# Opportunities for life course research through the integration of data across Clinical and Translational Research Institutes

Heidi A. Hanson[1*], William W. Hay Jr[2], Jonathan N. Tobin[3], Shari L. Barkin[4], Mark Atkins[5], Margaret R. Karagas[6], Ann M. Dozier[7], Cynthia Wetmore[8], Michael W. Konstan[9] and James E. Heubi[10]; For The CTSA Lifespan Domain Taskforce

[1] Department of Surgery, University of Utah School of Medicine, Population Sciences, Huntsman Cancer Institute, University of Utah, Salt Lake City, UT, USA
[2] Department of Pediatrics, University of Colorado School of Medicine, Aurora, CO, USA
[3] The Rockefeller University Center for Clinical and Translational Science and Clinical Directors Network (CDN), New York, NY, USA
[4] Department of Pediatrics, Vanderbilt University Medical Center, Nashville, TN, USA
[5] Department of Psychiatry, Institute for Juvenile Research, The University of Illinois at Chicago, Chicago, IL, USA
[6] Department of Epidemiology, Geisel School of Medicine at Dartmouth, Lebanon, NH, USA
[7] Department of Public Health Sciences, University of Rochester Medical Center, Rochester, NY, USA
[8] Department of Pediatrics, Emory University, Children's Health Care of Atlanta, Atlanta, GA, USA
[9] Department of Pediatrics, Case Western Reserve University School of Medicine, Cleveland, OH, USA
[10] Division of Gastroenterology, Pediatric Liver Care Center, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA

**Introduction.** Early life exposures affect health and disease across the life course and potentially across multiple generations. The Clinical and Translational Research Institutes (CTSIs) offer an opportunity to utilize and link existing databases to conduct lifespan research.

**Methods.** A survey with Lifespan Domain Taskforce expert input was created and distributed to lead lifespan researchers at each of the 64 CTSIs. The survey requested information regarding institutional databases related to early life exposure, child-maternal health, or lifespan research.

**Results.** Of 64 CTSI, 88% provided information on a total of 130 databases. Approximately 59% (n = 76/130) had an associated biorepository. Longitudinal data were available for 72% (n = 93/130) of reported databases. Many of the biorepositories (n = 44/76; 68%) have standard operating procedures that can be shared with other researchers.

**Conclusions.** The majority of CTSI databases and biorepositories focusing on child-maternal health and lifespan research could be leveraged for lifespan research, increased generalizability and enhanced multi-institutional research in the United States.

## Introduction

Health at any point across the life course is determined by a complex interplay of genetic and environmental exposures from gamete to grave [1–4]. Early life factors, such as in utero exposure to under-nutrition or toxins, may be particularly important because they have the potential to adversely alter short-term health and long-term trajectories of physical and mental health [5–7]. While basic science and epidemiological studies have shown the importance of considering the role of early life exposures on later life health outcomes, our understanding of these mechanisms needs to be expanded. However, the data requirements for a well-designed life course study may deter some investigators from adopting such a comprehensive approach to understanding health. Longitudinal studies are costly and time-consuming, and therefore most prospective data sources are constrained to specific geographic subpopulations and lack generalizability.

* Address for correspondence: H. A. Hanson, PhD, Huntsman Cancer Institute, #1501, 2000 Circle of Hope, Salt Lake City, UT 84112, USA.
(Email: Heidi.hanson@hci.utah.edu)

Life course research also requires a diverse set of data sources and analytic techniques because a combination of genetic, social, psychological, and environmental factors must be incorporated into the analyses. The interdependent role of these factors and timing of exposures, as well as cumulative effects over time, remains poorly understood. To address these concerns, we have compiled a list of available data sources across 64 research institutions. Leveraging data from multiple sources across a variety of subpopulations allows for the power necessary to further investigate the importance of timing of exposures and their later life health outcomes [8, 9]. However, there are few data catalogs that define data sources available for investigating how early life exposures affect later life health to conduct this type of lifespan research.

The US National Institutes of Health (NIH) designed a Roadmap for Medical Research with the purpose of improving the translation of research into practice by improving the understanding of complex biological systems, encouraging scientists to test multiple models for conducting research, and facilitating the efficient dissemination of research findings into clinical care [10]. Such a broad and lofty mission is essential for improving the health and well-being of the US population and requires the implementation of new forms of collaboration in the medical community. The Clinical and Translational Science Awards (CTSA) program of the NIH National Center for Clinical and Translational Sciences (NCATS) is a national network of institutions (Clinical and Translational Research Institutes (CTSIs)) designed to address this goal. Thus, the CTSA program creates a definable academic home designed to facilitate translational research and includes 64 medical research institutions in 31 states and the District of Columbia. Harnessing the data from these institutions with the goal of further elucidating links between early life exposures and later life health and using findings to inform focused interventions has the potential to affect the health of millions of the US population. The vast data sources that already exist to conduct lifespan research across all the CTSIs could be integrated to conduct lifespan research. Therefore, we conducted a survey to identify these resources and to begin to identify common data elements as well as linkages to established biorepositories.

The NCATS national CTSA organization created domain task forces (DTFs) to serve as the infrastructure for sharing ideas and collaborating to develop efficient and effective approaches to conducting and translating research into improved health. The Lifespan Domain Task Force is comprised of researchers across domains from preconception, infancy to geriatrics who examine ideas and conduct studies needed to advance lifespan research. A group of maternal and child health researchers and life course epidemiologists formed a sub-group of the Lifespan DTF, the Early Life Exposures Working Group (ELE WG), and identified the need to create a publicly available catalog of existing studies and cohorts that would broadly benefit investigators interested in ELE research. Developing a catalog of datasets from a national network of clinical research centers will provide a resource for future research examining the role of early life factors on later life health across the US population. It will also encourage collaboration between academic institutions and their community health partners and facilitate the future evaluation of programs aimed at integrating information about social, psychological, and environmental factors contributing short-term and long-term health outcomes. In order to address this objective, a survey was designed in REDCap and disseminated to all CTSIs with the goal of identifying potential resources that would benefit investigators interested in life course research with a special interest in early life exposures.

## Materials and Methods

The ELE WG created a RedCap survey designed by members of the ELE WG to be distributed to all CTSIs (n = 64). The REDCap survey requested information regarding institutional databases, such as cohorts or biorepositories from unique populations, related to early life exposure, child-maternal health, or lifespan research.

Surveys were sent to all CTSA Principal Investigators (PIs) who were asked to identify and send the survey to those in their institution with the greatest knowledge about lifespan research and/or existing data repositories. Reminder prompts were then sent to the PIs if there had been no initial response. Prompts were followed with personal appeals from members of the task force if the surveys had not been completed. If responses had not been received in a timely manner (2–3 months), follow-up emails were sent to each of the PIs by J.E.H. and thereafter his administrative assistant reached out to the PIs' administrative assistants to be certain that the PI had received and responded to the request.

Data collected from the RedCap survey were stored in the Early Life Exposure Database Repository and can be downloaded from the *Center for Leading Innovation & Collaboration* Web site (https://clic-ctsa. org/content/ele-redcap-table-resources). The full list of questions asked of participants is available in Supplementary Table S1.

## Results

The survey was completed by 56 of the 64 CTSA hubs for an overall response rate of 88%. All CTSA hubs completing the survey were academic centers and are widely dispersed across the United States (see Fig. 1a). There were 73 total respondents to the survey, with multiple respondents from 7 of the institutions. Nearly all respondents completed the survey, with an overall survey completion rate of 96%. In all, 90 completed surveys representing 130 lifespan related databases formed the basis of the result section.

Information on a total of 130 databases relating to early life exposures, maternal-child health, or life course research was collected from 49 of the participating CTSA centers. Fig. 1b shows the number of early life exposures, child-maternal health, or lifespan research databases by institution. The majority of CTSA hubs with a life course database had more than one database relating to early life exposures, child-maternal health, or lifespan research (n = 26), with the maximum of 10.

Table 1 provides a broad overview of the data collected from the RedCap survey (Supplementary Table S2 provides a detailed summary of each database). The reported databases contain information on cohorts ranging in size from 1–500 participants (n = 39 or 30.5%) to more than 100,000 participants (n = 13 or 10.2%), with cohort size being unknown for 18 of the databases. Cohorts included prenatal (n = 47), infants (n = 66), children (n = 55), young adults (n = 45), pregnant women (n = 46), adults (n = 53), older adults (n = 30). Longitudinal data, defined as having multiple measurements for a single patient over multiple time points, were available for 72% (n = 93) of the reported databases.

Approximately 59% (n = 76/130) of all reported databases have an associated biorepository, with multiple types of biosamples ($n_{Blood}$ = 58, $n_{Placenta}$ = 14, $n_{Tissue}$ = 28, $n_{Other/Unknown}$ = 23). Blood is the most commonly collected biosample. Examples of the other/unknown category of biosample include breast milk, fecal samples, umbilical cord, and omental adipose tissue. Nearly 57% of biorepositories were considered shareable (n = 43/76), which was defined as storing data on a platform that permits sharing and having an Institutional Review Board [IRB] protocol that facilitates sharing. Participants were asked to provide a brief description of how researchers can request biospecimen data and the responses ranged from contacting the PI to contacting specific NIH institutes that oversee the study. More than half of the biorepositories (n = 44/76; 58%) have standard operating procedures (SOP) that can be shared with other researchers. These procedures include the time between sample collection, collection
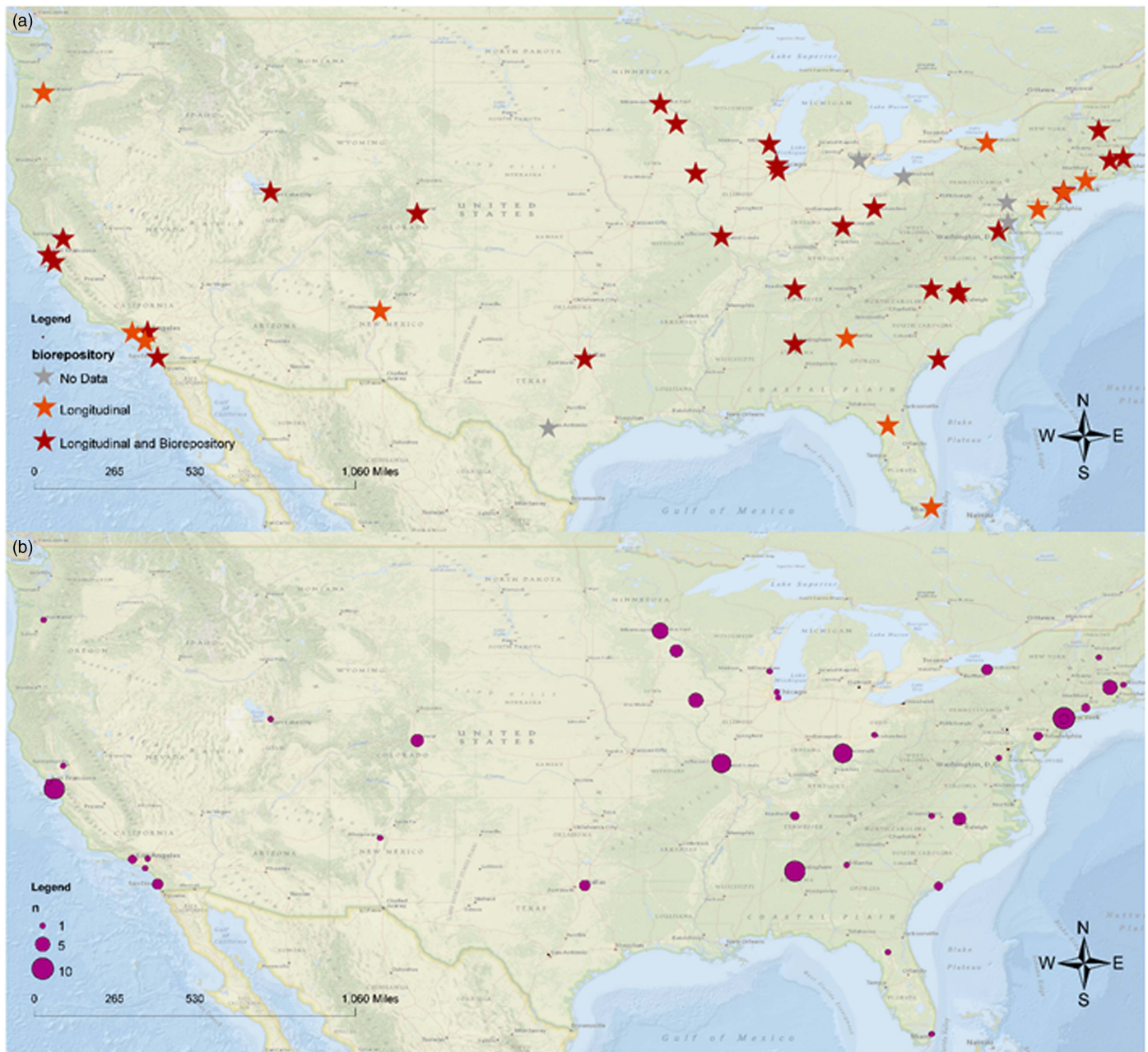
**Fig. 1.** Map of responding Clinical and Translational Science Awards institutions. (*a*) All participating institutions and the type of data stored in their database. (*b*) The number of early life exposures, child-maternal health, or lifespan research databases by institution.

method, and other SOPs. Of the biorepositories with SOPs, 64% (n = 28/44) have collection procedures that can be modified to accommodate prospective or new studies.

Most biorepositories have collected samples from subjects in both healthy and diseased states (n = 31/76; 41%). There are smaller numbers collected for disease-only (n = 11/76; 14%), healthy-only (n = 19/76; 25%), or unknown/other purposes (n = 15/76; 20%). The types of subjects that were classified as "other" include peri-menopausal women, children with lead poisoning, genetically at risk individuals, or pregnant women. The disease states reported include general disorders such as autoimmune diseases, autism, diabetes, preterm births, obese subjects, kidney disease, peripartum depressed women, and neurological disorders, as well as specific disorders such as Wolfram syndrome.

Data integrated with electronic medical records provide an exciting prospect for observing how early life exposures affect later life health trajectories. Nearly 70% of the biorepositories have been integrated with electronic medical records in some manner (n$_{Integrated}$ = 37/76; 49% and n$_{Somewhat/Maybe}$ = 16/76; 21%). Nearly all data that have been

linked to electronic health records (EHRs) have systems that are amenable to natural language processing (n = 49/53; 92%). In addition to administrative health care data, 49% of all biorepositories (n = 37/76) have laboratory results on tissues that are part of research and not medical practice. Another 21% (n = 16/76) may have these data available in partial form.

Figure 2 displays a summary of features of the databases with longitudinal and biorepository data by cohort size. A large proportion of longitudinal databases also have a biorepository (n = 57/93; 61%). The majority of the cohorts with biorepositories are smaller studies with under 5000 participants (n = 52/93; 56%). Three CTSA hubs (4 databases) reported having cohorts with over 100,000 participants and biorepository data. Biospecimens are available for the longitudinal studies over a range of cohort sizes, including cohorts with over 100,000 individuals. Blood is the most commonly available sample in databases with longitudinal data (n = 44/57; 77%), followed by tissues/fluids (n = 21/57; 37%). The data sets are not merely collections of diseased cohorts, with nearly half of the databases having subjects that are healthy and in a disease state (n = 25/57; 44%). Another 37% have

**Table 1.** *Characteristics of reported databases (n = 130)*

| | n | % |
|---|---|---|
| Database includes multiple measurements/multiple time points | 93 | 71.5 |
| Age at enrollment | | |
|     Prenatal | 47 | 36.2 |
|     Infant | 66 | 50.8 |
|     Child | 55 | 42.3 |
|     Young adult | 45 | 34.6 |
|     Adult | 53 | 40.8 |
|     Older adults | 30 | 23.1 |
|     Pregnancy | 46 | 35.4 |
|     Unknown | 8 | 6.2 |
|     Universal consent | 18 | 13.8 |
| Biorepository | 76 | 58.5 |
| Components of biorepository | | % among those with biorespository |
|     Blood | 58 | 76.3 |
|     Plasma | 14 | 18.4 |
|     Tissue/fluid | 28 | 36.8 |
|     Other | 12 | 15.8 |
|     Unknown | 11 | 14.5 |
| Type of subjects in biorepository | | |
|     Normal | 50 | 65.8 |
|     Disease state | 42 | 55.3 |
|     Unknown | 6 | 7.9 |
|     Other | 9 | 11.8 |
| Centralized repository | 68 | 89.5 |
| EHR integration | 62 | 81.6 |

EHR, electronic health records.

subjects that are all healthy or all in a diseased state ($n_{Healthy} = 11/57$; 19% and $n_{Diseased} = 10/57$; 18%) and the remaining 19% ($n = 11/57$) are unknown. The available databases also encompass many stages across the life course. Nearly all of the databases enrolled individuals between the prenatal period and young adulthood ($n = 51/57$; 89%), with the distribution by period of development as follows (categories are not mutually exclusive); prenatal ($n = 23/57$; 40%), infant ($n = 33/57$; 58%), childhood ($n = 27/57$; 47%), and young adult ($n = 23/57$; 40%).

One of the most exciting prospects for future life course research is the development of longitudinal databases that are linked to biorepository data and EHR. Fig. 3 displays a summary of features of the 40 CTSA databases from 22 CTSA hubs with all 3 components ($n_{Longitudinal} = 40/93$; 43% and $n_{Total} = 40/130$; 31%). The 4 large databases have been integrated with electronic medical records. Biospecimen data collected by longitudinal studies linked to EHR is available for a range of cohort sizes, health statuses, and age groups. Blood is the most commonly available biospecimen in databases with all 3 components ($n = 33/40$; 83%), followed by tissues/fluids ($n = 13/40$; 33%). Most of the longitudinal data with biosamples and electronic medical records have fewer than 5000 individuals enrolled ($n = 29/40$; 73%). Databases with all 3 components also span the entire life course, from prenatal to older adulthood, with 90% having enrolled individuals between the prenatal period and young adulthood ($n = 36/40$; 90%). The distribution of records by period of development is as follows (categories are not mutually exclusive); prenatal ($n = 16/40$; 40%), infant ($n = 24/40$; 60%), childhood ($n = 22/40$; 55%), and young adult ($n = 20/40$; 50%).

## Discussion

Life course methods conceptualize health as the dynamic interplay between biologic and environmental factors from conception to death

and has long been accepted by the World Health Organization [11–13]. Understanding factors that are amenable to intervention during early periods of development is particularly important because of its potential to improve health over an entire life course and possibly for future generations [14]. It may also prove useful for predicting the occurrence or progression of disease in current populations, allowing for a more targeted approach to disease specific surveillance and screening programs. Multiple databases and biorepositories focusing on maternal-child health and life course research are available to investigators within or outside the responding institutions that can be used to facilitate lifespan research.

Life course research is expensive. Utilizing the massive volume of research data and patient-specific information already being collected by health care systems to study the short-term and long-term effects of early life exposures may prove to be a cost-effective and powerful way to elucidate further factors that affect health during critical periods of development, and may reduce the selection biases inherent with recruiting research participants, and will contribute to the development of Learning Healthcare Systems. Combining research repositories with population-level data, such as vital records and EHR, makes it possible to quantify and potentially correct for the differences between the sample and overall population. Further, combining repositories that have been collected from multiple geographic locations and for diverse populations and purposes may result in a sample that is more representative of the larger population, as well as samples with larger sample sizes for subgroup analyses. Cataloging research databases and biorepositories across institutions that facilitate research on early life exposures and health across the lifespan is the first step in beginning to combine and analyze data that have already been collected. Linking clinical research records to administrative records within and between institutions could potentially revolutionize health care research by allowing individuals to be followed over longer periods of time. While challenging, successful examples exist on a smaller scale that demonstrate the feasibility of linking to records across institutions and to external data sources, such as vital statistics and Driver's License Data [15–18].

Synthesizing EHRs with data from external sources, such as population databases, biomonitors and environmental exposure data, would allow for investigations into the immediate and latent effects of risk factors over all ages. For example, individual level birth certificate and death certificate data can be linked to existing cohorts to increase the breadth and quality of measures relating to early life exposures [15, 19–21]. Combining these records also allows researchers to investigate dynamic health outcomes, such as how the relationship between changes in weight during mid-life affects later life disability [22] or how pregnancy outcomes affect trajectories of chronic conditions after the age of 65 [23]. Using geocoded data to link the databases and biorepositories identified in this study to other external data sets, such as environmental toxins and measures of the social determinants of health, also have great potential to improve our understanding of the long-term effects of early life exposures. One area that appears underrepresented in current databases is patient-reported measures, such as subjective well-being, which has been shown to be distinct from mental illness and predictive of long-term health and longevity [24, 25]. Whereas mental illness may be captured as diagnoses and prescriptions in electronic medical records, social well-being will not be captured, and thus adding brief indicators to existing databases could yield valuable information related to long-term health and disease prognosis, as well as patient centered outcomes [26].

Although combining data from multiple sources with computational, bioinformatics, and statistical methods allow us to observe previously unseen patterns in biomedical data, conceptual models, such as those used in life course epidemiology, can be used to provide the scaffolding for integrating scientific theory and approach to making sense of the patterns. There are multiple opportunities to utilize this framework in ongoing initiatives such as the Precision Medicine Initiative and the Environmental Influences on Child Health Outcomes program.
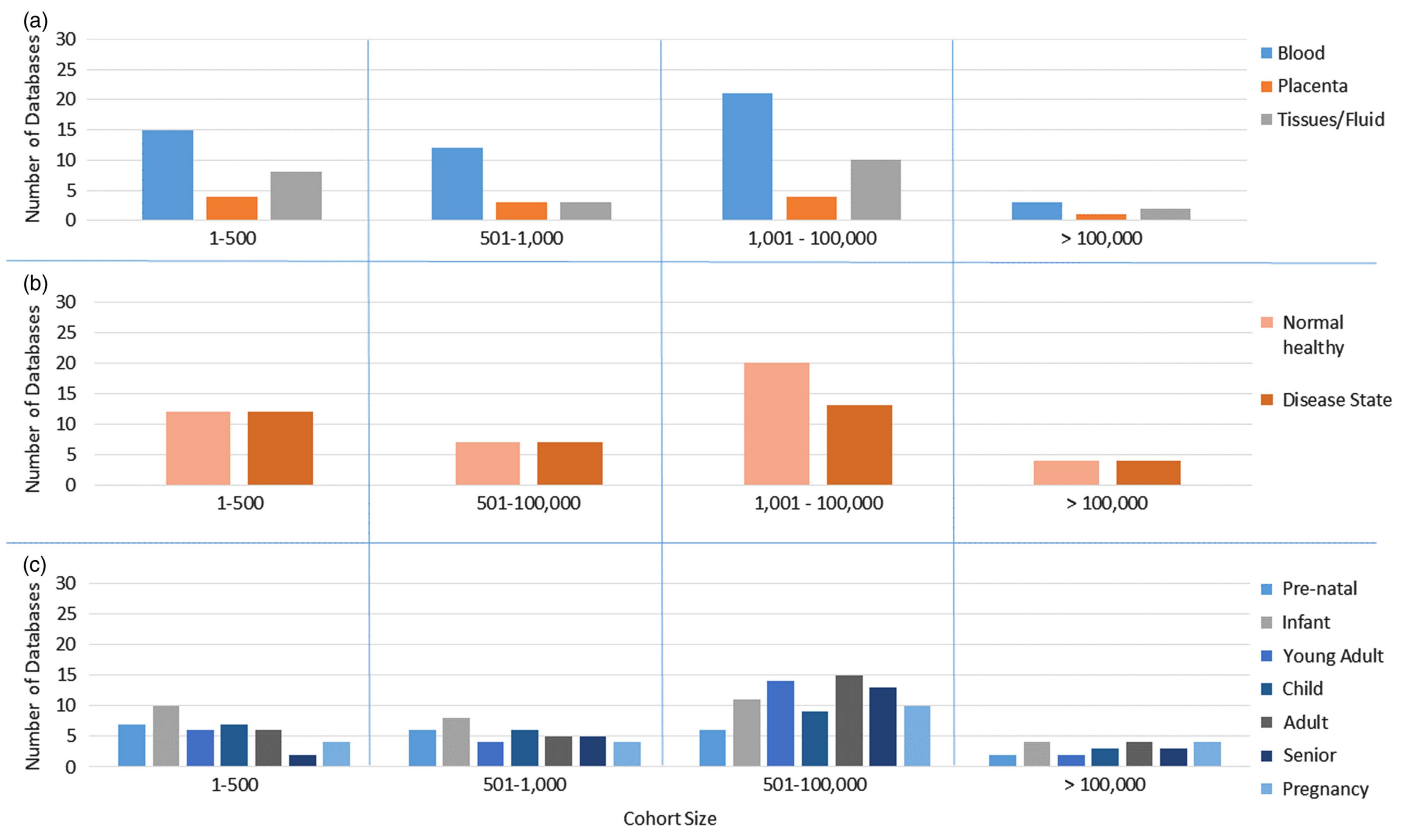
**Fig. 2.** Number of Clinical and Translational Science Awards databases with longitudinal data linked to a biorepository by cohort size. (*a*) The number of databases by sample type. (*b*) The number of databases with normal and/or disease state data. (*c*) The number of databases by the approximate age of the participant at the time of enrollment.
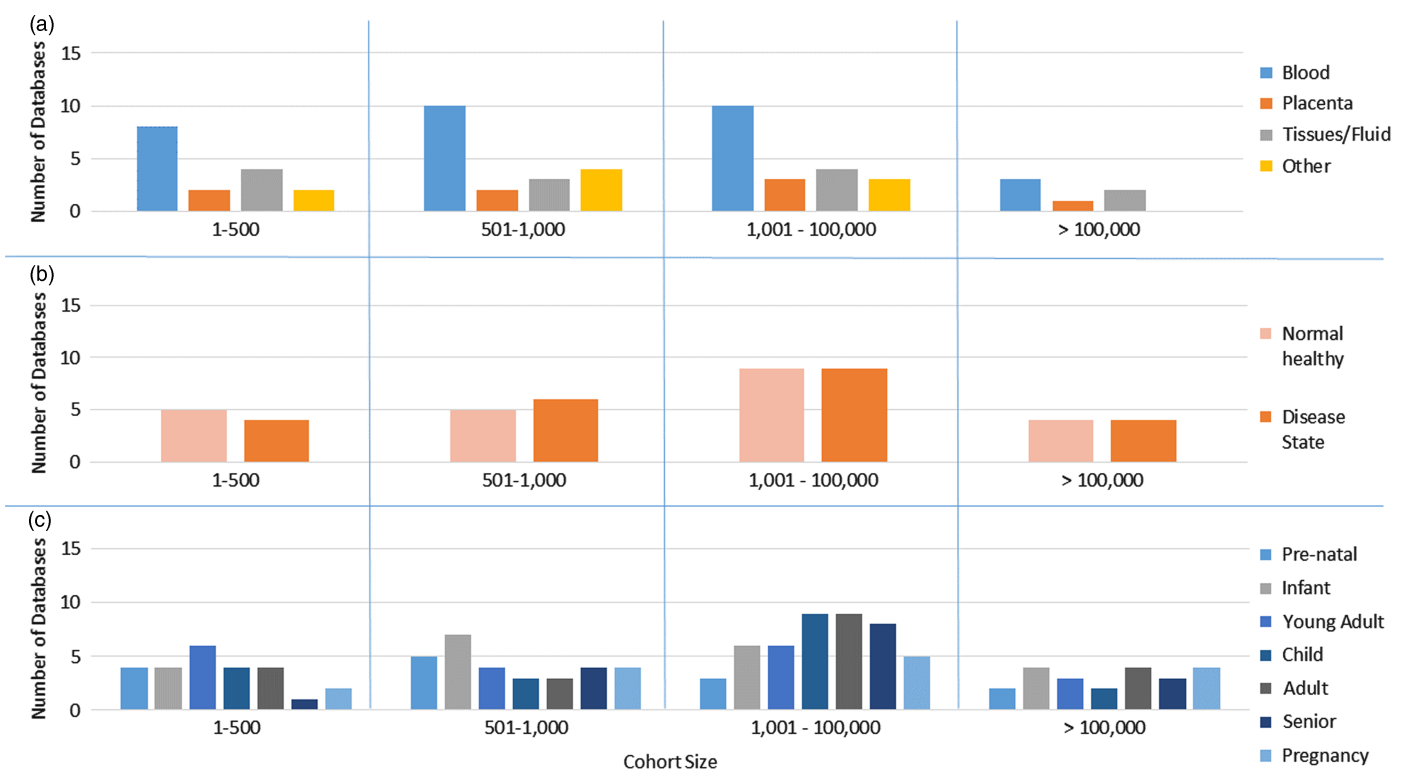


**Fig. 3.** Number of Clinical and Translational Science Awards databases with longitudinal data linked to a biorepository and electronic medical records by cohort size. (*a*) The number of databases by sample type. (*b*) The number of databases with normal and/or disease state data. (*c*) The number of databases by the approximate age of the participant at the time of enrollment.

Identifying factors early in life and across generations that affect health throughout the life course will facilitate the design of intervention and prevention programs that have the potential to optimize the health of an entire population. While this first attempt to catalog the data across institutions is valuable, more needs to be done to further this endeavor. First, more resources should be devoted to cataloging the data sets available for life course research. The Inter-university Consortium for Political and Social Research (ICPSR) is an

example of a successful data sharing resource that began archiving data in 1962 and currently holds over 68,000 data sets from more than 8000 studies [27]. A similar resource combining clinical and population health existing data sources housed across multiple institutions, guided by a conceptual model of life course research, and supported by the CTSI program across the United States would be a cost-effective way to further investigate the relationship between early life exposures and health. Second, to support reproducibility, data sharing across institutions should include sharing the protocols and methodologies used to collect, clean, analyze, and curate the data. Examples of online protocol repositories include Protocols.io [28] and Protocol Exchange [29]. Third, building off of the ICPSR model, training in data access, curation, and the analytic methods of life course research should be part of the life course data repository. Although there are many aspects, such as confidentiality and data sharing agreements that must be considered if such an endeavor were to be undertaken, these should not be seen as unsurmountable obstacles. Sensitive data sources could also be held by their respective intuitions and assigned a linkage id that would allow data sharing between groups that have gained the appropriate approvals from the relevant data contributors and IRB [15]. Insufficient time, lack of funding, and lack of data sharing platforms may also be prohibitive to the promotion of data sharing across institutions [30].

Other barriers also need to be addressed for large-scaled collaborations across institutions. For instance, data and biospecimens may only be internally available to researchers in the same institution. Thus, alternative strategies for collaborations across centers for replication of previous findings will be required. This includes concerns about confidentiality and privacy issues revolving around creating large databases with personal health information require pragmatic strategies that minimize the risk of loss of confidentiality while enhancing the opportunity to learn from real-world experience. One approach is to allow collaborators to perform analyses within their own institutional firewalls and share statistical estimates for pooling in collaborative analyses. Several approaches, from simple to complex, could be taken to achieve such collaborations. For example, a simple approach is to form cross-institutional research teams focused on a single research question, each with access to their own data sets and have them design and execute the study and analysis protocol simultaneously, and then combine summary data across sites. This model has also proven successful in the social sciences [31–33]. A more complex approach would be to develop a consortium of data science teams from participating institutions to develop common data elements and common procedures for life course research, also referred to as a "Federated Model." The National Patient-Centered Clinical Research Network Model and CTSA Informatics Domain Task Force is an example of this type of collaboration. It might be more successful, however, if the sometimes daunting task of sharing all data across institutions were focused on a smaller scale. This would circumvent the need for a data repository, which raises complex social, legal, and ethical challenges, and allow for the formation of cross-institutional research teams with common goals but independent data holdings. Further, the NCATS Streamlined, Multisite, Accelerated Resources for Trials IRB platform (SMART IRB) will help expedite multi-site clinical studies across CTSAs by providing a single IRB review process. Transforming such a platform from vision to reality, however, would require substantial support from multiple institutions and creative solutions for a complex problem.

There are noteworthy limitations to our study. Our survey was specific to CTSIs, and it is likely that the number of databases and biorepositories focusing on child-maternal health and lifespan research within CTSIs and available to investigators is under-reported. It is possible that the respondent at each CTSA was not fully aware of all related databases housed within each institution. Nonetheless, we see the development of our data catalog as a dynamic process and plan to incorporate other databases as we identify them. We also have considered updating the catalog to incorporate new or expanded databases. At the very least, this is a good start to which additional databases could be added in the future, and facilitate conversations and collaborations across multiple institutions.

## Acknowledgments

## Financial Support

## Disclosures

The authors have no conflicts of interest to declare.

## Supplementary material

To view supplementary material for this article, please visit https://doi.org/10.1017/cts.2018.29

## References

1. **Barker DJ, Martyn CN.** The maternal and fetal origins of cardiovascular disease. *Journal of Epidemiology and Community Health* 1992; **46**: 8–11.

2. **Kuh D, et al.** Life course epidemiology. *Journal of Epidemiology and Community Health* 2003; **57**: 778–783.

3. **Gluckman PD, et al.** Epigenetic mechanisms that underpin metabolic and cardiovascular diseases. *National Reviews. Endocrinology* 2009; **5**: 401–408.

4. **Ben-Shlomo Y, Cooper R, Kuh D.** The last two decades of life course epidemiology, and its relevance for research on ageing. *International Journal of Epidemiology* 2016; **45**: 973–988.

5. **Eskenazi B, et al.** In utero and childhood polybrominated diphenyl ether (PBDE) exposures and neurodevelopment in the CHAMACOS study. *Environmental Health Perspectives* 2013; **121**: 257–262.

6. **Almond D, Currie J.** Killing me softly: the fetal origins hypothesis. *The Journal of Economic Perspectives: A Journal of the American Economic Association* 2011; **25**: 153–172.

7. **Ben-Shlomo Y, Kuh D.** A life course approach to chronic disease epidemiology: conceptual models, empirical challenges and

interdisciplinary perspectives. *International Journal of Epidemiology* 2002; **31**: 285–293.

8.  **Liu C-Y, et al.** Design and analysis issues in gene and environment studies. *Environmental Health* 2012; **11**: 93.

9.  **Murcray CE, Lewinger JP, Gauderman WJ.** Gene-environment interaction in genome-wide association studies. *American Journal of Epidemiology* 2009; **169**: 219–226.

10. **Kantor LW.** NIH roadmap for medical research. *Alcohol Research & Health* 2008; **31**: 12.

11. **World Health Organization**. *The Implications for Training of Embracing: A Life Course Approach to Health.* Geneva, Switzerland: WHO, 2000.

12. **Hertzman C, Wiens M.** Child development and long-term outcomes: a population health perspective and summary of successful interventions. *Social Science & Medicine (1982)* 1996; **43**: 1083–1095.

13. **Marmot M.** Achieving health equity: from root causes to fair outcomes. *Lancet* 2007; **370**: 1153–1163.

14. **Pembrey M, Saffery R, Bygren LO.** Human transgenerational responses to early-life experience: potential impact on development, health and biomedical research. *Journal of Medical Genetics.* 2014; **51**: 563–572.

15. **DuVall SL, et al.** Evaluation of record linkage between a large healthcare provider and the Utah Population Database. *Journal of the American Medical Informatics Association: JAMIA* 2012; **19**: e54–e59.

16. **Littenberg B, Lubetkin D.** Availability, strengths and limitations of US State Driver's License Data for obesity research. *Cureus* 2016; **8**: e518.

17. **St. Sauver JL, et al.** Use of a medical records linkage system to enumerate a dynamic population over time: the Rochester epidemiology project. *American Journal of Epidemiology* 2011; **173**: 1059–1068.

18. **Newgard C, et al.** Evaluating the use of existing data sources, probabilistic linkage, and multiple imputation to build population-based injury databases across phases of trauma care. *Academic Emergency Medicine* 2012; **19**: 469–480.

19. **Edelman LS, et al.** Linking clinical research data to population databases. *Nursing Research* 2013; **62**: 438–444.

20. **Smith KR, et al.** Survival of offspring who experience early parental death: early life conditions and later-life mortality. *Social Science & Medicine* 2014; **119**: 180–190.

21. **Stroup AM, et al.** Baby boomers and birth certificates: early-life socioeconomic status and cancer risk in adulthood. *Cancer Epidemiology Biomarkers & Prevention* 2017; **26**: 75–84.

22. **Williams ED, et al.** The effects of weight and physical activity change over 20 years on later-life objective and self-reported disability. *International Journal of Epidemiology* 2014; **43**: 856–865.

23. **Hanson HA, Smith KR, Zimmer Z.** Reproductive history and later-life comorbidity trajectories: a medicare-linked cohort study from the Utah population database. *Demography* 2015; **52**: 2021–2049.

24. **Diener E, Chan MY.** Happy people live longer: subjective well-being contributes to health and longevity. *Applied Psychology: Health and Well-Being* 2011; **3**: 1–43.

25. **Westerhof GJ, Keyes CLM.** Mental illness and mental health: the two continua model across the lifespan. *Journal of Adult Development* 2010; **17**: 110–119.

26. **Keyes CLM.** Social well-being. *Social Psychology Quarterly* 1998; **61**: 121–140.

27. **ICPSR Data Management & Curation**. [Internet], 2018 [cited Mar 17, 2018]. (https://www.icpsr.umich.edu/icpsrweb/index.jsp)

28. **Protocols.io.** [Internet], 2018 [cited Mar 17, 2018]. (https://www.protocols.io/)

29. **Protocol Exchange**. Nature Publishing Group [Internet], 2018 [cited Mar 18, 2018]. (https://www.nature.com/protocolexchange/)

30. **Houtkoop BL, et al.** Data sharing in psychology: a survey on barriers and preconditions. *Advances in Methods and Practices in Psychological Science* 2018; **1**: 70–85.

31. **Dribe M, et al.** Socioeconomic status and fertility decline: insights from historical transitions in Europe and North America. *Population Studies* 2017; **71**: 3–21.

32. **Gagnon A, et al.** Is there a trade-off between fertility and longevity? A comparative study of women from three large historical databases accounting for mortality selection. *American Journal of Human Biology: The Official Journal of the Human Biology Council* 2009; **21**: 533–540.

33. **Lindahl-Jacobsen R, et al.** The male-female health-survival paradox and sex differences in cohort life expectancy in Utah, Denmark, and Sweden 1850-1910. *Annals of Epidemiology* 2013; **23**: 161–166.