Linking Maine Department of Education and Maine Department of Health and Human Services Early Childhood Data

Prepared by:

Craig A. Mason, Ph.D.

Shihfen Tu, Ph.D.

Quansheng Song, M.S.

June 2014

Maine Education Policy Research Institute
College of Education & Human Development
University of Maine
Orono, Maine

Linking Maine Department of Education and Maine Department of Health and Human Services Early Childhood Data

June 2014

Craig A. Mason, Ph.D.
Professor of Education and Applied Quantitative Methods

Shihfen Tu, Ph.D.
Associate Professor of Education and Applied Quantitative Methods

Quansheng Song, M.S.
Lead Database Administrator and Project Manager

Maine Education Policy Research Institute
A nonpartisan research institute funded by the Maine State Legislature, the University of Maine, and the University of Southern Maine.

Center for Research and Evaluation
College of Education and Human Development
University of Maine, 5766 Shibles Hall, Orono, Maine 04469-5766
(207) 581-2475

1865 THE UNIVERSITY OF MAINE

A Member of the University of Maine System

AUTHORS BIOGRAPHICAL INFORMATION

Craig A. Mason, Ph.D., is a Professor of Education and Applied Quantitative Methods at the University of Maine, where he also serves as the Director of the Center for Research and Evaluation and Co-Director of the Maine Education Policy Research Institute (MEPRI). For the past decade, Dr. Mason has also served as a methodological consultant to the U.S. Centers for Disease Control and Prevention. His research interests are in developmental growth models, parent-child relationships, informatics, and research methods. Dr. Mason has made numerous invited presentations to national groups on data linkage methodology and linking state population data systems. He has over 80 publications, and has been principal investigator or co-principal investigator on over $10 million in grants.

Shihfen Tu, Ph.D., is an Associate Professor of Education and Applied Quantitative Methods in the Department of Exercise Science and STEM Education at the University of Maine, where she is also affiliated with the Center for Research and Evaluation (CRE) and MEPRI. Dr. Tu has extensive experience in overseeing statewide projects in child health and development. She is the PI of a project that developed and currently maintains a statewide database system, *ChildLINK*, which manages early childhood screening data from programs in the Maine Center for Disease Control and Prevention (Maine CDC), Children with Special Health Needs (CSHN) Program. She is also involved in a longitudinal data analysis project with the Maine Department of Education State Longitudinal Data System.

Quansheng Song, M.S., is the Lead Database Administrator & IT Project Manager of the Center for Research and Evaluation (CRE) at the University of Maine. He has also serves as a database consultant to the U.S. Centers for Disease Control and Prevention, University of Guam, and Commonwealth Healthcare Corporation of Commonwealth Northern Mariana Islands. Mr. Song holds a B.E in computer science and technology from University of Science & Technology of China and a M.S. in computer science from the University of Maine.

EXECUTIVE SUMMARY

For the last several years, officials in Maine have discussed electronically linking child education data from the Department of Education with child health and developmental data from the Department of Health and Human Services. Sharing data will help programs improve the quality, timeliness, and efficiency of services, while simultaneously providing valuable information to inform policy decisions. Therefore, at the request of the Maine State Legislature, the Maine Educational Policy Research Institute (MEPRI) undertook a feasibility study to test and demonstrate the ability to link state birth/newborn records with state educational data. Specifically, the project sought to assess the degree to which the available data were sufficient to accomplish this goal given the absence of a shared unique identifier, the significant passage of time between birth and school data (five to ten years), and the existence of few potential identifying fields. If successful, a secondary goal was to use the linked information to illustrate how such data can provide additional information regarding long-term child outcomes. It should be noted that the MEPRI team has ongoing independent access to child-level data in both of the data systems used for this project. Consequently, it was possible to conduct this project without releasing data to anyone that did not already have access to child-level data within these systems. Nevertheless, approval was obtained from the University of Maine Institutional Review Board, and both the Maine CDC and the Department of Education.

The data used for this project were birth records from 2003 – 2005 (and select related health/development information), linked to 2010 school enrollment data and 2013 special education / state testing data. The linkage process involved a variety of iterative approaches described in detail in the full report. The process resulted in a final linked data file containing over 30,000 matched records that included both birth-related data from 2003 – 2005 and education-related data from 2010 – 2013. The pattern of data across records was somewhat complex, with children moving in and out of both the state and state education system.

The potential value of this type of linkage was illustrated using the Maine Newborn Hearing Program as an example[1]. The Maine Newborn Hearing Program promotes early hearing detection and intervention services in Maine, with the goal that all newborns are screened for hearing-loss prior to hospital discharge. Infants who do not pass their hearing screen are then to receive diagnostic testing by 3-months of age, and then to receive intervention services (such as hearing aids, sign language, etc.) by 6-months of age. Research suggests that this can be very valuable in promoting language and cognitive development in infants and young children with hearing loss – goals that are also particularly relevant for educators and education policy makers. While Maine has embraced this goal, there is no long-term data indicating how these children are developing and performing years later in school.

However, through this data linkage, 69 children born in Maine from 2003 – 2005 who had hearing loss that was screened and diagnosed through Maine's Newborn Hearing Program were subsequently linked to their education data in 2010 – 2013. This creates the possibility to see long-term educational outcomes for these children. For example, by linking these records it was found that among those students with hearing loss identified through the Newborn Hearing Program, 55% performed at the *proficient* or *proficient with distinction* level in reading in 2013. In regards to 2013 math proficiency, 49% performed at the *proficient* or *proficient with distinction* level – versus 37.5% for similar children whose hearing loss was ***not*** identified through the Newborn Hearing Program.

While the feasibility of conducting this type of data linkage was successfully demonstrated through this project, the experience suggests that including additional identifying information, in particular mother's name and date of birth, would be valuable for matching records that may include changes or errors. Also, when linking birth and education data, one anticipates that there will be a significant number of records in both systems that will not match simply due to in- and out-of-state migration over time. Including place of birth as an identifying field would allow one to automatically flag those records that cannot be matched with Maine birth records. Finally, it should be noted that MEPRI is uniquely qualified within Maine to assist in data linkage efforts across state agencies. MEPRI researchers have national reputations as experts in electronically

[1] Note that the Maine CDC Children with Special Health Needs (CSHN) Program provided permission to conduct these analyses.

linking population-based data systems, and have been invited to conduct workshops and trainings on record linkage for numerous national organizations including the U.S. Centers for Disease Control and Prevention, the U.S. Department of Health and Human Services, the Centers for Medicaid and Medicare Services, and others. MEPRI would be well positioned to further assist Maine in such efforts.

Contents

INTRODUCTION

Background: Why Link Records?

Like many other states, for the last several years, officials in Maine have discussed electronically linking child education data from the Department of Education with child health and developmental data from the Department of Health and Human Services. This reflects a recognition that by linking data, services and activities in both agencies can be strengthened in numerous ways. First, by sharing information, one can improve the timeliness, efficiency, and cost-effectiveness of services to children in need. For example, children with special health needs that are identified by the Maine Centers for Disease Control and Prevention would be able to access early intervention services more quickly if information regarding their cases were linked with Part C Services. Also, data from other agencies can be a powerful and cost-effective tool to assess and monitor the impact and effectiveness of programs or interventions. For instance, Maine's Newborn Hearing Program screens all newborns for hearing loss, with follow-up efforts made to have children accurately diagnosed and receiving services by 6 months of age in order to reduce their risk for language and cognitive delays. However, the long-term impact of this program on future cognitive and academic outcomes for these children is unknown. Furthermore, linking birth/early childhood data with education data can provide officials and policy makers with valuable information that can aid their decision making. For example, it can help education officials identify early childhood risk factors impacting student growth and achievement, and inform policy to better target valuable, limited resources in ways that maximize their potential benefit to students.

At the request of the Maine State Legislature, the Maine Educational Policy Research Institute (MEPRI) undertook a feasibility study of the possibility to link state birth/newborn records with state educational data for the same children when they were five to ten years old. As described in more detail later in this report, the project sought to assess the degree to which the available data were sufficient to accomplish this given both the absence of a shared unique identifier, the significant passage of time, and the existence of only a few potential identifying fields. If successful, a secondary goal was to use the linked information to illustrate how such data can provide additional information regarding long-term child outcomes.

The report begins with a description of data linkage methodology before reviewing the specific data systems that were used in this data linkage feasibility study. The methods and results of the linkage process are then described. Examples are given given illustrating how the linkage provided new information regarding the long-term educational outcomes for children identified with hearing loss through Maine's Newborn Hearing Program. The report concludes with thoughts and suggestions regarding future data linkage efforts.

Overview of Data Linkage Methodology

Data linkage involves connecting multiple records for the same individual across different data sources. It requires matching records based on certain identifying information, typically names, dates of births, or other demographic information. One begins by determining which of these variables exist in ***both*** data sets and can be used in combination to uniquely identify a person. These variables (referred to as *identifying fields*) are then used to match records in one data source with records in the other. As described below, records can be matched based on a deterministic or a probabilistic protocol.

Deterministic Linkage

A deterministic match requires records to be linked only when all identifying fields are identical in both records. If any of the identifying fields do not perfectly agree, the records are not linked. For example, a birth record for "Zbigniew Brzezinski" would not match with a school record for "Zbigniew Brzezinsky" because the last name is slightly different, even though few would doubt that it is likely the same person.

Deterministic matching is most effective when the linkage is done using a relatively small number of identifying fields and is applied to high quality and highly discriminating data. For instance, social security number is highly discriminating because in theory the number is unique to each person. No two people should share the same social security number. If two sources both include social security numbers, a deterministic match using social security numbers ***may***

be very effective and efficient. Gender, on the other hand, is not a particularly useful identifier. For a random person, he or she shares their gender with roughly half of the population.

Beyond its uniqueness, the value of an identifying field is limited by the quality of the data. If the data quality is poor, even a highly discriminating field is still of limited use. For example, if one is linking two sources based on social security number, but many of these numbers are missing or incorrect, then social security may be a poor field to use – particularly in a deterministic linkage that requires all identifying fields to match exactly.

Unique Identifier: A unique identifier is a single identifying field that uniquely identifies everyone in the database. For example, social security number may serve as a unique identifier. "ID" numbers are typically unique identifiers that have been created for a specific system. For example, the Maine Department of Education uses MEDMS IDs as a unique identifier for students. This allows one to easily match students across any two sets of records as long as both include MEDMS ID numbers.

Furthermore, the number of identifying fields used in a linkage can also create problems. Using too few identifying fields may mean that there is insufficient information to accurately discern between different individuals. For instance, a protocol that only uses First Name and Last Name may be able to uniquely match two records for "Zbigniew Brzezinski", but would likely be ineffective at linking records for someone named "John Smith". Increasing the number of identifying fields will help differentiate individuals, but will also increase the likelihood that two records for the same individual will not exactly agree on all fields. Consequently, too many identifying fields may result in increased missed matches unless the data quality is exceptionally high.

It should be noted that there are various strategies for addressing some of these issues. For example, if there are spelling errors in names, one approach is to truncate names and match on the first several letters in a name. In this case, a birth record for "Zbigniew Brzezinski" would match with a school record for "Zbigniew Brzezinsky" if one only matched on the first 4 letters of the last name. If there are many identifying fields that can be used, one may also conduct a series of several matches using different combinations of these identifying fields. Using this approach, one typically removes matched records at each step, creating a growing file of matched records and two shrinking files of unmatched records with each attempt.

Data Quality: One of the authors was previously involved in work outside of the State of Maine with an organization that was confident that the social security numbers in their system were unique and accurate. An examination of the data found hundreds of social security numbers similar to “123-45-6789” or “111-11-1111”. Because the social security number was required, if a person did not know his or her number, some workers in the agency would simply make one up so that the record could be processed.

As this suggests, deterministic matching can be a quick and efficient method of data linkage when the quality of the data is high, and/or when pre-existing unique identifiers are present across multiple datasets (e.g., MEDMS IDs). Some may suggest that one should always use a deterministic approach, arguing that because the two records must agree on identifying fields, it ultimately provides the best quality data. But it may also lead to correct records ***not*** being matched. Such non-links potentially create a systematic bias in the linked records. For example, ethnic groups that have uncommon names, non-standard letter combinations, or non-traditional spellings would be more likely not to be matched using a deterministic linkage. In such cases, a probabilistic linkage protocol may provide a powerful alternative tool.

Probabilistic Linkage

In contrast to a deterministic matching, probabilistic linkage does not require complete agreement on all identifying fields from both sources in order to conclude that the two records belong to the same individual. Instead, it statistically calculates a measure of the probability that two records belong to the same individual, ***even if they do not match*** on some fields. It does this by mathematically considering factors such as how common a name or value is, the quality of the data, and the expected number of matches.

Frequency of values. The more common the value in a field, the more likely it is that two records will agree on that field even if the records belong to different people. For example, consider a possible match where a birth record and an education record both have the first name “John”. Agreement on the name “John” does not provide much evidence that the two records belong to the same individual—there may be hundreds of other records with the first name John, and so this may match may not be the correct one. Alternatively, if a birth record and an education record both have the first name “Zbigniew” one is much more likely to conclude that the two records ***do*** belong to the same person. Statistically, when identifying fields agree on rare values, it is a stronger sign that the two records belong to the same person than when they agree on common values.

Quality of the data. The quality of a data field, defined as the accuracy and/or reliability of information contained in it, also influences the likelihood that two records belong to the same person. As noted previously, a data field is of poor quality if it contains many errors or incomplete information. Consider an example where a birth record and an education record agree on first name, middle name, and last name, but disagree on the date of birth. If one knows that the date of birth is very carefully recorded and almost always correct, disagreement on that field would be strong evidence that the two records are not a correct match. On the other hand, if it is known that date of birth is often entered wrong in one or both of the sources, disagreement may provide relatively little evidence that the match is incorrect. Statistically, disagreement on poor quality fields is less evidence of an incorrect match than is disagreement on high quality fields.

Number of expected matches. The third factor influencing probabilistic linkage is the actual number of matches that are expected to exist across the two sources. All things being equal, there is a greater probability that a potential match is correct when the two sources are known to contain records on exactly the same people. Consider the most extreme situation: Linking children born in 2013 with school records from 2010. In this case, the probability that ***any*** child born in 2013 correctly matches with a child attending school in 2010 is zero regardless of how well the records match on the identifying fields.

Computationally, the probabilistic approach is much more complicated and thus a more time-consuming and expensive method than the deterministic protocol. Nevertheless, probabilistic linkage provides an alternative to deterministic linkage when it is important to minimize the number of overlooked matches due to inconsistencies or errors in the data.

THIS PROJECT

As stated previously, the goal of this project was to conduct a feasibility study of the ability to link state birth and related records with state educational data for the same children five to ten years later. The project sought to assess the degree to which the available data were sufficient to accomplish this given both the absence of a shared unique identifier and the existence of only a few potential identifying fields. If successful, a secondary goal was to use the linked information to illustrate how such data can provide additional information regarding long-term child growth and developmental outcomes.

Therefore, MEPRI researchers linked data from two Maine state data systems: *ChildLINK* and the *State Longitudinal Data System*. *ChildLINK* is a partnership between the Maine Center for Disease Control and Prevention (Maine CDC) and the University of Maine. *ChildLINK* was used as the source for data on all births in Maine from 2003-2005, as well as newborn hearing screening results and diagnosis of hearing loss for children. The Maine Department of Education *State Longitudinal Data System* (SLDS) was used as the source for education data on all children born from 2003 to 2005 who were attending a Maine public school in 2010 and/or 2013. Both of these systems are described in more detail in the following section.

Note that the MEPRI team has had ongoing independent access to child-level data in both of these systems for several years prior to this project. Through its long-term partnership with the Department of Education and the Joint Standing Committee on Education and Cultural Affairs, MEPRI has access to the SLDS data in order to conduct policy analysis and answer education-related questions for the State. The same team also created *ChildLINK* and has operated it for the Maine CDC for over a decade. Consequently, MEPRI's unique position in Maine allowed it to conduct this project without releasing data to anyone that did not already have access to child-level data within these systems. Nevertheless, approval was obtained from the University of Maine Institutional Review Board, and both the Maine CDC and the Department of Education.

ChildLINK

ChildLINK is an integrated data system designed for early childhood health and development screening. It is a collaboration between the University of Maine and the Children with Special Health Needs (CSHN) Program within the Maine Center for Disease Control and Prevention.

First established in 2002, *ChildLINK* is a population-based data system linking records from various programs within CSHN. It includes information on all births in Maine (obtained from the state electronic birth certificate). This information is then linked with data for the Maine Newborn Hearing Program, Maine Birth Defects Program, Maine Newborn Bloodspot Program, and Maine Cleft Lip and Palate Program (see Tu and Mason, 2004; Tu, Mason, and Song, 2007 for more information on the design of *ChildLINK*). Furthermore, in collaboration with Maine Developmental Disabilities Council, a module for early childhood screening of autism spectrum disorders was recently developed for the *Maine Autism Spectrum Disorders Development Project* (MeASD). In addition, a module for screening critical congenital heart defects at birth is currently under development.

State Longitudinal Data System **(SLDS)**

While students are assessed throughout their academic careers, it has historically been difficult to track academic growth and experience over time due to the lack of a single, state-level system for organizing educational data from multiple sources across multiple years. The objective of the Maine Statewide Longitudinal Data System (SLDS) is to create a centralized data warehouse capturing Pre-K through higher education data that has existed in multiple, isolated state and district sources. Specifically, the SLDS…

> "…allows student data to be compiled over time, ensuring that each student has an accurate record regardless of transience across schools or districts. In addition, the SLDS will improve teachers' ability to access relevant data that pertains specifically to their students and will accurately align teachers, classes, and individual students."[2]

METHODS

This project involved linking data extracted from the Maine Department of Education State Longitudinal Database (SLDS) and the *ChildLINK* system. As noted previously, the MEPRI team conducting this research has access to SLDS data through MEPRI's long-term partnership with the Maine Department of Education, and the same MEPRI team also created and operates

[2] http://www.maine.gov/doe/excellence/resources/glossary.html

ChildLINK for the Maine CDC. Consequently, it was possible to conduct this project without releasing data to any entity that did not already have access to this data. Nevertheless, approval was obtained from the University of Maine Institutional Review Board, and both Maine CDC and the Department of Education.

Specifically, the records that were matched included *ChildLINK* data for children born from 2003-2005. This consisted of identifying fields for all children born in Maine during these years, as well as hearing screening results for all children screened and subsequent diagnostic testing results. SLDS data was drawn from 2010 Enrollment data, 2013 NECAP (state testing) data, and 2013 Special Education data and included identifying fields, special education status, special education placement, and NECAP proficiency data. Records from the SLDS were also restricted to 2003-2005 births. Note that MEDMS IDs were used to link data between the SLDS files. Additional detail regarding the data used in this project is presented in Table 1.

The linkage process involved a standard iterative approach in which records are first matched using highly restrictive or demanding criteria (e.g., records must exactly agree on all matching fields). Matches are then removed from both of the original datasets, and the remaining unmatched records are used in a second linkage attempt based on a different set of criteria. This process is then repeated, allowing each iteration to match a subset of records that may reflect different issues, errors, or missing data. For this project, eleven iterations were used – with nearly all matches occurring in the first iteration. After the first iteration (an exact match on name and date of birth), all subsequent possible matches were manually reviewed in order to determine if the records appeared sufficiently similar to be considered a match. The iterative model used in this project is summarized in Table 2

Table 1. Data fields used and source.

SOURCE: ChildLINK (2003-2005)

Identifying Fields:	Child First Name
	Child Middle Name
	Child Last Name
	Child Date of Birth
Additional Fields:	Screening Result
	Diagnostic Result

SOURCE: SLDS Enrollment Data (2010)

Identifying Fields:	Child First Name
	Child Middle Name
	Child Last Name
	Child Date of Birth
	MEDMS ID
Additional Fields:	Special Education Status
Additional Fields:	Special Education Category

SOURCE: SLDS NECAP (2013)

Identifying Fields:	Child First Name
	Child Middle Name
	Child Last Name
	Child Date of Birth
	MEDMS ID
Additional Fields:	Reading Proficiency
Additional Fields:	Math Proficiency

SOURCE: SLDS Special Education Data (2013)

Identifying Fields:	Child First Name
	Child Middle Name
	Child Last Name
	Child Date of Birth
	MEDMS ID
Additional Fields:	Special Education Status
	Special Education Category

.

Table 2. Iterative linkage protocol.

Iteration	SLDS	ChildLINK	Method	Special Requirement	m
1	FirstName	CFirst	Deterministic		0.98
1	birthdate	CDateofBirth	Deterministic		0.95
1	LastName	CLast	Deterministic		0.98
2	birthdate	CDateofBirth	Deterministic		0.95
2	FirstName	CFirst	Deterministic	First 4 Letters	0.98
2	LastName	CLast	Deterministic		0.98
3	LastName	CLast	Deterministic	First 4 Letters	0.98
3	FirstName	CFirst	Deterministic		0.98
3	birthdate	CDateofBirth	Deterministic		0.95
4	birthdate	CDateofBirth	Deterministic		0.95
4	FirstName	CFirst	Deterministic	First 4 Letters	0.98
4	LastName	CLast	Deterministic	First 4 Letters	0.98
5	birthdate	CDateofBirth	Deterministic	Month of the date	0.95
5	FirstName	CFirst	Deterministic	First 4 Letters	0.98
5	LastName	CLast	Deterministic	First 4 Letters	0.98
5	birthdate	CDateofBirth	Deterministic	Day of the date	0.95
6	birthdate	CDateofBirth	Deterministic		0.95
6	FirstName	CFirst	Deterministic	Last 4 Letters	0.98
6	LastName	CLast	Deterministic	Last 4 Letters	0.98
7	birthdate	CDateofBirth	Deterministic		0.95
7	LastName	CLast	Deterministic	First 4 Letters	0.98
7	FirstName	CFirst	Deterministic	First 2 Letters	0.98
8	FirstName	CFirst	Deterministic	First 4 Letters	0.98
8	birthdate	CDateofBirth	Deterministic		0.95
8	LastName	CLast	Deterministic	First 2 Letters	0.98
9	birthdate	CDateofBirth	Deterministic		0.95
9	LastName	CLast	Deterministic	First 4 Letters	0.98
9	MiddleName	CMiddle	Deterministic		0.95
10	FirstName	CFirst	Deterministic	First 4 Letters	0.98
10	MiddleName	CLast	Deterministic	Last 4 Letters	0.95
10	MiddleName	CMiddle	Deterministic	First 4 Letters	0.95
10	birthdate	CDateofBirth	Deterministic		0.95
11	birthdate	CDateofBirth	Deterministic		0.95
11	FirstName	CFirst	Deterministic	First 4 Letters	0.98
11	MiddleName	CMiddle	Deterministic	First 4 Letters	0.95
11	LastName	CMiddle	Deterministic	Last 4 Letters	0.95

This project consisted of three series of record linkages. The first linked *ChildLINK* data for 2003-2005 births with *2010 SLDS Enrollment* data. Once this was completed, *2013 SLDS NECAP* and *2013 SLDS Special Education* data were also linked to these matched records using MEDMS ID numbers. It is possible that some *ChildLINK* records that did not match a record in the *2010 SLDS Enrollment* data may nevertheless match a record in the *2013 SLDS NECAP* data. This would happen if a child born in Maine in 2004 (and thus in *ChildLINK*) was being home schooled in 2010, but was then subsequently enrolled in public school in 2013. Therefore, a second series attempted to match any *ChildLINK* records that were not linked to *2010 SLDS Enrollment* data to *2013 SLDS NECAP* data. For similar reasons, a third series attempted to match any remaining *ChildLINK* records to *2013 SLDS Special Education* data.

Data linkage was performed using software developed by MEPRI researchers at the University of Maine. This software includes a flexible, interactive tool that allows one to select from a variety of linkage tools using both deterministic and probabilistic techniques. The software has been used by government agencies and researchers in Maine, Iowa, Virginia, Florida, Guam and elsewhere.

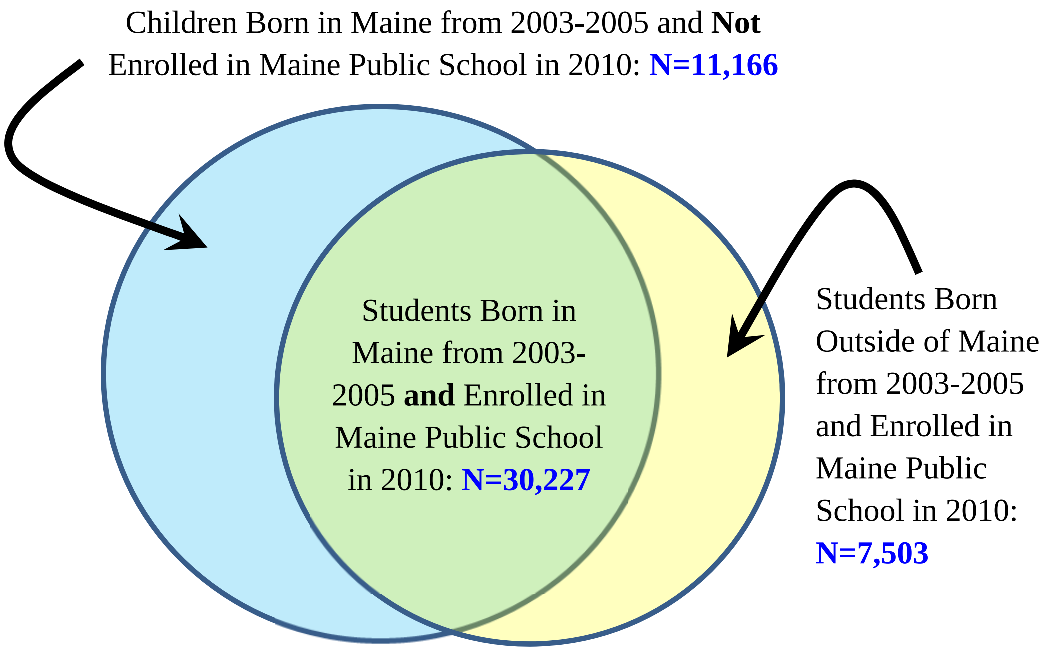
RESULTS

LINKING 2003-2005 *CHILDLINK* TO *2010 SLDS ENROLLMENT*

The process began by extracting records for all children born 2003-2005 from both *ChildLINK* and the *2010 SLDS Enrollment* data files. This resulted in 41,393 *ChildLINK* records and 37,730 Enrollment records. The previously described linkage algorithm was used to match these records based on first name, middle name, last name, and date of birth (in various combinations - see Table 2). As illustrated in Figure 1, the final linked file contained cases for 30,227 youth with records in both systems[3].

[3] Please note that while the SLDS contains data on all public schools in Maine, it also contains limited information on some non-public schools. Because this report is not examining public/non-public distinctions, for succinctness we will refer to the students in the SLDS as “public school students” and we will refer to schools in the SLDS as “public schools”. Technically, it would be more accurate to refer to these as “students attending schools reporting data to the SLDS” and “schools reporting data to the SLDS”. Alternatively, the authors could have excluded non-

Figure 1. Illustration of linkage between *ChildLINK* and *2010 SLDS Enrollment* data


Children Born in Maine from 2003-2005 and **Not** Enrolled in Maine Public School in 2010: **N=11,166**
Students Born in Maine from 2003-2005 **and** Enrolled in Maine Public School in 2010: **N=30,227**
Students Born Outside of Maine from 2003-2005 and Enrolled in Maine Public School in 2010: **N=7,503**

As reflected in Figure 1, the linked file does not represent all children born in Maine from 2003 – 2005, nor does it represent all public school students born from 2003 – 2005. It corresponds to a very specific, and in this case significant, subset of both of these groups: *Children born in Maine from 2003 – 2005 who **ALSO** were enrolled in Maine Public Schools in 2010*. Of the 41,393 children born in Maine during that time (*ChildLINK* records), 11,166 were not enrolled in a Maine Public School in 2010 and so are not part of the linked data. When using linked records, it is important to understand who is included in the final data and who is not. In this case, these 11,166 records reflect children who were born in Maine from 2003 through 2005, but who later moved out of state some time prior to 2010. It would also include children born in Maine from 2003 through 2005 who were either home schooled in 2010 or attending a private school that does not report into the SLDS.

public schools from all analyses; however, that would have defeated the purpose of assessing the degree to which ***all*** 2003 –2005 births could be linked with 2010 education information.

Identifying Fields: A review of 2003 – 2005 birth records in ChildLINK found no cases where a child had the same first name, last name, and date of birth. Nevertheless, additional identifying fields would help address errors or missing data. A manual review of the final results suggested that several hundred additional records might be matched with additional identifying fields.

Of the 37,730 *SLDS 2010 Enrollment* records belonging to children born from 2003 – 2005[4], 7,503 did not link with birth data in *ChildLINK*, suggesting that these are students attending Maine public schools who were not born in Maine.

EXPANDING LINKAGE TO *2013 SLDS NECAP* DATA

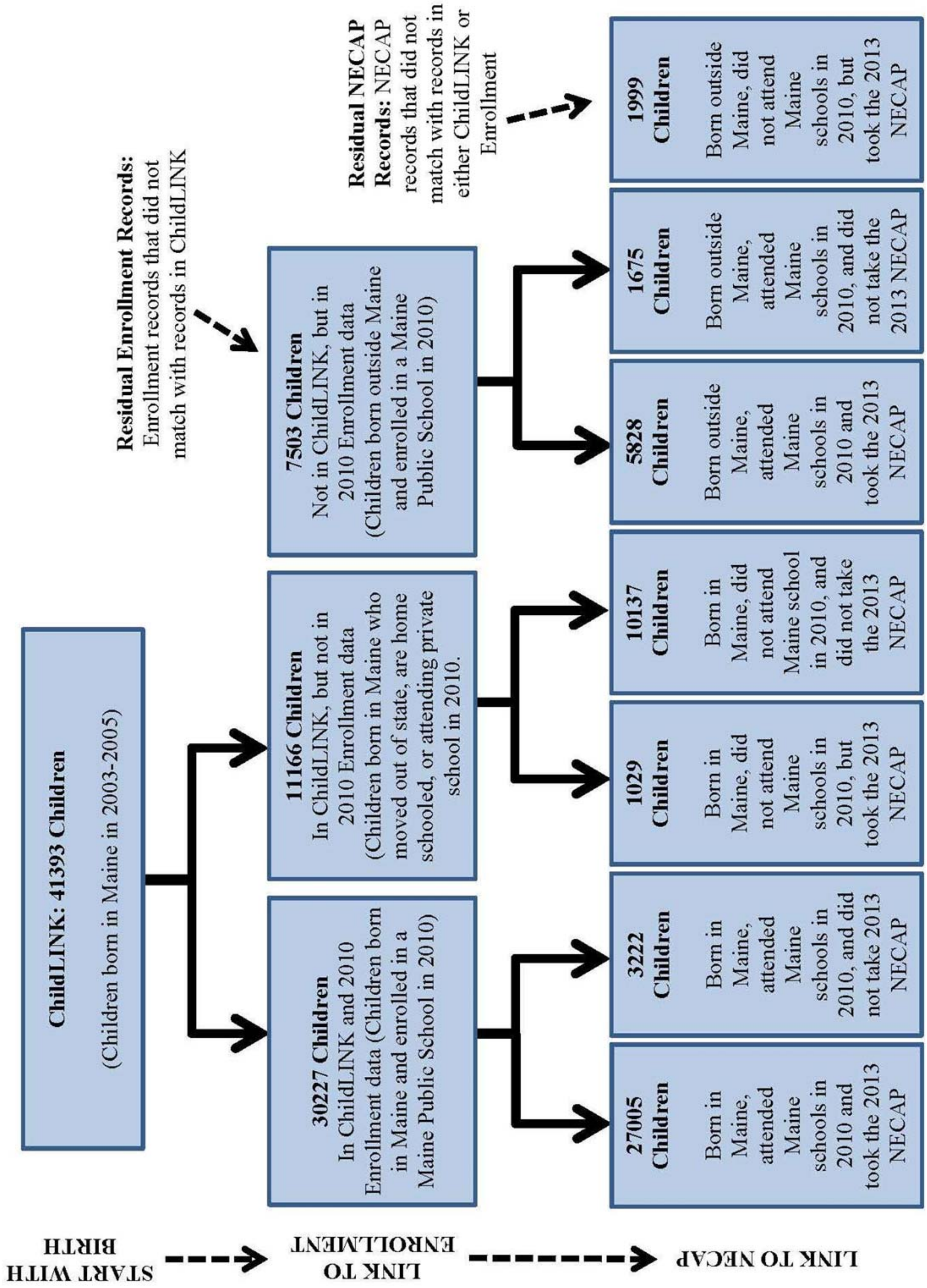
The three sets of records (*ChildLINK-Only*, *2010 Enrollment-Only* , and *Matched ChildLINK-2010 Enrollment*) were then linked to the *2013 SLDS NECAP* data. In essence, one would anticipate that each of these three groups will include some children who will appear in the *2013 SLDS NECAP* data and some children who will not. As this implies, as multiple data sources are sequentially linked, the matching process and nature of the resulting data can rapidly become more complex. The end-result of this process is summarized visually in Figure 2 and described in more detail below.

Linking *Matched ChildLINK-2010 Enrollment* Data to *2013 SLDS NECAP* Data

Fortunately, the SLDS data includes MEDMS ID numbers for all students enrolled in Maine public schools or included anywhere in the SLDS data system. MEDMS IDs uniquely identify these children and do not change over time, even if a child moves to a different school or district. Consequently, they can be used to link SLDS records across years throughout the state.

[4] Again, please note that it would technically be more accurate to refer to these as "schools reporting data to the SLDS". But given the purpose of this report is assess the ability to link ***all*** 2003 – 2005 births, rather than explore public/non-public distinctions, for succinctness we will refer these as "public schools" in this report.

Figure 2. Illustration of Linkage with *2013 SLDS NECAP* data.


START WITH BIRTH
LINK TO ENROLLMENT
LINK TO NECAP
ChildLINK: 41393 Children
(Children born in Maine in 2003-2005)
30227 Children
In ChildLINK and 2010 Enrollment data (Children born in Maine and enrolled in a Maine Public School in 2010)
11166 Children
In ChildLINK, but not in 2010 Enrollment data (Children born in Maine who moved out of state, are home schooled, or attending private school in 2010.
7503 Children
Not in ChildLINK, but in 2010 Enrollment data (Children born outside Maine and enrolled in a Maine Public School in 2010)
Residual Enrollment Records: Enrollment records that did not match with records in ChildLINK
Residual NECAP Records: NECAP records that did not match with records in either ChildLINK or Enrollment
27005 Children
Born in Maine, attended Maine schools in 2010 and took the 2013 NECAP
3222 Children
Born in Maine, attended Maine schools in 2010, and did not take 2013 NECAP
1029 Children
Born in Maine, did not attend Maine schools in 2010, but took the 2013 NECAP
10137 Children
Born in Maine, did not attend Maine school in 2010, and did not take the 2013 NECAP
5828 Children
Born outside Maine, attended Maine schools in 2010 and took the 2013 NECAP
1675 Children
Born outside Maine, attended Maine schools in 2010, and did not take the 2013 NECAP
1999 Children
Born outside Maine, did not attend Maine schools in 2010, but took the 2013 NECAP

The first of these analyses matched the records in the *Matched ChildLINK-2010 Enrollment* set to records in the *2013 SLDS NECAP* data file. Because of the existence of a common unique identifier (MEDMS IDs), this was performed using a single, straightforward deterministic match based on MEDMS IDs, and did not rely on the iterative approach summarized in Table 2.

As reflected in Figure 2, of the 30,227 children in the *Matched ChildLINK-2010 Enrollment* set, 27,005 also linked to a record in the *2013 SLDS NECAP* file. These are children who were born in Maine from 2003 – 2005, attended Maine public schools in 2010, and took the NECAP in 2013. The remaining 3,222 children were not linked to records in the *2013 SLDS NECAP* file, and reflect children who were born in Maine from 2003 – 2005, attended Maine public schools in 2010, but for some reason did not take the NECAP assessment in 2013. This would include children who moved out of state sometime between 2010 and 2013, as well as children who transitioned into home schooling or a private school that did not report data to the SLDS. It would also include those students enrolled in special education who did not take the NECAP assessment, as well as other possibilities such as death prior to 2013 or a serious illness in 2013 that prevented a student from taking the NECAP that year.

Linking *2010 Enrollment-Only* Data to *2013 SLDS NECAP* Data

The second of these analyses matched the records in the *2010 Enrollment-Only* set to records in the *2013 SLDS NECAP* file also using the student's MEDMS IDs.

As reflected in Figure 2, of the 7,503 records in the *2010 Enrollment-Only* set, 5,828 also linked to a record in the *2013 SLDS NECAP* file. These are children who were born from 2003-2005 in a state other than Maine, but attended a Maine public school in 2010, and took the NECAP in 2013 (i.e., students who were born outside of Maine, but moved to Maine prior to 2010 and then attended public schools in Maine). The remaining 1,675 children were not linked to records in the NECAP data, and reflect children who were born from 2003 – 2005 in a state other than Maine, attended Maine public schools in 2010, but for some reason did not take the NECAP assessment in 2013. This would include children who were born outside of Maine from 2003 – 2005, moved to Maine prior to 2010 and attended a public school, but then (a) moved out of state between 2010 and 2013, (b) transitioned into home schooling or a private school that did not report data to Maine DOE, (c) were enrolled in special education services and did not take the NECAP assessment in 2013, or (d) did not take the NECAP for other reasons.

Linking *ChildLINK-Only* Data to *2013 SLDS NECAP* Data

A third set of analyses matched the records in the *ChildLINK-Only* set to records in the *2013 SLDS NECAP* file. Given, there is no unique identifier across these systems (i.e., *ChildLINK* does not contain MEDMS IDs, and the SLDS does not include *ChildLINK* IDs), the same iterative approach was used in which records were first matched using highly restrictive or demanding criteria (e.g., records must exactly agree on all matching fields), and then those records that were not linked were sequentially re-matched using different sets of criteria (see Table 2). As before, all matches after the first iteration (an exact match on name and date of birth) were manually reviewed in order to determine if the records appeared sufficiently similar to be considered a true, correct match.

As reflected in Figure 2, of the 11,166 children in the *ChildLINK-Only* set, 1,029 linked to a record in the *2013 SLDS NECAP* file. These are children who were born in Maine from 2003 – 2005, did not attend a Maine public school in 2010, and yet took the NECAP in 2013. For example, this would include cases where a child's birthday was late in 2005 and the parents chose not to enroll him or her in public school in 2010. It would also include young children who were initially home schooled or attended a private school in 2010, and then enrolled in public school prior to 2013.

The remaining 10,137 children were not linked to records in the NECAP data. This reflects children who were born in Maine from 2003 – 2005, but did not attend Maine public schools in 2010, and did not take the NECAP assessment in 2013. This would include children who moved out of state prior to 2010, as well as children who are home schooling or attending private school throughout this period, as well as other possibilities, such as children who may have died prior to 2010.

Residual *2013 SLDS NECAP* Records

The process described above results in 6 groups of children based on (1) whether they appear in the *Matched ChildLINK-2010 Enrollment* data, the *2010 Enrollment-Only* data, or the

ChildLINK-Only data, and (2) whether they subsequently did or did not appear in the *2013 SLDS NECAP* data. However, as illustrated in Figure 2, there is also a seventh group that consists of 1,999 records in the *2013 SLDS NECAP* data that did not match to either the 2003 – 2005 *ChildLINK* birth data or the *2010 SLDS Enrollment* data. These are children born from 2003 – 2005 outside of Maine that were not enrolled in Maine public schools in 2010, but nevertheless took the NECAP assessment in 2013. This would include children born out of state from 2003 – 2005 who moved to Maine between 2010 and 2013 and subsequently enrolled in a public school and took a 2013 NECAP assessment. It would also include children born out of state from 2003 – 2005 who moved to Maine prior to 2010, but were initially home schooled, etc., during 2010, and then enrolled in a public school and took the 2013 NECAP assessment.

Expanding Linkage to *2013 SLDS Special Education* Data

This same process was then expanded and repeated a final time linking the seven sets of data created in the *ChildLINK* 2003 – 2005 birth data, *2010 SLDS Enrollment* data, *2013 SLDS NECAP* data linkage with *2013 SLDS Special Education* data. Following the same pattern, each of these seven sets of records can potentially be divided again into two smaller groupings, this time based on whether or not a matched record is found in the *2013 SLDS Special Education* data. This results in fourteen sets of possible matching combinations. There is also a 15th set that consists of *2013 SLDS Special Education* records that did not match any of the previously linked sources. This would reflect children who were born outside of Maine from 2003 – 2005 and were neither enrolled in Maine public schools in 2010 nor took the NECAP assessment in 2013, and yet were enrolled in special education in 2013. For example, this would include children born out of state from 2003 – 2005, who moved to Maine between 2010 and 2013, enrolled in public school and received special education services in 2013 and did not take the 2013 NECAP assessment. For succinctness this report will not detail the results of this final series of record linkages.

Illustrative Analyses Using Linked Data

The purpose of this report was to test and demonstrate the feasibility of linking birth/early childhood (i.e., newborn) records for Maine with Maine Department of Education records, rather than address a specific policy or research question. Nevertheless, the following results from the linked data set may help to illustrate some of the potential opportunities that such data provide policy makers and state officials.

For example, an ongoing data linkage such as this would provide health officials and policy makers in the Maine CDC and the Maine Department of Education with potentially valuable information regarding the long-term education-related effects of early intervention programs, such as the Maine Newborn Hearing Program[5]. The Maine Newborn Hearing Program promotes early hearing detection and intervention services in Maine, with the goal that all newborn babies are screened for hearing-loss prior to hospital discharge. Infants who do not pass their hearing screen are then to receive diagnostic testing by 3-months of age, and then to receive intervention services (such as hearing aids, sign language, etc.) by 6-months of age. Research suggests that early identification and intervention can be very valuable in promoting language and cognitive development in infants and young children with hearing loss (Joint Committee on Infant Hearing, 2007) – goals that are also particularly relevant for educators and education policy makers.

While Maine has embraced this goal, like many other states, Maine has no mechanism or system in place to assess the long-term impacts of these efforts on children. Linking Maine CDC data contained in *ChildLINK* with Maine Department of Education data contained in the *State Longitudinal Data System* would be a major step in accomplishing this objective.

Note that Maine began collecting data on newborn hearing screening in 2003, and the process of screening all infants, making referrals to audiologists, and obtaining diagnostic results was still in development from 2003 to 2005. Consequently, the newborn hearing screening and diagnostic data used in this illustration is limited[6]. Also, given this is a public report and hearing loss impacts a relatively small number of children, in order to guard confidentiality the following discussion focuses on general summaries rather than specific details.

[5] Note that the Maine Children with Special Health Needs (CSHN) Program provided permission to conduct the following analyses.

[6] For the core purpose of this project – testing and demonstrating the feasibility of linking birth records with education records — a later birth cohort could not be used as NECAP data would not be available (NECAP assessments begin in 3rd grade).

Ultimately, the data linkage described in this report found 92 children born in Maine from 2003 – 2005 who had hearing loss that was screened and diagnosed through Maine's Newborn Hearing Program. Without the record linkage, this is all that health or education officials would know about these children. However, through this test linkage between *ChildLINK* and the SLDS, 69 of these children were subsequently identified in Maine education data for 2010 and/or 2013 (appearing in one or more of the SLDS data files). This creates the possibility to see certain long-term educational outcomes for these children, such as special education placement and performance on state testing. For example, a higher percentage of these children (55%) were receiving special education services in 2010, versus similar children whose hearing loss was ***not*** identified through the Newborn Hearing Program (38%), a marginally significant difference ($\chi^2(1) = 3.032$, $p = .082$)[7].

Furthermore, in both 2010 and 2013, 55% of children whose hearing loss was identified through the Newborn Hearing Program were receiving special education services; although, these were not all the same children in both years. There was a degree of movement in and out of special education services, with some children receiving services in 2010 subsequently not receiving special education services in 2013 (but still enrolled in a Maine school), while others who were not receiving special education services in 2010 were receiving services in 2013.

Similarly, special education classifications for many of these students also changed between 2010 and 2013. Among those receiving special education services in 2010, half were identified with "hearing impairment", "deafness", or "speech and language impairment". Three years later many of these classifications had changed, and in 2013 the most common special education classification for this group was "multiple disabilities". Thirty-two percent of children whose hearing loss was identified through the Newborn Hearing Program and who were receiving special education services in 2013, were classified as having "multiple disabilities" — which is triple the rate for this category among other students receiving special education services. In regards to 2013 special education classroom placements, among the subset of these 69 students

[7] The small sample size for these initial years makes detecting statistically significant effects more challenging. With continued data linkage, this would be less of an issue and also allow for examining more subtle or complex effects.

who were receiving special education services, half were either in a regular classroom 80% or more of the time, or were attending a separate school.

Finally, 2013 NECAP assessments for the 69 students with hearing loss identified through the Newborn Hearing Program found 55% performing at the proficient or proficient with distinction level in reading. In regards to 2013 NECAP math proficiency, 49% performed at the proficient or proficient with distinction level – versus 37.5% for similar children whose hearing loss was ***not*** identified through the Newborn Hearing Program.

SUGGESTIONS FOR IMPROVING FUTURE RECORD LINKAGE

Finally, while the feasibility for conducting this type of data linkage was successfully demonstrated through this project, a few recommendations can be offered regarding possible ways to enhance future data linkage projects between state agencies in Maine.

Additional Identifiers

The set of common identifying fields between the two data sets was limited to only child first name, child middle name, child last name, and date of birth. Additional identifying fields would be valuable for matching records that may include changes or errors in one of these fields. For example, names can be misspelled—particularly non-traditional names, and dates can be transposed or confused. For those cases where there is a mistake or missing data in one field, a few additional identifying fields would allow for more powerful probabilistic matching, as well as more matching iterations using alternative identifying fields. Both would increase the number and accuracy of matches. Specifically, mother's name or maiden name, and mother's birthdate would be particularly valuable in uniquely identifying children and in linking with other child data systems.

Place of Birth

As demonstrated through this feasibility study, when linking birth and education data, one anticipates that there will be a significant number of records in both systems that will not match simply due to in-and-out of state migration over time. As seen in the narrative review of this linkage process, when an education record does not match to a birth record, the logical assumption is that the child was born out of state. However, this may not be true. Determining how many unmatched records is “acceptable” or how many is indicative of fundamental problems with the data or the linkage process can be difficult. Knowing that an education record is for a student born out of state would answer that question. In addition, it would allow one to automatically recognize and exclude records that cannot be matched from being included in the linkage process. This would prevent the creation of erroneous false-matches, while simultaneously reducing the uncertainty regarding unmatched records and providing greater confidence in the results.

MEPRI Assistance

Finally, as previously noted, MEPRI is uniquely qualified within Maine to assist in data linkage efforts across state agencies. MEPRI researchers have national reputations as experts in electronically linking population-based data systems and have been invited to conduct workshops and trainings on record linkage for groups including the U.S. Centers for Disease Control and Prevention, the U.S. Department of Health and Human Services, the National Association for Public Health Statistics and Information Systems, the Association of Maternal and Child Health Programs, the Centers for Medicaid and Medicare Services, and the Association of University Centers on Disabilities, as well as over a dozen other states. In short, MEPRI can function as a trusted third-party serving as an independent bridge between state programs—both public and private—with a proven record of linking, managing, and protecting sensitive health, development, and education data in Maine. MEPRI would be well positioned to further assist Maine in these efforts.

CONCLUSION

The goal of this project was to conduct a feasibility study of the potential to link state birth/newborn records with state educational data. Prior to conducting the record linkage, several factors suggested that accomplishing this may be difficult. Specifically, the absence of a shared unique identifier, the significant passage of time between birth and school data (five to ten years), and the existence of only a few identifying fields were seen as possibly limiting the ability to match records across these systems.

As described in the report, it was possible to successfully link birth records from 2003 – 2005 (and select related health/development information) with 2010 school enrollment data and 2013 special education / state testing data using a series of iterative matching approaches. The process resulted in a final linked data file containing over 30,000 matched records that included both birth and subsequent education-related data. The pattern of data across records was somewhat complex, with children moving in and out of Maine and in and out of the state education system. Nevertheless, the end product can be informative, as was demonstrated in the ability to – for the first time – examine long-term education outcomes for children served by a newborn health and development program.

In summary, with no fundamental technical challenges as barriers, the Maine Department of Education is in a prime position to link data with other agencies such as the Department of Health and Human Services or the Maine Center for Disease Control and Prevention. While doing so offers a number of benefits children, educators, and policy makers, there nevertheless exist other non-technical matters that must also be considered when making this type of programmatic decision. The Maine Education Policy Research Institute is uniquely positioned to help in this arena should the State decide to go in this direction.

REFERENCES

Joint Committee on Infant Hearing. (2007). Year 2007 position statement: principles and guidelines for early hearing detection and intervention programs. Pediatrics, 120, 898–921.

Mason, C.A., and Tu, S. (2008). Data linkage using probabilistic decision rules: A primer. *Birth Defects Research, Part A: Clinical and Molecular Teratology*, *82*, 812-821.

Tu, S. & Mason, C. A. (2004). Organizing Population Data into Complex Family Pedigrees: Application of a Second-Order Data Linkage to State Birth Defects Registries. *Birth Defects Research Part A* (formerly, *Teratology*), *70*, 603-608.

Tu, S., Mason, C. A., & Song, Q. (2007). Second-order linkage and family datasets. In L. M. Glidden (Series Ed.), R. C. Urbano & R. M. Hodapp (Vol. Eds.), *International Review of Research in Mental Retardation: Vol. 33*. *Developmental Epidemiology of Mental Retardation and Developmental Disabilities* (pp. 53-78). San Diego, CA: Elsevier.