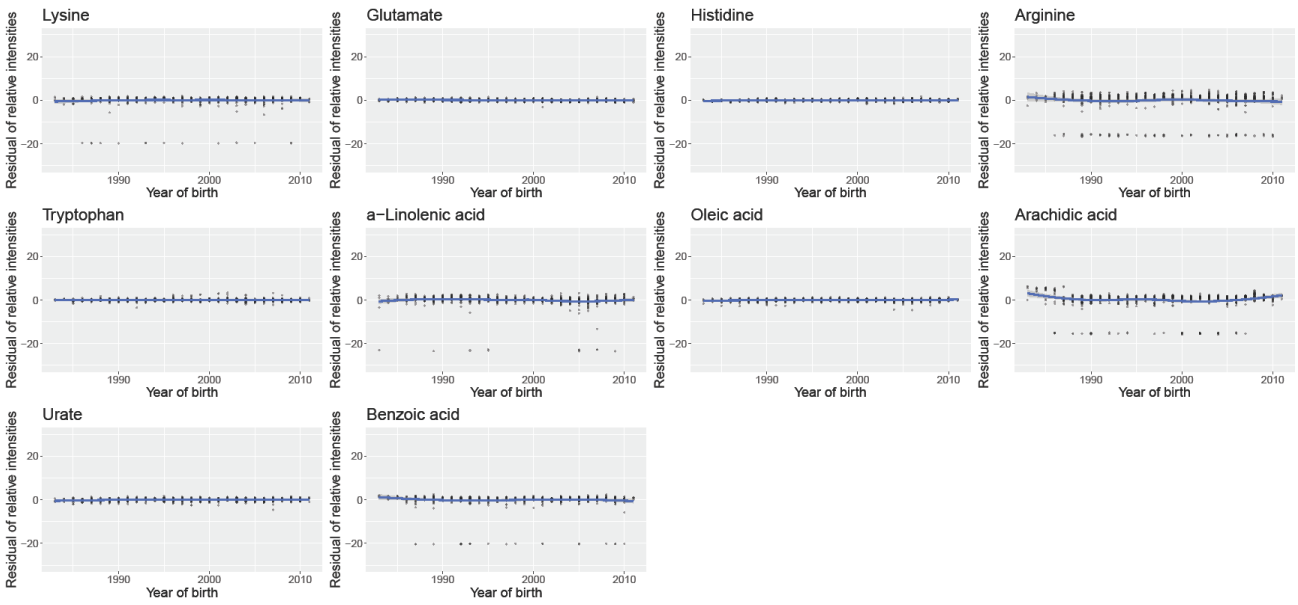


Figure 2-2. Log2-transformed residual of relative ion intensities of selected metabolites detected in C18 column, adjusted for Hispanic ethnicity and maternal age

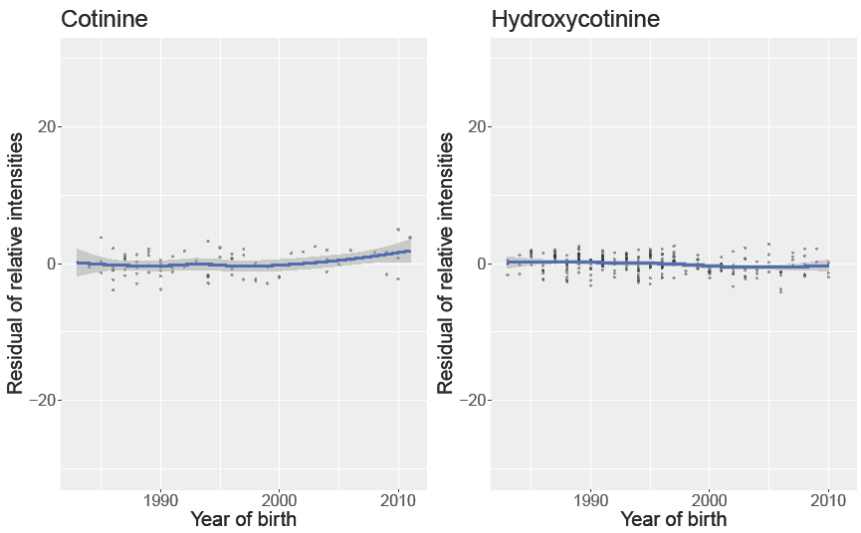


Figure 2-3. Log2-transformed residual of relative ion intensities of cotinine and hydroxycotinine, adjusted for Hispanic ethnicity and maternal age

2.7 Supplemental materials

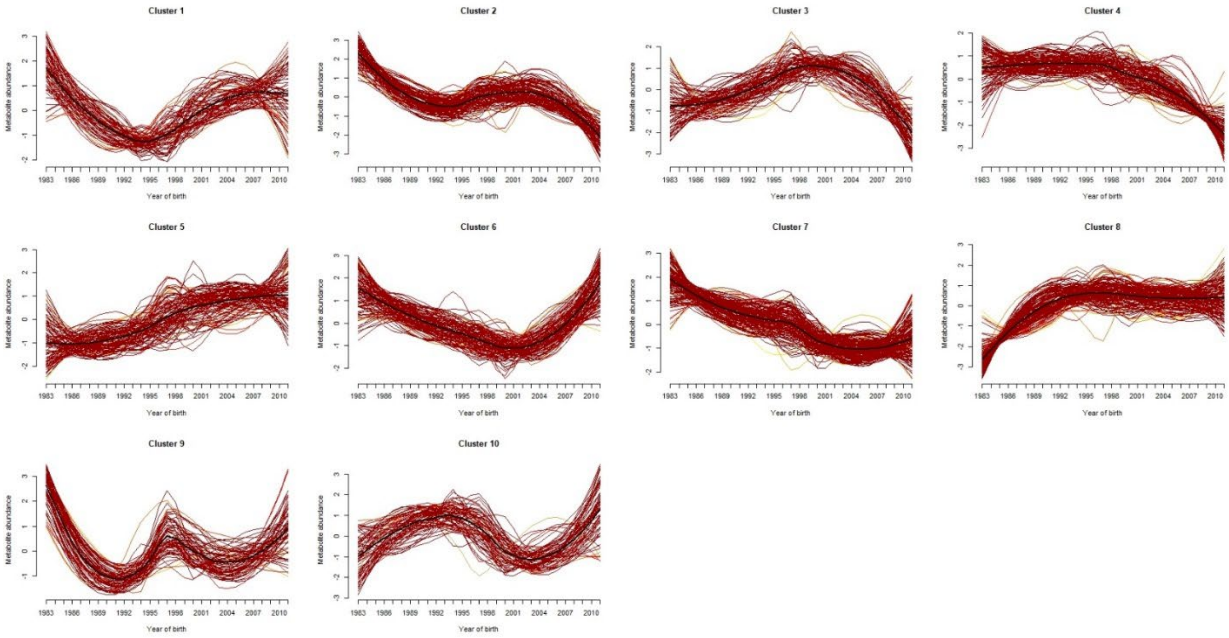
Supplemental Table 2-S1. Suggested m/z matches annotation for endogenous features (uploaded separately)

Supplemental Table 2-S2. Exogenous features detected in HILIC column with suggested m/z matches annotation and cluster IDs (uploaded separately)

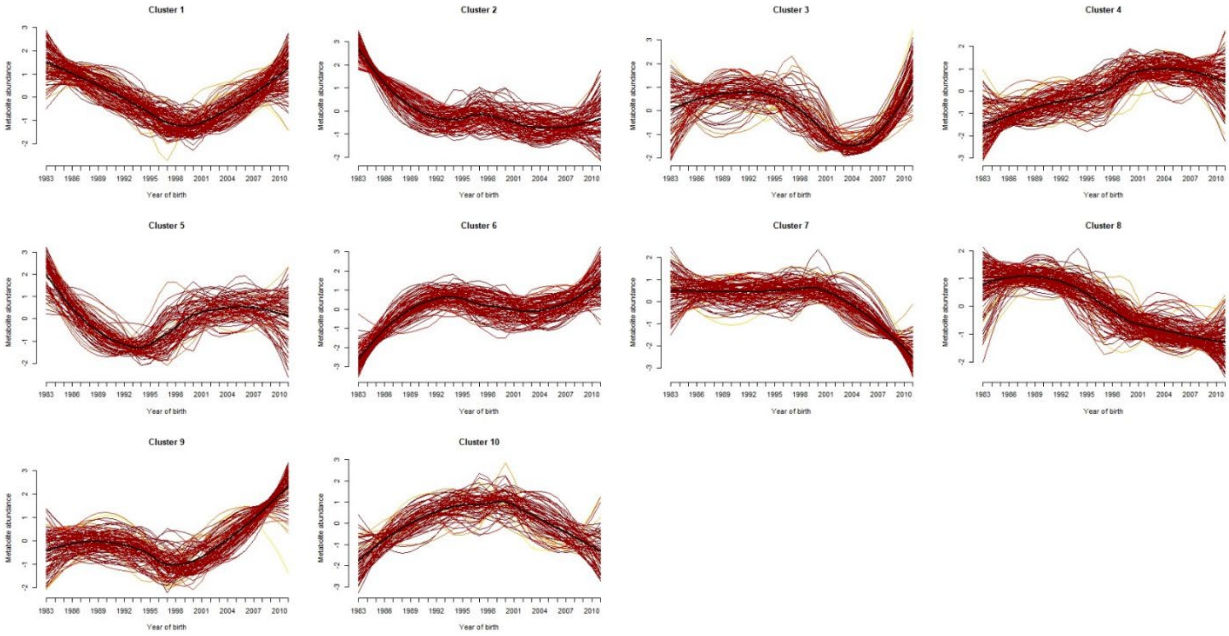
Supplemental Table 2-S3. Exogenous features detected in C18 column with suggested m/z matches annotation and cluster IDs (uploaded separately)

Supplemental Figure 2-S1. Endogenous Features in HILIC column with suggested m/z matches annotation (uploaded separately)

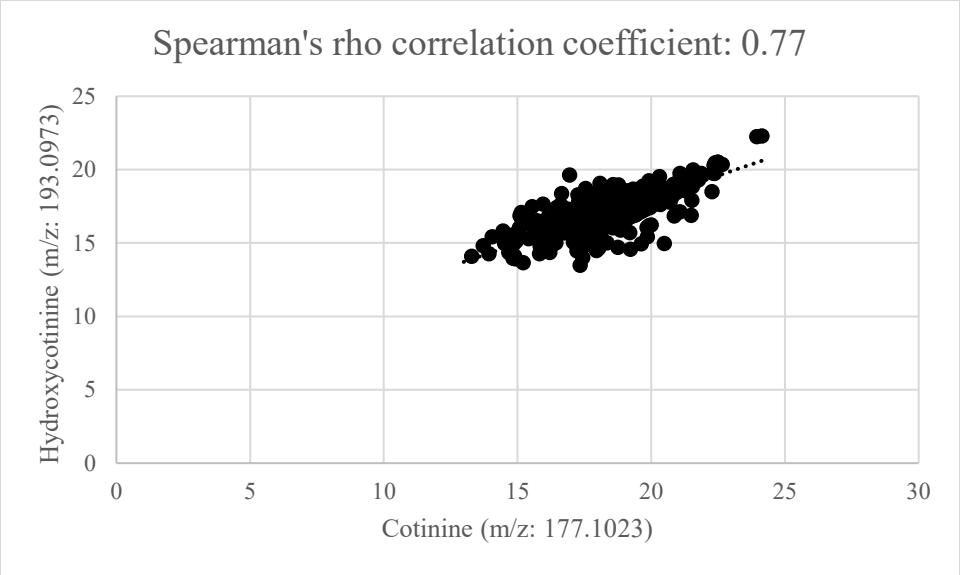
Supplemental Figure 2-S2. Endogenous Features in C18 column with suggested m/z matches annotation (uploaded separately)



Supplemental Figure 2-S3. FCM clustering for exogenous features in HILIC column. Features were fitted to a loess curve and Z-score scaled (red lines), adjusted for Hispanic ethnicity and maternal age, as a function of birth year. Averaged trend is shown in black lines.



Supplemental Figure 2-S4. FCM clustering for exogenous features in C18 column. Features were fitted to a loess curve and Z-score scaled (red lines), adjusted for Hispanic ethnicity and maternal age, as a function of birth year. Averaged trend is shown in black lines.



Supplemental Figure 2-S5. Spearman's rho correlation of log2-transformed cotinine and hydroxycotinine detected in the DBS

Chapter 3. A prediction model for classifying maternal pregnancy tobacco exposure using California state birth certificate information

3.1 Abstract

Background. Many studies have linked maternal smoking to adverse birth outcomes. However, assessing maternal smoking status in interviews or from birth certificates underestimates true smoking rates due to the stigma attached. Here, we built a prediction model for maternal tobacco smoking in pregnancy based on birth certificate information using a combination of self- or provider-reported smoking and biomarkers (smoking metabolites) in neonatal blood spots as the alloyed gold standard.

Methods. We relied on a population-based case-control study that included 894 cancer-free children by the age of 6 born in California between 1983 and 2011 to perform high-resolution metabolomics analyses and extract cotinine (*m/z* = 177.1023) and hydroxycotinine (*m/z* = 193.0973) from newborns' neonatal dried blood spots. Potential predictors of smoking were selected from the California birth certificate. Logistic regression with stepwise backward selection was used for prediction model building. Model performance was evaluated in the training sample, a bootstrapped sample, and an external validation sample.

Results. Out of seven predictor variables entered into the logistic model, five were selected by the stepwise procedure, including maternal race/ethnicity, maternal education, child's birth year, parity, and child's birth weight. We calculated an overall discrimination accuracy of 0.724 and an AUC of 0.805 (0.770-0.839) in the training set. Similar accuracies were achieved in the internal and external validation sets.

Conclusions. This easy-to-apply model may benefit future birth registry-based studies in California when there is missing maternal smoking information; however, some smoking status

misclassification remains a concern when only variables from the birth certificate are used to predict maternal smoking.

3.2 Introduction

Prevalence of current cigarette smoking among US adults has decreased across the past 5 decades, from 51% in 1965 to 31% in 2012 among males, and 34% to 23% among females.[107] For pregnant women, the prevalence of smoking decreased from 25.7% in 1985 to 10.1% in 2014.[108] Among California women, interview-based studies revealed that maternal pregnancy smoking prevalence declined from approximately 50% in the late 1970s [13] to 15% from 1995-2002 [14] and 5% from 2008-2018.[15] Tobacco smoke contains multiple toxic compounds that can cross the placenta after either direct or passive exposure to cigarettes.

As it is a risk factor for multiple pediatric diseases, maternal pregnancy smoking is an important confounding variable to be adjusted for in studies of maternal exposures and child outcomes, such as childhood cancers.[22] Record-linkage studies are advantageous due to their (often) population-based nature, larger sample size, and lower cost, but frequently have limited or no information on lifestyle factors such as smoking. Systematically recorded smoking data may not be available in vital statistics records for some periods or in some regions possibly resulting in uncontrolled confounding.

Even when records collect smoking status, the quality of the data might be a concern as self-reported data on pregnancy smoking may be unreliable due to social stigma, thus, missing data are not missing at random. Two studies in the United States in the 2000s and 2010s showed that approximately 14% of smokers deny smoking in interviews.[23,24] Compared with smoking data from other sources including CDC surveys and medical records, it has been reported that birth

certificates only capture 70.6-82.0% of prenatal smoking.[25-27] Maternal smoking may be underreported due to several reasons: women fail to disclose their smoking to their provider; providers failed to accurately record the information in the medical record; the medical record was not received by the hospital; or inaccurate input when clerks fill out the birth certificate.[26]

Given concerns about the accuracy of recorded smoking on the birth certificate, measurement of tobacco smoking using a biomarker of exposure is preferable in research studies. For pregnant women close to the time of delivery, cotinine levels in newborns' dried blood spots can be used as a biomarker of maternal smoking, with high sensitivity (92.3%) and specificity (99.7%).[28,29] Cotinine is the best marker of recent smoking including active and secondhand smoke.[30,31] However, although cotinine has a longer half-life than its parent compound, nicotine, it is only detectable for about 17 hours.[32] In adults, cotinine is the primary metabolite of nicotine and half of the cotinine is subsequently metabolized into hydroxycotinine.[109] While hydroxycotinine has a slightly longer half-life of 19 hours, it has other sources (e.g., dietary sources such as tomatoes and eggplant).[35,103]

Clinical prediction models have been used to predict tobacco-related disease,[62] calibrate self-reported maternal smoking using urinary cotinine,[63] and to characterize smoking patterns in older adults.[64] Here, we generated a clinical prediction model [61] for maternal late pregnancy smoking based on variables routinely collected on birth certificates and a combination of self- or provider-reported smoking information and metabolomics-derived smoking biomarkers together used as the alloyed gold standard for maternal smoking.

3.3 Methods

Study design and participants

We utilized data from the Smoking and Embryonal Tumor Study, a large population-based case-control study of childhood cancers, which ascertained cancer cases from the California Cancer Registry and included controls from among births in California between 1983 and 2011 (N = 1400). For the present study, we included 899 participants selected as controls and were cancer-free by the age of 6. Covariates were obtained from California birth certificates, including mothers' tobacco use information. From 1989 to 2005 medical providers collected some limited smoking information during pregnancy by answering to the question "Were there pregnancy complications due to tobacco use during pregnancy?". Starting in 2007, the "number of cigarettes per day (3 months before pregnancy and during each trimester)" was reported on the California birth certificate.[22]

We obtained neonatal dried blood spots from the Genetic Disease Screening Program (GDSP).[110] The GDSP collected blood samples from babies' heel-sticks between 12 and 48 hours after birth.[111] Blood samples were put on six small circles on a specialized filter paper, dried at room temperature for at least 3 hours, and shipped to the Neonatal and Prenatal Screening Laboratory within 24 hours of collection.[111] After screening, left-over specimens were packed and stored at -20°C.[111] Neonatal blood spots were analyzed using liquid chromatography with ultra-high resolution mass spectrometry.[112] Additional details concerning dried blood specimen and metabolomics analysis were previously described.[113] After excluding 6 samples considered outliers, 894 subjects were left for the analysis.

Outcome definition

We extracted cotinine (*m/z* = 177.1023) and hydroxycotinine (*m/z* = 193.0973) from the feature table generated by xMSanalyzer.[75] As a large volume of blood is needed to detect enough metabolites to discern secondhand smoke exposure,[29] we considered both metabolites' presence

to be evidence of maternal active smoking. We considered mothers to be active smokers if they were self- or provider-reported smokers on birth certificates or if we detected cotinine or hydroxycotinine in the newborn's blood at intensities greater than the 86th percentile, a cut-off chosen based on the prevalence reported by contemporaneous surveys of California women across our study period.[13-15,114] This yielded an overall mean prevalence of smoking of 17% from any smoking indicators.

Predictors

We identified potential predictors as those that were established or possible risk factors of maternal smoking and were associated with maternal smoking when evaluated in univariate analysis in our data. We excluded candidate predictors that with missing data over 10% such as factors that were only recorded during a short period to minimize possible misclassifications when imputing missing values for the predictors.[115] Also, to improve model performance for predicting a rare event, we followed the events per variable (EVP) rule that recommends at least 10 individuals need to have the event of interest for every predictor variable included in the model.[115] After excluding variables with missingness more than 10%, with very unbalanced distributions, only available for a short period, or highly correlated, we selected 7 variables as input predictors. These included maternal race/ethnicity (White non-Hispanic, Hispanic any race, Black, Asian/Pacific Islander, other), maternal age at childbirth (continuous), parity (0, 1, or 2+), gestational age in weeks (preterm birth ≤ 36 weeks, term births > 36 weeks), birth weight in grams (1499 or less, 1500-2499, 2500-3999, 4000+), child birth year (1983-1991, 1992-1996, 1997-2001, 2002-2011), and maternal education in years (≤8, 9-11, 12, 13-15, 16+).

Statistical analysis

Missing values of predictors were imputed with the R package "missForest". As the outcome is binary, we then utilized a stepwise logistic prediction model for model specification and estimation. We fitted a full model and then performed backward selection to generate a simplified model. We estimated the coefficients with maximum likelihood methods. Model performance was assessed using two criteria: calibration and discrimination. We fitted the calibration plot and used calibration plot intercept and calibration slope to describe model fit. We reported accuracy, area under the receiver operating characteristic (ROC) curve (AUC) and 95% confidence interval (CI), sensitivity, specificity, and sample-level positive predicted value (PPV) and negative predicted value (NPV) to evaluate the ability of our model to classify outcomes correctly. The threshold of classifying smokers vs. non-smokers was determined by the "closest.topleft" methods, which set the optimal threshold to be the point closest to the top-left part of the plot representing perfect sensitivity or specificity. Internal validation was performed with bootstrap methods and accuracy was reported.

To assess the external validity of the model, we obtained data from the UCLA Environment and Pregnancy Outcomes Study (EPOS).[116] EPOS is a case-control study nested within the 2003 birth cohort of all women residing in Los Angeles County. Cases of preterm birth and low birthweight were randomly selected from zip codes near an air pollution monitoring station. Controls were matched to cases by birth month and zip codes. A total of 2543 subjects were enrolled in the EPOS with a response rate of 40%. In addition to variables available on California birth certificates, the EPOS survey questionnaire collected information on lifestyle factors including smoking status (pregnancy smoker, former smoker, never smoker). After excluding subjects missing potential predictors or outcome information, there were 2493 subjects in the

EPOS population. We conducted sensitivity analysis limiting to children born in 2003 in our data and compared the model performances in the subset with that in EPOS data.

We performed additional sensitivity analyses to examine model performance in different settings. To assess whether additional eligible predictors available only during a shorter period would enhance the model performance, specifically, we added the month when mothers started prenatal care and neonatal intensive care unit admission as predictors and tested model performance. Given the large drop in smoking prevalence across the study period, we split the sample into two periods (birth year ≤ 1996 vs. > 1996) and tested how the model performed during the earlier period with much higher smoking rates versus during the later period with reduced rates.

3.4 Results

Overall, 149 women were defined as smokers, including 17 mothers identified by self-report or had tobacco-related pregnancy complications reported on the birth certificate. Specifically, 15 mothers were recorded to have pregnancy complications related with tobacco use (1989-2005) and 2 reported smoking cigarettes during pregnancy (2007-2011). Demographic characteristics of the study population are shown in Table 3-1. Compared to non-smoking mothers, smokers were more often White non-Hispanic (47.0% smokers vs. 31.3% non-smokers) or Black (10.1% smokers vs. 5.1% non-smokers) and less likely to be Hispanic (33.6% smokers vs. 49.9% non-smokers) or Asian (7.4% smokers vs. 11.0% non-smokers). Compared with non-smoking mothers, smokers were also more often younger than 29 years of age when the index child was born (69.8% smokers vs. 61.6% non-smokers), living in low to medium socioeconomic status (SES) neighborhoods (77.1% smokers vs. 70.6% non-smokers), and high school graduates (72.5% smokers vs. 55.3% non-smokers). Children whose mothers were smokers were less likely to be firstborn (37.6% smokers vs. 42.2% non-smokers) and born after the year 2002 (10.1% smokers

vs. 41.1% non-smokers), while they were more likely to be male (56.4% smokers vs. 47.4% non-smokers). Preterm births were similarly distributed among smoking mothers and non-smokers in our population (9.2% vs. 9.9%) with relatively small numbers.

Characteristics of potential predictors in the EPOS data are reported in Supplemental Table 3-S1. Questionnaire-based pregnancy smoking prevalence was approximately 5% in EPOS population.

Out of the 7 predictors entered into the logistic model for the outcome smokers defined as having any of the smoking indicators, five remained in the final model. Important predictor variables were identified based on the maximum likelihood ratio in the logistic regression. They were maternal race/ethnicity, maternal education, child birth year, parity, and birth weight group in a descending order of importance (Table 3-2).

We selected the optimal threshold to classify smokers vs. non-smokers and evaluated the model performance in the training, bootstrapped internal validation, and external validation data sets (Table 3-3 and Figure 3-1). The optimal thresholds approximately equalled the prevalence of the outcome in each set. The overall accuracy and AUC were approximately 0.7-0.8 in all 3 samples. The maximum sensitivity reached 0.91 when having a specificity better than chance (> 0.5), while the maximum specificity was 0.86 when sensitivity was required to be above 0.5.

Adding additional predictors into the model did not appreciably enhance the discrimination performance compared with the final stepwise model selected (AUC 0.811 vs. 0.805). Supplemental Table 3-S2 showed the performance of our model in 3 validation data sets containing children born in different time periods. We reached a higher AUC of 0.791 in a subgroup with lower smoking prevalence than during the period when heavy smoking was more prevalent (AUC=0.660). In the validation data set with births in 2003 only, two mothers were identified as

smokers and both were correctly classified by the model. Among non-smokers, 36 out of 46 were correctly classified, which yielded an AUC of 0.913 (95% CI 0.771-1.000).

3.5 Discussion

Our study utilized self- and provider reported birth registry data on maternal smoking in pregnancy in combination with smoking biomarker data from neonatal blood spots and built a maternal tobacco use in late pregnancy prediction model that uses variables from birth certificates. Five demographic characteristics predicted maternal pregnancy smoking in our sample. Compared with non-smokers, smoking mothers tended to be more often White non-Hispanic, less educated, had children in the early period of the study, were multipara, and gave birth to infants with lower birth weight. This model can be used to predict smoking in order to at least partially account for uncontrolled confounding due to missing maternal smoking information when conducting studies using California birth certificates.

Research studies of maternal and child health often rely on registry data, especially from state administered birth certificates. Nicotine metabolites, cotinine and hydroxycotinine, have been used as biomarkers of maternal smoking in newborn dried blood spots but these are resource intensive and costly to analyze. We combined the metabolite-derived biomarker measures and the data on birth certificates to build a prediction model that can help future studies using state-based birth registry data to predict maternal smoking status. The model can be applied to California state-based studies, even for years when maternal smoking was not collected on the birth certificate. In the United States and globally, pregnancy smoking prevalence and uptake are higher among disadvantaged groups who usually have low socioeconomic status (SES) and/or are related to racial/ethnic disparities.[117-119] In addition, smoking co-occurs with other environmental or lifestyle exposures or adverse health conditions, such as ambient air pollution, maternal underweight, small

gestational age, and drug abuse.[120,121] Thus, lack of ascertainment of smoking as well as other unhealthy lifestyle characteristics might lead to bias in effect estimates due to uncontrolled confounding. Therefore, our study may enhance the use of existing birth data by helping to predict maternal smoking status missing on many birth certificates due to changes in data collection or lack of reporting and to obtain a more reliable proxy for this important confounding variable needed for adjustment in many maternal and child health studies.

Moreover, our model is validated and easy to apply to other studies utilizing California birth certificates. We used stepwise logistic regression to predict a binary maternal smoking variable and obtained overall discriminant accuracy of around 70%-80% for apparent model performance, internal validation, and external validation. Previous studies identified pregnancy smoking or pregnancy smoking cessation predictors using univariate analysis usually aimed to identify clinically meaningful intervention groups. Some also had detailed smoking behavior information available as predictors such as the frequency, amount, and duration of smoking and even the second hand smoke exposure.[118,122,123] However, these models were not validated against internal and external testing data set and these smoking predictors are rarely if ever available in routine register based data, which limits the transportability of these models outside of the study population. This even applies to a study that randomly split the sample into a training and testing data set and reached an overall percentage of correct smoking status classification of 76%. Their model used predictors such as interview-based smoking behavioral variables and thus is not applicable to studies where there was missing maternal smoking information.[123]

Researchers have shown that demographic characteristics including maternal age, maternal race/ethnicity, maternal education, parity, prenatal care, household income, unemployment, and antenatal stress are related with maternal smoking.[117-119,124] In addition, maternal cigarette smoking

during pregnancy increases the risk for pregnancy complications and poor pregnancy outcomes, including preterm delivery, restricted fetal growth, and low birth weight.[119,125] Our model identified the same predictors available to us on birth certificates overlapping with this list and reached good prediction performance with the fewer number of predictors.

Our predictor selection process balanced adding all potential predictors and making the model easy to apply in future studies. We did not include SES or prenatal care information (month when prenatal care began, principal source of payment for prenatal care, number of prenatal care visits) in the model input. Although smokers were more likely to live in neighborhoods with lower SES (Table 1), census-based neighborhood SES was not originally available on the birth certificate and was created in our sample through a geocoding process. Moreover, SES of the neighborhoods changes over time and one would have to generate a time-varying SES variable for usage in future studies. Thus, we decided to leave out potential predictors like prenatal care, pregnancy complications and labor complications because they were only collected for a shorter period or had very few events recorded, including the complications placenta previa, abruptio placenta, premature rupture of membranes, and neonatal intensive care unit admission. We added prenatal care and neonatal intensive care unit admission as predictors in a sensitivity analysis as they had relatively few missing data points and an adequate number of events happened, but the model performance was very similar to our selected model. One variable that was eliminated during model simplification (from full model to final stepwise model) was preterm birth. This may be due to the fact that it is highly correlated with birthweight, a more complete variable.

Our model has a relatively good discrimination performance comparable with other smoking prediction models and is easy to apply in California birth data. A study that used a much more complex decision tree model and a multinomial logistic regression model to characterize the

smoking patterns among seniors obtained classification accuracies of 0.671 and 0.501, respectively.[64] Another study employed machine learning methods to predict smoking cessation and their best model reached an accuracy of 0.640 and an AUC of 0.660.[126] The performance of our model was the highest in the training set and slightly lower in the external validation set. Moreover, when comparing the model performance in subsets collected in earlier and later years, the model performed better among children born after 1996 when smoking became less prevalent compared to in earlier years, not only according to the negative but also the positive predictive value. This may reflect the fact that demographic variables can better predict maternal smoking status in more recent years when social unacceptability of smoking increased and smoking became more prevalent in socially disadvantaged individuals.

We acknowledge limitations in our study. First, although much more reliable than self-reported smoking, the gold standard for our smoking outcome still is an imperfect measure. Neonatal dried blood spots are typically collected from the infant during the postpartum hospital stay, or by trained midwives for out-of-hospital births, though the latter accounts for less than 1% of California births.[127,128] Only if the mother smoked close to giving birth or after birth and attempted breastfeeding would we have been able to capture smoking in neonatal biomarkers. In California, 67%-86% of mothers attempted any breastfeeding during 2004 to 2008, while as many as 90% initiated breastfeeding in 2010.[96,129] After smoking a cigarette, about 10% of its nicotine and nicotine metabolites (cotinine, hydroxycotinine, and cotinine *N*-oxide) pass through the epithelial cells of mammary glands into breast milk. Prolonged labor may also affect the transmission of nicotine metabolites from maternal blood to offspring due to its short half-life, but we expect that effect to be minimal as less than 1% of women in our sample had prolonged labor. Second, we had to eliminate variables that may predict pregnancy smoking for various reasons.

Adding more predictors might have helped the model to achieve a better AUC but the model would become less universally applicable. Third, the EPOS validation data set we relied on only captured births in 2003 in Los Angeles County and had a 40% response rate. However, we previously conducted extensive analyses comparing responders, non-responders, and the overall birth cohort and observed few differences in the characteristics of interest to this study.[116] In addition, the EPOS data is representative of most of the LA population as it was sampled from birth records in a population-based manner rather than being a select sample of women recruited from prenatal clinics and hospitals.[130]

In conclusion, we constructed a prediction model using birth certificate data to predict maternal smoking reaching an AUC of 0.805. The model may benefit future studies in the state that employ birth registry data. It can help to improve adjustment for maternal smoking and may also help guide clinical intervention to reduce smoking in pregnancy in California.

3.6 Tables and figures

Table 3-1. Demographic characteristics of the study population (N = 894)

	Non-smoker (N=745)	Smoker (N=149)
Birth year		
1983-1991	147 (19.7%)	84 (56.4%)
1992-1996	145 (19.5%)	35 (23.5%)
1997-2001	147 (19.7%)	15 (10.1%)
2002-2011	306 (41.1%)	15 (10.1%)
Sex		
Male	355 (47.7%)	84 (56.4%)
Female	390 (52.3%)	65 (43.6%)
Birth weight		
1499 g or less	10 (1.3%)	2 (1.3%)
1500-2499	26 (3.5%)	15 (10.1%)
2500-3999	624 (84.0%)	118 (79.2%)
4000 +	83 (11.2%)	14 (9.4%)
Missing	2	0
Maternal age		
<20	82 (11.0%)	18 (12.1%)
20-24	182 (24.4%)	42 (28.2%)
25-29	195 (26.2%)	44 (29.5%)
30-34	179 (24.0%)	35 (23.5%)
35+	107 (14.4%)	10 (6.7%)
Maternal race/ethnicity		
White non-Hispanic	233 (31.3%)	70 (47.0%)
Hispanic	372 (49.9%)	50 (33.6%)
Black	38 (5.1%)	15 (10.1%)
Asian/PI	82 (11.0%)	11 (7.4%)
Other	20 (2.7%)	3 (2.0%)
Parity		
0	314 (42.2%)	56 (37.6%)
1	230 (30.9%)	37 (24.8%)
2+	200 (26.9%)	56 (37.6%)
Missing	1	0
Preterm birth		
Preterm	71 (9.9%)	13 (9.2%)
Term	646 (90.1%)	129 (90.8%)
Missing	28	7
Census-based neighborhood SES level		
1 (low)	161 (21.7%)	34 (22.8%)
2	208 (28.1%)	40 (26.8%)
3	154 (20.8%)	41 (27.5%)
4	116 (15.7%)	20 (13.4%)
5 (high)	102 (13.8%)	14 (9.4%)
Missing	4	0
Maternal education		
8 or less years	80 (11.9%)	12 (11.8%)
9-11 years	137 (17.9%)	23 (22.5%)
12 years	195 (25.5%)	39 (38.2%)
13 to 15 years	134 (17.5%)	21 (20.6%)
16 more years	128 (16.8%)	7 (6.9%)
Missing	71	47

Table 3-2. Maternal smoking logistic regression prediction model using birth data, outcome defined as having any of the smoking indicators (N = 894)

Variable	Odds ratio (95% CI)	P value	Coefficient
Maternal Race/Ethnicity			
White non-Hispanic		Reference	
Hispanic	0.36 (0.20, 0.64)	0.001	-1.011
Black	1.14 (0.52, 2.41)	0.729	0.134
Asian/PI	0.58 (0.26, 1.20)	0.158	-0.543
Other	0.93 (0.20, 3.26)	0.918	-0.071
Maternal Education			
8 or less years		Reference	
9-11 years	1.38 (0.69, 2.80)	0.373	0.319
12 years	1.06 (0.50, 2.24)	0.880	0.057
13 to 15 years	0.73 (0.33, 1.64)	0.446	-0.313
16 more years	0.22 (0.07, 0.62)	0.006	-1.530
Birth year			
1983-1991		Reference	
1992-1996	0.46 (0.28, 0.74)	0.002	-0.780
1997-2001	0.18 (0.09, 0.33)	0.000	-1.724
2002-2011	0.09 (0.05, 0.16)	0.000	-2.428
Parity			
0		Reference	
1	0.90 (0.54, 1.46)	0.663	-0.110
2+	1.86 (1.17, 2.97)	0.009	0.620
Birth weight			
1499 g or less		Reference	
1500-2499	2.85 (0.48, 24.80)	0.284	1.049
2500-3999	0.99 (0.19, 7.77)	0.990	-0.012
4000 +	0.68 (0.12, 5.68)	0.684	-0.389

Table 3-3. Performance of the maternal smoking prediction model using the optimal threshold when classifying probabilities to binary outcome

Metrics	Training set	Bootstrapped set	External validation set (EPOS data)
Optimal threshold	0.17	0.17	0.05
Accuracy	0.724	0.724	0.667
AUC (95% CI)	0.805 (0.770, 0.839)	0.805 (0.770, 0.839)	0.689 (0.636, 0.743)
Sensitivity	0.759	0.759	0.648
Specificity	0.717	0.717	0.668
PPV	0.349	0.349	0.091
NPV	0.937	0.937	0.974

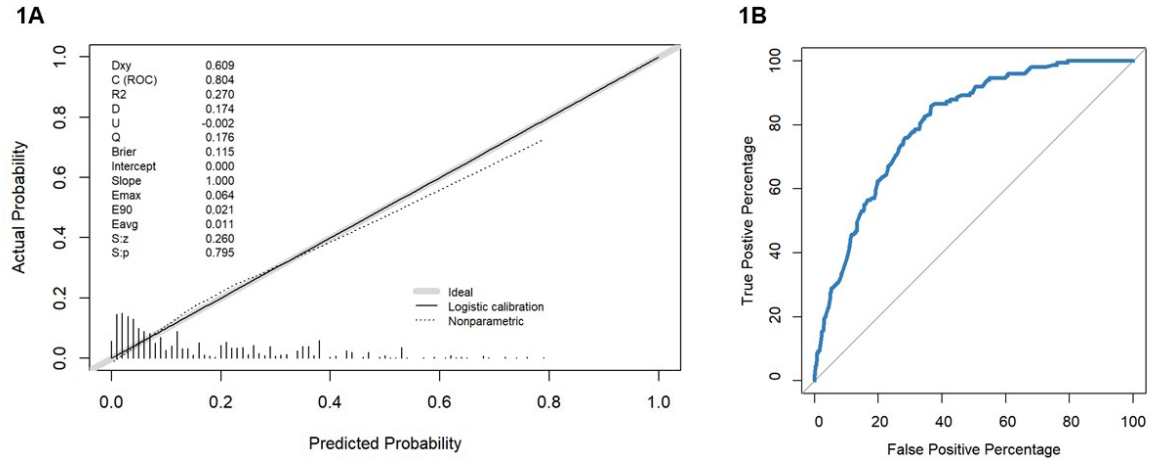


Figure 3-1. Model performance in the training set

3.7 Supplemental materials

Supplemental Table 3-S1. Demographic characteristics of predictors in the EPOS data (N = 2493)

	Non-smoker (N=2371)	Smoker (N=122)
Maternal age		
Mean (SD)	27.9 (6.39)	28.7 (6.54)
Median [Min, Max]	28.0 [13.0, 47.0]	28.5 [15.0, 43.0]
Maternal race/ethnicity		
White non-Hispanic	394 (16.6%)	34 (27.9%)
Hispanic	1625 (68.5%)	42 (34.4%)
Black	148 (6.2%)	32 (26.2%)
Asian/PI	172 (7.3%)	14 (11.5%)
Other	32 (1.3%)	0 (0%)
Parity		
0	948 (40.0%)	47 (38.5%)
1	727 (30.7%)	31 (25.4%)
2+	696 (29.4%)	44 (36.1%)
Preterm birth		
Preterm	957 (40.4%)	54 (44.3%)
Term	1414 (59.6%)	68 (55.7%)
Maternal education		
8 or less years	337 (14.2%)	4 (3.3%)
9-11 years	508 (21.4%)	28 (23.0%)
12 years	625 (26.4%)	51 (41.8%)
13 to 15 years	383 (16.2%)	19 (15.6%)
16 more years	518 (21.8%)	20 (16.4%)
Birth weight		
1499 g or less	68 (2.9%)	3 (2.5%)
1500-2499	417 (17.6%)	33 (27.0%)
2500-3999	1744 (73.6%)	76 (62.3%)
4000 +	142 (6.0%)	10 (8.2%)

Supplemental Table 3-S2. Performance of the maternal smoking prediction model in 3 validation sets containing children born in different periods

Metrics	Births ≤ 1996	Births > 1996	Birth in 2003 only
Optimal threshold	0.24	0.07	0.07
Accuracy	0.594	0.732	0.792
AUC (95% CI)	0.660 (0.590, 0.730)	0.791 (0.716, 0.865)	0.913 (0.771, 1.000)
Sensitivity	0.653	0.700	1.000
Specificity	0.576	0.734	0.783
PPV	0.324	0.151	0.167
NPV	0.842	0.973	1.000

Chapter 4. An untargeted metabolome-wide association study of maternal perinatal tobacco smoking in newborn blood spots

4.1 Abstract

Background: Maternal tobacco smoking in the perinatal period increases the risk for adverse outcomes in offspring. To better understand the biological pathways through which maternal tobacco use may have long-term impacts on child metabolism, we performed a high-resolution metabolomics (HRM) analysis in 899 newborns, following an untargeted metabolome-wide association study (MWAS) workflow.

Methods: The study population included 899 children without cancer diagnosis before age 6 and born between 1983 and 2011 in California. Newborn dried blood spots (DBS) were collected by the California Genetic Disease Screening Program between 12 hours and 48 hours after birth and stored for later research use. Based on HRM, we considered mothers to be active smokers if they were self- or provider-reported smokers on birth certificates or if we detected any cotinine or high hydroxycotinine intensities in newborn blood. We used partial least squares discriminant analysis and Mummichog pathway analysis to identify metabolites and metabolic pathways associated with maternal tobacco smoking.

Results: A total of 26,183 features (15,562 in HILIC column and 10621 in C18 column) were detected with HRM of which 1,003 were found to be associated with maternal smoking. Smoking affected metabolites and metabolic pathways in neonatal blood included vitamin A (retinol) metabolism, the kynurenine pathway, and tryptophan and arachidonic acid metabolism.

Conclusions: The metabolites and pathway perturbations associated with cigarette smoking that we identified suggested inflammatory responses and have also been implicated in chronic diseases of the central nervous system and the lung. Our results suggest that infant metabolism in the early

postnatal period reflects smoking specific physiologic responses to maternal smoking with strong biologic plausibility.

4.2 Introduction

Tobacco use is one of the most significant public health issues globally, with the number of global consumers of tobacco increasing over the past few decades.[6] In the U.S., the prevalence of current cigarette smoking among adults declined from 51% in 1965 to 31% in 2012 among males and from 34% to 23% among females.[11] It has also been reported that approximately 50% of California women smoked during pregnancy in the early 1980s [13] and this number declined to 15% in the 1990s [14] and 5% in the 2010s.[15] Despite the rate declining by more than half since 1964, cigarette smoking remains the most preventable cause of disease and death in the U.S. with disparities persisting across population groups.[16] In recent decades, smoking rates in California pregnant women have been approximately 5% lower than nationwide,[17] which is partly due to demographics as immigrant Mexican women make up about 25% of all pregnant women and most are non-smokers.[18] Components in cigarette smoke including polycyclic aromatic hydrocarbons (PAHs), N-nitrosamines, aromatic amines, aldehydes, volatile organic hydrocarbons, and metals have long been known or suspected carcinogens. Nicotine is also a modifier of cancer progression,[19] and thus may be contributing to childhood cancers due to prenatal and early childhood exposure from smoking mothers. Smoking in pregnancy has also been associated with adverse infant and child outcomes, including low birth weight, preterm birth, and congenital anomalies.[20,21]

For pregnant women close to the time of delivery, cotinine levels in their newborns' dried blood spots are a reliable biomarker of maternal smoking with high sensitivity (92.3%) and specificity (99.7%).[28,29] Cotinine best represents recent smoking,[30] with a half-life longer than its

parent compound, nicotine, and it is detectable for 15-20 hours after active smoking.[32,33] Hydroxycotinine, metabolized by the hepatic enzyme cytochrome P450 2A6, is the major metabolite of cotinine in most individuals,[33,34] and is detectable for a longer period with an approximated half-life of 18 hours.[35]

Tobacco smoking is related to metabolomic changes as recently documented for adult smokers, children exposed to passive smoke, and pregnant women.[56-58] A study of military personnel observed alterations in lipid and xenobiotic metabolism, and diverse effects on amino acid, sialic acid and purine and pyrimidine metabolism in tobacco smokers compared to non-smokers.[56] Another study among preschool children reported secondhand smoke-induced urinary metabolites including kynurenine, tyrosyl-tryptophan, and 1-(3-pyridinyl)-1,4-butanediol, peptides, and pyridines.[57] Utilizing second-trimester amniotic fluid, a third study found low-level maternal nicotine exposure from light smoking or secondhand smoke to be associated with dysregulated metabolic pathways in the fetus such as aspartate and asparagine metabolism, pyrimidine metabolism, and metabolism of other amino acids, and also saw decreases in acetylated polyamines.[58]

To better understand the biological pathways through which maternal tobacco use may be increasing the risk of disease, we performed a high-resolution metabolomics (HRM) analysis in 899 children born from 1983 to 2011 in California following an untargeted metabolome-wide association study (MWAS) workflow.

4.3 Methods

Study population

The study utilized the controls from a population-based case control study of childhood cancers [60] that enrolled 1400 children with cases ascertained from the California Cancer Registry. Population controls (20:1 matching rate) were randomly selected from California birth rolls and frequency-matched to cases by birth year. For this analysis, we only included 899 cancer-free controls.[131] Detailed demographic, socioeconomic, and gestational information for each child was collected from California birth certificates. In the years from 1989 to 2005 medical providers collected limited smoking information during pregnancy by answering the question "Were there pregnancy complications due to tobacco use during pregnancy?". Starting in 2007, the "number of cigarettes per day (3 months before pregnancy and during each trimester)" was reported on the California birth certificate.[22]

We obtained neonatal dried blood spots (DBS) for this analysis from the California Genetic Disease Screening Program. Nearly all newborns in the United States participate in a newborn screening program for genetic and metabolic diseases, and in California, the DBS are stored for research purposes after testing is complete.[36,37] Blood samples are collected from newborns by a heel-stick from the child between 12 and 48 hours after birth.[67] The blood fills six small circles on specialized filter paper and is dried at room temperature for at least 3 hours before shipment to the Neonatal and Prenatal Screening Laboratory within 24 hours of collection.[67] Since 1982, specimens left-over after the routine screening are packed and stored at -20°C. Additional details concerning dried blood specimen collection and storage are described elsewhere.[68,69]

High-resolution metabolomics

Neonatal blood spots were analyzed using liquid chromatography with ultra-high resolution mass spectrometry (LC-HRMS; Fusion, Thermo Scientific).[112] Samples were punched using a 5mm hole puncher and treated with 2:1 acetonitrile in water containing a mixture of stable

isotopic internal standards. Samples were mixed for 12 hours at 0-4 °C in the dark and then centrifuged to remove particulate matter. The resulting supernatant was analyzed in triplicate using hydrophilic interaction liquid chromatography (HILIC) with positive electrospray ionization (ESI) and C18 hydrophobic reversed-phase chromatography with negative ESI to enhance the coverage of metabolic feature detection.[132] The mass spectrometer was operated using ESI mode at a resolution of 120,000 and mass-to-charge ratio (*m/z*) range of 85-1275. Blood spot samples were analyzed in batches of 40. To evaluate system performance, we used two separate quality assessment methods. Our first Q.C. sample was NIST 1950 [133], which was analyzed at the beginning and end of the entire analytical run. The second Q.C. sample (Q-Std) included commercially purchased plasma pooled from an unknown number of males and females. Q-Std was analyzed at the beginning, middle, and end of each batch of 40 samples for normalization and batch effect evaluation. Raw data files were extracted and aligned using *apLCMS* [134] with modifications by *xMSanalyzer* [135]. Uniquely detected ions consisted of mass-to-charge ratio (*m/z*), retention time (rt), and ion abundance, referred to as metabolite features. Prior to data analysis, metabolite features were batch corrected using wavelet analysis [136]. For this analysis, we only included metabolic features with fewer than 30% missing values across all samples, with median coefficients of variation (CV) among technical replicates < 30% and Pearson correlation > 0.7.[137] Following quality assessment, replicate intensities were summarized using the median value, log2 transformed, and auto-scaled. Missing values were imputed using k-nearest neighbors (k=10) [138] imputed using the *impute* R package.

Exposure assessment

We extracted cotinine (*m/z* = 177.1023) and hydroxycotinine (*m/z* = 193.0973) from the feature table generated by xMSanalyzer. In order to indicate secondhand smoke exposure, larger

blood spots than were available to us would be needed to reach the detection limit for these metabolites.[29] Thus, we considered either metabolite to be evidence of maternal active smoking. Specifically, we considered mothers to be active smokers if they were self- or provider reported smokers on birth certificates or if the newborn blood had cotinine detected or if hydroxycotinine intensities detected were high i.e. greater than the 86th percentile chosen based on previous estimates of the prevalence of maternal smoking in pregnancy.[139]

Statistical analysis

After excluding 6 samples considered outliers and 10 samples with missing covariates, a total of 883 samples were left in the analysis. To control for potential confounding from maternal race/ethnicity (White non-Hispanic, Hispanic of any race, other), maternal age (<20, 20-24, 25-29, 30-34, >=35), birth year, infant's sex, and neighborhood socioeconomic status (SES)[140], we regressed the intensities of all metabolites other than cotinine and hydroxycotinine against confounders listed above in the following analyses.

We used a combination of univariate and multivariate analyses to identify neonatal blood spot metabolic features associated with maternal perinatal smoking. We adopted multivariate analysis partial least square - discriminant analysis (PLS-DA) to identify features that were significantly associated with maternal smoking status.[141,142] Features with Variable Importance in Projection (VIP) scores >= 2 were selected and fold changes for metabolites were calculated as the ratio of covariate-adjusted intensities comparing smokers and non-smokers. Logistic regression was then used to assess associations between discriminative metabolite features and maternal smoking. To evaluate the performance of selected features, we conducted a 10-fold cross-validation test utilizing support vector machine (SVM) and calculated the classification accuracy

of those selected features. All feature selection approaches were implemented within the R package *mixOmics* v6.3.1. We adjusted for multiple testing using false-discovery rate (FDR)-adjusted *p*-values.

Discriminative features selected by PLS-DA were first matched to a reference database of authenticated chemical standards (identification confidence level 1) previously analyzed using the same HRM.[70] The error tolerance was ±5 parts-per-million (ppm) and ±15 seconds (s) for *m/z* and retention time, respectively. Details about the reference database have been published previously.[112,143] Additional features were annotated by *xMSannotator*[144] which links to databases of metabolites (Human Metabolome Database; LipidMaps; KEGG; others). Accurate mass *m/z* for adducts formed under positive/negative ESI mode were matched to the Human Metabolome Database (HMDB), with a mass error threshold of 10 ppm. *xMSannotator* also took the correlation of intensities and retention time into consideration, which ensures accuracy of annotation. Only results with an annotation score >= 2 (scoring system 0-3, a higher score representing higher confidence) were kept.

In order to identify perturbed metabolism pathways, we conducted pathway enrichment analysis utilizing *mummichog* v 2.4.2 [145]. Features previously selected by PLS-DA VIP > = 2 and FDR-adjusted *p*-value < 0.05 were included in the pathway enrichment analysis. Pathways were considered as being statistically significantly enriched if gamma-adjusted *p*-values were smaller than 0.05.[146] We only reported pathways with a size of at least 3 metabolites per pathway detected. We performed a sensitivity analysis among children born at term only (gestational age >= 37 weeks) to determine whether there were similar perturbated pathways present among term birth neonates.

4.4 Results

The demographic characteristics of the 883 subjects included in this study are shown in Table 4-1. Compared to non-smoking mothers, smokers tended to be more often White non-Hispanic (48.4% vs. 30.8%), aged 25-34 years when the index child was born (56.9% vs. 49.7%), foreign born (62.7% vs. 52.1%), and high school graduates (61.3% vs. 49.0%). Children whose mothers were smokers were less likely to be firstborn (35.9% vs. 42.3%) and born after the year 2000 (15.7% vs. 44.9%), while they were more likely to be male (56.9% vs. 47.4%) and preterm births (11.9% vs. 9.5%).

In total, we detected 26,183 features (15,562 in HILIC column and 10621 in C18 column) with missing values in less than 30% of the samples. After discriminant analysis, we observed 520 HILICpos features and 483 C18neg features as being statistically significantly associated with maternal smoking (Figure 4-1). Among these features, we confirmed the identity of 17 metabolites using authentic standards (confidence level 1, Table 4-2). The pathway enrichment analysis selected 10 pathways that were associated with maternal smoking exposure (Table 4-3). Sensitivity analysis in which we restricted to children born at term corroborated the identification of vitamin A (retinol) and androgen and estrogen biosynthesis pathways and identified other amino acid-related pathways. (see Supplemental Table 4-S2) We listed the annotated metabolites within each enriched pathway in the Supplemental Table 4-S1.

4.5 Discussion

Our study is one of the first to examine metabolic signals in newborn blood due to maternal perinatal smoking, importantly we utilized a very large population-based sample of California children born over almost 3 decades. High-resolution metabolomic profiling methods allowed us

to identify newborn blood metabolome perturbations due to tobacco smoke exposure and our results suggest the involvement of vitamin A (retinol), tryptophan, and arachidonic acid metabolism pathways, as well as androgen and estrogen biosynthesis and metabolism. Some of these pathways such as the kynurenine pathway and tryptophan metabolism have previously been associated with an increased risk of chronic disease like lung cancer and neurological diseases.

We identified higher levels of vitamin A (retinol) in infants of smoking mothers and the vitamin A (retinol) metabolism pathways were enriched in both the HILICpos and C18neg columns. Retinoids, which include retinol, retinoic acid, and retinyl ester, regulate biologic processes and play a role in the prevention and treatment of several chronic diseases, including cancer.[147] Retinol and retinyl esters are the most abundant forms of retinoids present in the body and originate either directly from diet or are produced in the body through enzymatic cleavage of specific carotenoids.[147,148] Cigarette smoking exposure has been associated with the reduction of retinoic acid and may be responsible for an increased lung cancer risk in rats.[42] In addition, retinoic acid is crucial for regulating immune system homeostasis and for inflammatory responses.[149] We observed an alteration in vitamin A metabolism in newborn blood as a metabolic consequence of maternal tobacco smoking, suggesting a possible upregulation to counteract inflammatory actions.

Consistent with the existing literature, our results indicate an association between maternal smoking and elevated kynurenine in newborn blood as well as with tryptophan metabolism. Tryptophan is an essential amino acid used to build proteins and is a biosynthetic precursor for many neurologically active compounds. The kynurenine pathway plays a vital role in tryptophan metabolism and accounts for 95% of all tryptophan metabolism.[57,150] Activation of the tryptophan metabolism via the kynurenine pathway prevents hyperinflammation and has immunosuppressive effects. It also regulates energy homeostasis and brain function.[151] Tryptophan/kynurenine

metabolism has been linked with several diseases related to smoking. For example, a nested case-control study of 5,364 smoking-matched case–control pairs found that those measuring in the highest quintile of kynurenine were at a 20-30% higher risk of developing lung cancer with the strongest associations seen for current smokers, lesser among former smokers, and none among never smokers.[152] The kynurenine pathway has also been implicated in the pathophysiology of a range of neurodegenerative diseases including multiple sclerosis, Amyotrophic Lateral Sclerosis, Huntington's and Parkinson's disease, and Alzheimer's disease.[153] Low serum tryptophan can also contribute to immunodeficiency.[154] On the other hand, study results have been inconsistent with some finding a decreased level of kynurenine in smokers relative to non-smokers.[155] Inconsistent findings may be due to the nature of the disease investigated[155,156] or lifestyle factors and medications that are related to the disorder under investigation.

Steroid hormones play a major and complex role in mediating physiological and behavioral processes and cigarette smoking has been shown to affect the secretion, synthesis, metabolism, distribution, and excretion of hormones.[157] We identified enriched pathways related to the biosynthesis and metabolism of steroid hormones such as androgen, estrogen, and C21-steroid hormone. The sex hormones, androgens and estrogens, have a role in modulating several regulatory system and mediating sex-related differences in disease development.[157] Studies have demonstrated that smoking resulted in a decreased level of plasma estrogens in women of childbearing age, pregnant women, and newborns of smoking mothers, which was likely due to a smoking-induced hydroxylation and an increase in clearance.[157,158] In addition, a higher level of 17-OH pregnenolone, a pregnane (C21) steroid, was seen in smoking mothers with male fetuses compared to non-smokers.[158] Smoking during pregnancy induces changes in the production of steroid hormones in both mothers and newborns and these changes are seen to be differed by fetal

sex. These may indicate that smoking has an undesirable influence on adrenal steroidogenesis and the influence on male children may be associated with defects in adult reproductive function.[158]

We identified several other tobacco smoking-related pathways and metabolites that were previously mentioned in the literature in connection with smoking. Acetylcholine is a small molecule that plays an important role in maintaining homeostasis and brain functions, acting as a neurotransmitter in the peripheral nervous system.[159] Nicotinic acetylcholine receptor (nAChR), as a major class of acetylcholine receptors, is a typical ion channel type receptor. Nicotine in tobacco products, as an agonist of nAChR, evokes rapid depolarization responses to elicit neuronal excitation or skeletal muscle contraction.[159] Long-term exposure to nicotine exerts protection against neurotoxicity induced by toxic insults and the neuroprotective effects induced by long-term nAChR stimulation indicate that central nervous system nAChRs play crucial roles in promotion of neuronal survival under conditions such as brain ischemia and neurodegenerative diseases. [159] Arachidonic acid metabolites and enzyme transcripts involving both the lipoxygenase and cyclooxygenase pathways were reported to have different concentrations in smokers with asthma compared with never smokers, in multiple types of samples including urine, sputum, and blood.[160] Alanine aminotransferase is widely used as a clinical biomarker of hepatic health as it is involved in the transamination of alanine and presents in substantially higher concentrations in the liver than other organs.[161] Smoking has been shown to be associated with elevated alanine aminotransferase levels among anti-hepatitis C virus antibody-seropositive individuals.[162] Hippurate has also been shown to be increased with exposure to byproducts of cigarette smoke, such as phenolic compounds and toluene, in a study examining cigarette smoking associated vaginal tract metabolomic profile.[163] Itaconate was reported to be significantly increased with

cigarette smoking and has been implicated as an immune-response regulator in macrophages in a multi-omics study.[164]

Our sensitivity analysis that excluded preterm births infants found pathways previously also identified in a study that investigated maternal light active smoking and secondhand smoke exposures in amniotic fluid and maternal serum.[58] These pathways affected by maternal smoking exposure include vitamin A (retinol), lysine, tyrosine, porphyrin, and urea cycle/amino group metabolism. Smoking is a known risk factor for preterm birth [165] and the metabolic perturbations observed in the whole population but not in term birth may involve pathways related to preterm delivery. Out of the nine enriched pathways we identified as being different in smokers in analyses that included preterm births, four pathways (tryptophan metabolism, arachidonic acid metabolism, steroid hormone biosynthesis and metabolism, fatty acid oxidation and metabolism) were previously reported as preterm birth predictive pathways.[165]

Previous studies investigated tobacco metabolism and related pathways in adults -including pregnant women - and children, and our study affirms that several pathways previously identified are also perturbed in newborn infants whose mothers smoked in pregnancy. For example, the vitamin A (retinol) metabolism pathway, which plays a role in the development of the immune system, was reported to be perturbed by tobacco smoke exposure based on both second-trimester amniotic fluid and corresponding maternal serum analyses.[58] Similar to our findings, previous studies also identified kynurenine, a part of tryptophan metabolism, as smoking-related metabolites in adult smokers' serum and children's urine samples after secondhand smoke exposure.[57,166] Furthermore, five metabolic pathways (aspartate and asparagine metabolism and pyrimidine urea cycle/amino group metabolism, arginine and proline metabolism, and xenobiotics metabolism) reported as being perturbed by tobacco smoke exposure in both second-trimester

amniotic fluid and in adult serum of military personnel,[56,58] were not seen to be enriched in our study. The differences in findings across studies may be due to sample type and age, as well as differences in other demographic cofactors that influence metabolism and the kinetics of nicotine metabolism.[167] It has been suggested that neonates have a much slower nicotine metabolism such that nicotine has a 3 to 4 times longer half-life in newborns compared with adults.[167] Cotinine, on the other hand, is reported to have a similar half-life in neonates, older children, and adults of about 16.3 hours (95% CI, 12.4 to 23.9) in blood and 22.8 hours (95% CI, 19.5 to 25.8) in urine.[35,167] Despite having slightly longer half-life, this biomarker derived measure from neonatal blood spots is nevertheless imperfect and may underestimate maternal smoking due to a number of reasons such as women not initiating breastfeeding.[139] The metabolic signals we identified in this study in newborn blood may only reflect those metabolic pathways most strongly associated with chronic pre-and perinatal tobacco smoke exposure.

We acknowledge several limitations of this study. First, we were only able to control for potential confounders that are reported on birth certificates and with relatively low missingness. For example, we decided not to adjust for maternal and paternal education due to the large missingness of these variables. Second, due to the nature of untargeted metabolomics, we cannot obtain the absolute concentration of cotinine as a measure for current smoking status. Instead, we utilized cotinine and hydroxycotinine and selected a plausible cutoff according to reported maternal smoking in California during the years of the study. Thus, we might have misclassified exposure status, however, this would have been non-differential for exposure and the metabolomic profile. Our classification of smokers favored high specificity over sensitivity, therefore we may have called some smokers non-smokers. We would not have been able to identify tobacco metabolites in all exposed infants if mothers did not smoke within 20 hours of sample collection,

and in particular among neonates of mothers who never initiated breastfeeding. Nevertheless, we were able to identify a number of metabolites and pathways that were strongly associated with tobacco smoke and have been reported previously in relation to smoking. Another limitation is the limited ability to annotate and confirm metabolites using the untargeted platform. We were able to improve the coverage range and accuracy of the annotation by adopting pathway analysis. However, there may still be incorrect matches that could affect the interpretation of our results. MS/MS analysis is recommended for specific metabolites of interest.

In conclusion, we identified metabolites and pathway perturbations that have previously been associated with cigarette smoking, inflammatory responses, and diseases of the central nervous system and the lung in neonatal blood. Our results provided a global view of the newborn infant's metabolism in response to maternal smoking in the perinatal and early postnatal period. Our study suggests that maternal smoking during pregnancy has an impact on the child's metabolism during development and implicates pathways related to diseases previously associated with smoking such as lung disease and brain development.

4.6 Tables and figures

Table 4-1. Demographic characteristics of the study population

	Non-smoker (N=730)	Smoker (N=153)
	N = 883	
Maternal race/ethnicity		
White non-Hispanics	225 (30.8%)	74 (48.4%)
Hispanic of any race	370 (50.7%)	49 (32.0%)
Other/not specified	135 (18.5%)	30 (19.6%)
Birth Year		
1983-1990	121 (16.6%)	63 (41.2%)
1991-2000	281 (38.5%)	66 (43.1%)
2001-2011	328 (44.9%)	24 (15.7%)
Sex		
Male	346 (47.4%)	87 (56.9%)
Female	384 (52.6%)	66 (43.1%)
Maternal age		
<20	81 (11.1%)	18 (11.8%)
20-24	181 (24.8%)	38 (24.8%)
25-29	193 (26.4%)	44 (28.8%)
30-34	170 (23.3%)	43 (28.1%)
35+	105 (14.4%)	10 (6.5%)
Census-based neighborhood SES index		
1 (Low)	163 (22.3%)	30 (19.6%)
2	208 (28.5%)	40 (26.1%)
3	145 (19.9%)	47 (30.7%)
4	117 (16.0%)	18 (11.8%)
5 (High)	97 (13.3%)	18 (11.8%)
Parity		
0	309 (42.3%)	55 (35.9%)
1	226 (31.0%)	38 (24.8%)
2+	194 (26.6%)	60 (39.2%)
Missing	1	0
Preterm birth		
Preterm	67 (9.5%)	17 (11.9%)
Term	638 (90.5%)	126 (88.1%)
Missing	25	10
Maternal education		
Less than high school	82 (12.5%)	10 (9.0%)
High school graduate	321 (49.0%)	68 (61.3%)
Some college, college graduate or more	252 (38.5%)	33 (29.7%)
Missing	75	42
Foreign-born		
Yes	380 (52.1%)	96 (62.7%)
Missing	0	1

Table 4-2. Confirmed [a] chemical identity of metabolic features associated with maternal smoking status among studied infants

m/z	RT (s)	Adduct Form	Metabolite	Coefficient	FDR	VIP	Mode
90.055	52.4	M+H	Alanine	2.90E-01	2.93E-03	2.23	HILICpos
102.055	66.2	M+H	1-Aminocyclopropane-1-carboxylate	3.41E-01	1.22E-03	2.11	HILICpos
116.0344	78.1	M+H	Maleamate	3.25E-01	1.33E-03	2.15	HILICpos
132.0766	53.7	M+H	Creatine	2.51E-01	1.03E-02	2.01	HILICpos
146.0924	25.7	M+H	Guanidinobutanote	5.40E-01	8.05E-05	2.85	HILICpos
146.1176	36.4	M+H	Acetylcholine	4.75E-01	1.27E-04	2.73	HILICpos
180.0653	44.2	M+H	Hippurate	3.17E-01	1.45E-03	2.20	HILICpos
209.0922	51	M+H	Kynurenine	3.89E-01	2.11E-04	2.86	HILICpos
241.0308	183.1	M+H	Cystine	-3.36E-01	8.27E-04	2.34	HILICpos
269.2263	22.4	M+H	Vitamin A (Retinol)	3.73E-01	9.35E-04	2.29	HILICpos
129.0187	22.6	M-H	Itaconate	4.29E-01	1.37E-04	2.33	C18neg
180.0667	19.6	M-H	L-Tyrosine	4.22E-01	1.37E-04	2.18	C18neg
188.0566	17.8	M-H	N-Acetyl-Dl-Glutamic Acid	5.32E-01	2.93E-05	2.60	C18neg
195.0512	18.8	M-H	Gluconic Acid	3.46E-01	2.82E-04	2.72	C18neg
309.2803	291.2	M-H	FA 20:1 (Gondoic acid)	3.26E-01	3.76E-04	2.72	C18neg
311.296	291.3	M-H	Arachidic Acid	2.47E-01	1.14E-03	2.63	C18neg
588.076	17.2	M-H	Adenosine-5'-Diphosphoglucose	4.02E-01	2.06E-04	2.20	C18neg

[a] Chemical identification was conducted by matching peaks by accurate mass and retention time to authentic reference standards in an in-house library run under identical conditions using tandem mass spectrometry.

Table 4-3. Enriched metabolomic pathways associated with maternal smoking status among all infants (N=883)

Pathway	Overlap size	Pathway size	p-value	Mode
Vitamin A (retinol) metabolism	4	24	0.00992	HILICpos
Tryptophan metabolism	7	73	0.02596	HILICpos
Arachidonic acid metabolism	4	33	0.03218	HILICpos
C21-steroid hormone biosynthesis and metabolism	5	49	0.03823	HILICpos
N-Glycan biosynthesis	3	23	0.04579	HILICpos
Vitamin A (retinol) metabolism	3	22	0.01622	C18neg
Androgen and estrogen biosynthesis and metabolism	4	40	0.0184	C18neg
Fatty acid oxidation	1	3	0.04781	C18neg
Fatty acid metabolism	2	18	0.05966	C18neg

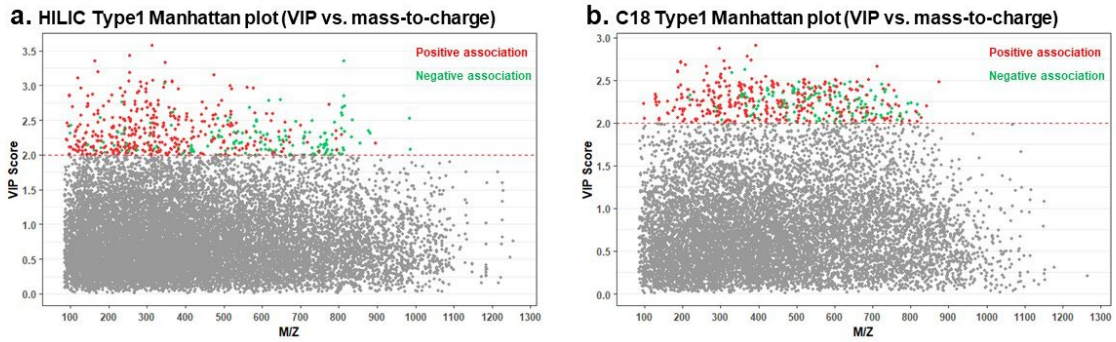


Figure 4-1. Identification of metabolic features associated with maternal smoking exposure among studied infants. A) Type 1 Manhattan plot for features in the HILIC column (positive ion mode), VIP score vs *m/z*. Red dots represent features that were positively associated with maternal smoking exposure and green dots represent features that were negatively associated with maternal smoking exposure; B) Type 1 Manhattan plot for features in the C18 column (negative ion mode), VIP score vs mass-to-charge.

4.7 Supplemental materials

Supplemental Table 4-S1. Mummichog Annotated Metabolites

Pathway	Name	KEGG ID	Ion	M/Z	Time	VIP	Coefficient	P vlaue	Adjusted p value	Fold change	Mode	Metabolite annotation confidence
Androgen and estrogen biosynthesis and metabolism	16alpha-Hydroxyandrost-4-ene-3,17-dione	C05140	M-H	301.1810	205.6	2.167	0.357	6.01E-04	1.21E-03	0.375	C18neg	4
	5beta-Dihydrotestosterone	C05293	M-H	289.2177	223.5	2.076	0.298	7.45E-04	1.36E-03	0.128	C18neg	4
	4-methoxyestrone	CE2189	M-H	299.1656	34.7	2.118	0.328	6.69E-04	1.26E-03	0.194	C18neg	4
Arachidonic acid metabolism	(5Z)-(15S)-11alpha-Hydroxy-9,15-dioxoprostanoate	C04671	M+H	352.2220	47.7	2.289	0.328	6.05E-04	1.13E-03	0.357	HILICpos	4
	10,11-dihydro-12-oxo-LTB4	CE5944	M+Na	335.2214	26.5	2.065	0.311	1.38E-03	1.85E-03	0.245	HILICpos	4
	15(R)-HEPE	CE7081	M+H	317.2109	24.0	2.569	0.366	1.33E-04	4.85E-04	0.441	HILICpos	4
	15-oxo-PGE1	CE7092	M+H	352.2220	47.7	2.289	0.328	6.05E-04	1.13E-03	0.357	HILICpos	4
C21-steroid hormone biosynthesis and metabolism	Androst-4-ene-3,17-dione	C00280	M+H	309.1825	49.1	2.287	-0.306	3.42E-03	3.84E-03	-0.382	HILICpos	4
	5beta-Androstane-3,17-dione	C03772	M+H	289.2154	26.8	2.061	0.301	9.94E-04	1.53E-03	0.326	HILICpos	4
	11-Dehydrocorticosterone	C05490	M+H	344.1981	55.7	2.426	0.257	5.16E-04	1.03E-03	0.020	HILICpos	4
Electron transport chain	Ubiquinone	C00399	M-H	299.1656	34.7	2.118	0.328	6.69E-04	1.26E-03	0.194	C18neg	4
Fatty acid oxidation	Ubiquinone	C00399	M-H	299.1656	34.7	2.118	0.328	6.69E-04	1.26E-03	0.194	C18neg	4
N-Glycan biosynthesis	Farnesyl diphosphate	C00448	M+H	383.1412	55.6	2.149	0.296	9.68E-04	1.51E-03	0.147	HILICpos	4
	Dolichyl beta-D-glucosyl phosphate	C01246	M+H	489.2238	64.2	2.350	-0.319	2.78E-04	7.41E-04	-0.630	HILICpos	4
	3'-UMP	C01368	M+Na	347.0250	197.4	2.140	0.527	5.02E-05	3.15E-04	0.239	HILICpos	4
Tryptophan metabolism	Alanine	C00041	M+H	90.0550	52.4	2.227	0.290	2.50E-03	2.93E-03	0.107	HILICpos	1
	Beta-alanine	C00099	M+H	90.0550	52.4	2.227	0.290	2.50E-03	2.93E-03	0.107	HILICpos	1
	Sarcosine	C00213	M+H	90.0550	52.4	2.227	0.290	2.50E-03	2.93E-03	0.107	HILICpos	1
	Kynurenine	C00328	M+H	209.0922	51.0	2.863	0.389	1.92E-05	2.11E-04	0.183	HILICpos	1
	Melatonin	C01598	M+H	255.1102	43.7	3.183	0.500	6.67E-07	4.56E-05	0.312	HILICpos	4
	4,6-Dihydroxyquinoline	C05639	M+H	162.0561	33.8	2.584	0.411	5.18E-05	3.21E-04	0.213	HILICpos	4
Vitamin A (retinol) metabolism	Vitamin A (Retinol)	C00473	NA	269.2263	22.4	2.289	0.373	4.28E-04	9.35E-04	0.743	HILICpos	1
	all-trans-4-Hydroxyretinoic acid	C16677	M+H	317.2109	24.0	2.569	0.366	1.33E-04	4.85E-04	0.441	HILICpos	4
	all-trans-4-Hydroxyretinoic acid	C16677	M-H	315.1966	203.6	2.021	0.338	9.49E-04	1.59E-03	0.204	C18neg	4
	9-cis-retinoate	CE1617	M+H-H2O	299.2005	24.0	3.021	0.442	3.85E-06	9.07E-05	0.424	HILICpos	4
	4-hydroxy-all-trans-retinyl acetate	CE2961	M-H	325.2176	227.5	2.218	0.328	5.96E-04	1.21E-03	1.116	C18neg	4
	5-oxo-EPE	CE7111	M+H	315.1953	23.9	3.076	0.405	4.81E-06	1.04E-04	0.533	HILICpos	4

* Schymanski EL; Jeon J; Gulde R; Fenner K; Ruff M; Singer HP; Hollender (2014) Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence. J Environ. Sci. Technol 48: 2097–8

Supplemental Table 4-S2. Enriched metabolomic pathways associated with maternal smoking status among term birth infants only (N=764)

Pathway	Overlap size	Pathway size	p-value	Mode
Vitamin A (retinol) metabolism	3	24	0.02277	HILICpos
Androgen and estrogen biosynthesis and metabolism	4	41	0.0258	HILICpos
Putative anti-Inflammatory metabolites formation from EPA	2	12	0.0279	HILICpos
Lysine metabolism	3	28	0.03428	HILICpos
Tyrosine metabolism	9	109	0.00235	C18neg
Porphyrin metabolism	3	22	0.01496	C18neg
Urea cycle/amino group metabolism	4	56	0.04605	C18neg
Parathion degradation	1	3	0.04798	C18neg
Vitamin B12 (cyanocobalamin) metabolism	1	4	0.0663	C18neg

Chapter 5. Conclusion and Public Health Implications

This dissertation investigated the measurement of maternal perinatal tobacco smoking and the associated metabolic signals in newborn blood spots, a validated biosample that is useful for epidemiological studies of rare diseases in childhood. In the first project (Chapter 2), we comprehensively assessed the stability of metabolites across all study years (1983-2011) and found that almost three decades of storage in -20 ˚C freezers did not affect the deterioration of metabolites that reflect the general human physiology or exogenous metabolites that are markers of common lifestyle-related exposures such as cotinine and caffeine. In the second project (Chapter 3), we utilized the combination of self- or provider-reported smoking and metabolites-derived smoking information and constructed a prediction model using birth certificate data to predict maternal smoking reaching an AUC of 0.805. The model may benefit future studies in the state that employ birth registry data and populations and would help to improve adjustment for maternal smoking and may also help guide clinical intervention to reduce smoking in pregnancy in California. In the third project (Chapter 4), we identified metabolites and pathway perturbations that have previously been associated with cigarette smoking, inflammatory responses, and diseases of the central nervous system and the lung in neonatal blood. Our results provided a global view of the newborn infant's metabolism in response to maternal smoking in the perinatal and early postnatal period.

Neonatal dried blood spots are valuable resources for epidemiological studies of child outcomes and childhood rare diseases. The collection and storage of DBS are minimally invasive and inexpensive and easy to multiplex and automate. The handling and logistics allows the deployment of DBS storage and shipment even in resource-poor settings. Moreover, DBS samples are compatible with many bioanalytical methods including chromatography, mass spectrometry, DNA, and immunoassays. With all the advantages, DBS applications are emerging in areas of

large scale neonatal screening, targeted and nontargeted metabolic profiling, and epidemiological disease surveillance.[168] Research targeting rare diseases must often rely on biospecimens that are collected routinely and have been stored for long time periods in order to accumulate a sufficient number of samples to address hypotheses with adequate statistical power. Results of this dissertation support the feasibility of conducting metabolomics studies using archived DBS stored for decades. Utilizing existing biospecimens from statewide newborn screening programs will provide a unique and important resource for future epidemiologic health and environmental exposure studies.

Although its prevalence drops over the past decades, maternal smoking remains to be prevalent and not only causes direct adverse health effects but also interferes with normal response to environmental exposures, increasing the susceptibility for disease. The developmental origin of the health and disease (DOHaD) concept proposes that early life exposures during critical windows can be persistent and may result in health consequences in later life. According to the DOHad principle, the health burden due to maternal smoking ranges in magnitude similar to ambient ozone exposure, which is recognized as a severe public health problem with high priority in research and policymaking. [169] Thus, the impact of maternal smoking remains a public health focus. Our results explore the effect of maternal smoking in offspring from a metabolic perspective and suggest that maternal smoking during pregnancy impacts the child's metabolism during development and implicates pathways related to diseases previously associated with smoking, such as lung disease and brain development. In addition, the application of the prediction model we built in the second project will at least partially account for uncontrolled confounding due to missing maternal smoking information when conducting studies on childhood are diseases using California birth certificates.

Complementing traditional epidemiological studies, our findings re-emphasize the importance of tobacco control, especially for vulnerable populations such as pregnant women and newborns. By combining large population-based data with metabolomics and prediction models approaches, our studies provide a valuable understanding of maternal smoking measurements and its health impact in vulnerable populations that may guide future research, disease surveillance, and policymaking.

Chapter 6. References

1. Tobacco. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/tobacco. Accessed July 16, 2021.
2. Warren GW, Cummings KM. Tobacco and lung cancer: risks, trends, and outcomes in patients with cancer. Paper presented at: American Society of Clinical Oncology educational book. American Society of Clinical Oncology. Meeting2013.
3. Henningfield JE. The tobacco endgame: it's all about behavior. *Preventive medicine.* 2014;68:11-16.
4. Samet JM. Epidemiology and the Tobacco Epidemic: How Research on Tobacco and Health Shaped Epidemiology. *American journal of epidemiology.* 2016;183(5):394-402.
5. Schüz J, Espina C, Villain P, et al. European Code against Cancer 4th Edition: 12 ways to reduce your cancer risk. *Cancer epidemiology.* 2015;39:S1-S10.
6. Saleheen D, Zhao W, Rasheed A. Epidemiology and public health policy of tobacco use and cardiovascular disorders in low-and middle-income countries. *Arteriosclerosis, thrombosis, and vascular biology.* 2014:ATVBAHA. 114.303826.
7. Taylor FR. Tobacco, nicotine, and headache. *Headache: The Journal of Head and Face Pain.* 2015;55(7):1028-1044.
8. Mahapatra S, Kamath R, Shetty BK, Binu VS. Risk of oral cancer associated with gutka and other tobacco products: A hospital-based case-control study. *Journal of cancer research and therapeutics.* 2015;11(1):199.
9. Amtha R, Razak IA, Basuki B, et al. Tobacco (kretek) smoking, betel quid chewing and risk of oral cancer in a selected Jakarta population. *Asian Pac J Cancer Prev.* 2014;15(20):8673-8678.

10. Jayalekshmi PA, Hassani S, Nandakumar A, Koriyama C, Sebastian P, Akiba S. Gastric cancer risk in relation to tobacco use and alcohol drinking in Kerala, India-Karunagappally cohort study. *World journal of gastroenterology.* 2015;21(44):12676.

11. Health UDo, Services H. The health consequences of smoking—50 years of progress: a report of the Surgeon General. In: Atlanta, GA: US Department of Health and Human Services, Centers for Disease …; 2014.

12. Cornelius ME LC, Wang TW, Jamal A, Homa DM. *Tobacco Product Use Among Adults — United States, 2020.* MMWR Morb Mortal Wkly Rep 2022;71:397–405. DOI: http://dx.doi.org/10.15585/mmwr.mm7111a1external2022.

13. Keyes KM, March D, Link BG, Chilcoat HD, Susser E. Do socio-economic gradients in smoking emerge differently across time by gender? Implications for the tobacco epidemic from a pregnancy cohort in California, USA. *Soc Sci Med.* 2013;76:101-106.

14. Mahadevan U, Sandborn WJ, Li DK, Hakimian S, Kane S, Corley DA. Pregnancy Outcomes in Women With Inflammatory Bowel Disease: A Large Community-Based Study From Northern California. *Gastroenterology.* 2007;133(4):1106-1112.

15. Sun Y, Li X, Benmarhnia T, et al. Exposure to air pollutant mixture and gestational diabetes mellitus in Southern California: Results from electronic health record data of a large pregnancy cohort. *Environ Int.* 2022;158:106888.

16. Newsroom C-. Smoking is down, but almost 38 million American adults still smoke. https://www.cdc.gov/media/releases/2018/p0118-smoking-rates-declining.html#:~:text=Cigarette%20smoking%20among%20U.S.%20adults%20has%20been%20reduced%20by%20more,death%20in%20the%20United%20States. Published 2018. Updated January 18. Accessed.

17. *Smoking prevalence among California and U.S. minus California adults, 1984-2010.* Sacramento, CA: California Department of Public Health, California Tobacco Control Program;2014.

18. Hoggatt KJ, Flores M, Solorio R, Wilhelm M, Ritz B. The "Latina Epidemiologic Paradox" Revisited: The Role of Birthplace and Acculturation in Predicting Infant Low Birth Weight for Latinas in Los Angeles, CA. *Journal of Immigrant and Minority Health.* 2012;14(5):875-884.

19. on Smoking O, Control CfD, Prevention. How tobacco smoke causes disease: The biology and behavioral basis for smoking-attributable disease: A report of the surgeon general. 2010.

20. Pollack H, Lantz PM, Frohna JG. Maternal smoking and adverse birth outcomes among singletons and twins. *Am J Public Health.* 2000;90(3):395-400.

21. Hackshaw A, Rodeck C, Boniface S. Maternal smoking in pregnancy and birth defects: a systematic review based on 173 687 malformed cases and 11.7 million controls. *Hum Reprod Update.* 2011;17(5):589-604.

22. Heck JE, Contreras ZA, Park AS, Davidson TB, Cockburn M, Ritz B. Smoking in pregnancy and risk of cancer among young children: A population-based study. *International journal of cancer.* 2016;139(3):613-616.

23. Himes SK, Stroud LR, Scheidweiler KB, Niaura RS, Huestis MA. Prenatal Tobacco Exposure, Biomarkers for Tobacco in Meconium, and Neonatal Growth Outcomes. *The Journal of Pediatrics.* 2013;162(5):970-975.

24. Pickett KE, Rathouz PJ, Kasza K, Wakschlag LS, Wright R. Self-reported smoking, cotinine levels, and patterns of smoking in pregnancy. *Paediatr Perinat Epidemiol.* 2005;19(5):368-376.

25. Ventura SJ, Hamilton BE, Mathews TJ, Chandra A. Trends and Variations in Smoking During Pregnancy and Low Birth Weight: Evidence From the Birth Certificate, 1990–2000. *Pediatrics.* 2003;111(Supplement_1):1176-1180.

26. Dietz PM, Adams MM, Kendrick JS, Mathis MP. Completeness of ascertainment of prenatal smoking using birth certificates and confidential questionnaires: Variations by maternal attributes and infant birth weight. *Am J Epidemiol.* 1998;148(11):1048-1054.

27. Howland RE, Mulready-Ward C, Madsen AM, et al. Reliability of Reported Maternal Smoking: Comparing the Birth Certificate to Maternal Worksheets and Prenatal and Hospital Medical Records, New York City and Vermont, 2009. *Maternal and Child Health Journal.* 2015;19(9):1916-1924.

28. Bardy AH, Seppälä T, Lillsunde P, et al. Objectively measured tobacco exposure during pregnancy: neonatal effects and relation to maternal smoking. *BJOG: An International Journal of Obstetrics & Gynaecology.* 1993;100(8):721-726.

29. Yang J, Pearl M, Jacob Iii P, et al. Levels of cotinine in dried blood specimens from newborns as a biomarker of maternal smoking close to the time of delivery. *American journal of epidemiology.* 2013;178(11):1648-1654.

30. Koskela P, Anttila T, Bjørge T, et al. Chlamydia trachomatis infection as a risk factor for invasive cervical cancer. *International journal of cancer.* 2000;85(1):35-39.

31. Sasaki S, Braimoh TS, Yila TA, Yoshioka E, Kishi R. Self-reported tobacco smoke exposure and plasma cotinine levels during pregnancy – A validation study in Northern Japan. *Sci Total Environ.* 2011;412-413:114-118.

32. Wall MA, Johnson J, Jacob P, Benowitz NL. Cotinine in the serum, saliva, and urine of nonsmokers, passive smokers, and active smokers. *American Journal of Public Health.* 1988;78(6):699-701.

33. Murphy SE, Wickham KM, Lindgren BR, Spector LG, Joseph A. Cotinine and trans 3′-hydroxycotinine in dried blood spots as biomarkers of tobacco exposure and nicotine metabolism. *J Expo Sci Environ Epidemiol.* 2013;23(5):513-518.

34. Jacob P, Yu L, Duan M, Ramos L, Yturralde O, Benowitz NL. Determination of the nicotine metabolites cotinine and trans-3′-hydroxycotinine in biologic fluids of smokers and non-smokers using liquid chromatography–tandem mass spectrometry: Biomarkers for tobacco smoke exposure and for phenotyping cytochrome P450 2A6 activity. *J Chromatogr B.* 2011;879(3):267-276.

35. Dempsey D, Jacob III P, Benowitz NL. Nicotine metabolism and elimination kinetics in newborns. *Clin Pharmacol Ther.* 2000;67(5):458-465.

36. Gonzales JL. Ethics for the Pediatrician. *Genetic Testing and Newborn Screening.* 2011;32(11):490-493.

37. Health CDoP. ABOUT Genetic Disease Screening Program (GDSP). https://www.cdph.ca.gov/Programs/CFH/DGDS/Pages/GDSP-Program-and-Services.aspx. Published 2019. Updated May 23, 2019. Accessed.

38. Wang Q, Sun T, Cao Y, et al. A dried blood spot mass spectrometry metabolomic approach for rapid breast cancer detection. *OncoTargets and therapy.* 2016;9:1389.

39. Kong ST, Lin H-S, Ching J, Ho PC. Evaluation of dried blood spots as sample matrix for gas chromatography/mass spectrometry based metabolomic profiling. *Analytical chemistry.* 2011;83(11):4314-4318.

40. Sharma A, Jaiswal S, Shukla M, Lal J. Dried blood spots: concepts, present status, and future perspectives in bioanalysis. *Drug testing and analysis.* 2014;6(5):399-414.

41. Jing Y, Wu X, Gao P, et al. Rapid differentiating colorectal cancer and colorectal polyp using dried blood spot mass spectrometry metabolomic approach. *IUBMB life.* 2017;69(5):347-354.

42. Petrick LM, Schiffman C, Edmands WMB, et al. Metabolomics of neonatal blood spots reveal distinct phenotypes of pediatric acute lymphoblastic leukemia and potential effects of early-life nutrition. *Cancer Lett.* 2019;452:71-78.

43. 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. *Leukemia research.* 2020;88:106268.

44. Murphy MSQ, Hawken S, Atkinson KM, et al. Postnatal gestational age estimation using newborn screening blood spots: a proposed validation protocol. *BMJ Glob Health.* 2017;2(2):e000365-e000365.

45. Petrick LM, Uppal K, Funk WE. Metabolomics and adductomics of newborn bloodspots to retrospectively assess the early-life exposome. *Current opinion in pediatrics.* 2020;32(2):300-307.

46. O'Broin SD, Gunter EW. Screening of folate status with use of dried blood spots on filter paper–. *The American journal of clinical nutrition.* 1999;70(3):359-367.

47. Zimmerman RK, Slater ME, Langer EK, Ross JA, Spector LG. Long-Term Stability of Folate in Dried Blood Spots Stored in Several Conditions. *The Journal of Pediatrics.* 2013;163(2):596-597.e591.

48. Prentice P, Turner C, Wong MC, Dalton RN. Stability of metabolites in dried blood spots stored at different temperatures over a 2-year period. *Bioanalysis.* 2013;5(12):1507-1514.

49. Li S, Dunlop AL, Jones DP, Corwin EJ. High-Resolution Metabolomics: Review of the Field and Implications for Nursing Science and the Study of Preterm Birth. *Biol Res Nurs.* 2016;18(1):12-22.

50. Nicholson JK, Lindon JC. Metabonomics. *Nature.* 2008;455(7216):1054-1056.

51. Walker DI, Lane KJ, Liu K, et al. Metabolomic assessment of exposure to near-highway ultrafine particles. *J Expo Sci Environ Epidemiol.* 2019;29(4):469-483.

52. Dénes J, Szabó E, Robinette SL, et al. Metabonomics of Newborn Screening Dried Blood Spot Samples: A Novel Approach in the Screening and Diagnostics of Inborn Errors of Metabolism. *Analytical Chemistry.* 2012;84(22):10113-10120.

53. Prentice P, Koulman A, Matthews L, Acerini CL, Ong KK, Dunger DB. Lipidomic Analyses, Breast- and Formula-Feeding, and Growth in Infants. *The Journal of Pediatrics.* 2015;166(2):276-281.e276.

54. Petrick L, Edmands W, Schiffman C, et al. An untargeted metabolomics method for archived newborn dried blood spots in epidemiologic studies. *Metabolomics.* 2017;13(3):27.

55. Koulman A, Prentice P, Wong MCY, et al. The development and validation of a fast and robust dried blood spot based lipid profiling method to study infant metabolism. *Metabolomics.* 2014;10(5):1018-1025.

56. Jones DP, Walker DI, Uppal K, Rohrbeck P, Mallon TM, Go Y-M. Metabolic pathways and networks associated with tobacco use in military personnel. *Journal of occupational and environmental medicine/American College of Occupational and Environmental Medicine.* 2016;58(8):S111.

57. Zhu H, Abdullah AS, He J, et al. Untargeted Urinary Metabolomics and Children's Exposure to Secondhand Smoke: The Influence of Individual Differences. *International journal of environmental research and public health.* 2021;18(2):710.

58. Fischer ST, Lili LN, Li S, et al. Low-level maternal exposure to nicotine associates with significant metabolic perturbations in second-trimester amniotic fluid. *Environment International.* 2017;107:227-234.

59. Montgomery SM, Ekbom A. Smoking during pregnancy and diabetes mellitus in a British longitudinal birth cohort. *Bmj.* 2002;324(7328):26-27.

60. Heck JE, Lombardi CA, Meyers TJ, Cockburn M, Wilhelm M, Ritz B. Perinatal characteristics and retinoblastoma. *Cancer Causes & Control.* 2012;23(9):1567-1575.

61. Steyerberg EW. *Clinical prediction models.* Springer; 2019.

62. Markaki M, Tsamardinos I, Langhammer A, Lagani V, Hveem K, Røe OD. A validated clinical risk prediction model for lung cancer in smokers of all ages and exposure types: a HUNT study. *EBioMedicine.* 2018;31:36-46.

63. Dukic VM, Niessner M, Benowitz N, Hans S, Wakschlag L. Modeling the relationship of cotinine and self-reported measures of maternal smoking during pregnancy: a deterministic approach. *Nicotine Tobacco Res.* 2007;9(4):453-465.

64. Moon SS, Kang S-Y, Jitpitaklert W, Kim SB. Decision tree models for characterizing smoking patterns of older adults. *Expert Systems with Applications.* 2012;39(1):445-451.

65. Heck JE, Lombardi CA, Cockburn M, Meyers TJ, Wilhelm M, Ritz B. Epidemiology of rhabdoid tumors of early childhood. *Pediatric Blood & Cancer.* 2013;60(1):77-81.

66. Heck JE, Park AS, Qiu J, Cockburn M, Ritz B. Retinoblastoma and ambient exposure to air toxics in the perinatal period. *Journal of Exposure Science and Environmental Epidemiology.* 2015;25(2):182.

67. Health CDoP. For Clinicians - Collection Information. Newborn Screening Program (NBS) Web site. https://www.cdph.ca.gov/Programs/CFH/DGDS/Pages/nbs/Collection-Information.aspx. Published 2018. Accessed.

68. Yang J, Pearl M, Jacob P, 3rd, et al. Levels of cotinine in dried blood specimens from newborns as a biomarker of maternal smoking close to the time of delivery. *American journal of epidemiology.* 2013;178(11):1648-1654.

69. Kharrazi M, Pearl M, Yang J, et al. California Very Preterm Birth Study: design and characteristics of the population- and biospecimen bank-based nested case–control study. *Paediatr Perinat Epidemiol.* 2012;26(3):250-263.

70. Liu KH, Nellis M, Uppal K, et al. Reference Standardization for Quantification and Harmonization of Large-Scale Metabolomics. *Anal Chem.* 2020;92(13):8836-8844.

71. Walker DI, Perry-Walker K, Finnell RH, et al. Metabolome-wide association study of anti-epileptic drug treatment during pregnancy. *Toxicology and Applied Pharmacology.* 2019;363:122-130.

72. Go Y-M, Roede JR, Orr M, Liang Y, Jones DP. Integrated Redox Proteomics and Metabolomics of Mitochondria to Identify Mechanisms of Cd Toxicity. *Toxicological Sciences.* 2014;139(1):59-73.

73. Liang D, Moutinho JL, Golan R, et al. Use of high-resolution metabolomics for the identification of metabolic signals associated with traffic-related air pollution. *Environment international.* 2018;120:145-154.

74. Yu T, Park Y, Johnson JM, Jones DP. apLCMS—adaptive processing of high-resolution LC/MS data. *Bioinformatics.* 2009;25(15):1930-1936.

75. Uppal K, Soltow QA, Strobel FH, et al. xMSanalyzer: automated pipeline for improved feature detection and downstream analysis of large-scale, non-targeted metabolomics data. *BMC Bioinformatics.* 2013;14(1):15.

76. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics.* 2007;8(1):118-127.

77. Accardi CJ, Walker DI, Uppal K, et al. High-resolution metabolomics for nutrition and health assessment of armed forces personnel. *Journal of occupational and environmental medicine.* 2016;58(8):S80-S88.

78. Liu KH, Lee CM, Singer G, et al. Large scale enzyme based xenobiotic identification for exposomics. *Nature Communications.* 2021;12(1):5418.

79. Schymanski EL, Jeon J, Gulde R, et al. Identifying small molecules via high resolution mass spectrometry: communicating confidence. In: ACS Publications; 2014.

80. Bunning BJ, Contrepois K, Lee-McMullen B, et al. Global metabolic profiling to model biological processes of aging in twins. *Aging Cell.* 2020;19(1):e13073.

81. Bakhireva LN, Shrestha S, Gutierrez HL, Berry M, Schmitt C, Sarangarm D. Stability of Phosphatidylethanol in Dry Blood Spot Cards. *Alcohol Alcohol.* 2016;51(3):275-280.

82. Strnadová KA, Holub M, Mühl A, et al. Long-Term Stability of Amino Acids and Acylcarnitines in Dried Blood Spots. *Clinical Chemistry.* 2007;53(4):717-722.

83. Li K, Naviaux JC, Monk JM, Wang L, Naviaux RK. Improved Dried Blood Spot-Based Metabolomics: A Targeted, Broad-Spectrum, Single-Injection Method. *Metabolites.* 2020;10(3).

84. Target Identification and Validation. https://www.genscript.com/target-identification-validation.html. Accessed.

85. Palmer EA, Cooper HJ, Dunn WB. Investigation of the 12-Month Stability of Dried Blood and Urine Spots Applying Untargeted UHPLC-MS Metabolomic Assays. *Anal Chem.* 2019;91(22):14306-14313.

86. Giordano G, Gucciardi A, Pirillo P, Naturale M. Quantification of Underivatized Amino Acids on Dry Blood Spot, Plasma, and Urine by HPLC-ESI-MS/MS. In: Alterman MA, ed. *Amino Acid Analysis: Methods and Protocols.* New York, NY: Springer New York; 2019:153-172.

87. Chace DH, Adam BW, Smith SJ, Alexander JR, Hillman SL, Hannon WH. Validation of accuracy-based amino acid reference materials in dried-blood spots by tandem mass spectrometry for newborn screening assays. *Clinical chemistry.* 1999;45(8):1269-1277.

88. Wu G. Amino acids: metabolism, functions, and nutrition. *Amino Acids.* 2009;37(1):1-17.

89. Nagahashi M, Abe M, Sakimura K, Takabe K, Wakai T. The role of sphingosine-1-phosphate in inflammation and cancer progression. *Cancer Sci.* 2018;109(12):3671-3678.

90. Zeisel SH, da Costa K-A. Choline: an essential nutrient for public health. *Nutr Rev.* 2009;67(11):615-623.

91. Jiang X, West AA, Caudill MA. Maternal choline supplementation: a nutritional approach for improving offspring health? *Trends Endocrinol Metab.* 2014;25(5):263-273.

92. Moya J, Phillips L, Sanford J, Wooton M, Gregg A, Schuda L. A review of physiological and behavioral changes during pregnancy and lactation: Potential exposure factors and data gaps. *Journal of Exposure Science & Environmental Epidemiology.* 2014;24(5):449-458.

93. Lamy S, Houivet E, Benichou J, Marret S, Thibaut F, for the Perinatal network of U-N. Caffeine use during pregnancy: prevalence of use and newborn consequences in a cohort of French pregnant women. *Eur Arch Psychiatry Clin Neurosci.* 2020.

94. Parsons WD, Neims AH. Prolonged half-life of caffeine in healthy term newborn infants. *J Pediatr.* 1981.

95. Wright AL, Schanler RJ. The Resurgence of Breastfeeding at the End of the Second Millennium. *The Journal of Nutrition.* 2001;131(2):421S-425S.

96. Breastfeeding Data and Reports. 2019. https://www.cdph.ca.gov/Programs/CFH/DMCAH/Breastfeeding/Pages/Data.aspx.

97. Pickett KE, Rathouz PJ, Kasza K, Wakschlag LS, Wright R. Self-reported smoking, cotinine levels, and patterns of smoking in pregnancy. *Paediatr Perinat Epidemiol.* 2005;19(5):368-376.

98. Walsh RA, Redman S, Adamson L. The accuracy of self-report of smoking status in pregnant women. *Addict Behav.* 1996;21(5):675-679.

99. Lindqvist R, Lendahls L, Tollbom Ö, ÅBerg H, Håkansson A. Smoking during pregnancy: comparison of self-reports and cotinine levels in 496 women. *Acta Obstet Gynecol Scand.* 2002;81(3):240-244.

100. Verification SSoB. Biochemical verification of tobacco use and cessation. *Nicotine & Tobacco Research.* 2002;4(2):149-159.

101. Vuckovic D. Current trends and challenges in sample preparation for global metabolomics using liquid chromatography–mass spectrometry. *Analytical and bioanalytical chemistry.* 2012;403(6):1523-1548.

102. Benowitz NL, Hukkanen J, Jacob P. Nicotine chemistry, metabolism, kinetics and biomarkers. *Nicotine psychopharmacology.* 2009:29-60.

103. Domino EF, Hornbach E, Demana T. The nicotine content of common vegetables. *New England Journal of Medicine.* 1993;329(6):437-437.

104. Sadok I, Gamian A, Staniszewska MM. Chromatographic analysis of tryptophan metabolites. *J Sep Sci.* 2017;40(15):3020-3045.

105. Dasgupta A. Chapter 2 - Biotin: Pharmacology, Pathophysiology, and Assessment of Biotin Status. In: Dasgupta A, ed. *Biotin and Other Interferences in Immunoassays.* Elsevier; 2019:17-35.

106. Koulman A, Prentice P, Wong MCY, et al. The development and validation of a fast and robust dried blood spot based lipid profiling method to study infant metabolism. *Metabolomics : Official journal of the Metabolomic Society.* 2014;10(5):1018-1025.

107. Reports of the Surgeon General. In: *The Health Consequences of Smoking—50 Years of Progress: A Report of the Surgeon General.* Atlanta (GA): National Center for Chronic Disease Prevention and Health Promotion Office on Smoking and Health; 2014.

108. Hansen AR, Akomolafe TO, McGalliard Z, Belle-Isle L, Zhang J. Striving to Meet Healthy People 2020 Objectives: Trend Analysis of Maternal Smoking. *Public Health Rep.* 2018;133(6):644-649.

109. Benowitz NL, Hukkanen J, Jacob P, 3rd. Nicotine chemistry, metabolism, kinetics and biomarkers. *Handbook of experimental pharmacology.* 2009(192):29-60.

110. Genetic Disease Screening Program. https://www.cdph.ca.gov/Programs/CFH/DGDS/Pages/GDSP-Program-and-Services.aspx. Published 2021. Updated November 17, 2021. Accessed.

111. Newborn Screening Collection Information. https://www.cdph.ca.gov/Programs/CFH/DGDS/Pages/nbs/Collection-Information.aspx. Published 2022. Updated January 26, 2022. Accessed.

112. Liu KH, Nellis M, Uppal K, et al. Reference Standardization for Quantification and Harmonization of Large-Scale Metabolomics. *Anal Chem.* 2020;92(13):8836-8844.

113. He D, Yan Q, Uppal K, et al. Metabolite stability in archived neonatal dried blood spots used for epidemiological research. *Am J Epidemiol.* 2022. (Manuscript submitted for publication).

114. Ebrahim SH, Floyd RL, Merritt RK, 2nd, Decoufle P, Holtzman D. Trends in pregnancy-related smoking rates in the United States, 1987-1996. *JAMA : the journal of the American Medical Association.* 2000;283(3):361-366.

115. Shipe ME, Deppen SA, Farjah F, Grogan EL. Developing prediction models for clinical use using logistic regression: an overview. *J Thorac Dis.* 2019;11(Suppl 4):S574-S584.

116. Ritz B, Wilhelm M, Hoggatt KJ, Ghosh JKC. Ambient Air Pollution and Preterm Birth in the Environment and Pregnancy Outcomes Study at the University of California, Los Angeles. *Am J Epidemiol.* 2007;166(9):1045-1052.

117. Kondracki AJ. Prevalence and patterns of cigarette smoking before and during early and late pregnancy according to maternal characteristics: the first national data based on the 2003 birth certificate revision, United States, 2016. *Reprod Health.* 2019;16(1):142-142.

118. de Wolff MG, Backhausen MG, Iversen ML, Bendix JM, Rom AL, Hegaard HK. Prevalence and predictors of maternal smoking prior to and during pregnancy in a regional Danish population: a cross-sectional study. *Reprod Health.* 2019;16(1):82.

119. Tong VT, Jones JR, Dietz PM, D'Angelo D, Bombard JM. Trends in Smoking Before, During, and After Pregnancy — Pregnancy Risk Assessment Monitoring System (PRAMS), United States, 31 Sites, 2000–2005. *Morbidity and Mortality Weekly Report: Surveillance Summaries.* 2009;58(4):1-31.

120. Boyd CJ, Esteban McCabe S, Teter CJ. Medical and nonmedical use of prescription pain medication by youth in a Detroit-area public school district. *Drug and Alcohol Dependence.* 2006;81(1):37-45.

121. Nohr EA, Vaeth M, Baker JL, Sørensen TIA, Olsen J, Rasmussen KM. Pregnancy outcomes related to gestational weight gain in women defined by their body mass index, parity, height, and smoking status. *The American Journal of Clinical Nutrition.* 2009;90(5):1288-1294.

122. Bailey BA. Factors predicting pregnancy smoking in Southern Appalachia. *Am J Health Behav.* 2006;30(4):413-421.

123. Woodby LL, Windsor RA, Snyder SW, Kohler CL, Diclemente CC. Predictors of smoking cessation during pregnancy. *Addiction.* 1999;94(2):283-292.

124. Härkönen J, Lindberg M, Karlsson L, Karlsson H, Scheinin NM. Education is the strongest socio-economic predictor of smoking in pregnancy. *Addiction.* 2018;113(6):1117-1126.

125. Jaddoe VWV, Troe E-JWM, Hofman A, et al. Active and passive maternal smoking during pregnancy and the risks of low birthweight and preterm birth: the Generation R Study. *Paediatr Perinat Epidemiol.* 2008;22(2):162-171.

126. Lai C-C, Huang W-H, Chang BC, Hwang L-C. Development of Machine Learning Models for Prediction of Smoking Cessation Outcome. *International Journal of Environmental Research and Public Health.* 2021;18(5).

127. Newborn Screening Process. https://newbornscreening.hrsa.gov/newborn-screening-process#:~:text=If%20your%20baby%20has%20an,before%20your%20baby%20is%20born. Published 2021. Updated December 2021. Accessed.

128. MacDorman MF, Mathews T, Declercq E. *Trends in out-of-hospital births in the United States, 1990-2012.* US Department of Health and Human Services, Centers for Disease Control and Prevention; 2014.

129. Racial and ethnic differences in breastfeeding initiation and duration, by state - National Immunization Survey, United States, 2004-2008. *MMWR Morb Mortal Wkly Rep.* 2010;59(11):327-334.

130. Ghosh JKC, Wilhelm MH, Dunkel-Schetter C, Lombardi CA, Ritz BR. Paternal support and preterm birth, and the moderation of effects of chronic stress: a study in Los Angeles County mothers. *Archives of Women's Mental Health.* 2010;13(4):327-338.

131. Ritz B, Yan Q, He D, et al. Child serum metabolome and traffic-related air pollution exposure in pregnancy. *Environ Res.* 2022;203:111907.

132. Liang D, Moutinho JL, Golan R, et al. Use of high-resolution metabolomics for the identification of metabolic signals associated with traffic-related air pollution. *Environ Int.* 2018;120:145-154.

133. Simon-Manso Y, Lowenthal MS, Kilpatrick LE, et al. Metabolite profiling of a NIST Standard Reference Material for human plasma (SRM 1950): GC-MS, LC-MS, NMR, and clinical laboratory analyses, libraries, and web-based resources. *Anal Chem.* 2013;85(24):11725-11731.

134. Yu T, Park Y, Johnson JM, Jones DP. apLCMS--adaptive processing of high-resolution LC/MS data. *Bioinformatics.* 2009;25(15):1930-1936.

135. Uppal K, Soltow QA, Strobel FH, et al. xMSanalyzer: automated pipeline for improved feature detection and downstream analysis of large-scale, non-targeted metabolomics data. *BMC Bioinformatics.* 2013;14:15.

136. Deng K, Zhang F, Tan Q, et al. WaveICA: A novel algorithm to remove batch effects for large-scale untargeted metabolomics data based on wavelet analysis. *Analytica Chimica Acta.* 2019.

137. Go Y-M, Walker DI, Liang Y, et al. Reference Standardization for Mass Spectrometry and High-resolution Metabolomics Applications to Exposome Research. *Toxicological Sciences.* 2015;148(2):531-543.

138. Troyanskaya O, Cantor M, Sherlock G, et al. Missing value estimation methods for DNA microarrays. *Bioinformatics.* 2001;17(6):520-525.

139. He D HX, Uppal K, Coleman AL, Walker DI, Ritz B, Jones DP, Heck JE. Biomarkers of Maternal Smoking and the Risk of Retinoblastoma in Offspring. 2022(Manuscript in preparation).

140. Yost K, Perkins C, Cohen R, Morris C, Wright W. Socioeconomic status and breast cancer incidence in California for different race/ethnic groups. *Cancer Causes Control.* 2001;12(8):703-711.

141. Liland KH. Multivariate methods in metabolomics – from pre-processing to dimension reduction and statistical analysis. *TrAC Trends in Analytical Chemistry.* 2011;30(6):827-841.

142. Gromski PS, Muhamadali H, Ellis DI, et al. A tutorial review: Metabolomics and partial least squares-discriminant analysis – a marriage of convenience or a shotgun wedding. *Analytica Chimica Acta.* 2015;879:10-23.

143. Go YM, Walker DI, Liang Y, et al. Reference Standardization for Mass Spectrometry and High-resolution Metabolomics Applications to Exposome Research. *Toxicol Sci.* 2015;148(2):531-543.

144. Uppal K, Walker DI, Jones DP. xMSannotator: An R Package for Network-Based Annotation of High-Resolution Metabolomics Data. *Anal Chem.* 2017;89(2):1063-1067.

145. Bonvallot N, Tremblay-Franco M, Chevrier C, et al. Metabolomics Tools for Describing Complex Pesticide Exposure in Pregnant Women in Brittany (France). *Plos One.* 2013;8(5):e64433.

146. Uppal K, Soltow QA, Promislow DE, Wachtman LM, Quyyumi AA, Jones DP. MetabNet: An R Package for Metabolic Association Analysis of High-Resolution Metabolomics Data. *Front Bioeng Biotechnol.* 2015;3:87.

147. Mernitz H, Wang X-D. Alcohol and tobacco smoke in retinoid metabolism and signaling: implications for carcinogenesis. In: *Alcohol, Tobacco and Cancer.* Karger Publishers; 2006:140-159.

148. O'Byrne SM, Blaner WS. Retinol and retinyl esters: biochemistry and physiology. *J Lipid Res.* 2013;54(7):1731-1743.

149. Erkelens MN, Mebius RE. Retinoic Acid and Immune Homeostasis: A Balancing Act. *Trends Immunol.* 2017;38(3):168-180.

150. Davis I, Liu A. What is the tryptophan kynurenine pathway and why is it important to neurotherapeutics? *Expert Rev Neurother.* 2015;15(7):719-721.

151. Sorgdrager FJH, Naudé PJW, Kema IP, Nollen EA, Deyn PPD. Tryptophan Metabolism in Inflammaging: From Biomarker to Therapeutic Target. *Front Immunol.* 2019;10.

152. Huang JY, Larose TL, Luu HN, et al. Circulating markers of cellular immune activation in prediagnostic blood sample and lung cancer risk in the Lung Cancer Cohort Consortium (LC3). *Int J Cancer.* 2020;146(9):2394-2405.

153. Lovelace MD, Varney B, Sundaram G, et al. Recent evidence for an expanded role of the kynurenine pathway of tryptophan metabolism in neurological diseases. *Neuropharmacology.* 2017;112:373-388.

154. Schröcksnadel K, Wirleitner B, Winkler C, Fuchs D. Monitoring tryptophan metabolism in chronic immune activation. *Clin Chim Acta.* 2006;364(1):82-90.

155. Naz S, Bhat M, Ståhl S, et al. Dysregulation of the Tryptophan Pathway Evidences Gender Differences in COPD. *Metabolites.* 2019;9(10):212.

156. Sorgdrager FJH, Vermeiren Y, Van Faassen M, et al. Age- and disease-specific changes of the kynurenine pathway in Parkinson's and Alzheimer's disease. *J Neurochem.* 2019;151(5):656-668.

157. Soldin OP, Makambi KH, Soldin SJ, O'Mara DM. Steroid hormone levels associated with passive and active smoking. *Steroids.* 2011;76(7):653-659.

158. Adamcová K, Kolátorová L, Chlupáčová T, et al. Changes to fetal steroidogenesis caused by maternal smoking. *Physiol Res.* 2017;66(Suppl 3):S375-s386.

159. Akaike A, Shimohama S, Misu Y. Nicotinic Acetylcholine Receptor Signaling in Neuroprotection. 2018.

160. Thomson NC, Chaudhuri R, Spears M, et al. Arachidonic acid metabolites and enzyme transcripts in asthma are altered by cigarette smoking. *Allergy.* 2014;69(4):527-536.

161. Gwaltney-Brant SM. Chapter 7 - Nutraceuticals in Hepatic Diseases. In: Gupta RC, ed. *Nutraceuticals.* Boston: Academic Press; 2016:87-99.

162. Wang C-S, Wang S-T, Chang T-T, Yao W-J, Chou P. Smoking and Alanine Aminotransferase Levels in Hepatitis C Virus Infection: Implications for Prevention of Hepatitis C Virus Progression. *Arch Intern Med.* 2002;162(7):811-815.

163. Nelson TM, Borgogna JC, Michalek RD, et al. Cigarette smoking is associated with an altered vaginal tract metabolomic profile. *Sci Rep.* 2018;8(1):852.

164. Titz B, Szostak J, Sewer A, et al. Multi-omics systems toxicology study of mouse lung assessing the effects of aerosols from two heat-not-burn tobacco products and cigarette smoke. *Computational and Structural Biotechnology Journal.* 2020;18:1056-1073.

165. Carter RA, Pan K, Harville EW, McRitchie S, Sumner S. Metabolomics to reveal biomarkers and pathways of preterm birth: a systematic review and epidemiologic perspective. *Metabolomics.* 2019;15(9):124.

166. Xu T, Holzapfel C, Dong X, et al. Effects of smoking and smoking cessation on human serum metabolite profile: results from the KORA cohort study. *BMC Medicine.* 2013;11(1):60.

167. Hukkanen J, Jacob P, Benowitz NL. Metabolism and disposition kinetics of nicotine. *Pharmacol Rev.* 2005;57(1):79-115.

168. Demirev PA. Dried Blood Spots: Analysis and Applications. *Anal Chem.* 2013;85(2):779-789.

169. Rumrich IK, Vähäkangas K, Viluksela M, Hänninen O. Chained Risk Assessment for Life-Long Disease Burden of Early Exposures–Demonstration of Concept Using Prenatal Maternal Smoking. *International Journal of Environmental Research and Public Health.* 2020;17(5):1472.