BMJ Nutrition, Prevention & Health

Obesity Prevention in Early Life (OPEL) study: linking longitudinal data to capture obesity risk in the first 1000 days

Erika R Cheng , Sami Gharbi, Tammie L Nelson , Sarah E Wiehe

To cite: Cheng ER, Gharbi S, Nelson TL, *et al*. Obesity Prevention in Early Life (OPEL) study: linking longitudinal data to capture obesity risk in the first 1000 days. *BMJ Nutrition, Prevention & Health* 2024;**7**:e000671. doi:10.1136/bmjnph-2023-000671

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjnph-2023-000671).

Department of Pediatrics, Indiana University School of Medicine, Indianapolis, Indiana, USA

Correspondence to
Dr Erika R Cheng;
echeng@iu.edu

Received 26 April 2023
Accepted 16 December 2023
Published Online First
4 January 2024

ABSTRACT

To develop robust prediction models for infant obesity risk, we need data spanning multiple levels of influence, including child clinical health outcomes (eg, height and weight), information about maternal pregnancy history, detailed sociodemographic information of parents and community-level factors. Few data sources contain all of this information. This manuscript describes the creation of the Obesity Prevention in Early Life (OPEL) database, a longitudinal, population-based database that links clinical data with birth certificates and geocoded area-level indicators for 19 437 children born in Marion County, Indiana between 2004 and 2019. This brief describes the methodology of linking administrative data, the establishment of the OPEL database, and the clinical and public health implications facilitated by these data. The OPEL database provides a strong basis for further longitudinal child health outcomes studies and supports the continued development of intergenerational linked clinical-public health databases.

INTRODUCTION

Overweight and obesity impact >40 million children under the age of 5.[1 2] The 'first 1000 days' from a woman's pregnancy to her child's second birthday is a critical period for addressing obesity.[3] Numerous risk factors for obesity exist during this time,[4] but little is known about the joint predictive performance of such factors, as population-based datasets often lack sociodemographic data alongside measured heights and weights across pregnancy and early childhood, and birth cohorts may not consistently capture maternal data. Further, while the built environment's influence on childhood obesity is recognised[5–7]; relatively little is known about these relationships because geographical data are often unavailable.

This paper presents the Obesity Prevention in Early Life (OPEL) database, a longitudinal, epidemiological data repository that combines birth certificate, contextual-level and health outcome data for children born in Marion County, Indiana between 2004 and 2019.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ While epidemiological studies have identified numerous risk factors for obesity during pregnancy and early life, risk prediction based on single factors is likely to be incomplete. A better approach would target multiple levels of influence, but existing population-based data sources tend to contain information on maternal risk factors separately from risk factors during infancy and from measures of height and weight across childhood.

WHAT THIS STUDY ADDS

⇒ This paper describes the development of a population-based database containing clinical data linked to children's birth certificates and geocoded area-level indicators created to study determinants of children's obesity risk in the first 1000 days.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ The Obesity Prevention in Early Life (OPEL) linked database provides a strong basis for longitudinal studies on children's early-life obesity risk and supports the continued development and use of linked clinical-public health-geographical databases.

⇒ This paper reports on: (1) the construction of the linked OPEL database; (2) the methodological assessment of the linkage and (3) the clinical, epidemiological and public health implications of the linked OPEL database.

METHODS

A complete list of variables available in the OPEL database is presented as online supplemental appendix A.

Data systems

1. The Child Health Improvement through Computer Automation (CHICA) system was a paediatric primary care clinical decision support system that operated in five Indianapolis community health centres from 2004 to 2019.[8] The database contains EHR data and comprehensive information collected from parents using a customised 20-item prescreening form. This information covers measured height and weight,

insurance status, demographics (eg, child sex, age and race/ethnicity) and social factors (eg, parent health literacy, food/housing insecurity, parental depression and infant feeding practices).
2. Information regarding live births in Marion County, Indiana is gathered through the Indiana Standard Certificate of Live Birth (ie, 'birth certificate) and stored at the Marion County Public Health Department (MCPHD). Birth certificate data provide legal and sociodemographic details about the child's mother and father, along with information on prenatal care, labour and delivery, neonatal conditions and procedures, and discharge.
3. The Social Assets and Vulnerabilities Indicators (SAVI) database, sponsored by the Polis Centre, compiles geocoded data on communities in the 11-county Indianapolis metropolitan statistical area. Data are drawn from >30 federal, state and local providers, and are linked to the smallest available geographical level of aggregation. Variables include welfare, education, health, public safety, housing, demographics and locations of health facilities.

Data sharing and security procedures

We developed a secure, multistep data sharing process, initiating with the transfer of files containing limited protected health information variables for linkage. Confidential data from MCPHD and the CHICA team were encrypted and transferred to a protected server using Slashtmp, an Indiana university-specific secure data transfer system. After the record linkage process, matched records were assigned a unique identifier. Core data files retained their unique link identifier and were only linked when generating analytical files.

Data were stored, linked and analysed on a password-protected and encrypted server with two-factor authentication. The server resides on a private IP address behind two layers of firewalls and network monitoring, with no internet access. Access to data is restricted to key personnel on this study.

Inclusion/exclusion criteria and study subjects

CHICA contains data for 63 741 children from birth to age 18, inclusive of 461 201 well-child visits conducted between 2004 and 2019. Focusing on the age group of 0–4 years, we identified 22 689 children with at least 2 well-child visits between 0 and 6 months, excluding those without a Marion County zip code (constituting <5% of the CHICA cohort). A well-child visit was defined based on the utilisation of an appropriate International Classification of Diseases diagnosis code indicative of preventive care.[9] This longitudinal dataset incorporates outcome data at ages 24, 36 and 48 months through sequential records for individual children. The MCPHD dataset contained 274 306 birth certificate records.

Dataset linkage using deterministic and probabilistic record linkage

Record linkage was performed using unique identifiers present across data sources, including social security number, first/middle/last name, first/middle/last initial, phonetic expression of first/last name (Soundex), month/day/year of birth, gender, race/ethnicity, street number of address, zip code (five digits) and phone number. Before linkage, these identifiers were cleaned and standardised, including removal of hyphens and parentheses from social security and phone numbers, elimination of prefixes, suffixes, hyphens, commas and other non-alphabetic characters from names, limiting zip codes to five digits, and standardising race/ethnicity. Frequencies were checked for unexpected observations such as numeric data for name.

Raw data from CHICA (N=22 689) and birth certificate records (N=274 306) were cleaned and coded using Stata/MP V.14.1.

We identified linkage pairs through deterministic matching, probabilistic matching and manual review (online supplemental appendix B).[10] The deterministic matching used a conservative automated approach that included three algorithms to exactly match different combinations of a subset of identifiers. We hand-validated a subset of matches using first/last name and date of birth (DOB), substituting mother's last name and father's last name in addition to the baby's last name. We refined algorithms to achieve the highest and most robust linkage rates. The first deterministic algorithm used last name, first name and DOB. Father's and mother's last names were used in subsequent deterministic algorithms.

We then used RecMatch software to enhance the likelihood of matching CHICA and birth certificate records using probabilistic algorithms.[11] We performed multiple probabilistic matches using a combination of different blocking and matching schemes to capture additional pairs.

The study team (ERC, SG and SEW) performed manual review using conservative thresholds for acceptance to identify true matches. True positives were determined by manual review, the gold standard.[12] Records representing the same individual were connected via transitive property across pairwise matches (eg, A=B and B=C would connect A, B and C), and assigned a unique study identifier.

Geocoding and geotagging procedures

We then linked individual-level health data from CHICA and birth certificates to SAVI. We cleaned (eg, removing apartment numbers, PO boxes, and nonsensical addresses) and geocoded addresses from the birth certificate. We then used ArcGIS[13] to geotag children's addresses with SAVI data at various levels (eg, street, 50 m buffer, zip code).

Starting with 17 712 unique addresses, we conducted geocoding in two runs: first, using street address and zip code and then street only. Run parameters included spelling sensitivities of 80, minimum candidate scores of 10 and minimum match scores of 80; parameters for Polis were set to 80/10/85, respectively.

Table 1 Sequential deterministic and probabilistic linking for the 22 689 records in Chica against 274 306 Marion County birth certificates

Matching phase	Type of match	# Linked	Cumulative # Linked	Cumulative % Linked	Remaining # Unlinked
1	Deterministic: LN, FN, DOB	17 240	17 240	76.0%	5449
1	Deterministic: Father LN, FN, DOB	350	17 590	77.5%	5099
1	Deterministic: Mother LN, FN, DOB	582	18 172	80.1%	4517
2	Probabilistic variations	1471	19 643	86.6%	3046
3	Manual review	214	19 857	87.5%	2832

DOB, date of birth; FN, first name; LN, last name.

RESULTS

Individual-level data linkage

We linked 87.5% (N=19 857) of patient records between data sources (table 1). Our deterministic algorithms linked 80.1% of all records; 94.9% of which were matched with the first algorithm using last name, first name and DOB. Subsequent algorithms contributed a few additional matches, 350 (1.8% of total records) and 582, respectively (2.9% of total records), with more success matching on mothers' names than on fathers'.

Probabilistic matching linked an additional 1471 records. We used 18 probabilistic algorithms, with the highest yield matching on last or first name, Soundex of last or first name and sex, when blocked by last or first name and DOB (online supplemental appendix B). Our manual review linked 214 more records.

Figure 1 presents the proportion of new matched pairs by linkage phase.

Contextual-level data geocoding and geotagging

We geocoded 92.5% (N=16 396) of unique addresses, or 94.5% of our cohort of babies (N=19 437), with 88.0% successfully geocoded in the first run using street address and zip code and an additional 4.5% geocoded in the second run. Remaining linkages (by transitive property) were determined via an in-house daisy chain process in

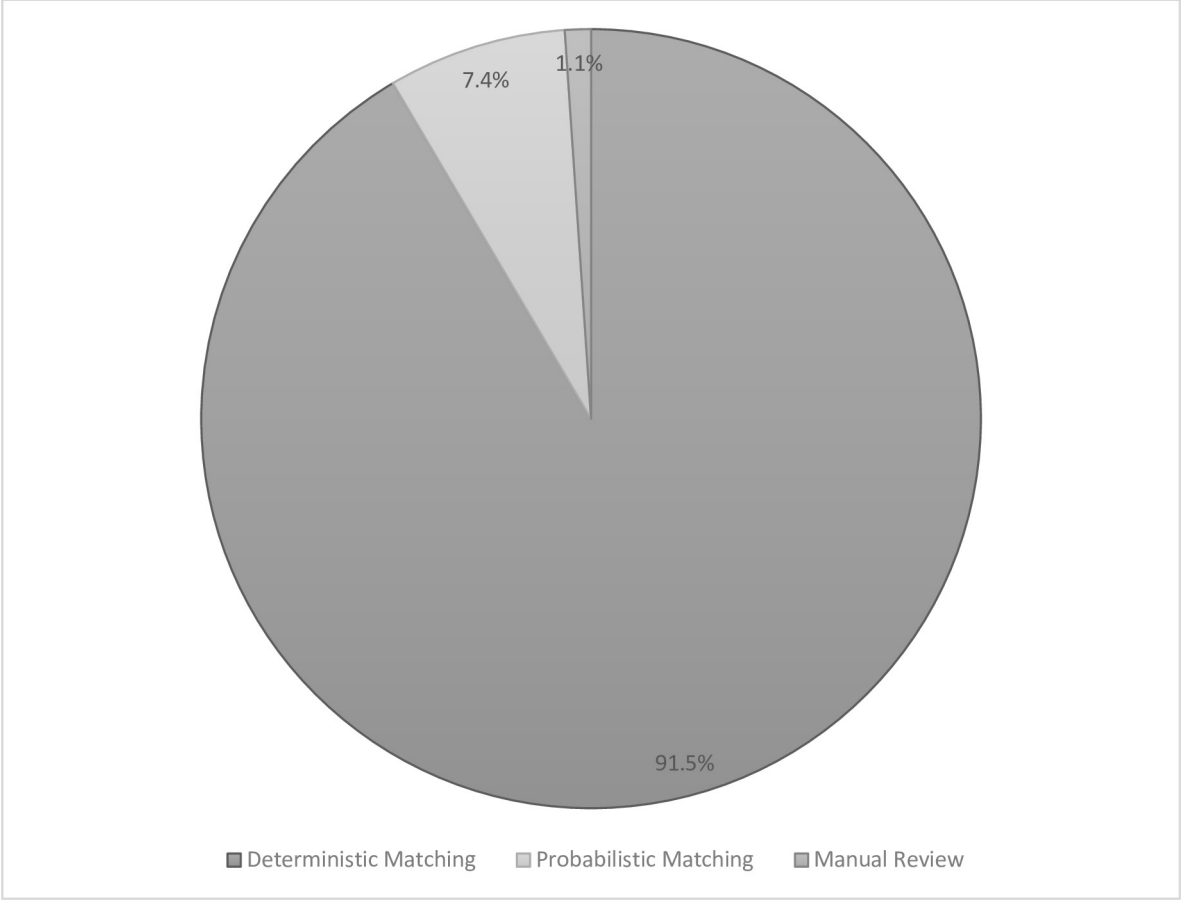

Figure 1 Proportion of matches from deterministic, probabilistic and manual record linkage processes.

STATA. The number of phases was determined by evaluated incremental increases in the number of pairs.

DISCUSSION

This paper introduces the OPEL database, a novel and comprehensive longitudinal data repository. This database, born out of the necessity to address the critical period of the first 1000 days, integrates birth certificate, contextual-level and health outcome data for children in Marion County, Indiana, born between 2004 and 2019. Linking data from CHICA, birth certificates and SAVI enables a multidimensional analysis, covering sociodemographic, clinical and geographic factors. Notably, the study achieves an impressive 87.5% linkage of clinical records and 95% successful geocoding, providing a robust foundation for future investigations into childhood obesity aetiology and related outcomes. If replicated in different states or health systems, this approach could offer valuable insights into the interplay of factors influencing obesity, informing research,[14 15] public health interventions and programming.

Applying our algorithms to different states and health systems may yield different linkage rates. Limitations include constraints of existing data (eg, missingness, data quality), potential misclassification from self-reported, and omission of other correlates of early child health (eg, genetics, paternal factors) not collected from these sources. Attrition may cause selection bias.

Strengths include the use of three matching processes to link more records than prior approaches and our use of primary data collection, EHR and contextual data sources. OPEL spans the first 1000 days and employs replicable methods, making it valuable for investigators in other geographic areas where similar data linkages are possible. OPEL serves as a foundation for additional longitudinal studies that link maternal and paternal information, public health programming and other contextual data that will allow us to comprehensively examine the aetiology of childhood obesity and to track obesity-related outcomes.

Contributors ERC helped conceptualise the study, interpreted data and drafted the manuscript. SG carried out the analyses and interpretation of results. TLN helped with interpretation of data and drafting of the manuscript. SEW provided feedback on the research question, analysis plan and interpretation of the data. All authors critically reviewed the manuscript for important intellectual content and approved the final manuscript as submitted and agreed to be accountable for all aspects of the work.

Funding This research was funded by the National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health (K01DK114383).

Disclaimer The sponsors had no role in the study design; collection, analysis and interpretation of data; writing of report; or decision to submit for publication.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval This study was conducted according to the guidelines laid down in the Declaration of Helsinki and all procedures were approved by the Indiana University Institutional Review Board.

Provenance and peer review Not commissioned; externally peer reviewed by Dr. Emma Derbyshire, Nutritional Insight, UK.

ORCID iDs

Erika R Cheng http://orcid.org/0000-0001-8289-5087
Tammie L Nelson http://orcid.org/0000-0003-4204-7555

REFERENCES

1 Friedrich MJ. Global obesity epidemic worsening. *JAMA* 2017;318:603.
2 Collaborators GO. Health effects of overweight and obesity in 195 countries over 25 years. *N Engl J Med* 2017;377:13–27.
3 Brown T, Moore TH, Hooper L, *et al*. Interventions for preventing obesity in children. *Cochrane Database Syst Rev* 2019;7:CD001871.
4 Woo Baidal JA, Locks LM, Cheng ER, *et al*. Risk factors for childhood obesity in the first 1,000 days: a systematic review. *Am J Prev Med* 2016;50:761–79.
5 Papas MA, Alberg AJ, Ewing R, *et al*. The built environment and obesity. *Epidemiol Rev* 2007;29:129–43.
6 Dunton GF, Kaplan J, Wolch J, *et al*. Physical environmental correlates of childhood obesity: a systematic review. *Obes Rev* 2009;10:393–402.
7 Lovasi GS, Hutson MA, Guerra M, *et al*. Built environments and obesity in disadvantaged populations. *Epidemiol Rev* 2009;31:7–20.
8 Anand V, Biondich PG, Liu G, *et al*. Child health improvement through computer automation: the CHICA system. *Stud Health Technol Inform* 2004;107:187–91.
9 Texas Children's Health Plan. HEDIS quick reference for well-child visits. Secondary HEDIS quick reference for well-child visits. 2020. Available: https://www.texaschildrenshealthplan.org/sites/default/files/pdf/PR-2005-028_HEDIS_Wellchild.pdf
10 Dusetzina SB, Tyree S, Meyer A-M, *et al*. Linking data for health services research: a framework and instructional guide. Rockville, MD Agency for Healthcare Research and Quality (Prepared by the University of North Carolina at Chapel Hill under Contract no.290-2010-000141); 2014.
11 Grannis S, Egg J, Ribeka N. RecMatch: probabilistic patient record matching; 2008.
12 Grannis SJ, Overhage JM, McDonald CJ. Analysis of Identifier performance using a deterministic linkage algorithm. Proceedings / AMIA… Annual Symposium. AMIA Symposium; 2002:305–9
13 Esri Inc. ArcGIS pro 2.8 version. program; 2021.
14 Cheng ER, Cengiz AY, Miled ZB. Predicting body mass index in early childhood using data from the first 1000 days. *Sci Rep* 2023;13:8781.
15 Cheng ER, Steinhardt R, Ben Miled Z. Predicting childhood obesity using machine learning: practical considerations. *BioMedInformatics* 2022;2:184–203.