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

Heidi A. Hanson1*, William W. Hay Jr2, Jonathan N. Tobin3, Shari L. Barkin4, Mark Atkins5, Margaret R. Karagas6, Ann M. Dozier7, Cynthia Wetmore8, Michael W. Konstan9 and James E. Heubi10; 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 Public Health Sciences, University of Rochester Medical Center, Rochester, NY, USA
7 Department of Pediatrics, Emory University, Children’s Health Care of Atlanta, Atlanta, GA, USA
8 Department of Pediatrics, Case Western Reserve University School of Medicine, Cleveland, OH, USA
9 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.

Key words: Early life exposure, lifespan, life course, clinical and translational science, biorepository.

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 undernutrition 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.

© The Association for Clinical and Translational Science 2018. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike licence (http://creativecommons.org/licenses/by-nc-sa/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the same Creative Commons licence is included and the original work is properly cited. The written permission of Cambridge University Press must be obtained for commercial re-use.

Downloaded from https://www.cambridge.org/core, IP address: 54.70.40.11, on 28 May 2021 at 01:00:49, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms . https://doi.org/10.1017/cts.2018.29
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 CTSIs 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://ctic.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 CTSAs hubs for an overall response rate of 88%. All CTSAs 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 data bases 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 CTSAs centers. Fig. 1b shows the number of early life exposures, child-maternal health, or lifespan research databases by institution. The majority of CTSAs 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...
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 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)

<table>
<thead>
<tr>
<th>n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Database includes multiple measurements/multiple time points</td>
<td>93</td>
</tr>
<tr>
<td>Age at enrollment</td>
<td></td>
</tr>
<tr>
<td>Prenatal</td>
<td>47</td>
</tr>
<tr>
<td>Infant</td>
<td>66</td>
</tr>
<tr>
<td>Child</td>
<td>55</td>
</tr>
<tr>
<td>Young adult</td>
<td>45</td>
</tr>
<tr>
<td>Adult</td>
<td>53</td>
</tr>
<tr>
<td>Older adults</td>
<td>30</td>
</tr>
<tr>
<td>Pregnancy</td>
<td>46</td>
</tr>
<tr>
<td>Unknown</td>
<td>8</td>
</tr>
<tr>
<td>Universal consent</td>
<td>18</td>
</tr>
<tr>
<td>Biorepository</td>
<td>76</td>
</tr>
<tr>
<td>Components of biorepository</td>
<td></td>
</tr>
<tr>
<td>Blood</td>
<td>58</td>
</tr>
<tr>
<td>Plasma</td>
<td>14</td>
</tr>
<tr>
<td>Tissue/ fluid</td>
<td>28</td>
</tr>
<tr>
<td>Other</td>
<td>12</td>
</tr>
<tr>
<td>Unknown</td>
<td>11</td>
</tr>
<tr>
<td>Type of subjects in biorepository</td>
<td></td>
</tr>
<tr>
<td>Normal</td>
<td>50</td>
</tr>
<tr>
<td>Disease state</td>
<td>42</td>
</tr>
<tr>
<td>Unknown</td>
<td>6</td>
</tr>
<tr>
<td>Other</td>
<td>9</td>
</tr>
<tr>
<td>Centralized repository</td>
<td>68</td>
</tr>
<tr>
<td>EHR integration</td>
<td>62</td>
</tr>
</tbody>
</table>

EHR, electronic health records.

subjects that are all healthy or all in a diseased state ($n_{health} = 11/57$; 19% and $n_{disease} = 10/57$; 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 as well as 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 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, biomarkers, 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.
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 for 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 CTSI 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 support the creation of the REDCap Survey and collection of data. However, it is possible that the respondent at each CTSI 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

The authors thank Cindy Pastern and Leah Dunkel for their valuable help coordinating the project and their role in the support of the Life Span Domain Task Force and Early Life Exposure working group.

Financial Support

This publication was made possible by the following CTSI grants from the National Center for Advancing Translational Science (NCATS), National Institutes of Health: Rockefeller University Center for Clinical and Translational Science (UL1TR001886), Vanderbilt Institute for Clinical and Translational Research (UL1TR002243), UIC Center for Clinical and Translational Science (UL1TR002003), Clinical and Translational Science Collaborative of Cleveland (UL1TR000439), University of Rochester’s Clinical and Translational Science Institute (UL1TR002001), Cincinnati Center for Clinical and Translational Sciences and Training (UL1TR001425), University of Utah Center for Clinical and Translational Science (UL1TR001067). Support for the work described in this publication was also provided by the CTSI Consortium Coordinating Center (C4) and C4 REDCap (UL54TR000123) award from the NCATS at the NIH. C4 staff assisted with the creation of the REDCap Survey and collection of data. H. Hanson is also partially funded by Utah Building Interdisciplinary Research Careers in Women’s Health Career Development Program (1K12HD085852).

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/ccts.2018.29

References

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


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


