Integrated Databases of Administrative Data

From Rhode Island to Israel: Lessons for Creating Data-Driven Policy

January 2021
Ilana Pinshaw & Hadas Gabay Larom
Nova

JDC JointElka
משרד ראש הממשלה
Prime Minister's Office
הלשכה המרכזית לסטטיסטיקה
Central Bureau of Statistics
دائرة الإحصاء المركزية
NOVA
DATA STRATEGY FOR SOCIAL CHANGE

Contents

Background to this document

In April 2019, Nova, with the assistance of the State Department, led a study trip to the US with representatives of the Prime Minister's Office and several other ministries . During this trip, participants met with Professor Justine Hastings, founder and leader of RIPL (Research Impacting People's Lives) to learn of the innovative project that RIPL led in partnership with the State Government of Rhode Island to establish an integrated database that would serve the government in policy design and evaluation. The meeting demonstrated the huge benefits of the database to Rhode Island policymakers and provided inspiration to the Israeli government.

Nova, as an advisor to the government on this project since its inception, received support from the US State Department to write and publish insights from the RIPL project and its application in the Israeli context. This report is based on a series of webinars held in Israel with Professor Hastings, members of the Rhode Island Administration, and representatives from the Israeli government from October – December 2020. RIPL publications, and international research conducted by Nova were also used in the writing of this document.

About Nova

Nova is a nonprofit whose aim is to turn Israel's public and social sectors into results-based entities by improving the ability of social organizations, government ministries, and local authorities to use data in planning, ongoing administration, and evaluation processes.

Nova works collaboratively with government ministries, the philanthropic community, and the business sector to remove barriers to data use, develop access to administrative data, embed tools for data collection and analysis, and increase overall the resources dedicated to the issue.

Introductory Messages by Project Partners

Prof. Danny Pfeffermann

Chief Statistician and Director of the Israeli Central Bureau of Statistics

In recent years, the last year in particular, there has been a growing recognition of the importance of using up-to-date, high-quality and detailed data in order to improve policy and increase the effectiveness of government actions. Until recently, however, insufficient attention has been given to the infrastructure required in order for data to be used for these purposes.

As the authority responsible for official statistics in Israel, the Central Bureau of Statistics (CBS) collects and maintains a great deal of data, which it processes and makes accessible to decision-makers, researchers, and the general public. However, there are still significant gaps between what the CBS currently holds and what is needed. As a result, government policy planning is often carried out without the support of up-to-date and reliable information, and policymakers are left with limited tools to test the effectiveness of their decisions.

As part of the CBS' process of digital transformation, and in parallel with the implementation of Government Resolution 4753, we have over the past year invested intensely in finding answers to the questions surrounding the establishment of a government data lake in the CBS. The lake will be built in a way that will provide an effective and quick response to the growing need for up-to-date and comprehensive data from a variety of content areas.

I will note in this context that we learned a lot from the RIPL program in Rhode Island in the United States, run by Professor Justine Hastings.

In conclusion, this is a huge, resource-intensive project, with clear and significant benefits, that will require the close cooperation of all government ministries. As it is written in the book of Psalms, we will be like dreamers.

Oren Cohen

Deputy Director of Government and Society Department, the Prime Minister's Offic

Are children placed in foster families enrolled in higher education? Are they succeeding in the job market at the same rate as children from similar backgrounds who were not placed in foster families? How many of them have turned to crime?

The foster care service is a social service provided by the state to an extremely vulnerable population. It is thus our duty to know whether the service achieves its long-term goals and succeeds in integrating foster children into society in the best possible way. However, though most of the data needed to answer these questions is already held by the state, our ability to use this data is limited

Living in the information age, it is clearer to us than ever that the accessibility and availability of data is critical to the planning, policy-making and efficient and effective management of government operations.

In light of this, on November 24, 2019, the government adopted a resolution on "Increasing the use of government data to improve government policy and increase the effectiveness of government actions". The resolution appointed a committee to examine the establishment of an integrated interdisciplinary database (information lake) at the CBS. The establishment of the lake is intended to create a broad, secure and accessible data infrastructure.

We hope that as the process matures, the foster care service and the hundreds of other services provided by the government will be able to benefit from this infrastructure in order to make data-based decisions, manage the services better and improve the quality of services offered to citizens.

Dr Sigal Shelach
JDC Executive Director

Proper, thoughtful, responsible and ethical use of data by social and public organizations can substantially advance the quality of social services, the efficiency of systems and can significantly impact target populations. Through our work, we at the JDC see how using data can improve and enhance our work in the field - helping to base decision making on data and understand decisions in depth, reduce risks and increase trust and transparency. Proper use of data by decision makers promotes better decision making, leads to more relevant changes, emphasizes results and creates more focused and accurate responses and services for the individual.

Data use preoccupies us today at the JDC and it is clear to us that in order to get the most out of existing data, internal and inter-sectoral collaborations are critical. The data itself is a basic necessity, but the ability to integrate the data, create data sharing agreements and partnerships - are all necessary conditions for success and to lead smarter and deeper change processes in the state and in Israeli society.

Hadas Gabay Larom
Nova Director

About three years ago Nova, an organization that promotes data use in the public and social sector in Israel, identified that without a national data infrastructure that will integrate data from a variety of sources, with clear and structured privacy restrictions, Israel will be unable to realize the potential of data in planning and evaluation processes. We were delighted to meet with and learn from Prof. Hastings, a change leader who realized in Rhode Island the reality that we envision for Israel. Fortunately, the call for change was embraced by the CBS and the Prime Minister’s Office, and in the last two years substantial efforts have been made to promote this idea.

We are encouraged by the progress to date, and are focused now on the future, the question of how this policy will come to fruition. From our experience on the ground, working with government services, local government and NGOs, we see how great the need for data is across a range of fields. Our focus today is on building knowledge of how to build an integrated administrative database that will truly serve the broad public interest of its many stakeholders. The decisions we make today will affect what data will be in the database, who will have access to it and how policymakers and practitioners will be encouraged to use the data. As a civil society organization, we will continue to advise and assist in the efforts to build a data lake in Israel, while emphasizing leading practices from around the world to create the data infrastructure that will serve as great a variety of stakeholders and social issues as possible.

This report was made possible by the generous support of the American people through the United States Department of State / U.S. Embassy in Israel. The contents are the responsibility of Nova and do not necessarily reflect the views of the Department of State or the United States Government.

The Value of Data-Driven Decision Making

Billions of shekels are spent every year in Israel on social services and various intervention programs, while in fact very little is known about what actually brings about change in the lives of the target population. Integrated data systems provide policymakers and program designers with access to high quality data to better define the problem to be addressed, plan the intervention, and evaluate its long-term impact on participants.

Integrated databases of administrative data enable the creation of data-based policies: policies based in evidence, that work better for the people they are designed to serve.

History has shown that policies that make intuitive sense, or were created due to transient or political needs, do not necessarily contribute to the goal that they were designed to serve, and indeed may end up harming more than they help.

Data can help policymakers choose between policy options based on a cost/benefit analysis. Government can also play an important role in providing citizens with the data that they need to make informed decisions about their lives.

Many policymakers today understand the importance of making decisions based on data. The challenge comes, however, in collecting the data needed to inform policy decisions in a timely manner. Policy makers at all levels often find that the evidence that they need to inform decisions is unavailable, or can take months or years to collect, by which point the decision is no longer relevant.

A Counter-Intuitive Result Uncovered through Measurement

In the social services, often our instincts can betray us. Programs designed to reduce teen pregnancy by having teenagers spend a weekend looking after an animatronic doll have been in use for decades, and in 2016 were used by more than 40,000 institutions worldwide. A study was designed to test the effectiveness of the program: teenage school students were randomly assigned to one of two groups; the trial group received the electronic dolls, the control group did not, and both groups received a series of education sessions. The study then used medical records to follow the incidence of pregnancy among the two groups until age 20. Contrary to expectations, the group that used the dolls was found to have higher pregnancy rates than the control group – that is, not only did the program not achieve its goal, it appeared to actually have raised the likelihood that participants would get pregnant during their teenage years. This case emphasizes the importance of an evidence-based approach[1]

[1] https://www.nicswell.co.uk/health-news/baby-doll-simulators-may-actually-increase-teen-pregnancy-rates
https://www.smh.com.au/healthcare/they-really-liked-it-failed-baby-robot-program-increased-teen-pregnancies-20160825-gr0pzn.html

Integrated Databases of Administrative Data: A Tool for Data-Driven Decision Making

An integrated database of administrative data (also known as a “data lake”) utilizes data collected as part of the government’s operational demands to create a permanent resource for data-driven decision making, rather than relying on ad-hoc data collection based on changing policy or project needs.

The state collects extensive information about citizens for different purposes and in discrete data systems (birth and death data, education, health, social services provided by the state, income, employment, etc.) An integrated data system combines data on the relevant population from multiple resources, is governed by a regulated privacy policy, and allows access to the information needed for policymaking, measurement and evaluation.

This is particularly important for social service provision, where the state is often the primary provider of these services and responsible for ensuring their effectiveness

Evaluating program impact using integrated administrative data: a theoretical case

A program targeting teenagers from the periphery that provides preparation and support in the transition to higher education, could request information segmented by criteria such as gender, parental characteristics, socioeconomic status, etc., on the status of these teenagers several years afterwards. The integrated data infrastructure would then be able to provide long term information on outcomes, including questions such as: Did they serve in the army/perform national service? Did they participate in higher education? Are they currently participating in the labor market? What salary level have they reached? The information is aggregated at the group level, so no individual can be identified. This information can then be compared to the outcomes of a comparison group of teenagers with similar characteristics who did not participate in the program. The comparison between the two groups elucidates the actual contribution of the program to the education and employment of these youth over time.

RIPL – A Pioneer in Data-Driven Decision Making for Government

In 2015, Professor Justine Hastings, a professor of Economics and International and Public Affairs at Brown University and a Research Associate with the National Bureau of Economic Research, established RIPL, then the Rhode Island Innovative Policy Lab, to explore how data and science could be used more rigorously and enduringly to make public policy more efficient and effective.
A partnership was established with the Office of the Governor of Rhode Island, led by Governor Gina Raimondo, to develop a public sector equivalent of a model that was by that point well established in the technology sector: in-house insight labs built in partnership with top academic

researchers to collect and analyse data and drive business decisions.

The Rhode Island government set a goal of "making government more responsive to the people it serves" by focusing on "problem solving, implementing innovative solutions, and improving agency performance and outcomes."[2] Utilizing RIPL's expertise in how data and science can be used to improve policy and people's lives, the partnership determined that the major point of collaboration should be to establish a "Data Lake," an integrated database of administrative data, that would collect and store administrative data from a variety of state agencies to help support data-driven decisions by Rhode Island policy makers.

The process developed by RIPL (now known as Research Improving People's Lives) in Rhode Island is being used in states across the US and around the world. Policymakers collaborate with scientists, economists and data engineers to create the foundations to enable ongoing data-driven policy design, improvement and evaluation. Users found that integrated databases, with their standardized data management and powerful analysis tools, improved government departments' ability to work with their own data, and allowed them to combine the data with data from other sources.

An International Perspective

Brief Background on the Establishment of National Integrated Databases

In a number of countries, an integrated database was established in response to a specific (or a series of specific) policy questions. For example, Sweden faced rising levels of long-term sick leave, resulting in rising costs and an increase in early retirements. The government found that responsibility for statistics, research, and policy on the relationship between the health of the working population and the labor supply was divided between authorities, resulting in an absence of information and in problems falling between the cracks. In response, an intergovernmental committee created "LISA," the Longitudinal Integration Database for Health Insurance and Labor Market Studies.

In other countries, the establishment of integrated databases was driven by a desire to benefit from the value of integrated administrative data, without a particular policy goal in mind. For example, in France, a national strategy on artificial intelligence led to the establishment of the Health Data Hub, and the expansion of SNDS, the National Health Data System, with the goal of putting health data "at the service of the greatest number while respecting the ethics and rights of our fellow citizens." In particular, the law aimed to support innovative projects in the service of research, on patients, healthcare professionals and the health system.

In New Zealand, the Cabinet agreed that a cross-agency data-sharing solution would enable the government to deliver better public services, and so established the Integrated Data Infrastructure (IDI). Data holders from within or outside government then apply to have their data added to the database in order to benefit from connecting their data to other data held in the IDI, while an inter-sectorial committee sets priorities.

Some countries benefit from a mix of the two approaches, such as in the UK, where the Digital Economy Act was passed to formalize researchers' access to government data. The Digital Economy Act also established Administrative Data Research UK (ADR) to coordinate and promote the integration of administrative data for research and policy development. ADR works with government departments to establish integrated databases around policy focus areas, for example the connection between mental health and educational outcomes.

[2] RIPL website: https://www.ripl.org/our-story/

An Israeli Integrated Database of Administrative Data

Over the past few years, policymakers across the Israeli government have expressed a growing understanding of the importance of an evidence-based approach to policy. In April 2019, representatives of the Israeli government participated in a study trip to the US, organised by Nova and the State Department, in partnership with the Prime Minister's Office, to learn of the innovative project to establish an integrated database that RIPL led in partnership with the State Government of Rhode Island. Following that formative visit, the Israeli government set a goal: to establish a national data infrastructure that would support policy decisions on two fronts. First, in the planning stages, to create data-based policy and then in the evaluation of projects, to test the efficacy and impact of the policies.

On November 24, 2019, Government resolution no. 4753 entitled

"Increasing the use of government data to improve government policy and increase the effectiveness of government actions" [3] was passed. This decision resolved that an inter-ministerial committee would plan and assess the establishment of a governmental database of integrated administrative data as critical research infrastructure to support data-based policymaking and evaluation of the efficacy of government activities. It was determined that the integrated database ,or "data lake," would be held and managed by the Central Bureau of Statistics (CBS). The inter-ministerial committee would form recommendations as to ways to improve the way that governmental information is collected and organized to serve policy needs, while reducing the exposure to cyber threats and maximizing the protection of data subjects' right to privacy.

Current timelines for data access – Real world examples

Policy Question	Requesting Authority	Time to collate and receive the data
How many Holocaust survivors currently receive housing subsidies from the Housing Ministry?	Prime Minister's Office, Israel	**4** MONTHS
How many aged care workers are eligible for a "work grant" (negative tax), and how many take advantage of the benefit?	Prime Minister's Office, Israel	**1** YEAR
What are the outcomes for participants in socialinterventions?	Various ministries, Israel	**3+** YEARS
Which sub-parts of the "Real Jobs Rhode Island Workforce Development Program" were effective and which weren't?[4]	Rhode Island Department of Labor and Training	**20** MINUTES

[3] הגברת השימוש במידע ממשלתי לצורך שיפור המדיניות הממשלתית והגברת האפקטיביות של פעולות הממשלה

[4] Subsequent to the inclusion in the data lake of data on the general effectiveness of the workforce development program as compared to control groups.

Government Resolution 4753: Progress to Date

The inter-ministerial committee appointed by the resolution has worked over the course of 2020 to find solutions to some of the practical challenges of establishing a data lake in Israel. In particular, the focus has been on the basic infrastructure of the lake – the technological infrastructure, data security and privacy, and formalizing data collection processes with data providers (ministries). The committee is expected to publish its findings and a new proposed government resolution shortly. During the recent webinars, which took place as part of Prof. Hastings' virtual visit to Israel, committee representatives shared a few conclusions to date:

1. The integrated database will begin with data already held by the CBS, but will be designed to be responsive and flexible to changing needs, include updated data, and integrate information across content areas.

2. One of the goals is to increase the frequency and timeliness of data transfers from ministries to the CBS. One of the major acts of the implementation team will be to formalize data agreements with ministries regarding the frequency of transfers, data cleaning methods and agreements on purposes to which the data can be used. The process may also include creating unique agreements for data transfers from ministries to the CBS (to replace the current committees for data transfers (ועדות להעברות מידע) that are used today).

3. The committee set a goal to improve users' access to data through a variety of access routes including virtual research rooms that can be accessed remotely, and greater support for Customized Data & Statistics support (עיבודים מיוחדים)

4. While there is a strong focus on technological infrastructure and privacy concerns, the greatest challenge to the establishment of the data lake will be to build the ecosystem that will support the data infrastructure, which will include: enabling data holders to manage their data in a way that supports data sharing, developing ministerial capabilities to produce insights from the data, and gaining buy-in and support from various levels of government.

Establishing a Data Lake: Key Principles

Three fundamental principles critical to the establishment of an integrated database are outlined below, drawn from Professor Hastings' experience and from the establishment of several integrated databases worldwide.

Principle 1: Integrated administrative databases should break silos

One of the goals of integrated databases is to break down the silos that currently exist in government work. Integrated data from several agencies may be needed to assess the impact of one agency's policies on another. For example, assessing the needs of small businesses may require combining information from municipalities on municipal taxes and business registration data with data from the tax authorities on earnings, employees and industry. Data on certain government programs, for example interventions working to increase access to higher education, may be stored by implementing organizations which could be commercial entities or NGOs. Creating an integrated database requires a rethinking of data "ownership".

In Rhode Island, this rethinking was done as policymakers defined their goals and data needs, and could discuss those needs with the heads of other Rhode Island departments. Sitting around the same table, data holders could see how interwoven their needs were with those of other agencies, and how they themselves would benefit from sharing their data. This work emphasized the interconnected nature of government programs – the need to integrate data across agencies and even from non-government sources, in order to create a full picture of the influences and outcomes of various policies and programs.

In New Zealand, Statistics New Zealand plays this role; in order to enrich their data by combining it with data from other agencies, agencies must add their data to the national integrated data infrastructure (IDI).

Minority Representation

Getting insights from groups representing minorities or underrepresented groups can be particularly important at this stage. For example, ADR Northern Ireland worked in partnership with Stronger Together, a network for racial equality in Northern Ireland, to gain insights from migrant and ethnic minority communities on mental health services and accessibility. The consultation helped to identify gaps, such as cultural differences and barriers to health access that may result in immigrant communities receiving less mental health care support or in lower levels of uptake of disability benefits. Administrative data was used to identify gaps in prescriptions and differences in outcomes among the different populations; the consultation helped to connect the raw data to impact in the community, while building public support for the use of data for these issues.

Deciding what data will be included in the integrated database

Governments collect vast amounts of data for a range of purposes. Each variable added to the database has a cost, in terms of the time and resources involved in cleaning the data and connecting it with other data. Because an integrated database is founded on the principle of longstanding connections replacing ad-hoc project-based solutions, the decision of what to include in the database is an important one, and determines what types of insights it is possible to develop.[5]

The process of selecting the most important variables for research is most effectively conducted by an interdisciplinary, inter-sectorial group that can pinpoint the various uses of data. Expanding the diversity of representatives included in this group was a critical step in the development of integrated databases in France, the UK and New Zealand.[6]

In Rhode Island, key variables were chosen through a multi-faceted consultation involving policymakers, economists and data engineers. Later policy leaders and scientists came together with community groups to discuss the data and analysis. The state found that these discussions enabled them to undertake comprehensive interagency policy approaches not previously possible.

France – multisector consultation

The expansion of the French integrated health database was based on the recommendations of a working group led by three experts from government, academia and the private sector. The working group based their recommendations on interviews with more than a hundred public and private actors working in the health data sector. The goal of the integrated database was to promote the use of health data for the public good, and so the consultation included potential users of the database from all sectors.

[5] This is of course a question of priorities rather than absolute decision making – which data will be included in the first year(s) of the project, and which will be included only in subsequent years.

[6] Chapter 1: “Selecting the Content of the Data Lake,” International Benchmark (2020)

Principle 2: The data lake must be founded on secure infrastructure with strong privacy protections

The first step to establishing an integrated database in Rhode Island was to create the necessary infrastructure to support secure hosting of data with strong cyber security and privacy protections. This infrastructure provides the relevant government authority with full control over the data, while access to particular data sets data is granted based on an approved project-based need.

While many of these arrangements are technical, requiring specialised knowledge in cyber-security, statistical methods or privacy law, there are some relevant general principles.

Limited or no access to identifiable personal data

The first principle is that identifiable personal data (data that is linked to a person's name or ID number) is available to only a handful of authorized users responsible for managing the database.[7] Data to be accessed within the database has all personally-identifiable data removed. Instead, identifying criteria (such as name, national identity number) are deleted or encrypted, enabling records to be connected using a global anonymized identifier (or pseudo-identifier).

This produces what is known as "pseudo-anonymised data," data that has no obvious identifiers but contains enough information that it could still be tied to its original source by combining with other known data. Due to the risk of re-identification, this data cannot be opened freely to the public. Instead, authorities use a system that combines organisational, legal, technical and statistical methods, most commonly known as the "five safes."

Rhode Island: A cloud-based secure system

In Rhode Island, RIPL developed a secure cloud-based system to hold and manage the data, an automated pipeline for integrating and anonymizing data, and a process to ensure security and anonymity in joined administrative records. Though the data is held in the cloud, the government retains full ownership over the data and full control over access. Users are able to access only that data relevant to their project, and the data never leaves government control; it cannot be downloaded or removed from the secure space without express permission and approvals.

In Rhode Island, the government found that the services offered by a secure cloud (in this case, Amazon AWS) were actually able to strengthen their security protocols. The security hub automatically checks compliance with controls every 4 hours and highlights any deviance from protocols (for example if an employee were to accidentally change the controls over the data). The system logs every instance of data access, enabling remote monitoring, and alerts can be set for certain behaviours or types of data. The cloud-base system also has additional benefits, such as simple remote access that preserves cyber security controls and stronger computing capabilities.

[7] In the case of Rhode Island, they were able to reduce this number to zero by generating the pseudo-identifier through an automated process.

Question: Doesn't combining complex data increase the risk of unauthorized access to personal data or reidentification of pseudo-anonymised data?

Answer: On the one hand, the integrated database will indeed contain far more linked data than that held by any individual ministries. On the other, the risk of unauthorized access to personal data or reidentification exists widely wherever data are held on local systems. Often, sensitive data are held on personal computers of public workers without proper encryption or access protections, and may not be properly disposed of at end-of-life. The data lake, with its strong cyber security and access protections, offers much higher levels of security than can be offered by individual departments or ministries. There are a number of protections in place to ensure that all data access is properly governed. To access any data in the database, you must have an approved project and be able to justify why that particular data is needed for that project. All outputs must be approved before they can be taken out of the system. And if it is determined that someone is using the data inappropriately, their access permissions can be easily withdrawn.

Managed access to data

The system used by most authorities worldwide to manage access to private data after all identifiers have been removed is the "five safes" model.[8] The Five Safes Model breaks down the decisions surrounding data access and use into five related but separate dimensions: safe people, safe projects, safe settings, safe data, and safe output. National authorities worldwide provide researchers with access to private integrated data only if all 'five safes' conditions are met.

Five Safes

Safe people
Researchers are vetted and must commit to using data safely before they can access the data.

Safe projects
To gain access to integrated data, researchers must have a project they can demonstrate is in the public interest and meets all ethical and legal requirements.

Safe settings
A range of privacy and security arrangements keep data safe that manages access and limits unauthorized use or mistakes.

Safe data
Data has had identifying information removed, and researchers only get access to the data they need.

Safe output
All statistical results are checked to ensure that they do not contain any identifying results before they can be released.

[8] The UK and New Zealand explicitly state that they use the "five safes" model. Other countries, while they do not name this model, use the same principles.

These principles guide current access to administrative data, and will continue to guide access following the establishment of the integrated database.

For more information on the five safes, and for considerations of how access to data can be expanded while maintaining privacy principles, please see Chapter 2, “Access to the Data Lake” in International Benchmark (2020).

Data must be updated regularly, with clearly stated agreements (SLAs)

One of the challenges facing decision makers today is the reliance on outdated data. Current time frames mean that for many types of data, the most recent data available to researchers and policy makers is a year or two old.[9] Ensuring access to timely data requires creating the infrastructure and agreements that will enable data to be regularly and automatically updated.

Worldwide, the frequency of data updates varies with the data type – for some data it is important (and possible) to update the data every month or even every day, while other types of data are updated only annually. In many cases, data can and should be updated more frequently, but data updates have simply not been properly integrated into public agencies’ work methods. Regardless, it is important that the frequency of the updates is clearly stated to all potential users. The Israeli Inter-Ministerial Committee found that current agreements on data transfers to the CBS are insufficient, and determined that establishing clear agreements on data updates should be part of the process of establishing the data lake.[10]

> “Developing insights on a policy timeline requires data resources that are already well documented, easy to use, and accessible to research teams. This avoids lengthy delays which can result from building up new resources from scratch for each new project or purpose.”
>
> - Justine Hastings (2019) p. 8

[9] For example, in 2020, the CBS’s most recent available data on higher education is from the academic year 2018/19.

[10] Webinar on Legal Issues related to the Data Lake held on 10 November, 2020

Principle 3: Policy insight development benefits from a collaborative approach

When initiating the establishment of an integrated database, the focus is naturally on creating the data infrastructure necessary to support the collected data. Nonetheless, even at this early stage, the question of how to transform the raw data into useable insights to facilitate evidence-based decision making in public policy should be addressed.

The first step in organizing for data-driven decision making is to transform policy goals into measureable objectives. Experts then work with policy makers to determine what data is needed to measure these outcomes. In Rhode Island, as the RIPL team led this discussion, policy makers found that critical data for insights often came from outside their agency, reinforcing the need for integrated interdepartmental data infrastructure.

The process of transforming raw data into insights was undertaken by an interdisciplinary team. Engaging economists and other academic researchers to help with the interpretation of the data was of particular importance, in particular to understand and identify causality in the relationships between variables and to help policymakers consider the intended and unintended consequences of different decisions. Policymakers and practitioners, with their knowledge of the needs and constraints of implementation, provided context to the various data fields, evaluation measures and recommendations. The teams then designed summary tables of the data with the variables that were most critical to decision making, in order to make it easy to quickly generate insights on a policy timetable.

Once established, the data lake was able to support a variety of policy questions, such as questions into the effectiveness of existing social interventions, developing more detailed and accurate understandings of target populations so as to better design future interventions; predicting future at-risk behaviors, etc.

Case Study: Evaluating the Effectiveness of a Government Intervention

Rhode Island hospitals invest $4,000 more in health inputs for infants if the child weighs just less than 1,500 grams, which classifies them as "very-low birthweight" (VLBW). To understand how early-life interventions can improve outcomes for at-risk children born at LBW and VLBW, the state partnered with RIPL to measure the long-term impact of the intervention on the outcomes of the child. To conduct this research, it was necessary to combine data on birth weight, program participation, expenditure on social services, test scores and college enrollment – an immense amount of inter-governmental data that would be all but impossible without the existence of the data lake.

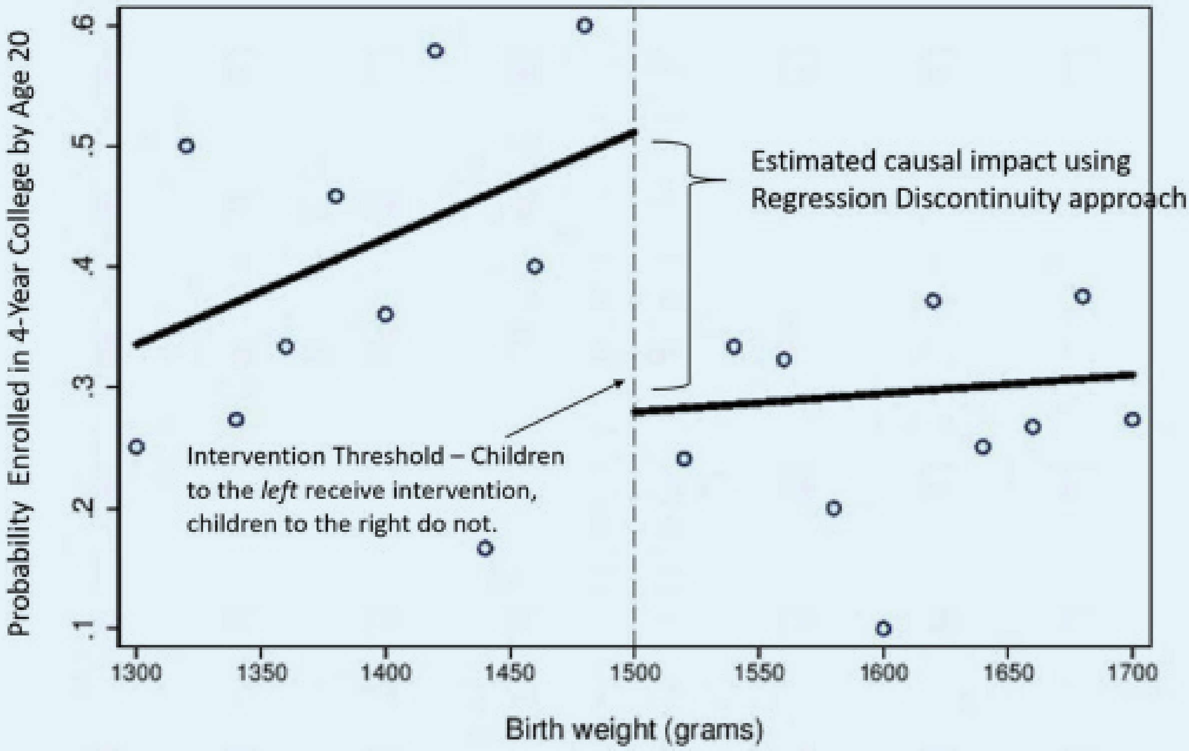
The research compared the outcomes of children with very low birthweight (<1500g) to those with slightly higher birth rate (1500-1700g) and found that the program:

- Raised test scores 0.3-0.4 standard deviations from 3rd-8th grade
- Increased college going by 17 percentage points off a base of 30%.
- Reduced total social program expenditures by $68k by age 14

Rhode Island now has important evidence that it can use to inform potential expansions of health investments for low-birthweight infants.

Rhode Island, RIPL, as presented by Professor Hastings

Causal Impact of Low Birthweight Intervention on Eventual College Enrollment



Case Study: Predicting Future Expenses by Identifying At-Risk Populations

Rhode Island policymakers set a goal to reduce Medicaid expenses without harming health outcomes. Analysis of previous research found that a significant portion of rising costs resulted from a higher number of visits to the emergency room among Medicaid recipients, and that many of these visits could be prevented through greater investment in primary care (family health physicians). Machine-learning analysis of Rhode Island's integrated data generated predictions of which users were likely to become high-cost in the following year, and singled-out four types of preventable emergency department costs. As a result, the team were able to refine the general policy question of 'How can we reduce Medicare costs without compromising health care?' to 'Which social assistance programs can be used to reach Medicare recipients predicted to become high-cost and effectively intervene to reduce future emergency department visits?'. They were able to propose a field trial centered around partnering with nonprofit organizations to provide coordinated care. Results could be easily measured using data from the lake.

Data must be well described to support insights

By definition, use of administrative data for research is a secondary use of data that was collected for operational purposes. Administrative data may not be well labelled or may be labelled using terms that are clear to those using the data on a daily basis, but which may be difficult to comprehend for users outside the original agency. Therefore, in order for the data to be useable for research and evaluation, information on it must be collected that will enable it to be understood (for example, clear field labels, data creation date, data source, purpose of the data, etc.).

Prof. Hastings found that practitioners were critical to understanding data fields when the data was poorly documented. The team was able to develop codebooks (also known as metadata) describing the data based on discussions with individual practitioners. Good documentation of variables was essential to enabling analysts and research partners to correctly and reliably interpret the data and to draw the appropriate policy conclusions from their analyses

Too often this information is treated as proprietary information that should remain in the province of the data holder, but in order to be used effectively by different agencies, it needs to be shared.

For data already collected and processed by the CBS, this work has generally already been completed. But as the lake expands, and as ministries or local authorities begin to create their own integrated databases using internal data, the recording of metadata will become more important.

Adapting operational data to enable insights will also sometimes demand its reorganization. For example, the link between a student and teacher, or child and social worker, may be stored separately from that child's outcomes. Connecting the data can provide new insights about the impact of a new teacher training program on children's outcomes, for example, that would be very difficult to achieve using traditional data management systems.

The importance of collaboration when developing insights

The primary purpose of public data lakes is to serve public policy. In today's reality, where the implementation of many social programs is outsourced or conducted in collaboration with NGOs and private companies, developing policy insights is most effective when it is collaborative – both interdisciplinary and inter-sector.

RIPL set as a principle in their collaboration with Rhode Island that the government should bring the best people together to solve problems; researchers, subject experts and behavioural economists worked together with policy makers, sharing the data for diverse needs. This was important for a number of reasons:

1. Not every government agency has the capabilities to work with data; the partnership with RIPL provided these capabilities and strengthened agencies' abilities to work with data, benefit from insights, and integrate those insights into their work processes.
2. Ministries are naturally siloed: their focus is on the outcomes and regulation for which they have statutory responsibility. The Rhode Island government found that outsiders were better at looking across agencies, seeing the interconnectivity of issues.

For example, before the establishment of the data lake, the Department of Corrections in the State of Rhode Island could only review the impact on recidivism of interventions implemented in prisons. However, recidivism programs have been found to be more effective when they address the root issues that drive people towards crime, such as employment or drug treatment, and they may continue outside the prison system. Through the involvement of non-government organizations and external experts, the government was able to look at this issue more holistically and develop programs that took a range of causes into account to more effectively address crime.

3. When social services are outsourced, is it not enough to provide infrastructure that serves only the central government. Local government, NGOs and service providers must also have the necessary data to make informed decisions when designing and evaluating the interventions they run. In order to facilitate this use of data, the Rhode Island data lake used an approach of "science as a service." Permissions are granted to external uses by government owners of the data lake on a per-project basis, using data to enable the evaluation of programs and development of effective policy. The system maintains transparency of data use so that who uses what data can easily be tracked.
4. Professor Hastings found that teams often benefit from insights that other researchers have generated. In order to build knowledge across sectors and projects, the team implemented a process to standardize knowledge and documentation across projects. These standards were shared in information about how to use the data lake.

Working in collaboration, the Rhode Island team were able to evaluate the effectiveness of programs, identifying the areas of highest return. It is important to stress that the process was not about giving "good grades" as to who is doing well but rather focused on problem solving, improving the effectiveness and efficiency of interventions.

A look to the future

The data lake was originally established to enable easy access to data for analysis to support policy making. Once established and in regular use, however, the users in Rhode Island found that the infrastructure created could support much more sophisticated decision making, utilizing advanced data tools such as machine learning and artificial intelligence to help citizens make better decisions about their futures.

Case Study: Helping low-income students get to college

Many qualified low-income students don't apply for college. To help encourage low-income students to make informed choices about their college education, RIPL used the data lake to assess the return on investment for college choices for a variety of career options. The information included the cost of college compared to expected future earnings, based on historical data of college graduates. This information was provided to high school students to help them make informed decisions about where to apply for college and what to study.

Taking this model a step further, an AI chat-bot then offered students cash incentives to students at key milestones to stay on track for their college application. The cost per student was $700, with potential additional earnings of $11,000+ at age 28.

The state is now using the lake to verify whether participating students actually registered for college, in order to assess the impact of the program on college registration.

Case Study: Reskilling during a pandemic / "Netflix for retraining"

Rhode Island decided to treat the unprecedented levels of unemployment in the state resulting from the pandemic in a revolutionary way. Understanding that many of the jobs lost would not be coming back, they developed a system based on big data and machine learning to better prepare people for a return to the workforce and to help individuals find in-demand careers in keeping with their skills and interests. The data lake in this instance is not only being used for analysis of past programs, but to facilitate future decision making. The lake collected information on past career decisions by thousands of people, using those decisions to produce recommendations for jobseekers. For many, the idea of choosing a new career out of an array of options, taking into account potential earnings, necessary training and market demand, is so daunting that they stay where they are, even given a dearth of opportunities. Making this decision easier, the system, nicknamed the "Netflix for Retraining," produces suggestions for users of possible careers based on decisions made by people similar to them. They found that as a result, users found more long-lasting careers and higher earnings after their period of unemployment. Skills RI, a nonprofit public-private partnership, then works with job seekers to provide career assistance, find training programs with proven effectiveness, and connect them to best-fit jobs. The system also matches employers with best-fit candidates, and feeds metrics on success back into the engine to create better predictions in the future.

Sources

Justine Hastings (2019) ‘Fact-Based Policy: How Do State and Local Governments Accomplish It?”,
The Hamilton Project

Webinar on Government Data Lakes held on 22 September, 2020, with Prof. Danny Pfeffermann, Chief Statistician and Director of the Israeli Central Bureau of Statistics; Oren Cohen, Deputy Director of the Prime Minister’s Office; Prof. Justine Hastings, Professor of Economics and International and Public Affairs at Brown University and Director of RIPL; Scott Jensen, Director of Labor and Training, Rhode Island; Dr Sigal Shelach, JDC Executive Director and Hadas Gabay Larom, Nova Director.
Available at https://www.youtube.com/watch?v=hT5vUTUfNT8

Webinar on Legal Issues related to the Data Lake held on 10 November, 2020, with Oren Cohen, Deputy Director of the Prime Minister’s Office; Prof. Justine Hastings, Professor of Economics and International and Public Affairs at Brown University and Director of RIPL; Mark Howison, Director of Research & Technology, RIPL; Matt Gee, CEO Brighthive; Michal Horin, CBS Legal Counsel

Webinar on Social Aspects of the Data Lake, held on 19 November, 2020, with Oren Cohen, Deputy Director of the Prime Minister’s Office; Prof. Justine Hastings, Professor of Economics and International and Public Affairs at Brown University and Director of RIPL; Eric Bean, former secretary of the Rhode Island Executive Office of Health and Human Services, currently VP of regulatory & government affairs at Unite Us; and Ronit Dudai, National Rehabilitation Commissioner, Mental Health Division, Ministry of Health

“An International Benchmark: Government Policy on Creating Access to Integrated Administrative Data” (2020), Ilana Pinshaw, Nova, at the request of Digital Israel (in Hebrew)

“Baby doll simulators may actually increase teen pregnancy rates” (August, 2016), NHS UK
https://www.nicswell.co.uk/health-news/baby-doll-simulators-may-actually-increase-teen-pregnancy-rates

“Teen pregnancy program had a reverse effect: study” (August, 2016), Kate Aubusson,
The Sydney Morning Herald
https://www.smh.com.au/healthcare/they-really-liked-it-failed-baby-robot-program-increased-teen-pregnancies-20160825-gr0pzn.html