Discussion Paper

Methods for a National Birth Cohort Study

Greg Duncan, Virginia Lesser, Barbara Entwisle, Graham Kalton, Andy Shih, Elaine Faustman, Rosalind Wright, Kerry Anne McGeary, Richard Gershon, Steven Wysmuller, Shelley Merritt, and Charlee Alexander

May 31, 2016
About this Paper

This paper is part of the National Academy of Medicine Perspectives series, *Lifelong Impact: Why the United States Needs a National Birth Cohort Study*. To download both papers in the series, visit www.nam.edu/lifelonghealth.

**Background:** With funding from the Robert Wood Johnson Foundation, the National Academy of Medicine (NAM) and the Division of Behavioral and Social Sciences and Education of the National Academies of Sciences, Engineering, and Medicine held two expert meetings to identify the reason why the US needs a longitudinal birth cohort study and the methods that can be used to make its implementation successful. The meetings drew on the National Research Council/Institute of Medicine report *The National Children’s Study 2014: An Assessment*. In addition to engaging a diverse array of stakeholders from academia, philanthropy, local communities, industry, and government, the meetings included discussions about how to design a national longitudinal cohort study that is multidisciplinary, focuses on the main drivers of health, engages communities, employs a diverse set of data sources, and includes innovative techniques in data analysis. Discussions included how to use the findings from such a study to improve and direct resources toward improvements in the drivers of health.

The final product of this effort is *Lifelong Impact: Why the United States Needs a National Birth Cohort Study*, a two-paper NAM Perspectives series. The goal of the series is to provide input into and catalyze action toward improvements in the design of future longitudinal birth cohort studies.

**About NAM Perspectives:** NAM Perspectives are expert commentaries and discussion papers by leading voices in health and health care. They are not reports of the National Academy of Medicine or the National Academies of Sciences, Engineering, and Medicine and therefore are not subject to their review processes. Learn more at www.nam.edu/perspectives.
Methods for a National Birth Cohort Study

Greg Duncan, University of California, Irvine; Virginia Lesser, Oregon State University; Barbara Entwisle, University of North Carolina at Chapel Hill; Graham Kalton, Westat; Andy Shih, Autism Speaks; Elaine Faustman, University of Washington; Rosalind Wright, Icahn School of Medicine at Mount Sinai; Kerry Anne McGeary, Robert Wood Johnson Foundation; Richard Gershon, Northwestern University; Steven Wysmuller, IBM; Shelley Merritt, Booz Allen Hamilton; Charlee Alexander, National Academies of Sciences, Engineering, and Medicine

As we outlined in the previous paper in this series (Riley et al., 2016), our nation needs foundational data in order to understand how social, physical, chemical, and nutritional environments interact to impact how Americans grow, live, and prosper. To satisfy this need, we propose a nationally representative birth cohort study beginning in the prenatal period and following the children through adulthood. Existing research efforts are inadequate because their data are not sufficiently comprehensive and representative to identify both positive and negative factors affecting children’s health or to fully understand health inequities in the United States.

A crucial element of the proposed study is a well-designed national probability sample from which conclusions can be drawn to the larger population from which the sample was randomly selected. In contrast, self-selection sampling consists of volunteers who elected to be part of a study. This technique introduces self-selection bias and can lead to a sample that is not representative of the population being studied. In fact, a report by the National Research Council (NRC) and the Institute of Medicine (IOM), *The National Children’s Study 2014: An Assessment*, endorsed a probability sample design for a future national longitudinal birth cohort study (NRC/IOM, 2014; Riley et al., 2016). In this paper, we provide an overview of a feasible sample design, methods for stakeholder engagement, tools for data collection, and approaches for providing access to the data that would maximize its value.

**SAMPLE DESIGN AND SAMPLE SIZE**

The area-based sample design used in the National Children’s Study (NCS) in 2009–2010 proved to be problematic and unsustainable as implemented. As a result, beginning in 2010 several variations of the initial design were tested or explored (Baker et al., 2014; Hirschfeld et al., 2011, Robbins et al., 2015; Trasande et al., 2011). One of the tested designs involved sampling pregnant women directly via prenatal care providers, rather than sampling women via the geographic areas in which they lived. A version of this approach was piloted from 2012 to 2014 in three NCS locations: Jefferson County, KY; Harris County, TX; and Worcester County, MA. The piloted version involved constructing lists of prenatal providers who served pregnant women in these counties, selecting samples of providers within each county, and then, for each selected provider, sampling time periods during which all pregnant women making their first prenatal visits would be invited to enroll in the NCS (National Children's Study Program Office, 2015).

For the future national birth cohort study we propose in this series, we suggest an alternative version of this approach for establishing the birth cohort. This alternative version was discussed by the National Institute of Child Health and Human Development for use in the NCS around 2014 but was not implemented (NRC, 2014). The approach involves

- selecting a nationally representative sample of hospitals and birthing centers in the United States;
- obtaining lists of prenatal providers who have birthing privileges at the sampled hospitals;
- sampling providers from those lists; and
- sampling time periods to enroll pregnant women visiting the providers’ offices for their first prenatal visits, as was done in the version used in the NCS pilot study.

There are two lists of U.S. hospitals—one maintained by the American Hospital Association and the other by Verispan L.L.C.—that can serve as the basis for the sample frame construction of all hospitals with birthing facilities in the United States. Both frames have high coverage and contain valuable data on the hospitals for use in sampling, thus supporting the selection of an efficient hospital sample. In order for a study to make reliable estimates for certain demographic groups, those groups often need to be oversampled. The hospital-based sample design would allow for oversampling. For example, oversampling public hospitals located in low-income areas in order to over-represent disadvantaged women in the cohort.

Within selected hospitals, a probability sample of prenatal providers with birthing privileges at the hospital would be invited to participate. One advantage of this approach is that the hospital connection will potentially increase prenatal provider participation. Another considerable advantage is that it restricts the collection of the birth biological specimens to the sampled hospitals.

We believe this design will yield a cost-effective and logistically feasible study. There are of course minor limitations. Data are not collected from sampled women during the preconception stage and the early stages of pregnancy before they seek prenatal care (although retrospective reports and record data may sometimes be useful). However, a sizable proportion of the sampled women would likely be enrolled within the first trimester of their pregnancies. In 2007, more than 70 percent of pregnant women who delivered a baby received prenatal care in the first trimester (HHS, 2016). Some pregnant women do not receive prenatal care, but they could be enrolled by study staff soon after the birth of the child (Kozinetz et al., 2016). Pregnant women who neither seek prenatal care nor give birth in a hospital are not covered by the design, but this non-coverage is minimal—in 2012, births outside of hospitals represented less than 1.4 percent of all births in the United States (MacDorman et al., 2014).

A solution that can address the limitation of lack of direct data before the first prenatal visit, including during the preconception period, is to include subsequent pregnancies within a specified period from families already enrolled in the study. Although such a sample is clearly not representative of all births, including the fact that the environment will be different for subsequent pregnancies, this approach would be very cost-effective in that information about the enrolled families would already be collected with respect to the first enrolled child.

The large sample size of the NCS was identified by the NRC/IOM report as being a very costly goal. The NCS called for a sample size of 100,000 births in order to enable evaluation of outcomes with relatively low prevalence on the order of 2 per 1,000 (0.2 percent), such as cerebral palsy and type 1 diabetes (Branum et al., 2003). As the first paper in this series describes (Riley et al., 2016), there are many important outcomes that are more widespread in our nation, such as preterm birth (9.6 percent of births [March of Dimes, 2015]) and childhood asthma (10 percent of the population [Bloom, 2013]). A much smaller sample size would suffice for studying these higher-prevalence outcomes. For a future study, we suggest that the sample size could be reduced to as few as 30,000 to 50,000 births (Barksdale Boyle et al., 2015; Duncan et al., 2015). This smaller sample size sacrifices very few possible study goals and would make for a more manageable study as well as a very sizable reduction in cost. An analysis presented in the NRC/IOM report indicated a roughly proportional relationship between 21-year project costs and sample size, implying that a sample size of 50,000 births would cost roughly half as much as a study with a sample of 100,000 births (Duncan et al., 2015; NRC/IOM, 2014).
STAKEHOLDER ENGAGEMENT

Studies of this magnitude have many stakeholders: funders, participants, the research community, and the public at large. Engagement of these stakeholders and sensitivity to the sociopolitical context are critical to study execution, as well as the level and duration of political support and financing. For longitudinal studies in particular, the duration and ongoing resource requirement pose additional challenges to maintain stakeholder enthusiasm and support, and speak to the need for innovative approaches in funding, implementation, and sustainability (Bennett et al., 2011).

To begin designing engagement strategies, the study should aim to gather information on factors that resonate, as well as elements that cause concern, among members of the communities served by the selected hospitals. The engagement strategies should consider common barriers to successful community engagement in research—geography, culture, and socioeconomic status (Sapienza et al., 2007). Community input can be gleaned from a variety of methods, such as in-person community events, focus groups, social media campaigns, etc. The specific methods used to obtain community input should be consistent with how each community typically engages with those outside the community. Information gained from communities should be used to develop study recruitment and retention materials to improve community members’ trust of the researchers, which in turn may increase participation rates (Booker et al., 2011).

To increase engagement of study participants, the research community, funders, and the population at large, one strategy is to deliver general, easily digestible, aggregated descriptive results of the study in a timely manner. However, within longitudinal studies, there are concerns related to reporting results. One is that reporting results could lead to behavior change among study participants. Yet, if the behavior change improves health, it would be unethical not to share the information. This concern could be balanced by determining how information is presented to stakeholders—whether it is individual results delivered to study participants, aggregate results for a national or regional demographic group, or a general national data brief of study findings (e.g., a national study on post-partum depression and child development).

Communication of study findings can be enhanced by clear, proactive, and high-impact messaging and communication practices culled from the fields of science and risk communication (Fischhoff, 2013). Robust communication of research results will allow stakeholders throughout the nation to develop a vested interest in the study outcomes, which will in turn increase the impact of the science by increasing uptake and sustainability of the public health agenda and promoting evidence-based decision making in policy (Woolf et al., 2015).

DATA COLLECTION METHODS

Since the initial planning of the NCS in the early 2000s, data collection methodologies have evolved rapidly. Today, methods are available to collect survey data more cost-effectively and in ways that are less burdensome to respondents. For example, respondents can complete questionnaires online rather than scheduling a home interview, which holds promise for a longitudinal study with geographically mobile participants. When designing the specifics of a particular protocol, researchers could leverage features like these to reduce study costs and respondent burden. The study could collect various types of data, including survey questionnaires, physical measurements, administrative records, and neighborhood conditions, from existing social and environmental data. Although challenging now, it will be possible in the not-too-distant future to substitute administrative data for some of the information now collected in interviews. The collection of global positioning system (GPS) coordinates for sample participants is common and will fulfill a number of important needs for data collection and
interpretation now and in the future (Lioy et al., 2009). Smartphones are now routinely used to passively collect data on movement, exercise, and standing activity (del Rosario et al., 2015) and define exposures throughout the day (Browning and Soller, 2014).

Next, we briefly outline three key dimensions of the data that need to be collected for a national longitudinal birth cohort study.

### Measurement of Health

The study should include measures of the physical and mental health, as well as the well-being, of the study participants. The study should rely on outcome measurement tools that have been developed through the National Institutes of Health in recent years for children and adults, such as the Patient-Reported Outcomes Measurement Information System (PROMIS). To assess child outcomes, PROMIS pediatric self-reported instruments are available for children ages 8–17, and parent proxy reports are available for children ages 5–17. PROMIS covers the physical health domains—fatigue, mobility, asthma, and pain behavior—and the mental health domains—anxiety, depression, and anger (NIH, 2016).

The study should also collect blood and urine samples from participants frequently. These samples can be analyzed for a variety of measures, including genetics, epigenetics (DNA methylation and histone modification), and common biomedical indicators, such as glucose, lipids, and hormones, etc. Given that the study will have already partnered with hospitals and prenatal providers for recruitment, it may also be possible to extend that partnership to pediatricians, who could report common health measures such as height, weight, pulse, blood pressure, vision, hearing, other developmental measurements, and assessments to monitor well-being. Finally, while not quite a reality today, within the timeframe of the study we propose, it is probable that the complete medical history of each child will be available to researchers through electronic medical records (EMRs). Sampled persons’ EMRs will be a vital data resource within a cohort study because they allow for phenotyping and public health surveillance without undue burden on the study participants. EMRs hold great promise for epidemiological research, although they have not yet been used in a national cohort study across hospital systems (Hruby et al., 2016).

### Social Environment Tools

As discussed in the previous paper in this series, the study we propose will need to collect extensive data on the social environment, including the family environment, accessibility of health care, cultural factors, experiences of violence, social networks and support, media exposure and use, screen time, and the built environment.¹ A growing literature suggests that, individually and together, aspects of the social environment have consequences for the health and well-being of children and adults (Duong and Bradshaw, 2016; Golding et al., 2009; Wright et al., 1998).

Some data on the social environment are routinely collected in many surveys. For example, household roster data can be used to measure the presence and characteristics (age, occupation, education, etc.) of the biological mother and father, social mother and father, siblings, and other adult relatives, including grandparents. As part of this, it is important to measure not only race and ethnicity, but also the immigrant status and home country of the parents and grandparents, as well as languages spoken. It is also important to measure the resources available, including family income, parental education, labor force status, and availability and type of health insurance. Life events—such as the death, serious illness or injury, or incarceration of a family member—are socially patterned and help to shed light on

¹ The man-made parts of the environment where respondents live and work or attend school.
health disparities. Social networks can be evaluated by leveraging information about contacts that are provided to the study for tracing purposes change over time, as was done in the Framingham Heart Study.

Another class of important social factors, such as parenting style, interpersonal relationships, family conflict, control, and organization, are less commonly measured. However, because they are most proximal in children's lives, these factors are often the most powerful (Bronfenbrenner and Morris, 2007). When determining what tools to use to measure these important social factors, we can look to studies such as the Early Childhood Longitudinal Study--Birth Cohort (ECLS-B). There are well documented measures of social factors for the neighborhood—such as cohesion and social control, crime and safety, concentrated disadvantage and affluence, ethnic and racial segregation, neighborhood turnover, availability of green spaces, walkability, food store availability, and accessibility of health care and other services. Some neighborhood measures require input from the respondent, but many draw on external resources, such as street maps and other data sources, like the American Community Survey, which serves to reduce respondent burden.

Recently, studies are also beginning to measure the impact of digital technologies on child health and development. Access to and use of these technologies can be measured via questionnaire, as is already being done in many studies (Duch et al., 2013). These questionnaire-based measures can be complemented by observations of devices in the home.

**Chemical, Physical, and Nutritional Environment Tools**

Measures of the chemical, physical, and nutritional environments of children and their parents are necessary to understand health outcomes across development, socioeconomic, and geographic boundaries. Highly promising technologies provide an unprecedented measurement platform for collecting these data. Examples include more efficient dietary intake assessments, remote sensing, passive personal monitors, improved analytical chemistry, and new types of biological samples. Dietary intakes can now be measured with reasonable accuracy using web-based 24-hour recalls or mobile phone–based food diaries (Diep et al., 2015; Ravi et al., 2015; Thompson et al., 2015). Remote sensing enhances our ability to assess exposure to air toxicants because it can fill in gaps in time and place left by traditional ground-based monitoring systems (Al-Hamdan et al., 2014). As an example, remote sensing and available data have been used to develop well-established models to estimate exposures to air toxicants in populations across the globe (Geddes et al., 2016; van Donkelaar et al., 2015). Passive personal monitors, such as silicone wristbands, can be used to measure exposures from a suite of environmental chemicals. These wristbands do not require calibration, batteries, or extensive preparation and can provide individualized exposure data for both parents and children (O’Connell et al., 2014a,b; Kile et al., 2016). With recent improvements in analytical chemistry, targeted and untargeted analytical chemistry approaches can be used to obtain measurements of hundreds of chemicals in relatively few assays.

New types of biological samples, such as deciduous teeth, hair, and blood spots, can be used for exposure assessment purposes. For example, after deciduous teeth fall out, participating parents can mail them to the study coordinating center, where they can be sent to a laboratory and analyzed with sophisticated methodologies that combine histological and chemical analysis to precisely sample tooth layers that correspond to specific life stages. These analyses have the potential to reconstruct exposure in the second and third trimesters of prenatal development and during early childhood (Arora and Austin, 2013). Much like tooth analysis, shafts of hair can provide a temporal record of exposures and response to metals and other chemicals, and can also record biological responses, such as cortisol content in the hair (Kirschbaum et al., 2009; Moro et al., 1992; Raeppel et al., 2016). Deciduous teeth and hair samples allow for more specific characterization than can be produced using other relatively
inexpensive measures, such as questionnaires. They also reduce the number of samples needed, as exposures from more than one time point can be measured in a single sample.

Other Data Collection Considerations

It is very likely that available data collection technologies will expand and improve during the course of following a birth cohort through infancy, childhood, adolescence, and adulthood. Therefore, the study we propose will need to be flexible and able to adopt new technologies as they become accepted for use in research. Numerous technologies undergoing development and testing will expand the study’s ability to collect data in more cost-effective and less burdensome ways.

Two more examples that have been tested in smaller-scale or cross-sectional studies are: GPS tracking of participants and mobile phones that are enhanced with applications and sensors. These can collect a range of exposure and outcome information and have the potential to provide major advancements in understanding how children and adolescents move through and interact with environments and exposures. Tracking this movement with GPS-enabled devices will not only provide more precise exposure measurements, but will also shed light on socioeconomic and racial/ethnic differences in those exposures. Pilot studies collecting this kind of data are currently under way in different locations across the country. While it is not yet possible to bring this work to a national scale, it will be possible to do so within the foreseeable future (Elgethun et al., 2003; Browning and Soller, 2014; Fenske et al., 2005; Viet et al., 2013).

Studies have used mobile phones to monitor noise (Neitzel et al., 2016) and exposure to air pollutants (Nieuwenhuijsen et al., 2014, 2015). By monitoring changes in behavior, phones can be used to monitor daily stress levels (Bogomolov et al., 2014a), social networks (Onnela et al., 2014), and, when coupled with demographics, possibly crime risk (Bogomolov et al., 2014b). As with GPS technologies, these tools show great promise, but work is still needed to bring them to a national scale. Thus, although valid and reliable measures to capture key aspects of the social and physical environment are available now, we can reasonably expect measurement opportunities to improve in the near future.

DATA ORGANIZATION, MANAGEMENT, AND RELEASE

Good data management practices will help the proposed birth cohort study maximize its benefits to the research community and public. The study will repeatedly collect substantial amounts of data on study participants through a variety of means, including questionnaires, diaries, biological specimens and environmental samples, direct monitoring, EMRs, GPS coordinates, and links to administrative data. Providing security for collection and initial transfers of data—and organizing, documenting, storing, and releasing appropriately protected versions of these data to the research community—requires considerable effort, diverse expertise, and advance planning. Fortunately, solutions to these problems have been developed in a number of national studies over the past several decades. Further, the experience and expertise needed to adapt these solutions to the evolving landscape with respect to data collection modality and participant engagement are already available.

The two most important components to consider in a data organization plan are (1) standardization of protocols and data management procedures and (2) data release to and dialogue with the general research community. The most successful national data collection studies are highly centralized, with a central office overseeing the design of the data collection instruments, field data collection operations, data coding and cleaning, linkages across data components and to external data, and documentation and release of the collected data with appropriate levels of protection for the confidentiality/privacy of the respondents. Distributing these duties to regional centers can create inefficiencies and introduce opportunities for data
non-comparability. Specialized tasks, such as assembling administrative records and performing specimen assays, may be subcontracted to third parties, but it is vital to monitor these processes centrally.

Timely release of data to the research community increases the scientific benefits generated from large public investments in studies such as the one proposed. We recommend that the data be released to the research community as soon as they are cleaned and documented. Existing national studies such as the Panel Study of Income Dynamics have been releasing data shortly after cleaning for more than 40 years. In addition, the National Longitudinal Study of Adolescent to Adult Health shares many of the characteristics of the study we propose, such as data collected from questionnaires, biological, and administrative data. This program has applied a policy for immediate data release over the past 20 years.

In planning for data release, it is vital to maintain the confidentiality of the study participants. This is difficult to do in longitudinal studies, but various data access modes (statistical enclaves, restricted use and public use files), and a variety of statistical disclosure control methods can be employed to protect participant privacy while allowing researcher access to the data (Domingo-Ferrer and Muralidhar, 2016; Duncan et al., 2011; Hundepool et al., 2012).

Different organizational structures could be considered to manage the proposed study. In the United States, some large-scale longitudinal studies have been managed by statistical agencies with extensive statistical and survey experience, but these have often included collaborations with survey organizations for data collection. Other such studies have been managed by university-based groups with one or more principal investigators who bring scientific and survey knowledge to the project with funding obtained from a federal grant or cooperative agreement. A university research group can collaborate with or subcontract to a survey research organization for data collection, training, and management. While there are advantages and disadvantages to both approaches, considerations should be made to allow for flexibility, bureaucracy, and regulatory burdens (e.g., Office of Management and Budget clearance, compliance with the Federal Information Security Modernization Act).

CONCLUSION

The NCS, as well as a number of similar studies conducted both in the United States and abroad, have produced a wealth of information regarding methodologies that can be used to design a nationally representative birth cohort study. A study of this magnitude requires a breadth of expertise and ample time for design. In terms of content, the designers of this new study can tap into lessons learned from the NCS and other studies to make well-informed decisions about measures to be included—balancing the scientific importance of the measures with their logistical feasibility, costs, and respondent burden. Designers should engage stakeholders early in the implementation phase to get input for recruitment and as results are being released to show funders, other stakeholders, and society at large the benefits of the research.

In this paper, we provide reasonable and actionable guidance regarding what can be moved forward from the NCS and other experiences and outline how that work may be coupled with improvements in technology for data collection to make a new birth cohort study logistically feasible and cost-effective. This guidance, coupled with recommendations for study management and oversight in the NRC/IOM review of the NCS (2014), provide a solid platform on which such as study could be built.

Our nation desperately needs data that can be collected only in a nationally representative birth cohort study of the type we have outlined. Such data will allow policy makers, researchers, and others to understand how the social, physical, chemical, and nutritional environments all interact to impact how Americans grow, live, and prosper. Such data
are critical to identify those at heightened risk so that interventions may be applied early to promote optimal development and help end health disparities.

Authors

Greg Duncan, PhD, is a distinguished professor of education at the University of California Irvine, in Irvine, California.

Virginia Lesser, DrPH, is a professor and chair of the department of statistics at Oregon State University, in Corvallis, Oregon.

Barbara Entwisle, PhD, is a distinguished professor of sociology and vice chancellor for Research at the University of North Carolina-Chapel Hill, in Chapel Hill, North Carolina

Graham Kalton, PhD, is a senior vice president at Westat in Rockville, Maryland.

Andy Shih, PhD, is the senior vice president, scientific affairs at Autism Speaks in New York, New York.

Elaine Faustman, PhD, is a professor of environmental and occupational health science at the University of Washington, in Seattle, Washington.

Rosalind Wright, MD, MPH, is the Dean for Translational Biomedical Sciences, Icahn School of Medicine at Mount Sinai, New York, New York.

Kerry Anne McGeary, PhD, is a senior program officer for health economics at the Robert Wood Johnson Foundation, in Princeton, New Jersey.

Richard Gershon, PhD, is the vice chair for research in medical social sciences at Northwestern University, in Chicago, Illinois.

Steven Wysmuller, MS, is the sustainability and emergency management leader for IBM’s cognitive solutions team in Somers, New York.

Shelley Merritt, MS, is a senior associate at Booz Allen Hamilton, in McLean, Virginia.

Charlee Alexander is an associate program officer with the Forum on Investing in Young Children Globally and the Innovation to Incubation (i2I) program at the National Academies of Sciences, Engineering, and Medicine.

References


Suggested Citation


Sponsor

Support for this Perspective was provided by the Robert Wood Johnson Foundation.

Disclaimer: The views expressed in this Perspective are those of the authors and not necessarily of the authors’ organizations or of the National Academy of Medicine (NAM). The Perspective is intended to help inform and stimulate discussion. It has not been subjected to the review procedures of, nor is it a report of, the NAM or the National Academies of Sciences, Engineering, and Medicine. Copyright by the National Academy of Sciences. All rights reserved.