HHS Public Access

Author manuscript

Matern Child Health J. Author manuscript; available in PMC 2019 April 01.

Published in final edited form as:

Matern Child Health J. 2018 April ; 22(4): 485–493. doi:10.1007/s10995-017-2414-9.

Implementation of a Regional Perinatal Data Repository from Clinical and Billing Records

Eric S. Hall, PhD,
Perinatal Institute, Cincinnati Children's Hospital Medical Center, Department of Pediatrics, University of Cincinnati College of Medicine, Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio United States, (513) 803-2083

James M. Greenberg, MD,
Perinatal Institute, Cincinnati Children's Hospital Medical Center, Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, Ohio United States, (513) 636-3149

Louis J. Muglia, MD, PhD,
Division of Human Genetics, Cincinnati Children's Hospital Medical Center, Perinatal Institute, Cincinnati Children's Hospital Medical Center, Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, Ohio United States, (513) 803-7902

Parth Divekar,
Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio United States, (513) 636-1004

Janet Zahner,
Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio United States, (513) 803-1946

Jay Gholap, MS,
Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio United States, (513) 803-6088

Matt Leonard, and
Perinatal Institute, Cincinnati Children's Hospital Medical Center, Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio United States, (513) 636-0235

Keith Marsolo, PhD
Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, Ohio United States, (513) 803-0333

Abstract

Please Direct Correspondence to: Eric S. Hall, PhD, Perinatal Institute, Cincinnati Children's Hospital Medical Center, 3333 Burnet Ave, ML 7009, Cincinnati, OH 45229, Telephone: 513.803.2083, Fax: 513.803.0968, Eric.Hall@cchmc.org.

Objectives—To describe the implementation of the first phase of a regional perinatal data repository and to provide a roadmap for others to navigate technical, privacy, and data governance concerns in implementing similar resources.

Methods—Our implementation integrated regional physician billing records with maternal and infant electronic health records from an academic delivery hospital. These records, representing births during 2013–2015, constituted a data core supporting linkage to additional ancillary data sets. Measures obtained from pediatric follow-up, urgent care, emergency, and inpatient encounters were linked at the individual level as were measures obtained by home visitors during pre- and postnatal encounters. Residential addresses were geocoded supporting linkage to area-level measures.

Results—Integrated data contained regional billing records for 69,290 newborns representing approximately 81% of all regional live births and nearly 95% of live births in the region's most populous county. Billing records linked to 7,293 infant delivery hospital records and 7,107 corresponding maternal hospital records. Manual review demonstrated 100% validity of matches among audited records. Additionally, 2,430 home visiting records were linked to the data core as were pediatric primary care, urgent care, emergency department, and inpatient visits representing 42,541 children. More than 99% of the newborn billing records were geocoded and assigned a census tract identifier.

Conclusions—Our approach to methodological and regulatory challenges affords opportunities for expansion of systems to integrate electronic health records originating from additional medical centers as well as individual- and area-level linkage to additional data sets relevant to perinatal health.

Keywords

Infant; Medical Record Linkage; Mothers; Population Health; Pregnancy

Introduction

Efforts to improve maternal and child population health outcomes require reliable and timely data to inform planning, resource allocation, testing of key relevant hypotheses, and efficient operation of clinical care programs. For decades, integration of relevant perinatal data has proven difficult due to barriers created by regulatory concerns, privacy issues, questions related to data ownership, technical limitations, and lack of sustainable funding (1–5). Electronic health record (EHR) systems have improved the integration of maternal and newborn records within the labor and delivery setting (6, 7), yet remaining barriers prevent seamless integration of records needed for effective care transitions between providers, agencies, or health care institutions. The many challenges associated with connecting mother-infant dyads, as well as disparate, administratively fragmented data sources restrict our collective ability to define gaps in service, measure the effectiveness of current programs, and to design and test new strategies to reduce infant mortality and improve maternal and infant health.

Previous efforts to establish population-based, integrated data resources such as the Pregnancy to Early Life Longitudinal (PELL) system have been largely driven by health

departments, utilizing vital records and administrative data as core data sets for linkage purposes (8, 9). Through a public-private collaborative partnership PELL has enabled researchers in the Maternal and Child Health Department at the Boston University School of Public Health, the Massachusetts Department of Public Health, and the Centers for Disease Control and Prevention to investigate a broad range of topics related to pregnancy, premature birth, and infant mortality (8, 10, 11), demonstrating the potential for a regional perinatal data repository.

Unlike previously implemented systems that use vital records to establish a core data set, we sought to develop a regional perinatal data system leveraging electronic clinical and billing records. Current study team members have previously evaluated linkage approaches and utilized probabilistic and deterministic strategies to link maternal and child clinical and program-based data to evaluate specific study hypotheses (12–15). Despite a significant overlap of the data sources utilized for these evaluations, each study was approached as an independent investigation necessitating redundant data management, regulatory, and linkage efforts. We aim to develop a sustainable infrastructure to support future evaluations and minimize the need for duplicative efforts. In this report, we present our system architecture as well as our approach to privacy and data governance concerns. We also provide initial system metrics and describe opportunities facilitated by the inclusion of rich clinical data in the integrated Maternal and Infant Data Hub (MIDH).

Materials and Methods

Project Initiation

The project was initially conceived by members of Cincinnati Children's Hospital Medical Center (CCHMC). A team of clinicians, informaticians, and programmers proposed a three year development plan and were granted funding through the CCHMC Research Foundation as well as the Center for Clinical and Translational Science and Training at the University of Cincinnati to support initial efforts. Development milestones were outlined as follows: Year 1) Partner with the region's academic delivery hospital and one home visiting agency, address data sharing and regulatory issues, evaluate potential data sources, and initiate database design and development; Year 2) Extract data sets from each partner institution, load geocoded and formatted data into an MIDH database, and implement an automated data linking strategy; Year 3) conduct several demonstration use cases to be leveraged in applications for additional funding intended to provide ongoing project sustainability. This report was drafted at the close of the second year efforts.

Population

Physicians at CCHMC provide nearly comprehensive clinical coverage for neonates born throughout the greater Cincinnati, tristate region as they are contracted to direct newborn care in each of the region's 14 delivery hospitals. Although an overwhelming majority of the newborn encounters with CCHMC neonatologists and pediatricians occur in the delivery hospital setting (~98%), these encounters generate physician billing records maintained by the children's hospital EHR system, forming a regional data set. As a consequence, CCHMC newborn billing records (representing approximately 23,000 newborns annually, including

all infants admitted to regional neonatal intensive care units), may function as a backbone for data linkage serving a role typically reserved for population-based vital birth records. Our implementation integrated physician newborn billing records with EHRs from a single delivery hospital, the University of Cincinnati Medical Center (UCMC), an academic medical center delivering approximately 2,500 infants each year. The EHRs provided a richer source of clinical measures for the subset of regional newborns encountered at the UCMC delivery hospital setting. Our study includes data representing births in 2013–2015.

Data sets

For this implementation, MIDH data core elements originated from regional newborn billing records as well as corresponding maternal and newborn medical records generated during the delivery encounter. Newborn billing records included billable diagnoses, addresses, and identifiers used for linkage. Data abstracted from the delivery hospital EHR system included coded diagnoses, procedures performed, medication administrations, vital signs, laboratory results (including maternal toxicology representing intrauterine exposures to substances of abuse) infant gestational age and birth weight, and length of hospital stay for both mother and baby. Rather than maintaining a single “current address” variable, dates and residential addresses were captured with each patient encounter. Identifier fields including names, dates of birth, sex, and addresses were also shared to enable linkage within the MIDH.

Each newborn encounter with a CCHMC physician generated data captured within the regional physician billing data set including an indicator of the encounter location. Thus, a corresponding newborn medical record was expected to found for each physician billing record generated at the UCMC location. Conversely, a fraction of UCMC newborn medical records (~5%) did not have a corresponding physician billing record in cases when infants were transferred to the children’s hospital immediately after delivery (where physician billing charges were initiated) or in other rare cases when newborn care was provided exclusively by non-CCHMC physicians (16).

The linked data core enables association with ancillary data sets at the individual or area level (Figure 1). For example, using individual identifiers or geospatial data, records can be linked to community-based home visitation records, pediatric primary or specialty care, urgent care, and emergency department visit EHR records, research biorepository records, vital birth and death records, environmental exposure data including measures of airborne particulate matter (17), or other databases measuring sociodemographics or social determinants such as the American Communities Survey (18). Our demonstration includes integration of 1) neonatal as well as inpatient and outpatient pediatric medical records generated at the children’s hospital, or a CCHMC clinic subsequent to the delivery hospital transfer or discharge, and 2) records for participants in home visiting services for high-risk, first time mothers that were collected as part of the Every Child Succeeds program (19). Neonatal and pediatric medical records were generated during children’s hospital encounters subsequent to the infant’s discharge from the delivery hospital. Data shared from these EHRs includes diagnoses, procedures, medications, and laboratory results and identifier fields to facilitate record linkage. Additionally, measures of healthcare utilization including dates and types of encounters (e.g. emergency department, urgent care, primary care, etc.)

were captured as were measures of child growth and development. Home visiting data contain measures representing high-risk maternal-infant dyads including sociodemographics, content and frequency of home visits, and measures of maternal mental health, substance use, and parenting environment attained during pre- and postnatal time points. Only a subset of patients represented in the data core were expected to have a corresponding record in these ancillary data sets.

Data Acquisition and Preparation

The MIDH utilizes the Observational Medical Outcomes Partnership (OMOP) data model as its core table structure providing standard representations for many common healthcare data domains, including diagnoses, demographics, laboratory results and other observational data (20). Additionally, it maintains a set of vocabulary tables that provide mappings between vocabularies of the same domain (e.g., ICD-9, ICD-10, SNOMED), and a common target vocabulary for each domain, facilitating data integration from multiple sources. MIDH source data are delivered via secure file transfer and refreshed quarterly. After linking, records are de-identified, though a mapping table is retained to allow re-identification by authorized honest brokers.

Linkage Approach

We faced many of the traditional challenges of linking multiple data sets, including incomplete or incorrect identifiers, as well as issues unique to birth records (21). In some cases infants are assigned temporary first names (e.g., “InfantGirl”, “BoyA”) until a final name is given (22). Also, typically infant last names are entered into the delivery hospital record consistent with the mother’s last name; however, in many cases the official last name on subsequent medical records (or official birth records) may not match the initial assignment.

We used an iterative approach with deterministic and probabilistic components to link records. Records of all patients, both mothers and infants, from all sources were stored in a common staging table. Fields in this table include first and last name, sex, date of birth, residence address, and in the case of infants, birth weight and available parental names. Due to data entry workflows for each source, it is likely that one or more of these fields would be missing values for any given patient. For the deterministic method, we first removed all non-alphabetic characters (e.g., spaces, numbers, hyphens, etc.) from the name fields. We then compared the first and last name, date of birth, and sex of each patient across the different sources. A patient was considered to be a “match” if date of birth, first name, and last name had perfect agreement and the sex field was not a non-blank mismatch. Just 35.7% of records met these criteria (16). After deterministic matching, linked records were removed from the unmatched pool which was then subjected to a previously detailed probabilistic matching approach in which records were linked when a likelihood score threshold was exceeded (12, 16). Along with patient names, dates of birth, and sex, additional variables were utilized in the probabilistic matching approach including extracted street number and street name components of the address, zip code, birth weight, and similarity of parental names as well as similarity of infant and parental surnames across data sources. All matching patient records were assigned the same unique MIDH master person identifier.

Twins or other multiple gestation infants were identified using common elements found in the newborn billing records including dates of birth, birth address information, and shared parental names. Multiple gestation siblings were differentiated from one another using conflicting birth weight, sex, or infant first names. Following the record linking process, all linked data core records as well as a 10% random sample of linkages to ancillary records were manually reviewed to verify accuracy of matches. Linkage between maternal and infant delivery hospital EHR data was facilitated by a built-in reference identifier indicating the relevant corresponding delivery hospital EHR.

Geocoding and Census Tracts

Using available home address information from EHR and billing records, a latitude and longitude coordinate pair was generated corresponding to each encounter. The geocoding process enabled linkage of individuals to area-level measures including public data sets provided by the Environmental Protection Agency and the American Communities Survey. Previous exercises in linking EHRs to area-level measures have identified community-level and environmental factors associated with risk factors for poor pregnancy outcomes such as obesity (23), as well as preterm and stillbirth outcomes (24, 25).

To translate patient addresses into their estimated latitude and longitude coordinates, we used an internally developed geocoder that assigns coordinates based on the 2015 TIGER/Line Shapefiles (26, 27). This program parses addresses attempting to fill in missing information using common postal abbreviations. The program performs a fuzzy lookup against the database derived from TIGER/Line Shapefiles and produces geographic coordinates using address interpolation as well as precision and score values for assessing the accuracy of estimated geocodes. The entire geocoding process is HIPAA compliant, running on a server within the CCHMC network, removing the need to transmit patient addresses to a third party.

To determine the census tract for a given geocode, we used the United States Census relationship data files (28), which include mappings required for census tract assignment. Using census tract definitions, which include the interpolated latitude and longitude per census tract, we assigned the nearest census tract to a given geocode by finding the shortest distance between that geocode and corresponding census tract using the Haversine formula (29).

Data Governance

Data governance remains a critical consideration in developing any shared data system, as agreements must be in place to satisfy privacy, compliance, and other data collaboration requirements. We first met with key stakeholders based at institutions contributing data to the MIDH, including privacy officers and members of the Institutional Review Board (IRB). These discussions led to a consensus that the MIDH would be best served by an honest broker organizational structure. The CCHMC IRB approved the establishment of the MIDH with a waiver of informed consent for the inclusion of any data collected during the normal course of clinical care. No data were collected solely for the purposes of the MIDH. Additionally, IRBs at each institution contributing data to the MIDH also approved the study

through reliance on the CCHMC IRB. The approved protocol, allowed for the distribution of de-identified data sets (as per HIPAA and HITECH legislation (30)) without the need for additional approvals. However, additional review and approval is required to distribute any data set containing protected health information. While the MIDH could link to an inventory of available biospecimens in the CCHMC biobank, we agreed that no samples or results of any genetic tests would be distributed without additional protocol review and approval. A governance board was established which includes a member from each institution contributing data to the MIDH. Upon receipt of a request for data, unanimous approval is required from the representatives of each institution from which data is requested. The purpose of the board is to protect the institutional interests of those that contribute data to the MIDH, minimize redundant or competing data requests, protect data integrity, and provide additional oversight to ensure adherence to relevant patient privacy regulations.

Results

Pre-linked, aggregated data contained regional newborn billing records for 69,290 infants (7,404 who received care at the UCMC location), as well as delivery hospital records representing 7,573 mothers and 7,792 infants (also from UCMC) (see Figure 2).Within the data core, delivery hospital medical records representing 7,293 infants were linked to 7,404 infants in the newborn billing record set indicating the UCMC location. As stated previously, we had expected all 7,404 newborns with billing records at the location to have a corresponding medical record resulting in a linkage rate of 98.5% (7,293/7,404).The 7,293 infants with delivery hospital records were each linked to a corresponding maternal delivery hospital record representing 7,107 mothers. Manual review of agreement between identifier fields demonstrated 100% validity of matches among audited records (16). Demographic characteristics for each population comprising the data core are listed in Table 1.

In Table 2 we present the number of data core records linked to home visiting records, children's hospital (post-delivery hospital discharge) records, and census tract identifiers. Data are presented for the entire set of 69,290 infants represented by the data core as well as for the subset of 7,293 infants which also had linked EHR records from the UCMC location.

Of the children represented by the data core 62,360 (90.0%) resided within the eight-county greater metropolitan region constituting the CCHMC primary market region. Through coordination with health departments, we determined the number of resident live births in the same tristate region over the study period was 77,114 (CCHMC Perinatal Institute, unpublished data, July 2017) resulting in a capture rate of 80.9% (62,360/77,114) of live born infants. Within Hamilton County, Ohio, (in which CCHMC and the city of Cincinnati are located) 94.7% (31,083/32,823) of resident births were captured.

Discussion

The MIDH offers a unique, clinically-focused resource for the conduct of regional research, including efforts in which data span institutions, state and county jurisdictions, or time periods. Successful implementation required that the study team address numerous technical and regulatory challenges to enable data integration. Regional expansion of the project is

currently underway with the goal of incorporating EHR data representing each delivery hospital in the greater metropolitan region of Cincinnati. Incorporating additional delivery hospitals involves engagement with each relevant healthcare organization to address their regulatory and technical considerations. Following the best practices established during our initial implementation efforts with UCMC, we have approached privacy officers and IRBs representing four additional regional healthcare organizations. We are working with each group to assure data ownership and privacy concerns are adequately addressed. The data governance board will expand to include representation from each additional hospital network to continue to ensure that the interests of each institution are well-represented. Once regulatory concerns are addressed, a technical point of contact will be identified at each contributing institution who will provide regular EHR extracts.

Efforts are currently planned to utilize the MIDH resource for a variety of purposes including support of 1) census tract-level surveillance of perinatal conditions including preterm birth and intrauterine exposures to substances of abuse, 2) modeling neonatal healthcare utilization patterns, 3) identifying area-level environmental and community factors influencing emergency department utilization and hospital readmission, and 4) monitoring pediatric development for infants receiving home visiting services compared to eligible, but unserved children. As the MIDH system matures, we anticipate increasing relevance to a broad range of stakeholders as an instrument for testing research hypotheses, a tool for informing allocation of scarce resources, and a mechanism for evaluating the effectiveness of community-based programs and public health initiatives.

We propose the following key activities as a roadmap to establishing a population-based resource: 1) Identify data sets that will serve as the foundational data core, such as vital birth or EHR records, representing a defined target population. 2) Obtain institutional buy-in from the health systems from which data will be provided. This includes engagement with privacy officers and institutional review boards and will require the establishment of data sharing/data use agreements either directly or with a third party acting as an honest broker for data containing protected health information. 3) Identify technical leads who will provide raw data from the source institutions and provide them with a data dictionary describing the desired data elements. 4) Select a common model for representing data elements from disparate sources within the repository and the process for harmonizing data that correspond to the same domain, but are represented in different terminologies. 5) Determine a technical approach for record linkage, the development of a unique master person identifier, and a process for geocoding address information. 6) Formalize a process for submission, review, and distribution of data requests.

Challenges and Limitations

Despite the potential of the system described in this report, several challenges remain. Best practices must be formalized and protocols must be established to assess data quality, particularly in handling conflicting information received from different sources, such as inconsistent demographics or identifiers for a single individual (e.g. differing sexes, races, or dates of birth for the same individual). In these cases, it must be determined which data source is the “most trusted,” which may differ depending on the data element or patient type.

Also, when multiple addresses or geocodes are present, investigators must thoughtfully consider which best meets their specific research need (e.g. most current, first available, etc.). Another current limitation of this initial implementation is the lack of data from more than one delivery hospital. As demonstrated in Table 1, the racial and ethnic composition of patients seen at the delivery hospital are not entirely representative of the greater regional population. We hope to overcome this limitation though expansion to additional regional delivery hospitals. Also, the current version of the MIDH only approximates a population-based representation as it is biased toward the sub-population of infants cared for by CCHMC physicians (~81% of regional births). Nevertheless, the architecture developed for this implementation provides a framework for the incorporation of additional data sources. Specifically, future development plans aim to integrate vital records supporting the realization of a truly population-based system. Either a live birth or fetal death vital record should be generated for every regional birth event for which there is a corresponding maternal medical record (9), whereas the current system does not include any measure of fetal deaths. Of course, these efforts will necessitate additional agreement with state or local health departments.

Conclusion

Efforts to improve child and maternal health outcomes require access to maternal health records, data from the perinatal period, and the child's subsequent heath records. It is rare that a single institution would have access to all of this information for a broad regional population. The creation of a linked data core is one potential solution this problem. A linked data core provides an added benefit by allowing the subsequent incorporation of environmental and community variables, which provides an even more comprehensive view of the population. Our integrated MIDH will provide investigators in the greater Cincinnati area with a novel opportunity for studying perinatal health at the population level by enabling precise phenotyping and the comparison of clinical treatments and outcomes supporting more rigorous evaluations of public health interventions.

Acknowledgments

This work was supported by the National Center for Advancing Translational Sciences of the National Institutes of Health through the Center for Clinical and Translational Science and Training at the University of Cincinnati [5UL1TR001425-02] and the Cincinnati Children's Research Foundation Academic and Research Committee. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Pallotto EK, Hunt PG, Dykes FD, et al. Topics in Neonatal Informatics: Infants and Data in the Electronic Health Record Era. Neoreviews. 2013; 14(2):e57–e62. DOI: 10.1542/neo.14-2-e57
2. Neutel CI, Johansen HL, Walop W. 'New data from old': epidemiology and record-linkage. Prog Food Nutr Sci. 1991; 15(3):85–116. [PubMed: 1784735]
3. Scheuren F. Methodologic issues in linkage of multiple data bases. Vital Health Stat. 1988; 4(25): 75–95.
4. Roos LL, Wajda A. Record linkage strategies. Part I: Estimating information and evaluating approaches. Methods Inf Med. 1991; 30(2):117–23. [PubMed: 1857246]

5. Herrchen B, Gould JB, Nesbitt TS. Vital statistics linked birth/infant death and hospital discharge record linkage for epidemiological studies. Comput Biomed Res. 1997; 30(4):290–305. [PubMed: 9339323]
6. Dufendach KR, Lehmann CU. Topics in Neonatal Informatics: Essential Functionalities of the Neonatal Electronic Health Record. Neoreviews. 2015; 16(12):e668–e73.
7. Meghea CI, Corser W, You Z. Electronic Medical Record Use and Maternal and Child Care and Health. Matern Child Health J. 2016; 20(4):819–26. DOI: 10.1007/s10995-015-1912-x [PubMed: 26676978]
8. Shapiro-Mendoza CK, Tomashek KM, Kotelchuck M, et al. Risk factors for neonatal morbidity and mortality among "healthy," late preterm newborns. Semin Perinatol. 2006; 30(2):54–60. DOI: 10.1053/j.semperi.2006.02.002 [PubMed: 16731277]
9. McLaughlin, M.Weiss, J.Kotelchuck, M., et al., editors. Improving the linkage of deliveries across time using vital records and hospital discharge data; American Public Health Association Annual Meeting; November 6, 2006; Boston, MA.
10. Barfield WD, Clements KM, Lee KG, et al. Using linked data to assess patterns of early intervention (EI) referral among very low birth weight infants. Matern Child Health J. 2008; 12(1): 24–33. DOI: 10.1007/s10995-007-0227-y [PubMed: 17562149]
11. Kotelchuck M, Hoang L, Stern JE, et al. The MOSART database: linking the SART CORS clinical database to the population-based Massachusetts PELL reproductive public health data system. Matern Child Health J. 2014; 18(9):2167–78. DOI: 10.1007/s10995-014-1465-4 [PubMed: 24623195]
12. Hall ES, Goyal NK, Ammerman RT, et al. Development of a linked perinatal data resource from state administrative and community-based program data. Matern Child Health J. 2014; 18(1):316–25. DOI: 10.1007/s10995-013-1236-7 [PubMed: 23420307]
13. Goyal NK, Hall ES, Meinzen-Derr JK, et al. Dosage effect of prenatal home visiting on pregnancy outcomes in at-risk, first-time mothers. Pediatrics. 2013; 132(Suppl 2):S118–25. DOI: 10.1542/peds.2013-1021J [PubMed: 24187113]
14. Goyal NK, Hall ES, Jones DE, et al. Association of maternal and community factors with enrollment in home visiting among at-risk, first-time mothers. Am J Public Health. 2014; 104(Suppl 1):S144–51. DOI: 10.2105/AJPH.2013.301488 [PubMed: 24354835]
15. Seske LM, Muglia LJ, Hall ES, et al. Infant Mortality, Cause of Death, and Vital Records Reporting in Ohio, United States. Matern Child Health J. 2017; 21(4):727–33. DOI: 10.1007/s10995-016-2159-x [PubMed: 27456308]
16. Hall ES, Marsolo K, Greenberg JM. Evaluation of identifier field agreement in linked neonatal records. J Perinatol. 2017; doi: 10.1038/jp.2017.70
17. United States Environmental Protection Agency. [Accessed December 2, 2016] AirData. 2016. Available from: http://www.epa.gov/airdata/
18. United States Census Bureau. [Accessed February 3, 2017] American Community Survey (ACS). 2016. Available from: https://www.census.gov/programs-surveys/acs/
19. Ammerman RT, Putnam FW, Kopke JE, et al. Development and implementation of a quality assurance infrastructure in a multisite home visitation program in Ohio and Kentucky. J Prev Interv Community. 2007; 34(1–2):89–107. DOI: 10.1300/J005v34n01_05 [PubMed: 17890195]
20. Stang PE, Ryan PB, Racoosin JA, et al. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. Ann Intern Med. 2010; 153(9):600–6. DOI: 10.7326/0003-4819-153-9-201011020-00010 [PubMed: 21041580]
21. Weber GM, Mandl KD, Kohane IS. Finding the missing link for big biomedical data. JAMA. 2014; 311(24):2479–80. DOI: 10.1001/jama.2014.4228 [PubMed: 24854141]
22. Adelman J, Aschner J, Schechter C, et al. Use of Temporary Names for Newborns and Associated Risks. Pediatrics. 2015; 136(2):327–33. DOI: 10.1542/peds.2015-0007 [PubMed: 26169429]
23. Roth C, Foraker RE, Payne PR, et al. Community-level determinants of obesity: harnessing the power of electronic health records for retrospective data analysis. BMC Med Inform Decis Mak. 2014; 14:36.doi: 10.1186/1472-6947-14-36 [PubMed: 24886134]

24. DeFranco E, Hall E, Hossain M, et al. Air pollution and stillbirth risk: exposure to airborne particulate matter during pregnancy is associated with fetal death. PLoS One. 2015; 10(3):e0120594.doi: 10.1371/journal.pone.0120594 [PubMed: 25794052]

25. DeFranco E, Moravec W, Xu F, et al. Exposure to airborne particulate matter during pregnancy is associated with preterm birth: a population-based cohort study. Environ Health. 2016; 15(1):6.doi: 10.1186/s12940-016-0094-3 [PubMed: 26768419]

26. United States Census Bureau. [Accessed July 10, 2017] 2015 TIGER/Line Shapefiles (machine-readable data files). 2015. Available from: https://census.gov/geo/maps-data/data/tiger-line.html

27. Schuyler, E., Brokamp, C., Chapman, K., et al. geocoder: v2.1. 2016.

28. United States Census Bureau. [Accessed July 10, 2017] Relationship Files. 2010. Available from: https://www.census.gov/geo/maps-data/data/relationship.html

29. Rosetta Code. [Accessed December 26, 2016] Haversine formula. 2016. Available from: https://rosettacode.org/wiki/Haversine_formula

30. Office for Civil Rights. [Accessed November 29, 2016] Health Information Privacy. 2015. Available from: http://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/#datause

Significance

What is already known on this subject?

Integration of perinatal data to facilitate planning, allocation of resources, testing of key hypotheses, and efficient operation of clinical care programs has proven difficult due to barriers created by regulatory concerns, privacy issues, questions related to data ownership, technical limitations, and lack of sustainable funding.

What this study adds?

We demonstrate a pilot implementation of a regional perinatal data repository supporting individual and area-level linkage to ancillary data sets. We provide development details and outline a roadmap to aid others in overcoming technical and regulatory hurdles.

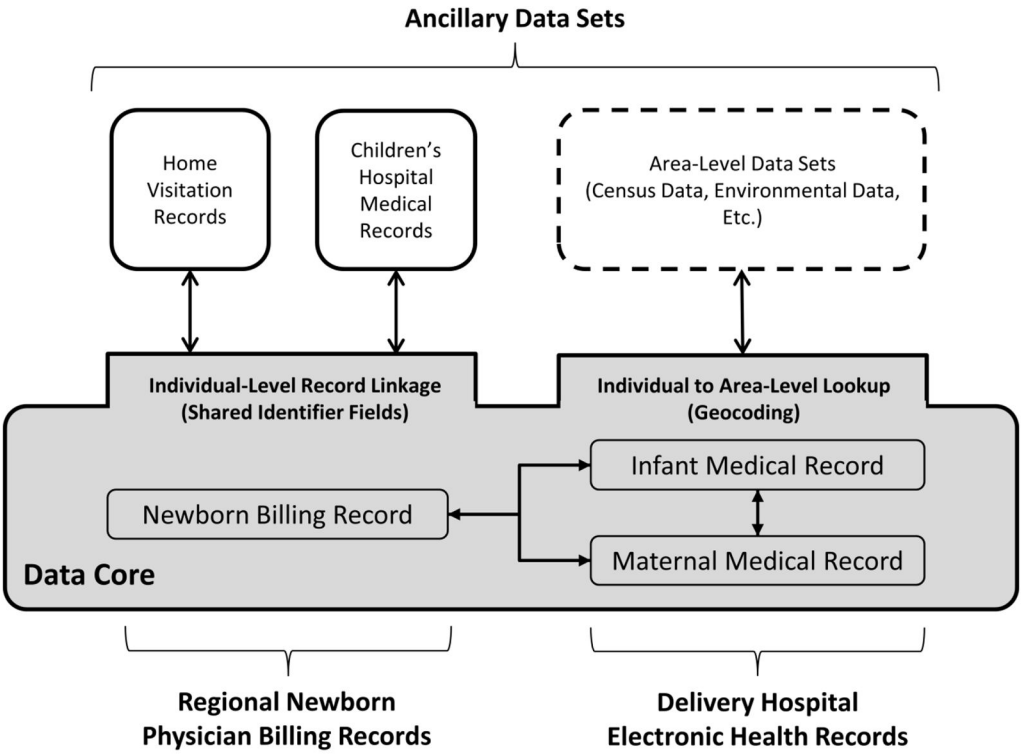

Figure 1.

Maternal and Infant Data Hub structure enabling individual-level and area-level linkage between the data core and ancillary data sets.

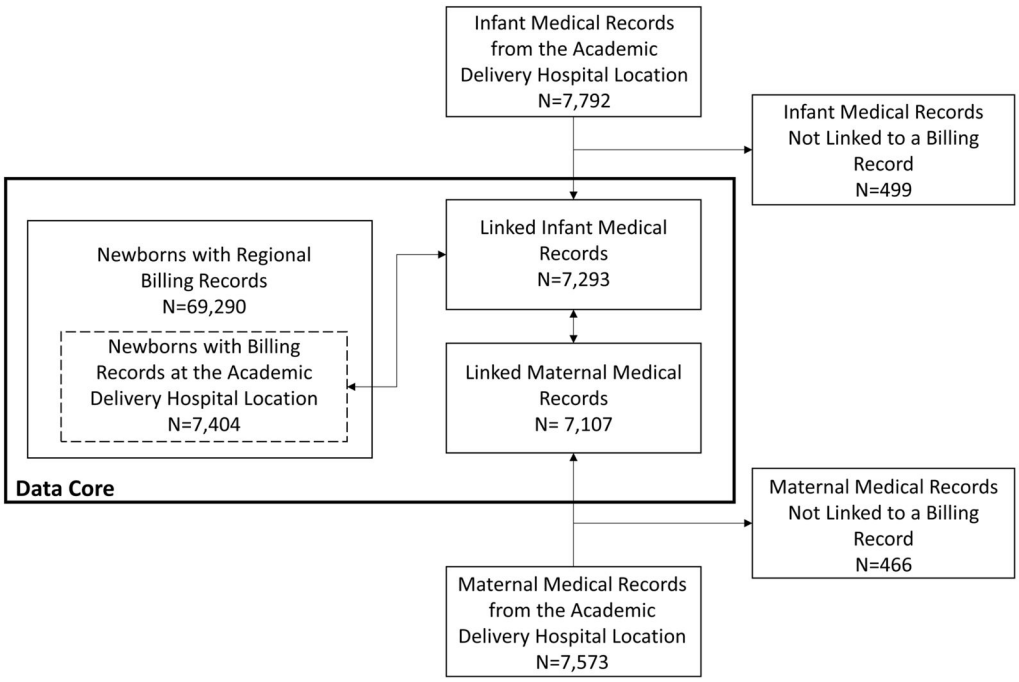

Figure 2.
Records comprising the Maternal and Infant Data Hub data core.

Table 1

Demographic characteristics of individuals represented by data core records

	Regional newborn billing records, N (%)	Delivery hospital infant medical records, N (%)	Delivery hospital maternal medical records, N (%)
Race			
Black	8,651 (12.5%)	3,227 (44.3%)	3,062 (43.1%)
White	27,090 (39.1%)	2,532 (34.7%)	2,652 (37.3%)
Other	3,920 (5.7%	484 (6.6%)	676 (9.5%)
Multiracial	3,777 (5.5%)	167 (2.3%)	74 (1.0%)
Missing race	25,873 (37.3%)	909 (12.5%)	745 (10.5%)
Ethnicity			
Hispanic	3,005 (4.3%)	882 (12.1%)	817 (11.5%)
Non-Hispanic	54,323 (78.4%)	6,243 (85.6%)	6,109 (86.0%)
Missing ethnicity	11,964 (17.3%)	168 (2.3%)	181 (2.5%)
Insurance status			
Public insurance	29,727 (42.9%)	5,172 (70.9%)	4,355 (61.3%)
Private insurance	32,376 (46.7%)	2,080 (28.5%)	2,280 (32.1%)
Uninsured	7,195 (10.4%)	44 (0.6%)	0 (0.0%)
Insurance missing	175 (0.3%)	7 (0.1%)	472 (6.6%)
Sex			
Male	35,523 (51.3%)	3,691 (50.6%)	0 (0.0%)
Female	33,767 (48.7%)	3,601 (49.4%)	7,107 (100.0%)
Sex-missing	0 (0.0%)	1 (0.0%)	0 (0.0%)
Twin or multiple	3,016 (4.4%)	308 (4.2%)	
Total	69,290 (100.0%)	7,293 (100.0%)	7,107 (100.0%)

Individuals may be represented with conflicting race, ethnicity, or insurance information by records representing the same individual at different encounters within the same data set. Consequently, an individual may be counted in more than one row for each demographic category and the sum of subcategories may exceed the total number of records.

Table 2

Metrics of data core records linked to ancillary data sets.

	Infants in the Data Core (All Infants with a Regional Newborn Billing Record)	Subset of Infants in the Data Core with a Linked Electronic Health Record
Individuals represented, N	69,290	7,293
Individuals with a linked home visiting record, N (%)	2,430 (3.5%)	859 (11.8%)
No. of home visits, N	48,557	8,802
Individuals with a linked children's hospital record, N (%)	42,541 (61.4%)	5,063 (69.4%)
No. of pediatric primary care visits, N	8,295	1,941
No. of urgent care visits, N	13,337	1,916
No. of emergency department visits, N	18,214	2,663
No. of inpatient admissions, N	26,788	3,438
Individuals with a linked census tract using geocoding, N (%)	68,884 (99.4%)	7,282 (99.8%)