Secondary Use Data Project

Demonstration Project #5

Final Report

Date: 2018-11-02

Version: 1.0

With additional support from:

I. Document Versioning

Date	Document Version	Contributor	Description of Changes/Additions
2018-04-19	V0.1	Alberta Innovates	Initial Working Draft
2018-08-13	V0.2	PolicyWise	Draft for feedback
2018-11-02	V1.0	PolicyWise	Full report

Acknowledgements

Demonstration Project 5 could not have been possible without the support and contributions from Alberta Innovates, Alberta Health Services including the Maternal, Newborn, Child and Youth Strategic Clinical Network, and the Government of Alberta. The Women and Children's Health Research Institute, Stollery Children's Hospital Foundation, and the Lois Hole Hospital for Women also generously supported the project through the Maternal, Newborn, Child and Youth Strategic Clinical Network.

II. Table of Content

III. Executive Summary

The Secondary Use Data Project (SUDP) is a provincially led, multi-partner project to facilitate the enhanced and advanced use of secondary use health and social data for the health and socioeconomic benefit of Albertans. The SUDP initiative identified five demonstration projects that aim to examine Alberta's capabilities and challenges related to secondary data access and use; findings from the projects will be used to determine the tools and expertise required to further improve Alberta's use of secondary data. The five projects are being completed by health system partners (Alberta Health Services, Alberta Health, PolicyWise for Children and Families, Community and Social Services, Canadian Primary Care Sentinel Surveillance Network) along with Alberta Innovates as a coordinating agency. This is the final report of the SUDP Demonstration Project #5.

The Maternal, Youth, Children and Infant Linked Data Alberta Initiative (MyCHILD) is a collaboration between the Maternal, Newborn, Child and Youth Strategic Network of Alberta Health Services, Alberta Community and Social Services, PolicyWise for Children and Families (PolicyWise), and Alberta Innovates. The initiative is a Demonstration Project for the Secondary Use Data Project that aims to integrate health and health-related data to facilitate research and quality improvement efforts.

Of interest is the integration of data on children with neurodevelopmental disabilities, as well as those who experience complex medical conditions. Such children are served through many social services including health, children's services, disability services, and others. With better integration of data between these diverse sectors, novel insight can be gained into their experiences and outcomes, thereby informing improvement in care and efficiencies of service. Data for this project includes clinical data, service administration data, and where available, financial data from Alberta Health Services and the disability programs of Alberta Community and Social Services.

Overall, the goal of Demonstration Project #5 was to identify and address barriers to the linkage of data and facilitate access to researchers, clinicians, policy analysts and other stakeholders. MyCHILD would become a linked data repository that could be leveraged to conduct analysis on children with neurodevelopmental disabilities, complex medical conditions, and many other health and social challenges. The project was not able to reach the state of extraction and linkage of data between the public body partners, but many accomplishments were achieved, and lessons were learned.

Key accomplishments include:

- A Scientific Advisory Committee and a Policy Advisory Committee were struck to identify priority research and policy questions that could be addressed by integrating the linking of data from health and non-health sources. Two proposals were put forward to take advantage of the MyCHILD linked data.
- With secondary use and linkage of administrative data, concerns were raised regarding compliance to privacy legislation in Alberta. PolicyWise worked with privacy experts from Alberta Health Services and Alberta Community and Social Services to develop a draft privacy impact assessment to identify risks and propose mitigation strategies.
- PolicyWise also leveraged an in-house software linkage solution that would comply with the existing privacy frameworks in Alberta. This method uses hashed identifiers for linkage to avoid the need for disclosure of direct identifiers from the public bodies.
- A key requirement for the successful functioning of the MyCHILD repository was development of a governance structure that would oversee the use of the linked data. Although some discussions are ongoing, it was not possible to complete this within the project timeframe.
- During the course of Demonstration Project #5, significant changes occurred to some of the partnering organizations, as well as to the landscape of data integration in Alberta. Reorganizations contributed to

some delay and discontinuity in stakeholder representation. How this Demonstration Project fit within a broader picture of other data integration projects also represented a challenge to stakeholders.

Ultimately, accomplishments from this Demonstration Project can be leveraged in future data integration initiatives. There is a clear will in the public sector to take a cross-sector approach to data. The learnings presented here will help lower barriers to this approach.

IV. Project Background and Overview

Alberta currently pays more per capita for health care than any other Canadian province (except Newfoundland)[1], yet Albertans do not enjoy substantially better health outcomes than peer provinces[2]. At the same time, there is increasing recognition that health outcomes are determined by factors beyond health care such as early life events, life-style choices, family education, and income[3]. One of the goals for the Secondary Use Data Project (SUDP) is to integrate health and health-related data to facilitate research and quality improvement efforts supporting delivery of personalized medicine as well as enhanced social, educational, and justice services for Albertans. The SUDP is a provincially led, multi-partner project to facilitate the enhanced and advanced use of secondary use health and social data for the health and socioeconomic benefit of Albertans. There are five demonstration projects being completed by health system partners (Alberta Health Services, Alberta Health, PolicyWise for Children and Families, Community and Social Services, Canadian Primary Care Sentinel Surveillance Network) along with Alberta Innovates as a coordinating agency.

Children with neurodevelopmental disabilities (NDD), including those related to prenatal alcohol exposure (fetal alcohol spectrum disorder, FASD), attention deficit hyperactivity disorder (ADHD) and Autism Spectrum Disorder (ASD), receive supports through Children's Services, Health, Alberta Health Services, Community and Social Services, and other ministries. The prevalence of NDD's ranges from 10-15% and costs associated with government supports for these children is unknown. 'Medically complex' children are those with multi-system disease, have some degree of developmental delay, and/or have a physical disability that is technology dependent. Some medically complex children may also have an NDD. Medically complex children represent less than one percent of all children but consume about one-third of the total health care dollars spent on all children. Costs and outcomes associated with social service ministry supports for medically complex children are also unknown.

There is little information about how use of services in one area or Ministry may lead to a reduction of services in another, and the relationship between service use and child outcomes. For example, costs in Alberta Children's Services to address the needs of children with pervasive developmental disorder such as Autism Spectrum Disorder, may result in decreased use of emergency services with overall improvement in health and functioning among these children. It is important to explore the trajectory of service use among children with NDD's and medical complexity across sectors, and describe health outcomes and estimated costs.

The Maternal, Youth, Children and Infant Linked Data Alberta Initiative (MyCHILD) is a collaborative project led by the Maternal, Newborn, Child and Youth Strategic Network (MNCY SCN) with the primary purpose to improve the quality of care for children with complex health needs and the services provided to their families through enhancing data access, linkage, and analytics.

Demonstration Project #5 will incorporate the MyCHILD project as a use case for secondary use data integration, linkage and access for quality improvement. The project scope will focus on testing the linkage of health and health-related data, building on the experience of the partnership project that established the Child and Youth Data Lab. The stakeholders of this project will include SUDP, Alberta Health Services (AHS), Alberta Community and Social Services (CSS), PolicyWise for Children & Families (PolicyWise), and the MNCY SCN.

[1] https://www.cihi.ca/en/spending-and-health-workforce/spending/national-health-expenditure-trends/nhex2015-topic6
[2] http://www.conferenceboard.ca/hcp/provincial/health.aspx
[3] https://www.cma.ca/En/Pages/health-equity.aspx Canadian Medical Association

The project will work to develop data integration and linkage as well as access strategies for researchers in order to support the secondary use data needs of the MyCHILD project:

1. Pair quality and financial data[4] to enhance value-for-money in healthcare provided to Alberta's mothers-to-be and their children. This will include creation of a data asset accessible by clinicians, administrators, and researchers. With access to appropriate data, these groups will be able to accurately determine the health care resources consumed by each health problem they examine, as well as how much the related resources cost. Further, it will stimulate collaborations between front-line clinicians and researchers, allow rigorous evaluation of quality improvement projects, and allow projects to be carried out for much lower costs. Lastly, it will facilitate recruitment and retention of cutting-edge clinical and health service researchers.
2. Pair health and other public social sector data to not only better understand the effect of social determinants of health on health outcomes but also optimize public sector policies. This will allow more rigorous evaluations of existing Alberta health prevention and public sector strategies, and provide public policy collaboratives with data that supports the development of higher quality public sector initiatives.

Existing technologies, infrastructure, relationships, processes, and expertise from partner organizations were leveraged to strengthen the Demonstration Project:

- PolicyWise experience with linking provincial health and social determinant datasets including the relationships established with multiple government ministries.
- The Demonstration Project aligns well with the vision of the Secondary Analysis to Generate Evidence (SAGE) Initiative at PolicyWise. SAGE provides the mechanism, infrastructure, process, and expertise in working with all partners to link data and support the access and analysis of linked data.
- The MyCHILD project is being organized to facilitate data literacy and practice, policy evaluation/development, and improvement components that will use data integrated and linked as part of this Demonstration Project.
- In addition, in-kind support is being provided from Women and Children's Health Research Institute (WCHRI)/Stollery Children's Hospital Foundation and MyCHILD, to work on linkage of the quality/finance datasets and to coordinate stakeholder engagement, privacy impact assessment, and any governance activities required to facilitate integration and access to data assets.

Providing potentially identifiable data and linkage to accomplish these aims will require a privacy and legal review and assessment and related analysis of the need for health data de-identification.

Access to and use of the linked data will be an important feature of MyCHILD. Demonstration Project #5 provides the foundation for improved health and health-related data integration, linkage, and access. PolicyWise will provide the initial infrastructure and expertise to link the data and to house and support the use of the linked data. Where SUDP Demonstration Projects #1 – 4 are focused on health data, the outputs of Demonstration Project #5 will assist SUDP to document the required system components that will support of research and quality improvement efforts that make use of social determinants of health data. Development of a use case and

[4] The data being considered for inclusion in this project goes beyond health data to include other government ministry data e.g. Human Services; however, all non-health data is located within GoA i.e. no data sources external to government will be used.

roadmap will greatly improve the collective understanding of stakeholders regarding what is possible, what remains to be determined, and what we envision for secondary use data integration, linkage and access.

The approach for the project involved three broad streams of work (workplan in Appendix 2):

1. Legal/Privacy: to work with stakeholder organizations to devise a schema for linkage, analysis and reporting that would satisfy existing privacy laws and ethical considerations.
2. Scientific: to work with the MNCY SCN, CSS, and other information users to determine high priority questions to be answered with the most appropriate analytic methodologies, such as the trajectory of service use among children with NDD's and medical complexity across sectors, and describe health outcomes and estimated costs.
3. Data/Technical: to work with AHS and CSS to determine the technical details of data transfer, linkage and security.

This report focuses on PolicyWise's objectives for this project (detailed objectives and deliverables in Appendix 1):

- Determine MyCHILD requirements related to the integration, linkage, and access of health and health-related data.
- Identify and implement (as is feasible and reasonable) strategies to meet MyCHILD health and health-related data requirements gaps:
    - Recommend and initiate processes that support health and health-related data integration, linkage, and access activities.
    - Align data integration and linkage efforts with other secondary use data initiatives in the province, harnessing existing processes and assets as appropriate e.g. Provincial Health Analytics Network (PHAN) Portal, PHAN De-identification Project, Enterprise Data Analytics Platform (EDAP, Service Alberta), and SAGE.
    - Determine ongoing operational needs required to sustain the resulting data asset outside the preliminary scope of SUDP Phase 2 project[5].
    - Document the behaviors and experiences of secondary use data users as they work to initiate the integration, linkage, and access strategies.
    - Identify and document challenges and opportunities related to linking health and health-related data in Alberta for the purposes of research and quality improvement.

[5] Depending on the outputs and outcomes related to the project, the long-term goal of including the MyCHILD^{ALBERTA} data assets as part of an eventual secondary use data solution for Alberta may or may not be in scope for this Demonstration Project.

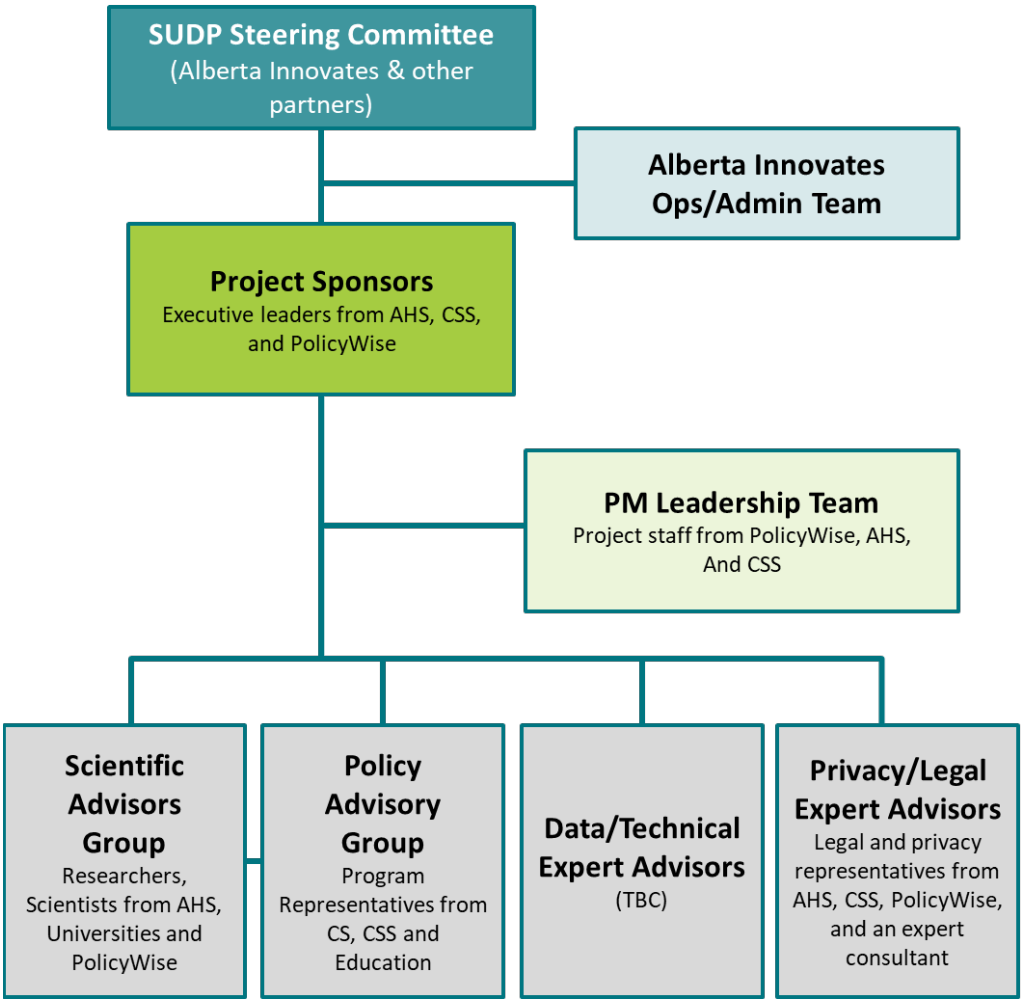
V. Project Results

A. Project Performance

Not all components of the project plan were implemented within the project timeframe. This final report will focus on describing the approach, documenting challenges and lessons, and outlining considerations for the future. This final report will focus on issues germane to the linkage, access and analysis of administrative data.

a. Project Governance

With the complexity inherent in linkage of sensitive data across multiple public agencies, creation and maintenance of proper governance to oversee the project was crucial to success. The following structure was devised:



Group	Function	Time Involvement (May 2018 to Dec 2018)
Project Sponsors:	High level decisions and corporate support	Minimal/Receive periodic updates
PM Leadership Team:	Overall management of the Project	Monthly Meetings
Scientific Advisory Group:	Provide content expertise and design/conduct analysis	Concentrated work upfront and when data is linked
Policy Advisory Group:	Provide policy context and set priorities for analysis	Two meetings/workshops and review of documents as needed
Data/Technical Expert Advisors:	Design and carry out processes for data transfer and linkage	Two month concentrated work (weekly involvement)
Privacy/Legal Expert Advisors:	Advise on privacy considerations of data linkage and sharing	Periodic involvement for drafting PIA, legal agreements and governance design

The Demonstration Project was sponsored by executive-level representation from AHS, CSS and PolicyWise, and reports up to the SUDP Steering Committee. The day-to-day coordination was performed by the PM Leadership Team, with assistance from the Alberta Innovates Ops/Admin Team. Four key advisory committees were planned to provide expert advice on various aspects of the Demonstration Project.

The Policy Committee consisted of leaders and policy staff in the Government of Alberta (GoA) Ministries; namely, Children`s Services, CSS, and Education. These represent key Ministry stakeholders in the support of children with disabilities. Although not all Ministries or programs are contributing data for the Demonstration Project, broad expertise is required to inform the direction of the analysis and evaluation of the data. The MNCY SCN provides perspectives on AHS delivery policies and fosters cross-sector discussions on policies that have broad impacts. The Policy Committee was asked to:

- Determine policy & program priorities;
- Develop relevant research questions;
- Identify sources of relevant data, including costing data;
- Identify and support engagement of organizational leadership;
- Identify outcome measures for children and families; and,
- Aid in assessment of impact and evaluation of MyCHILD.

The Policy Committee met as a group twice to agree on general parameters of the Demonstration Project. To limit complexity in governance, it was agreed that the initial focus should be on the CSS disability programs (Persons with Developmental Disabilities [PDD], Assured Income for the Severely Handicapped [AISH], and Family Support for Children with Disabilities [FSCD]). Involvement of Children's Services and Education programs would be important in the future to provide a full picture of services for children, but starting with one Ministry would be more feasible as a Demonstration Project.

The Scientific Committee consisted of members from AHS, MNCY SCN, and faculty from the Universities of Alberta, Toronto, and Calgary. This committee is made up of experts in their respective fields relating to children with disabilities, pediatrics, health economics and social health science. The committee would:

- Review and define algorithms to identify children with disabilities;
- Assess best practice models for cost evaluation of health and social service use; and,
- Provide guidance on analytic activities for MyCHILD.

The Scientific Committee met numerous times to discuss research questions, definitions of the population(s) of interest, and analytical methodologies. By September, 2018, this group is still active and will continue to conduct research in the areas of medical complexity and disabilities through other means.

A group of privacy and legal experts were engaged through AHS, CSS, the Office of the Information Privacy Commissioner (OIPC), and an expert consultant through Alberta Innovates. This group was not struck as a committee to streamline processes and limit the burden on participants. Advice from this group was critical in the progress made as described in the next section.

PolicyWise started preliminary discussions with data and technical experts in AHS and CSS to determine how data could be extracted. Some discussions were held with Alberta Health in testing the linkage methodology.

A memorandum of understanding was signed between AHS and PolicyWise in November, 2017. It laid out the following intent for the partnership between AHS and PolicyWise:

- AHS would share data with PolicyWise for the purposes of integration and linkage with other non-health data.
- AHS would disclose non-identifying information to PolicyWise.
- PolicyWise would help AHS anonymization of the data.
- It is understood that the de-identified data would be accessible with proper governance for research and evaluation.
- PolicyWise would implement appropriate safeguards to protect privacy and confidentiality of patients.

PolicyWise has worked extensively with GoA to develop governance and stewardship models for handling individual-level data. This has been demonstrated through implementation of the Child and Youth Data Laboratory (CYDL). PolicyWise will apply lessons learned from CYDL to guide activities of MyCHILD. Progress in determining appropriate governance for the access to linked data from MyCHILD yielded important lessons, but were not completed before September of 2018. These challenges are discussed below.

Enablers of success:

- MNCY SCN acting as a strong champion within AHS, and overall driver for the project.
- PolicyWise's experience in working with multiple ministries in the generation and mobilization of knowledge.
- Progressive AHS leadership receptive to liberating data.
- Project management and financial support from Alberta Innovates.
- Support from forward thinking staff at all partnering organizations.

Key accomplishments:

- Project governance designed and implemented with partnering organizations.
- Gained support for project from executive leadership from partner organizations, including the Deputy Minister of Human Services.
- Memorandum of understanding was signed between PolicyWise and AHS.
- Project teams, Scientific Committee and Policy Committee assembled and operational.

Key next steps (incomplete):

- Determine appropriate access governance structure acceptable to all stakeholders (see *Legal and Privacy* section below).
- Determine appropriate quality assurance processes.
- Critical resources:
    - Engaged leadership and program staff at all stakeholder organizations will be required to maintain momentum. Representatives must also have clear roles, objectives, and mandate within the project.

b. Legal and Privacy

In general, three key pieces of legislation to consider in the Demonstration Project are the Health Information Act (HIA), the Freedom of Information and Protection of Privacy (FOIP) Act and the Children First Act (CFA). In this case, AHS is an HIA custodian, and CSS is a FOIP body. PolicyWise is a non-profit entity whose activity, in general, does not explicitly fall under either pieces of legislation. However, PolicyWise has been designated by the CFA to have a special role:

> *As of 2013, through the Children First Act, Alberta Government departments are authorized to share anonymous health and social data with PolicyWise to facilitate research to support, (i) the development of effective programs and services for children; (ii) the integration of policies affecting children; and (iii) the co-ordination of programs and services for children.*

In discussion between AHS, CSS and the OIPC of Alberta, it was determined that the best way to ensure that privacy risks are mitigated to the satisfaction of all stakeholders is to create and submit a Privacy Impact Assessment (PIA) for review by the OIPC. PolicyWise led development of a draft PIA in collaboration with an expert consultant from Alberta Innovates. A more detailed discussion of how privacy legislation applies to similar projects are available in a separate report: *Post-Initiative Data Integration, Linkage, and Access Assessment & Gap Analysis*.

A key innovation for mitigating risk was to employ Privacy Preserving Record Linkage (PPRL). Using this methodology, identifying data would be encrypted with a one-way hashing algorithm at AHS and CSS before disclosure to PolicyWise. Therefore, data shared between partner organizations will not include direct identifiers such as names, full dates of birth, phone numbers or addresses. The method of encryption will make it nearly impossible to back-compute identifying information. PPRL will allow linkage of individuals with such hashed information in a probabilistic manner from the hashed output.

In the context of the HIA, this activity is interpreted as fitting under Section 32(1), disclosure of non-identifying data for any purpose. As the subsequent linkage of data will not create individually identifying health information, it is interpreted that this activity does not fit within the definition in Section 1(1)g for data matching. In the context of the FOIP Act, the data disclosed by a FOIP body would not readily support re-identification. In the context of any Ministries covered under the CFA, this disclosure is also consistent with Section 5(2).

The linked dataset would be housed at PolicyWise inside the same environment as the CYDL, which is already entrusted by the GoA to hold anonymous linked datasets. Integration and linkage of data would occur on an ongoing basis so that data is kept up to date, ensuring relevance to current policy and practice. PolicyWise will administer access to this data for research and evaluation purposes by staff of public bodies and other researchers.

The specifics of the request process and access governance were in discussion at the end of the Demonstration Project. It was contemplated that:

1. Requests to access project-specific extracts of the linked data will be made to PolicyWise.
2. A multi-stakeholder Data Stewardship Committee will review the proposals to determine factors such as feasibility to answer the proposed question, scientific merit, and protection of privacy.
3. Research proposals would need to be reviewed by a Research Ethics Board in accordance with HIA or institutional requirements where applicable.
4. Upon approval by the Data Stewardship Committee, a Data Access Agreement would be executed between the requestor and PolicyWise.
5. PolicyWise staff would create a data extract of the data elements approved for access by the Data Stewardship Committee and Research Ethics Board (where appropriate) and assess for re-identification risk.
6. This data extract would be made available through the remote desktop infrastructure at PolicyWise. Analytic software is provided inside the isolated desktop.
7. Results, such as graphs, tables and figures are manually reviewed by PolicyWise staff to assess re-identification risk prior to release to the requestor.
8. Access may be provided in a cost-recovery model to ensure sustainability of the resource long-term.

The Data Stewardship Committee would be composed of representatives from PolicyWise and the public bodies involved in the data request. How the committee would operate, criteria it would use to assess requests, and level of AHS and CSS staff involvement were yet to be determined.

Enablers of success:

- PolicyWise's experience dealing with data privacy, especially with linked data.
- Early and ongoing engagement with the OIPC.
- Willingness of privacy staff in AHS and CSS to participate.
- Availability of expert consultant from Alberta Innovates.

Key accomplishments:

- Development of draft PIA.
- Buy-in from AHS of overall information flow and linkage process.

Key next steps (incomplete):

- Submit PIA to OIPC for review and feedback.
- Execution of Data Stewardship Agreements between AHS – PolicyWise, and CSS – PolicyWise.
- Critical resources:
    - Legal and privacy experts are critical to the success of this type of project. Such experts from each stakeholder organization must also work well together to ensure consistent interpretation of relevant legislation and policies.

c. Scientific

MyCHILD aims to create an analytic data sharing platform that will enable access and use by policymakers, administrators, researchers and practitioners to investigate relevant questions using the data relating to children with disabilities. The intent of MyCHILD is to gain insight into the trajectories of care for Albertan children and

particularly children with NDD's and medical complexity. This project would also inform future work that may examine outcomes for children with other disabilities or conditions. Data linked across these sectors will allow for the exploration of a) profiles of care and service use across sectors and b) assessment of how costs associated with those patterns of service use are related to health and social service outcomes. Data from healthy children in Alberta requiring less resource-intensive and coordinated services will also be collected so as to create a baseline from which to compare children with disabilities.

Access, analysis and interpretation of data from the MyCHILD data sharing platform would be undertaken in collaboration with service providers, decision makers and academics to ensure results are interpreted within the current service environment. Key messages derived from the data will facilitate effective knowledge translation and uptake. The resulting information will enable policymakers to develop relevant and up-to-date policies that improve the lives of children with disabilities, and their families and caregivers. Additionally, the service programs will be provided with information from MyCHILD and encouraged to dialogue about the potential relevance to their patterns of care, and to use this to inform decisions on resource allocation to meet the needs of children with disabilities.

The Scientific Advisory committee met numerous times during the project to discuss potential avenues of investigation. The goal was to prepare analyses that would take advantage of the linked dataset. Some parameters were discussed, such as years of data necessary to build a relevant longitudinal dataset, ages of children and youth to include, and datasets to include. However, many of these discussions were tentative, awaiting resolution of governance and privacy issues.

The intent was to include data from PDD, AISH, and FSCD from CSS. From AHS, the intent was to include data from physician claims, the Discharge Abstract Database, the National Ambulatory Care Reporting System, diagnostic imaging, laboratory, Pharmaceutical Information Network, health registry and vital statistics.

Broadly, other preliminary analysis questions included:

- What are the characteristics of those with NDD's and/or medical complexity?
- What is the trajectory of service use across sectors for those who have one or more NDD's or experience medical complexity? Can these trajectories be evaluated during transition from paediatric to adult care?
- What are the total (and average per person) costs of government services for those who have one or more NDD's or experience medical complexity? What are these costs per annum and over time? Would increases in costs in one service area be associated with cost change in other services areas?
- What proportion of children with NDD's or medical complexity is accessing the supports for themselves and their families as defined by program eligibility criteria? What are the characteristics of those who are not accessing available supports and services?
- Are there regional variations in access to services, support and outcomes?
- What are the gaps in data that limit the assembly of comprehensive and meaningful profiles?
- Is it possible to examine child to adult trajectories in terms of service utilization and outcomes?

Two research proposals detailing specific questions and rationale are included in Appendix 3. These first projects were designed to take advantage of the linked MyCHILD administrative data to address gaps in evidence that are difficult or costly to fill through primary data collection.

Enablers of success:

- Involvement of academic researchers and clinicians who have a clear use case for the linked data.

- Involvement of policy and program staff from multiple ministries that can speak to how results can impact coordination of services.
- MNCY SCN as a multi-stakeholder driver of innovation.

Key accomplishments:

- Significant work to define populations of interest using administrative data.
- Development of plans for initial analyses.

Key next steps (incomplete):

- Finalize definition of population(s) of interest. Determine methodology for analyses.
- Determine methodology for economic analyses.
- Conduct analyses.
- Knowledge mobilization (clinicians, policy makers, etc.), which may include internal reports, peer-reviewed publications or other learning activities.
- Critical resources:
    - Research, clinical and policy personnel are important to the analysis phase of the project. The primary goal for MyCHILD is to improve outcomes of children. This requires a cross-disciplinary approach to addressing knowledge gaps and mobilizing knowledge. As familiarity with administrative data was important, it was proposed that analyst staff at PolicyWise, AHS and CSS would be involved with the analysis of the data and supporting researchers, clinical and policy staff.

d. Data and Technical

Learning from the CYDL example, there would be significant technical work for the Demonstration Project. Assuming governance and legal issues are addressed, technical steps include:

1. Specific data elements, datasets and their business owners must be defined. These would be driven by the questions to be asked, and stipulated in the agreements between AHS, CSS and PolicyWise.
2. Data integration within organizations. For example, data in AHS is housed in many data systems across the organization. Before extraction, these must be linked internally.
3. Data to be anonymized using PolicyWise-provided PPRL methodology.
4. De-identified data to be extracted and transferred from AHS and CSS.
5. De-identified data to be linked using PPRL methodology.
6. Conduct quality control steps such as linkage quality assessment. Data cleaning where necessary.
7. Data workbook and metadata to be developed.
8. Setup the access infrastructure within the SAGE system at PolicyWise.

The MNCY SCN led significant work to link data within AHS. The data required for the project exists across several different data systems within AHS. An analyst within the MNCY SCN linked these datasets together in preparation for transfer to PolicyWise. This is an important achievement as this AHS data is still of great utility without the linkage of non-health data.

Significant progress was made in adapting the PPRL methodology. In collaboration with Curtin University in Australia, PolicyWise developed a novel PPRL system which uses a specific algorithm to hash identifying information that can be subsequently matched. Individual level data are available to answer population level questions; however, only aggregate level data will be reported.

LinkWise is a PPRL software solution created by PolicyWise to link population-level data. It follows a probabilistic data linkage model and supports clear-text and PPRL. LinkWise was built to be very user-friendly. The software pre-processes and cleans data, splits the identifiers from service-related fields, and calculates all required linkage parameters automatically. It uses published Bloom filter algorithms to conduct PPRL hashing and linkage. Most critically for the Demonstration Project and beyond, it allows incremental linkage, whereby new data can be added to existing data. This is important for long-term updates of a linked dataset. It is able to handle millions of records and distribute load across multiple processors to reduce run time. In testing with mock data against comparator software, LinkWise was able to provide high specificity and sensitivity in linkage accuracy (Appendix 4).

The software consists of two parts: the client side and the server side. The client side software would run at AHS or CSS and conduct the one-way hashing of identifiable data. First, it splits each dataset into two files: the research file and the linkage file. The research file contains all content fields (e.g. visit dates, service rendered, outcomes), and the linkage file contains all identifiers (e.g. date of birth, gender, address). The hashed linkage files generated by the client side of the software are then submitted to the server side to generate linkage keys. Linkage keys are then used to link research files between different datasets, AHS and CSS in this case. The server side software is maintained at PolicyWise.

Once data is linked, it was anticipated that the remote research environment at PolicyWise would be used to facilitate access of the linked datasets for researchers, AHS, CSS, and others. The SAGE Analytic Environment is a virtual desktop system which allows secured, remote access to data and relevant analytical tools. The individual desktops have no internet access. Specific data extracts containing only elements approved by the Data Stewardship Committee and research ethics board (where applicable) are placed on the remote desktop so that no direct access to the whole linked dataset is possible. Access is secured through industry-standard technologies include two-factor authentication and Secure Sockets Layer (SSL) encryption. To protect participant privacy and confidentiality, data cannot leave the SAGE Analytic Environment, and outputs are reviewed by SAGE staff prior to export.

Enablers of success:

- PolicyWise's expertise and experience in data linkage and analysis of administrative data.
- Significant capacity within AHS to link and manage health data.

Key accomplishments:

- Significant progress in linkage of datasets within AHS.
- Testing of novel linkage software for the MyCHILD scenario.

Key next steps (incomplete):

- Further testing and validation of linkage method with AHS and CSS.
- Identification of specific data elements, databases, and business owners.
- Data integration within organizations.
- Data anonymization, transfer, and linkage.
- Quality control and data documentation.
- Deploy remote desktop infrastructure to facilitate access.
- Critical resources:

- To move forward with the project, data management staff at each organization must be engaged to determine parameters around the extraction, transfer and linkage of data. Such staff would also be involved in the quality assurance processes.
- For linkage, it was tentatively decided that LinkWise would be the solution. PolicyWise maintains the expertise around the operation of this software. No specialized hardware is required beyond what stakeholder organizations already have.
- For data access, it was tentatively decided that the SAGE remote research environment would be used for the pilot phase of MyCHILD. PolicyWise maintains the expertise for the ongoing operation of this infrastructure. This infrastructure may need to be expanded to accommodate additional users from the MyCHILD project.

B. Challenges and Lessons

a. Stakeholder Reorganization

There was significant momentum for the Demonstration Project in late 2016. Under MNCY SCN leadership, buy-in within AHS was already well-established by this point, but more engagement of the social service ministry was required. A high-level meeting was held with the Deputy Minister of Human Services, along with several Assistant Deputy Ministers and leaders with the Ministry in November of 2016. The Deputy Minister supported the work and discussions began on project governance and next steps.

In January of 2017, Human Services began a significant reorganization into the Ministries of Children's Services and Ministry of CSS. Many organizational changes were made to create the two new Ministries. Executive leadership, including Deputy Ministers also changed during this period. Time was required to determine where previous champions and responsible business units landed within the new structure and to re-engage.

Lessons learned:

- Continuity in leadership support is critical for a project that requires multiple stakeholders to be aligned on a long-term basis. While large-scale reorganizations are difficult to predict, one way to mitigate disruption is to develop leadership support at multiple levels and to ensure the project is well-aligned to broad organizational goals. It was difficult for the Demonstration Project as changes occurred at both the Deputy Minister and Assistant Deputy Minister levels during the split.
- Long-term initiatives will often need to deal with leadership changes. This requires developing new relationships with new leaders. There is also a need to convey the value proposition of the initiative within a background of many initiatives that are also jockeying for support from new executive leadership.
- A Memorandum of Understanding with CSS may have helped to maintain continuity by enshrining the original intent to collaborate.

b. Related Data Integration Initiatives

When conveying the value of an initiative to stakeholders, it is also important that its place within the big picture be clear. Concurrent data-related initiatives, each with their own brands, caused confusion within stakeholder organizations.

- Alberta Innovates was leading the SUDP initiative. It was not well-understood that this Demonstration Project was a part of the broader SUDP initiative, and for stakeholders not familiar with SUDP, it represented another order of complexity in positioning where the Demonstration Project sits within the bigger picture. It was also branded as MyCHILD during discussions.

- Service Alberta was working on the EDAP, which may be a central repository for all GoA data. This was perceived as a competing interest by some stakeholders. As both EDAP and the Demonstration Project/SUDP had a vision of data integration, there was confusion over the subtle differences in the goals between the two. Some also felt they were conflicting priorities; one coming from within the GoA and one coming from an external source. For some, it was difficult to reconcile the two and how limited resources should be spent.
- PolicyWise were also in discussion with the GoA about other data-related work. PolicyWise was in discussion with partnering Ministries for the refresh/expansion of CYDL data. While SAGE did not directly involve government, it was another data-related initiative that was discussed with GoA staff, especially as it pertained to privacy issues.
- PolicyWise was also supporting the FASD Data Integration project, which involved many of the same Ministries as the CYDL and the Demonstration Project. This project is aimed at integrated data from FASD-serving agencies. It further muddied the water when the FASD Data Integration project was designated as a demonstration project for the EDAP.

Towards the summer of 2018, momentum was building for EDAP and it was decided that the CYDL-related functions would be brought internal to government to focus resources on the EDAP. This internal focus in the GoA further limited opportunities to advance other data integration initiatives that were perceived to be overlapping.

Lessons learned:

- Messaging must be clear on how an initiative is related to other, similar initiatives. Even if they are distinct, it is often difficult for stakeholders to understand the subtle differences between highly technical projects. At a high level, they all fall under the theme of data integration, which became a crowded area. The value of each must also be clear. Related to the above, this is critically important when eliciting support from executive leaders new to the area/project.
- Related initiatives need to be well-connected through communication and collaboration. The only constant is change and projects and priorities inevitably evolve over time. It is important that initiatives evolve with stakeholder needs so that they are not perceived as redundant, competing or conflicting.

c. Overall Staff Changes

The project experienced some staff changes that required time to rebuild momentum and continuity. As indicated in the original proposal from PolicyWise to Alberta Innovates, a Project Lead was budgeted for and hired. Discontinuity in this role led to disruption in day-to-day operations of the project in early 2017. Changes with privacy experts within the stakeholder organizations also occurred. For AHS, the first privacy expert engaged retired, and the subsequent one was seconded to work on Connect Care. For CSS, the first privacy expert was seconded to another Ministry. As privacy issues were key to the project and the plans to mitigate risk was detailed and nuanced, these changes led to some delay in getting new representatives up to speed. More broadly, staff transitions following the Human Services reorganization also represented a challenge.

Along with staff change is sometimes uncertainty over specific responsible personnel at each stakeholder organization. In some instances, responsibility for the project was not well-articulated between leadership and program staff. Some staff felt they had to work on the project as a side-of-desk project, having no official executive mandate.

Lessons learned:

- Momentum must be fostered in close collaboration between all stakeholders, and at all levels of the organizational hierarchy. Change is inevitable so multiple champions at each stakeholder must be engaged to carry forward the vision of the project even through organizational upheaval.

VI. Conclusion

Increased capacity in the public sector to collect and analyze data has opened up new avenues of how such organizations can advance practice and improve outcomes. In the social sector, healthcare has been a leader in leveraging such capabilities to improve the lives of patients. However, with the recognition that health is only a part of overall well-being, it has become necessary to take more cross-disciplinary approaches to the collection and analysis of data. This Demonstration Project was an effort to probe how this can be accomplished within the current legislative and policy context in Alberta.

While this integration was not completed within the timeframe of the Demonstration Project, many accomplishments were achieved, and lessons were learned. In bringing the leadership of the partnering organizations together, it was recognized that the will to integrate data was strong within the public sector. While multiple concurrent data integration projects being in-progress across the partnering organizations caused some conflict, it showed that data integration was a priority for the public sector and resources were being dedicated to such projects. Through this project, questions regarding how data can be integrated in a way that complies with privacy legislation were raised and explored. While open questions regarding privacy remain, this discussion will no doubt be useful to the continuation of this and other data integration projects. Methodology was also developed to address privacy issues. The availability of a PPRL solution will be of great utility to any linkage project moving forward.

In terms of change management, a key lesson of the Demonstration Project was the importance of continuity at all levels of the stakeholder organizations. Bold projects require bold leaders to drive them and support from staff. Such work rarely succeeds as side-of-desk projects. Clear resources need to be committed, with strong mandates supporting them. Another key success factor for the future is clear and transparent coordination between initiatives that aim to improve data integration across the public service. While each has its role, this must be clearly articulated and understood by others involved. This way, such work can be aligned and focused towards a common goal.

The knowledge, processes and technology leveraged as part of this Demonstration Project will inform future efforts to integrated health and non-health data. This Demonstration Project has moved stakeholders one step closer to realizing the benefits to well-being and system efficiency that comes from a cross-sector approach to data. While the data integration contemplated was not achieved by end of the project timeframe, it has stimulated collaboration between stakeholders and will fuel the desire for other data integration projects.

VII. Glossary

ADHD	Attention deficit hyperactivity disorder
AHS	Alberta Health Services
AISH	Assured Income for the Severely Handicapped
ASD	Autism Spectrum Disorder
CFA	Children First Act
CSS	Alberta Community and Social Services
CYDL	Child and Youth Data Laboratory
EDAP	Enterprise Data Analytics Platform
FASD	Fetal alcohol spectrum disorder
FOIP	Freedom of Information and Protection of Privacy Act
FSCD	Family Support for Children with Disabilities
GoA	Government of Alberta
HIA	Health Information Act
MNCY SCN	Maternal, Newborn, Child and Youth Strategic Clinical Network
MyCHILD	Maternal, Youth, Children and Infant Linked Data Alberta Initiative
NDD	Neurodevelopmental disabilities
OIPC	Office of the Information and Privacy Commissioner
PDD	Persons with Developmental Disabilities
PHAN	Provincial Health Analytics Network
PIA	Privacy Impact Assessment
PPRL	Privacy Preserving Record Linkage
SAGE	Secondary Analysis to Generate Evidence
SSL	Secure Sockets Layer
SUDP	Secondary Use Data Project
WCHRI	Women and Children's Health Research Institute

VIII. Appendix 1: Detailed Project Goals, Objectives and Deliverables

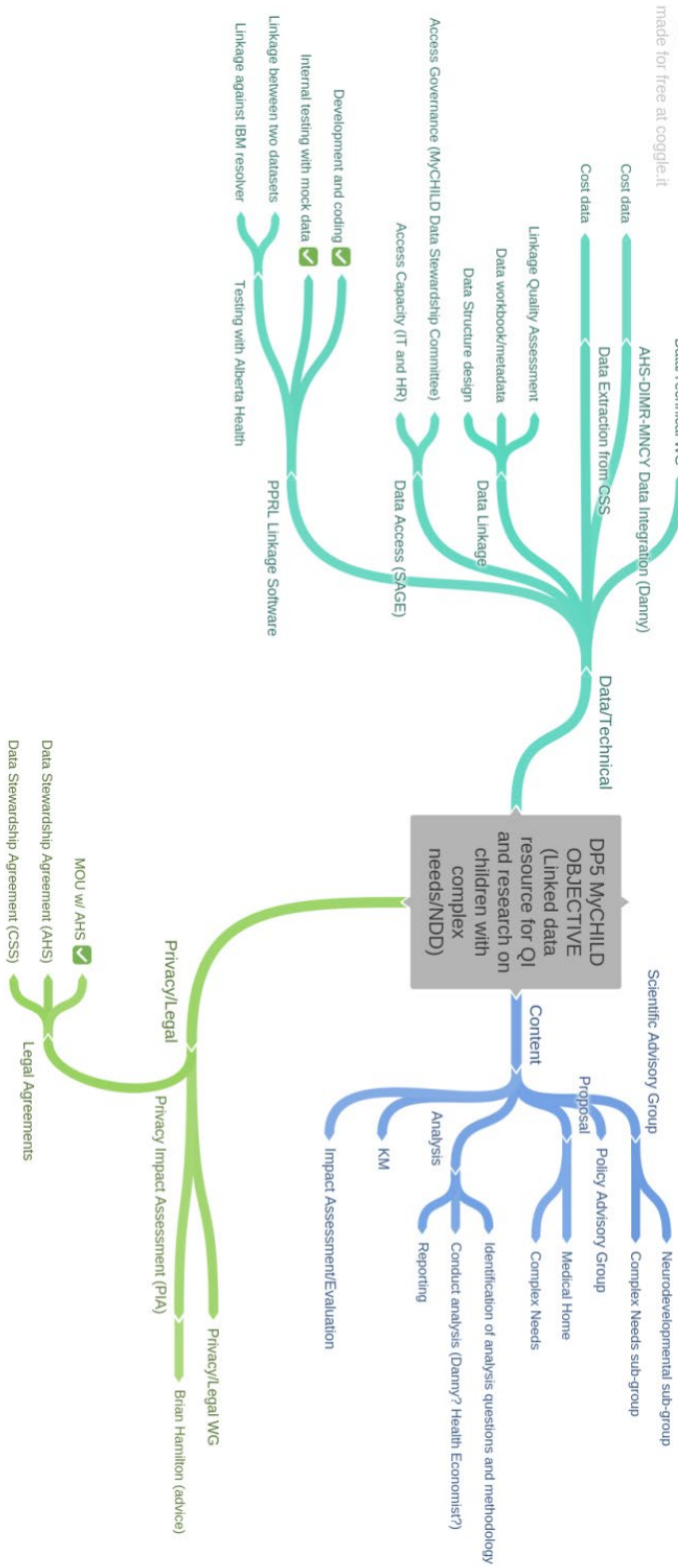
- Conduct a review of initiatives previously completed and/or underway that link health and health-related data identifying and documenting data integration, linkage, and access successes and challenges.
- Determine *MyCHILD* requirements related to the integration, linkage, and access of health and health-related data.
- Complete gap analysis for *MyCHILD* health and health-related data requirements
- Identify and implement (as is feasible and reasonable) strategies to meet *MyCHILD* health and health-related data requirements gaps:
    - Recommend and initiate processes that support health and health-related data integration, linkage, and access activities.
    - Align data integration and linkage efforts with other secondary use data initiatives in the province, harnessing existing processes and assets as appropriate e.g. Provincial Health Analytics Network (PHAN) Portal, PHAN De-identification Project, and SAGE.
    - Determine ongoing operational needs required to sustain the resulting data asset outside the preliminary scope of SUDP Phase 2 project.
    - Document the behaviors and experiences of secondary use data users as they work to initiate the integration, linkage, and access strategies.
    - Identify and document challenges and opportunities related to linking health and health-related data in Alberta for the purposes of research and quality improvement.
- Provide a roadmap and recommendations for enhanced and advanced data integration, linkage, and access based on Demonstration Project findings:
    - Identify and document the resources, processes, skills, and infrastructure required for data access, integration, and linkage as well as analytics support for both longitudinal research efforts and continuous reporting (quality improvement and quality assurance activities that support Alberta government programs and services).
    - Identify, document, and validate an efficient and scalable methodology for linking health and health-related data in Alberta.
    - Investigate and evaluate methods for updating data sets on an ongoing basis once integration and linkage have occurred.
- Determine end user requirements for a portal that provides self-serve access to health data for researchers
- As appropriate, implement a portal for self-serve access to health data for researchers

As appropriate, create awareness and build relationships with non-Health government departments regarding the benefits of integrating health and health-related data, including relationships with business intelligence programs and open data/open government initiatives.

Deliverable	Description	Notes
Work Plan	Includes the work breakdown structure and estimated hours per resource assigned to the project. Identifies the main dependencies between critical tasks and strategies for remaining on course given the project's short time period. The work plan will include activities that ensure regular status reporting is provided to the SUDP Project Team	Appendix 2
Demonstration Project Evaluation Plan	An overall Evaluation Framework guides the SUDP; the Demonstration Project Evaluation plan will contribute standard evaluation data and findings as required by the Framework. In addition, the Demonstration Project Evaluation Plan identifies methods, measures, and processes specific to the Demonstration Project's context, scope, and objectives.	Evaluation deemed not necessary at the current time as project did not reach completion.
Organizational Change Management Plan	An overall Organizational Change Management Plan guides the SUDP.	Deemed not necessary as SUDP does not have immediate plans to implement, however, suggested next steps are in this report.
Final Report	This document will provide an overview of the Demonstration Project activities.	This report.
Post-Initiative Data Integration, Linkage, and Access Assessment	This document will include information regarding the past experiences of provincial and local initiatives as they relate to integrating, linking, and accessing health and health-related data. The document will include a list of findings and preliminary recommendations for use in Demonstration Project #5.	In separate report: Post-Initiative Data Integration, Linkage, and Access Assessment & Gap Analysis
MyCHILD Integration, Linkage, and Access Requirements	This deliverable will document the detailed health and health-related data integration, linkage, and access needs of the MyCHILD project.	Requirements partially defined, and included in final report, Appendix 5.
Gap Analysis	This deliverable will document the gaps facing the MyCHILD project as it seeks to integrate, link, and access the data assets it requires for analysis. Special attention will be paid to the long term/ongoing access and use of the data assets once they are linked and integrated.	In separate report: Post-Initiative Data Integration, Linkage, and Access Assessment & Gap Analysis
Self-Serve Portal Requirements	This document will include a detailed description of both business and technical requirements for the self-serve secondary use data portal in the form of a table.	Project did not progress far enough to determine whether a self-serve portal was feasible or appropriate. A remote research environment (SAGE) at PolicyWise was contemplated to be

Deliverable	Description	Notes
		the near-term technical solution to access.
Privacy Impact Assessment	Work to complete PIAs and related data sharing agreements will be documented and provided to Alberta Innovates at project end.	Draft included in final report.
Self-serve Secondary Use Data Portal	The data portal that provides self-serve access of health and health-related data for researchers.	Project did not reach implementation stage.

IX. Appendix 2: Workplan



MyCHILD Workplan	COMPLETED	May	June	July	August	Sept	October	November	December	January & beyond
Data/Technical										
Data/Technical WG										
AHS-DIMR-MNCY Data Integration (Danny)										
Data Extraction from CSS										
Data Linkage										
Data Access (SAGE)										
Access Governance (MyCHILD Data Stewardship Committee)										
PPRL Linkage Software										
Content										
Scientific Advisory Group										
Policy Advisory Group										
Proposal										
Analysis										
KM										
Impact Assessment/Evaluation										
Privacy/Legal										
Privacy/Legal WG										
Privacy Impact Assessment (PIA) Development										
PIA submission										
Legal Agreements										

X. Appendix 3: Proposals for initial analyses on the MyCHILD linked dataset

Proposed Project #1

Objectives: The overarching objective is to describe cross-ministerial trajectories of service utilization for children and youth with disabilities in Alberta. Specific objectives include a description of timing, frequency and intensity of services accessed, patterns of utilization of physician and hospital services, Family Support for Children with Disabilities (family-focussed, child-focussed and specialized services), and access to Alberta Income for the Severely Handicapped (AISH) and Persons with Developmental Disabilities (PDD) funding. Access to services and funding will be analyzed by specific diagnostic groups and by region to determine population and regional variations in the extent and timing of access to the various services.

Sample: All Albertans aged 0-24 years of age between the years of 2004 and 2016 will be selected for possible inclusion. The sub sample of children with medical complexity will be identified using the approach outlined by Cohen et al that utilizes a well-defined set of ICD10 codes which identifies children with neurologic impairment, complex chronic conditions and/or medical technology assistance (Berry et al. 2015, Cohen et al 2012).

Outcomes of interest:
Stratified by Zone, by urban/regional/rural, by First Nations status, by SES (est. by Census DA/Pampalon Index)

1. *Acute health care utilization* (number of emergency department visits, hospitalized days, ICU admissions and total days) by children and young adults with medical complexity

2. *Evidence of 'medical home'*
   2.1. The average or median number of visits to this 'coordinating' physician, i.e. the physician most commonly visited?

   2.2. The proportion of children with a 'coordinating' physicians remaining constant over the study period.

   2.3. The ratio of visits to the 'coordinating' health care provider relative to the total number of visits to <u>all</u> providers involved in these children's care?

3. *Optimal coordination of visits and services* – The average number of health care encounters per week? [To determine the numerator, each physician and allied health visit, laboratory test, diagnostic image test, visit to the pharmacy and any other health care services will count as 'one' health care encounter.]

4. *Engagement with Social Services (Social services defined through AISH, FSCD, PDD, child support services, and child care subsidy)*

   4.1. Proportion of medically complex children accessing social services and specific types of social services

   4.2. Timing of social services engagement potentially classified as early or late (defined by timing from initial diagnosis and first service utilization)

4.3. Relationship between amount of engagement (numbers of encounters and health outcomes (as defined by service utilization ICD codes for acute care, emergency and primary care physician services).

4.4. Relationship between early and late engagement with social services and health outcomes (as defined by service utilization ICD codes for acute care, emergency and primary care physician services).

5. Costs

5.1. AHS/acute care costs including hospitalizations, emergency department, urgent care visits (AHS only), day surgery, medical day procedures; for hospitalizations, estimate costs utilizing DRG and RIW available for urban and regional hospitals

5.2. Ambulatory Care costs including physician office visits, outpatient diagnostic imaging, outpatient laboratory, outpatient rehabilitation, public health visits (including childhood vaccinations, public health nurses visits, etc.)

5.3. Social Services costs based on services accessed and availability of costing data

Analysis: Descriptive statistics will be calculated stratified by Zone, by urban/regional/rural, by First Nations status, by SES (est. by Census DA/Pampalon Index) and will include: numbers, proportions, and ratios based on outcomes of interest. Differences in means and proportions will be analysed using regression analysis with dependent variables determining regression approach used for analysis. Time to event (survival analysis) will be considered to examine the relationship between diagnosis and access to services.

Implications: Information about service utilization and integration or lack thereof across Health and Social Services Systems as well as timing of access to various services and supports is essential to identify gaps and opportunities for improvement that will provide decision makers with data to inform more coordinated and cost efficient cross-ministerial services in Alberta.

Proposed Project #2

The Journey Through the Alberta Health and Social Services Systems for Children and Youth with Disabilities in Alberta

Objectives: The overarching objective is to describe cross-ministerial trajectories of service utilization for children and youth with disabilities in Alberta. Specific objectives include a description of timing, frequency and intensity of services accessed, patterns of utilization of physician and hospital services, Family Support for Children with Disabilities (family-focussed, child-focussed and specialized services), and access to Alberta Income for the Severely Handicapped (AISH) and Persons with Developmental Disabilities (PDD) funding. Access to services and funding will be analyzed by specific diagnostic groups and by region to determine population and regional variations in the extent and timing of access to the various services.

Sample: All Albertans aged 0-18 years of age within the two-year period of 2004-2005 will be selected for possible inclusion. Individuals with disabilities will be identified using a validated algorithm based on International Classification of Disease – 9th edition (ICD-9) (Chien et al., 2015). ICD-9 codes submitted with physician claims and ambulatory care data will be used to identify the sample and individual data will be matched to individual FSCD/AISH and PDD data.

Outcomes of interest: Proposed outcomes of interest are: extent and timing of access to physician and hospital services, CSS services and supports, variations in extent and timing of access by region and diagnostic group (including cost analysis).

Analysis:. Descriptive statistics will be used and will include 1) time from diagnosis, or first use of relevant ICD-9 code to initial access to FSCD services, and 2) frequency of timing and extent services are accessed by region and diagnostic group. ANOVAs will be used to analyze differences in mean service and resource utilization by region and diagnostic groups.

Implications: Information about service utilization across Health and Social Services Systems as well as timing of access to various services and supports is relevant for coordination of cross-ministerial services in Alberta.

XI. Appendix 4: Evaluating LinXmart and LinkWise using Alberta Health scrambled data

Introduction

To unleash the power of data to make evidence-based decisions, linking data (record matching) is a critical step. Linking two datasets means matching records from the first dataset to the second dataset so that the matched records indicate the same individual. Although analysing each separate dataset can be useful for policy making, however linking independent datasets provides us with an incredible amount of information which can be used for evidence-based policy making. For example, by linking health data to education data the government is able to reveal various health-related problems among students in various schools and help students through some new policies etc.

Data linkage is a difficult problem because in most real-world datasets there is no a unique identifier (like personal health number, SIN numbers etc.) shared among various datasets. As a result, using identifiers like Name, Address and Date of Birth is the only option to link datasets. Linking data using those identifiers can be challenging because datasets are collected using different standards, they might be noisy, and contain error.

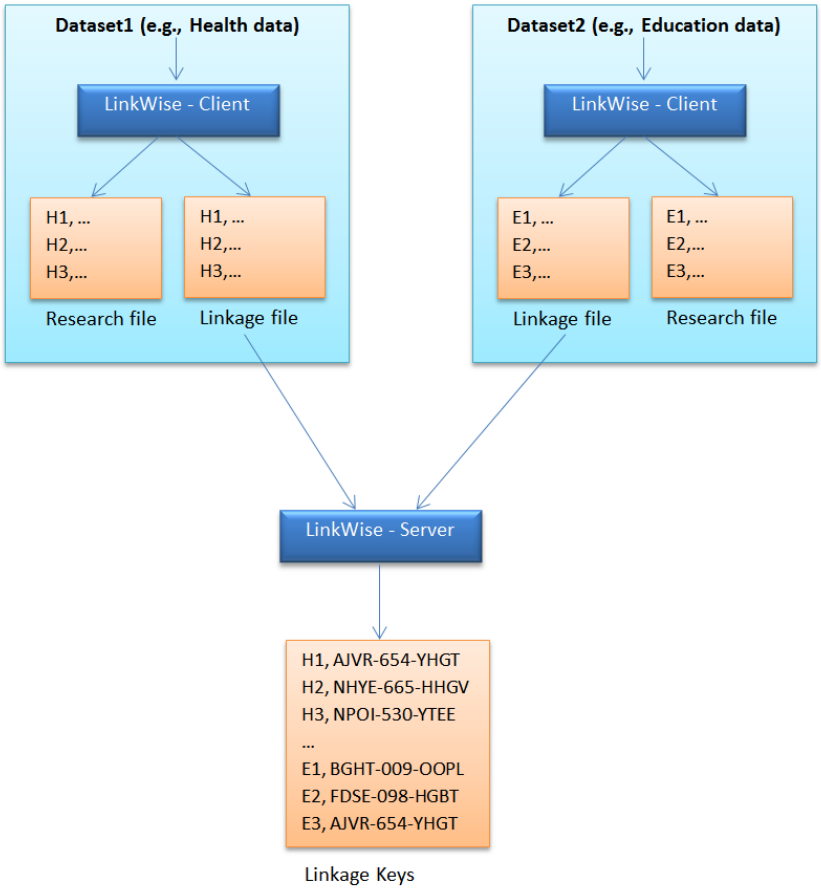
Techniques applied for record matching can be divided into two categories: deterministic methods and probabilistic methods. Deterministic linkage technique is the simplest way to match records. It considers two records as match if all or a subset of their identifier is identical. Deterministic record linkage is a good option when there is no error in data which is not the case in real-world datasets. On the other hand, probabilistic record linkage is a more complicated technique. It assigns a weight to each identifier based on its ability to identify a match or non-match. By comparing the identifiers, the technique calculates a matching weight and if the matching weight was equal or higher than a pre-defined threshold, those records will be considered as match. In real-world datasets, probabilistic record linkage usually surpasses the deterministic methods over various evaluation criteria.

According to the type of data being used in the linkage process, the linkage techniques can be divided into clear-text methods and Privacy Preserving techniques (PPRL). In clear-text method, the identifiers are the same as their original form at the time of data collection. They are human readable and can be used to identify individuals. On the other hand, in PPRL method, the identifiers have to be encoded in a non-identifiable format using hashing techniques. The hashed data has to be used in the linkage process.

To the best of our knowledge, LinXmart is the first software which is able to link data using either a clear-text method or PPRL. The software is open-source and built by the Centre for Data Linkage, at the Curtin University, Australia. At the time of writing this report, the software does not have a proper documentation and it is not user friendly. To run data linkage using LinkXmart, a project definition file has to be defined in JSON format. The match/non-match weights have to be assigned to identifiers manually. The match/non-match threshold has to be defined manually too. The software is not able to load files if the data error is higher than 5% of records (e.g., wrong date format). In case of error, the software is not able to provide useful information to help the end-user to resolve the problem. In many cases, the dataset has to be pre-processed before going to LinXmart (e.g., correcting date format, sex, assigning a unique identifier to each record, etc.). The software has around 12 GB of hard derive

capacity and runs on a virtual machine (VM). Using this software in ministries with difficult software installation policies is not possible.

LinkWise is software created by PolicyWise for Children and Families to address problems in LinXmart. The software is user friendly and does not require background knowledge to use it. The following picture shows the procedure of record linkage in LinkWise:



As shown in this picture, the software has two pieces: a client-side which splits a dataset into linkage part and research part, and a server-side which links multiple linkage files. The output will be a linkage keys file which assigns a key value to each record. Records with the same key values are matches. Using the key file, the researchers are able to marge research datasets (research files in the picture) without having access to any identifier.

Datasets

For linkage evaluation, Alberta Health provided a scrambled dataset with the following identifiers: Given name, middle name, family name, and address. The data is scrambled so that the individuals in the provided dataset are not identifiable. Using the provided dataset and some synthetic data we have created a base dataset with the following identifiers:

- PHN,
- Fname,
- Mname,
- Lname,
- City,
- Prov,
- postalCode,
- DOB,
- Gender

PHN is used as a unique identifier for each individual to evaluate the linkage performance and it is not considered for record matching. The base dataset contains 100,000 records and around 30 percent of those records are duplicates. After generating the base dataset, three new datasets are created by adding missing values and data errors to the base dataset. According to the amount of assigned missing values and generated errors, the new datasets are called easy, medium and hard. The following table shows the percent of assigned missing values as well as generated errors to each dataset:

	Easy data		Medium data		Hard data	
Identifiers	% missing	% error	% missing	% error	% missing	% error
PHN (only for evaluation)	0	0	0	0	0	0
Fname (given name)	10	20	15	30	20	40
Mname (middle name)	45	20	67.5	30	90	40
Lname (family name)	5	20	7.5	30	10	40
City (a Canadian city)	15	10	22.5	15	30	20
Prov (a Canadian province)	15	10	22.5	15	30	20
postalCode (Canadian)	30	30	45	45	60	60
DOB (date of birth)	15	20	22.5	30	30	40
Gender (sex)	10	10	15	15	20	20

Linkage results

We evaluated the performance of LinkWise and LinXmart in terms of precision, recall, and F-Measure. These three criteria are the most popular ones in evaluating classifiers (here linkage software). Precision is the rate of detected true positives (specificity), and recall is the rate of detected matches among all available matches (sensitivity). F-measure is a combination of recall and precision. Higher F-Measure indicates a higher performance in terms of sensitivity and specificity. The two software was used to match records in the three generated datasets. In fact, here the record matching was done on each single dataset. The following table shows the performance of record matching in terms of the three evaluation criteria.

		Easy	Medium	Hard
LinkWise-ClearText	Precision	95.6%	96.5%	96.7%
	Recall	71.2%	44.1%	23.1%

	F-Measure	81.6%	60.5%	37.2
LinkXmart-ClearText	Precision	99.75%	99.7%	99.94%
	Recall	40.07%	14.41%	3.22%
	F-Measure	57.18%	25.18%	6.24%
LinkWise-PPRL	Precision	99.5%	99.1%	98.6%
	Recall	60.4%	31.5%	14.6%
	F-Measure	75.1%	47.7%	25.4%
LinkXmart -PPRL	Precision	100%	99.96%	100%
	Recall	35.32%	11.22%	3.12%
	F-Measure	52.2%	20.18%	6.05%

Comparing clear-text LinkWise to clear-text LinXmart, LinXmart has a slightly higher precision, however it has a very lower recall compared to LinkWise. The reason is that matching threshold in LinkWise is slightly lower to accept some linkage error but increase recall. Considering PPRL in these two, both software have a very high precision and LinkWise has a higher recall than LinXmart. In overall, considering the F-measure as the final comparison criterion, the LinkWise surpasses the LinXmart in all cases.

XII. Appendix 5: MyCHILD Integration, Linkage, and Access Requirements

MyCHILD: Requirements for Integration, Linkage & Access					
	Gov't of Alberta	Alberta Health	Alberta Health Services	PolicyWise	Technology Concept
Process	**Introduction to MyCHILD**: critical to provide an overview of project with stakeholders **Discussion of data needs:** equally important for data procurer to understand the needs of stakeholders **Engagement:** maintain an open channel for communication to answer questions, report progress, share comments and ideas **Identification of program area:** program areas are identified through the iterative process **Dev't of research questions:** ensures that the right questions are being asked	**Introduction to MyCHILD**: critical to provide an overview of project with stakeholders **Engagement:** start and maintain an open channel of communication with periodic updates and opportunities to clarify processes	**Introduction to MyCHILD**: critical to provide an overview of project with stakeholders **Engagement:** start and maintain an open channel of communication with periodic updates and opportunities to clarify processes **Recruitment of staff:** funding supports an analytic team (x4) to undertake analysis of data deposits **Exploring intervention programs:** since AHS owns their data (custodians), intervention programs can	**Engagement:** as procurer and host of pending data, PolicyWise's responsibility is to establish and maintain engagement with stakeholders throughout the duration of the project **Sustainability:** how is the project and core concepts (data sharing platform) envisioned in the long-term? Development of a sustainable roadmap/plan can inform this process **Cost recovery:** will there be a fee for access to repository as a means to maintain	**Research PPRL methods:** an assessment of current methods are limiting: 1. algorithm used to data match is not obtainable/sharable; 2. tool used is dated and expensive; 3. inability to add new data (year and programs) without risk of losing existing data or first purging existing data Researching current best practices in PPRL may address above limitations **Linking data:** requires internal testing and re-testing of functions via deterministic technique ○ Technique minimizes error rate ○ Data scientist

			be explored after reviewing results from analysis of research questions	upkeep of the system?	builds an AI that conducts matching ○ Method requires very little human interaction
Governance	**Policy Committee:** establish diverse committee with various expertise and management level to participate in the iterative process [helps with overall buy-in of a project & may yield learnings to strengthen and provide credibility to project] Committee members provide a social perspective approach to project	N/A	**Scientific Committee:** establish committee with clinical expertise in/with project cohort Committee members develop research and outcomes questions and measures from a clinical perspective	**Establish Committees:** convene individuals with expertise in each of the identified areas in an advisory board(s) to help guide the project and generate evidence-informed processes	N/A
Privacy	**Conversations with privacy personnel:** provide an overview of project; discuss how project relates to current legislation and leadership understanding and interpretation of legislation. Continue discussions to map out if the project can be approved and if so, what type of agreements are needed	**Conversations with privacy personnel:** provide an overview of project; discuss how project relates to current legislation and leadership understanding and interpretation of legislation. Continue discussions to map out if the project can be approved and if so, what type of agreements are needed	**Conversations with privacy personnel:** provide an overview of project; discuss how project relates to current legislation and leadership understanding and interpretation of legislation. Continue discussions to map out if the project can be approved and if so, what type of agreements are needed	**Privacy Impact Assessment:** complete and submit for review. Although a PIA is not necessary for SUDP DP#5/MyCHILD, one is being done to make assurances to stakeholders that reasonable security controls are set in place to protect the	Develop data flow diagram mapping ▪ How data will be received ▪ Who will receive the data ▪ Where data will be stored ▪ Who will access the data (internal) ▪ How data will be accessed ▪ Who will access

				public's data and to demonstrate that privacy protocols are/were followed	the data (external) ▪ Management plan on the how and who will access data ▪ Outline potential risks and mitigation plans
Legal	**Conversations with legal personnel:** approval is necessary from both the stakeholders' privacy and legal sectors. The legal sectors' understanding of legislation and what is allowable within project scope may differ. It is critical to start and continue conversations until agreements are inked	**Conversations with legal personnel:** approval is necessary from both the stakeholders' privacy and legal sectors. The legal sectors' understanding of legislation and what is allowable within project scope may differ. It is critical to start and continue conversations until agreements are inked	**Conversations with legal personnel:** approval is necessary from both the stakeholders' privacy and legal sectors. The legal sectors' understanding of legislation and what is allowable within project scope may differ. It is critical to start and continue conversations until agreements are inked	**Privacy Impact Assessment:** complete and submit for review. Although a PIA is not necessary for SUDP DP#5/MyCHILD, one is being done to make assurances to stakeholders that reasonable security controls are set in place to protect the public's data and to demonstrate that privacy protocols are/were followed and that the project follows current legislation	Develop data flow diagram mapping ▪ How data will be received ▪ Who will receive the data ▪ Where data will be stored ▪ Who will access the data (internal) ▪ How data will be accessed ▪ Who will access the data (external) ▪ Management plan on the how and who will access data ▪ Outline potential risks and mitigation plans
	Assess technology capacity at the Ministry	**PPRL demonstration tests:** a demonstration of PPRL functions and provision for	**PPRL demonstration tests:** a demonstration of PPRL functions and provision for	Research tools → develop or modify tools → consultation	**PPRL demonstration tests:** a demonstration of PPRL functions and provision for

Technology	Can the linkage technology be uploaded without challenges or will staff need training and/or are upgrades needed?	Q& A Test PPRL against existing methodologies and discuss pros/cons of each method Select the best matching method with minimum error rate	Q& A Test PPRL against existing methodologies and discuss pros/cons of each method Select the best matching method with minimum error rate	→ test tool → consultation → re-testing: repeat cycle as many times as needed until the tool function as planned	Q& A Test PPRL against existing methodologies and discuss pros/cons of each method Select the best matching method with minimum error rate
Security	The privacy and legal sectors would want to review the security management plan for the project along with periodic monitoring and evaluation of risks	The privacy and legal sectors would want to review the security management plan for the project along with periodic monitoring and evaluation of risks	The privacy and legal sectors would want to review the security management plan for the project along with periodic monitoring and evaluation of risks	**Management Plan:** development of security protocol can have a third party review (OIPC via PIA) to ensure that risks are identified and appropriately mitigated	Develop security protocol and test the process and system