NITI Aayog

DEVELOPMENT MONITORING AND EVALUATION OFFICE

# Compendium of Case Studies on Using ADMINISTRATIVE DATA FOR EVIDENCE-BASED POLICY MAKING

J-PAL
ABDUL LATIF JAMEEL POVERTY ACTION LAB
SOUTH ASIA AT IFMR

CLEAR
South Asia Center

NITI Aayog DEVELOPMENT MONITORING AND EVALUATION OFFICE

CLEAR South Asia Center

J-PAL ABDUL LATIF JAMEEL POVERTY ACTION LAB SOUTH ASIA AT IFMR

NITI Aayog

DEVELOPMENT MONITORING AND EVALUATION OFFICE

# Compendium of Case Studies on Using

# ADMINISTRATIVE DATA FOR EVIDENCE-BASED POLICY MAKING

J-PAL
ABDUL LATIF JAMEEL POVERTY ACTION LAB
SOUTH ASIA AT IFMR

CLEAR
South Asia Center

The authors are  Dr. Radha R. Ashrit (Dy. DG), Ms. Gujan Saini (Research Officer) and Ms. Ankita Gupta (Young Professional) at DMEO, NITI Aayog and Ms. Bhavya Pandey, Mr. Mayank Bhushan, Mr. Evan Williams, Ms. Harini Kannan, Ms. Mrignaina Tikku, Mr. Nishant Lodha, Mr. Ramakrishnan N, Ms. Sohini Mookherjee, and Ms. Aparna Krishnan at CLEAR/J-PAL South Asia.

Suggested Citation: Administrative Data Toolkit (2024), Development Monitoring and Evaluation Office, NITI Aayog

# Table of Contents

NITI Aayog DMEO DEVELOPMENT MONITORING AND EVALUATION OFFICE

CLEAR South Asia Center J-PAL ABDUL LATIF JAMEEL POVERTY ACTION LAB SOUTH ASIA AT IFMR

# Acknowledgements

The Compendium was a joint effort by DMEO, NITI Aayog and CLEAR/J- PAL South Asia based on secondary research as well as inputs provided by key stakeholders.

This work would not have been possible without the incessant efforts put in by the team at DMEO, NITI Aayog especially Dr. Radha R. Ashrit (Dy. DG), Ms Gujan Saini and Ms Ankita Gupta for providing valuable inputs and suggestions to structure the report and facilitating State department interactions. We would like to thank CLEAR/J-PAL South Asia for collating all the information, providing technical expertise on the subject, and writing the report, especially Ms. Bhavya Pandey, Mr. Mayank Bhushan, Mr. Evan Williams, Ms. Harini Kannan, Ms. Mrignaina Tikku, Mr. Nishant Lodha, Mr. Ramakrishnan N, Ms. Sohini Mookherjee, and Ms. Aparna Krishnan.

Sincere thanks to the Department of Governance Reforms and Public Grievances, Punjab and National Health Mission, Tamil Nadu for engaging in our exercise on culling out best practices on administrative data in India.

Nidhi Chhibber
Director General
Development Monitoring and Evaluation Office
NITI Aayog, New Delhi 110001

# Preface

Administrative data is among the most granular and potentially robust data available to evaluate the effectiveness of programs and schemes. Relative to survey data, administrative data is not prone to social, survey, or recall biases, and therefore becomes a powerful and crucial tool in evidence-based decision-making. However, its use may be constrained due to issues related to data quality, completeness, accessibility, and usability for analysis, interpretation, and research purposes. This paper attempts to highlight the various use cases of administrative data, showcasing both effective use and challenges, along with suggestions to enhance its calibre in informing policy design and implementation.

The wave of digitisation in government ministries and departments has further aided in the timely accessibility, granularity, and quality of administrative data available. Further, it has enabled the potential for interlinkages between different datasets through several MIS systems established across sectors such as health, education, agriculture, water, etc, facilitating complex analytical research in novel ways. While there are limitations on the nature of data that is captured (especially related to individual behavioural outcomes and experience), the potential for the use of such data to design/target programmes and measure critical outcome indicators is still enormous. It accounts for universal coverage while maintaining accuracy and the long-term availability of data. It also allows policymakers to swiftly identify problems using the data at their disposal.

Administrative data can be prone to human errors – mostly within data collection, accuracy, and exclusion. Therefore, it is important to account for and manage these aspects by establishing data audits, validation checks in-built within the data collection system, and high-frequency data checks. Further, it is pivotal to practise robust data storage practices that include thorough documentation and follow data protection and data security protocols to protect the personal information of the beneficiaries.

The near real-time collection of administrative data, along with minimised errors, sound coverage, and interpolation will help in identifying areas for intervention, designing policies, and proficiently implementing them.

This compendium is intended for use by officials in government ministries and departments, as well as staff of research institutions and civil society organisations collaborating to strengthen the use of administrative data for decision-making and improving social welfare. It serves as reference material and as a starting point for government agencies interested in systematically developing a data strategy and enabling the necessary institutional and infrastructure measures and processes to adopt a data and evidence-informed approach to policy making. The case studies highlight initiatives and best practices that may be replicable in other contexts and locations. The technical sections outline a roadmap with guidelines for integrating data to design, pilot, test, and scale public programmes and services to address critical policy priorities and socio-economic development challenges.

# Section 1:
# Introduction to Administrative Data

Administrative data are those that are routinely collected for operational purposes (such as administering public services) rather than for a specific research objective. The majority of administrative datasets are data about individual entities. These individual entities may be individual persons (e.g. students, farmers, patients, scheme beneficiaries, etc.), entities (e.g. households, firms, service delivery centres), properties (e.g. land, school buildings, public assets), or events (e.g. distribution of rations, tax collection transactions, issuing of certificates for birth, property registration, income, or caste). Administrative data typically refers to descriptive, socio-economic, geographic or other information about these individual entities. Such administrative data is routinely collected as part of the delivery of public services by government agencies to eligible beneficiaries through their bank accounts, or the provision of public services. Thus, to describe it in its simplest form, **administrative data consists of information created and collated when people interact with the government in the course of accessing public services and benefits.**

Most administrative datasets are either collected directly by the government or collected on behalf of the government by frontline workers or external agencies. Digitised records, in addition to improving the efficiency of service delivery, are also becoming increasingly important as possible sources of information by expanding the volume and granularity of data that would otherwise have to be collected through sample surveys, which are time-consuming and could be difficult to administer (e.g. because of declining response rates in surveys).

In the last decade or so, digitisation initiatives have led to the timely availability of data and have thereby improved the ability of governments to monitor trends and detect and course-correct deviations in programme implementation and service delivery. The meaningful organisation and systematic analysis of data, in line with program or policy goals, help identify gaps and areas of intervention. The process of routine data collection also makes certain analytical approaches possible. For instance, administrative data is often suitable for constructing statistics at a small area level, which may be cumbersome to collect through social surveys – which need to follow accurate sampling strategies to be representative.

Data and its analysis by itself are insufficient to address gaps in the delivery of government services or improve the welfare of citizens. Data analyses need to be combined with the interpretation of insights using both **theory** and **local context**. This helps identify specific challenges, and accordingly, design and test interventions to address these gaps and ultimately adopt the most cost-effective solutions at scale. Impact evaluations and randomised controlled trials have improved the quality and breadth of evidence used to inform better policymaking. Several of these studies also make use of existing data sources, typically administrative databases. Yet, more often than not, this type of research is complex and costly – as it involves primary data collection. Relevant, reliable, and comprehensive administrative data that can be accessed and used by researchers can promote new studies at lower costs and consequently improve the adoption of evidence-informed policy systematically and frequently across governments.

There are several advantages to using administrative data for routine analysis and research to generate insights:

- Administrative data can **measure certain indicators more objectively** and therefore help avoid social desirability or recall biases of survey data. The usage of technology like biometric capture or automatic geotagging can make administrative data more reliable and accurate than self-reported information.
- **New types of data**, used in conjunction with administrative data, opens exciting areas of research that can improve policies and programs. For example, utility billing, point of sale data, or phone usage data can provide insights into the behaviour of people over time at a granular level.

- The availability of large volumes of granular data enables analysis using advanced statistical techniques **leveraging machine learning and artificial intelligence** – expanding the nature of insights that can be generated from data. This can lead to a whole new class of insight and inference.

In this compendium, we summarise the advantages of using administrative data and discuss some of the best practices that can ensure better quality and usability of administrative data. We also highlight how government institutions have adopted some of these best practices and developed effective and replicable solutions to use administrative data. This paper highlights the potential for the use of data beyond the immediate administrative purpose for which it is collected. Hence, it does not focus on aspects such as the design of management information systems (MIS) or associated data dictionaries, structures, or technology systems which are fundamental to the generation of administrative data.

The focus of the paper is on how such data generated through the routine course of government functioning can be used for purposes of analysis and research, aligned with the goal of evidence-informed decision-making. While there are observations and guidelines on data quality, storage and sharing, it is in the context of the **use of data for decision-making.** The case studies and examples included are representative of good and innovative uses of data and digitisation. However, unless otherwise specified through references to impact evaluations, the case studies do not make any claims to the effectiveness of the underlying schemes or programmes themselves. They only seek to illustrate systematic efforts to integrate the use of data in routine planning and decision-making, and the progress achieved on the same. We hope that this compendium can provide actionable recommendations for central and state government agencies to develop robust administrative data systems and use this data for adopting an evidence-based approach to policymaking.

# Section 2: How can administrative data be useful for policy and research?

Digitisation initiatives have greatly improved the granularity and frequency of availability of administrative data in India. For example, mobile health applications (m-health apps) can provide real-time information on when immunisation services are delivered at health sub-centres in towns and villages. Similarly, point of sale (PoS) devices record every transaction at the fair price (ration) shop. School-level management information systems (MIS) capture information at the student level on their attendance, tests, and socio-economic characteristics. However, the availability of such data in digitised formats and in near real-time does not automatically translate into its use for planning and decision-making (i.e. through advanced analysis or use in research). Barring some exceptions, the use of data has been mostly limited to its primary purpose of aggregation for reporting purposes or for an investigation into any issues that emerge. However, there are a few recent examples of how the availability of granular and real-time data has helped provide targeted services effectively or improve the efficiency of services delivered.

### Example 1: Use of data for management of Covid-19 pandemic

*During the Covid-19 pandemic, a large volume of data was generated and used by governments, social enterprises, private organisations, and citizens for different purposes. This was largely enabled by the daily reports on Covid-19 RT-PCR tests and outcomes published by the Indian Council of Medical Research (ICMR) in a standard format, irrespective of whether the tests were conducted by a private or public lab anywhere in the country. These were helpful in estimating prevalence and transmission rates up to the district level. The Ministry of Health and Family Welfare, Government of India released information pertaining to the number of active cases, number of deaths, number of people vaccinated, etc. A few local city governments such as Delhi and Mumbai were also releasing information on tests, prevalence, availability of hospital resources, and hospitalisations at the ward level, enabling the localised management of Covid-19 hotspots.*

*New forms of data from private businesses, such as cell phone records combined with surveys, helped analyse and inform the riskiness of certain activities. A machine learning algorithm used mobile phone data in Togo to identify the company's poorest subscribers, who were then given humanitarian aid towards Covid-19 relief. When Covid-19 vaccinations became available in India, the CoWin platform helped governments track the take-up and coverage of vaccinations at the individual, vaccination centre, and district levels, grouped by demographics (gender, age) and geographical area (rural/urban, district). Such real-time and granular information led to the development of many targeted interventions to improve the coverage of vaccinations. Thus, only eleven months after the commencement of multiple vaccination drives, close to 90% of the eligible population across the country had received its first vaccine dose.*

There are many other instances, nationally and internationally, that demonstrate how good quality data can be used meaningfully to design relevant and quick solutions in a crisis. Similarly, routine monitoring and analysis of beneficiary and transaction-level data can help improve the efficiency of government programmes (i.e. ensuring services are delivered in a timely manner) and their effectiveness (i.e. ensuring benefits reach the intended beneficiaries) in normal times as well. For example, the Ministry of Tribal Affairs, Government of India has developed a Direct Benefit Transfer (DBT) Tribal Portal for the efficient disbursement of scholarship benefits. The digitisation of this process, standardisation of data collection, and streamlining of benefit transfers have helped reduce delays and leakages in the distribution of scholarship funds, as described by the Ministry (Case Study Box 1).

# Case Study 1:
# Using Direct Benefit Transfers to digitally disburse Scholarship Benefits

## Pre-Matric and Post-Matric Scholarship Schemes, Ministry of Tribal Affairs

### Background

The Ministry of Tribal Affairs (MoTA) implements five scholarship schemes for students pursuing studies from class nine to Post-Doctoral levels in India and abroad. The Pre-Matric Scholarship and Post-Matric Scholarship Schemes are Centrally Sponsored Schemes and are implemented through states with an annual budget of about INR 20 billion.

Until 2018, the scholarship schemes were implemented manually by the states. The applications were through physical forms and were verified manually by institutes/ authorities. It was a cumbersome and time-consuming process, resulting in delayed payments of scholarships and a higher possibility of fake beneficiaries and fraudulent institutes availing the benefits of these schemes. There was also a high pendency of arrears and a marked absence of robust communication or grievance redressal mechanisms. While some states had unit-wise data of beneficiaries in their IT systems, there was no mechanism to share this data with the MoTA, due to differences in agencies, platforms, and data structures. Even after the scholarships were disbursed, states sent paper-based utilisation certificates (UC) for the funds spent – which were not supported with data on how the funds were utilised. Despite the relatively huge expenditure of the scholarship schemes, the availability of information was low and process monitoring was plagued with sequential difficulties and delays owing to the manual nature of the process.

### Use of digitised administrative data

MoTA developed the Direct Benefit Transfer (DBT) Tribal Portal for Pre- and Post-Matric Scholarships to digitise the process of applying for and disbursing scholarships to students. The portal is a Management Information System (MIS) which captures unit-wise beneficiary data from all the states mandated by the DBT Mission. Each state now has its own portal or is using the National Scholarship Portal developed by the Ministry of Electronics & Information Technology (MeiTY) for the identification of beneficiaries, verification of students and institutes, and disbursal of scholarships. Since these portals are on different platforms, using different databases and different formats, a common 30-field format has been designed to facilitate data sharing, capturing beneficiary details, bank details, location of the school, course details, and transaction details. The MIS draws critical information from these portals. The states can share beneficiary data online via an adaptive data-sharing module. The data collected is used to generate state-wise, institute-wise, gender-wise, and stream-wise reports for monitoring and coordination with universities and students regarding the disbursement and utilisation of the scholarship. They are also now quickly shared with states online so that they can clean and report correct data in case of discrepancies.

States are given grants through the Public Finance Management System (PFMS) system and states further release these funds to students through DBT. For this, states with PFMS were mapped out. States which have not yet shifted to PFMS are required to share expenditure data from their treasury to PFMS and report their expenditures in the PFMS portal. Further, configuring the DBT to disburse the scholarship to the beneficiaries is to be done in the PFMS. The DBT portal allows states to directly upload queries, Utilisation Certificates and Statements of Expenditure, and communicate with Ministry officials. Since these documents are directly accessible to officials of the Scholarship Division of the Ministry, they can upload online deficiencies in the proposal, upload sanction orders, and send SMS/email notifications to states and beneficiaries.

**Key outcomes**

With the implementation of the DBT Tribal Portal, MoTA is one of the first ministries to capture beneficiary data of centrally sponsored scholarship schemes as directed by the DBT Mission. The DBT ecosystem developed by MoTA is not only a great example of data collection and standardisation, but also enables data use for program monitoring at the appropriate administrative levels.

The DBT Portal has helped strengthen the verification process, prevented fake and fraudulent cases, and has resulted in saving time and a speedy and robust grievance redressal mechanism. Information on the performance of all stakeholders, district-wise details of the beneficiaries, and the manner in which funds have been utilised has been placed in the public domain through the Performance Dashboard. The Portal has enabled the adoption of a single-point collection and monitoring of data in a secure environment.

As Central and State governments emphasise adopting an evidence-based approach to policymaking, the availability of granular and quality administrative data offers tremendous potential for use in advanced statistical analysis and data-sharing research. Beyond ensuring the efficiency of scheme implementation, community-level sample surveys of youth (collecting information on socio-economic status, access to education, and whether or not they receive scholarships) could help identify whether the scholarship scheme is reaching the intended beneficiaries. In other words, when surveys can be used to match administrative data from scheme implementation, inclusion and exclusion errors can be identified, further improving scheme effectiveness.

### Example 2: Development of social registries for improving scheme delivery in Chile

*The Ministry of Social Development and Family in Chile has set up a Social Information Registry (RIS), an integrated information system that pools data on social, civil, and socioeconomic attributes of the country's population from municipalities, public, and private institutions that manage social benefits. This data bank serves as a central platform, which then helps the institutions in the selection of eligible beneficiaries easily and improves the process and efficiency in the delivery of benefits and services. The RIS in Chile encourages proposals for high-quality research projects that can contribute to evidence-based policymaking. Through institutionalised collaborations and strong principles of privacy, data security and ethical research, projects were selected across sectors of education, labour, housing, healthcare etc.*

Similarly, large-scale randomised impact evaluations have been able to leverage existing administrative data to identify gaps, design innovative solutions, and pilot, test and scale such innovations. Insights from existing data, when combined with rigorous research, can generate evidence that is relevant and timely for policymaking. Some examples include the usage of administrative data for program monitoring of agricultural subsidy programs in Telangana and Odisha; to ensure better coverage of potential scheme beneficiaries receiving maternal and child health services in Tamil Nadu; to study the impact of support services and programs on health indicators in South Carolina, United States; and to evaluate the effects of residence locality on residents' well-being in cities across the United States. Case study 2 exhibits how multiple administrative datasets can be linked and used effectively by researchers for evaluating the day-to-day implementation of public services such as road safety, which can help decision-makers in correcting deficiencies in areas such as the deployment of human resources.

# Case Study 2: Using integrated administrative datasets for an efficient deployment of police resources

## Randomised evaluation of drunk driving crackdown efforts by Rajasthan Police, Government of Rajasthan

J-PAL affiliated researchers Abhijit Banerjee (MIT), Esther Duflo(MIT), and Daniel Keniston (Louisiana State University) with co-researcher Nina Singh (Government of Rajasthan) worked with the Rajasthan Police in 2010-2011 to evaluate an anti-drunk driving programme using sobriety testing checkpoints to determine what solutions can be most effective at preventing drunk driving and reducing traffic accidents. Over two years, researchers used administrative data such as court records, breathalyser data, and supplementary surveys from a set of randomly selected checkpoint locations to gather information on road accidents, deaths, and police performance.

“The integrated administrative datasets were used to synthesise reports and generate vital statistics such as the number of accidents in a city area (within the jurisdiction of the police station) every week, accidental deaths during a particular time of the day in a month, the density of accidents around a designated check-point and the changes required in the deployment of personnel and the planning to relocate check-points.” (Banerjee et al. 2019).

The study’s main results on accident and death rates were drawn from administrative accident reports by the police. For each accident, data was collected properly on the police station, the date and time of the incident, the number of individuals killed or injured, and the types of vehicles involved. This was supplemented with additional survey data on police activity at the checkpoints, collected by surveyors sent to monitor randomly selected checkpoint locations.

The connections between the datasets from police, judiciary, and on-site surveys enabled the police to deploy innovative strategies to crack down on accidents caused due to drunken driving by testing the impact of having either fixed or rotating checkpoint locations.

The results suggested that people quickly learn about police interventions and adjust their behaviour in response; hence police strategies targeting a single high-crime area are rapidly undone by the diversion of criminal activity to other areas. This made a fixed checkpoint at the single highest-potential location less effective at reducing road accidents and deaths than checkpoints across multiple initially less-promising locations which were harder to predict and avoid. However, drivers also learned when the checkpoint enforcement period was over, and slowly reverted to their original behaviours.

## 2.1 Advantages of administrative data

It is important to analyse the factors that make administrative data a viable and attractive option for use in policy research and decision-making. These advantages enhance the ability of the data to inform integral policies and decisions.

### 2.1.1 Near-universal coverage:

Administrative data typically covers an entire population of beneficiaries when collected through the delivery of routine government services and transactions. For example, this could be the population eligible for taxation, all births and deaths in a country, all vulnerable persons eligible for specific social protection benefits such as universal rations or insurance, etc. These datasets provide better coverage than even large-scale surveys, which rely on sampling. This is because near-universal coverage ensures that almost all beneficiaries are included in the data, even those who may not typically constitute a part of the survey sample due to various reasons.

During large-sized surveys, many groups could be left out due to sampling errors or time/budget constraints. An example of such groups could be individuals/households in small or remote settlements. Respondents could be clustered from certain blocks or districts for ease of survey operations. This underrepresentation of groups often restricts researchers or policymakers from identifying variegated trends at the district level or below. Administrative data (assuming the services do indeed reach all intended) helps tackle these problems since it contains in-depth information relevant to a given program, thus providing better coverage of the individuals in question and not excluding any beneficiary groups.

However, the same issues of representativeness are crucial for programs that are not universal. For targeted programs, it is essential to verify that all the persons who are “eligible” for government services and benefits can avail of them. It must be noted that administrative data only collects information about the individuals who interact with the government system to access benefits. Many who are eligible may not be availing of them because they may not be aware of their entitlements or may be deterred by the tedious processes required to access benefits. As a result, marginalised and vulnerable populations may not be represented in the data if they are not accessing the benefits or services. Thus, attention must be paid to understanding the population covered in the administrative data before using it. This is done by combining scheme-specific datasets with universal datasets, such as the census.

A good way to combine the power of administrative and survey data would be to conduct a survey in a few locations to identify if the eligible population in the sample are accessing benefits. The self-reported and observed data on their socio-economic characteristics can be compared to the data captured for the same persons in the administrative records to determine the accuracy of the data at the individual level, and the errors of exclusion/ inclusion at the dataset level. This helps validate and identify the percentage of exclusion or inclusion errors in the larger administrative dataset of the scheme/program.

Let’s take a hypothetical example of an unemployment insurance scheme available to individuals aged 19-45 who have completed Class XII and live in households with an annual income less than Rs. 75,000 in a state in western India. The scheme could use an underlying dataset which can be compiled by inviting applications from unemployed youth or collating information through other existing records such as the BPL family database combined with Class XII exam pass data, etc.

When implementing any large-scale scheme/programme, errors can creep in due to lack of updation of records, migration, etc. Routine surveys, say in one district, can identify unemployed youth in the area within the eligible age group, their educational status, and their annual household income (reported or as verified through their BPL card status or other proxy measures). This data can then be compared with administrative records on persons receiving unemployment insurance in that district. Comparison of the number at the aggregate level indicates coverage, and at the individual level would identify inclusion and exclusion errors. That is, are there individuals who are no longer eligible still receiving benefits, such as persons over 50 years, or not having completed Class XII (inclusion errors); or persons who meet all eligibility criteria but are not receiving the benefits (exclusion errors) due to lack of awareness of the scheme details, difficulty in securing proof of eligibility criteria, or discretion exercised at the field level by lower level functionaries.

Another benefit of using administrative data is its ability to quantify impact across smaller-subsets of beneficiaries due the large amount of data available for analysis. This is because, in addition to the data of beneficiaries/citizens, there may be data available on their socio-economic outcomes over time.

### 2.1.2 Accuracy:

While collecting data for research, a variety of information is required from respondents which are often difficult to remember – such as immunisation history, ante-natal care visits, etc, or require the respondents to make subjective judgements – such as the distance between two places, time taken to travel, quantities of items purchased, etc. A key question to evaluate the quality of the implementation of a program is whether beneficiaries are receiving intended benefits in a timely manner, evaluating if the program is functioning as prescribed in the protocols, and what hindrances, if any, have affected or may affect the process in the future. Such information may be readily available in administrative data.

For instance, it is much easier to observe the different vaccinations given to children as per the routine immunisation schedule by collating the data from the health MIS used by front-line workers, rather than rely on parents' abilities to recall specific vaccinations and distinguish them from other health services received. Similarly, certain health outcomes such as blood pressure, weight, and anaemia are better measured through equipment used during routine health checks rather than self-reported by respondents in a survey.

### Example 3: PROGRESA – An early example of using administrative data

*PROGRESA is a national Conditional Cash Transfer (CCT) program introduced by the Government of Mexico covering a third of rural families across the country (~2.6 million families by the year 2000) to improve the health and nutrition outcomes of children in these families. The transfers were conditional on their undertaking certain health behaviours to access preventive health care and visit clinics. An impact evaluation of this programme (Gertler and Boyce 2001) was conducted in Mexico (1997 - 2000) on a wide range of health outcomes, for which a combination of administrative records, household surveys, and a large-scale census was used to monitor and measure critical outcomes.*

*PROGRESA first conducted a census to collect information on the socio-economic conditions of rural families, providing the basis to identify eligible communities and households. The programme, adopting an experimental phased rollout, also planned for and undertook a baseline survey of a sample of eligible households including both beneficiaries who received benefits (treatment) and those households that didn't (control). Four follow-up surveys, undertaken every six months after the start of the programme, captured data on key socio-economic and demographic indicators, as well as extensive data on beneficiaries' health status and their use of healthcare facilities. Further, administrative records of public clinics were used to understand details on visits to the clinics (a necessary condition for the transfer but also an important direct outcome of the programme). The data on visits also validated the self-reported data by households through the surveys. For instance, by comparing and triangulating data from the administrative records with survey data, it was observed that in the first year in which PROGRESA was operational in all treatment localities, there were 2.09 more visits per day to clinics in PROGRESA areas than in non-PROGRESA areas.*

*The review of administrative records (on preventive care utilisation, clinic visits and child, adolescent and adult health) in tandem with the routine primary data collection exercise allowed for a rigorous impact analysis across an array of outcomes, namely the impact on preventive care utilisation, the impact on child, adolescent and adult health, as well as the impact on the incidence of serious illnesses.*

Such instances are not limited to health data collection alone. Technology can be leveraged to collect and process data on pollutants more accurately – as evidenced by a study conducted by researchers from J-PAL, who used data on emissions collected through electronic devices installed by the Gujarat Pollution Control Board to estimate the impact of an emissions trading program (Case Study Box 3).

# Case Study 3:
# Using administrative data to implement a pollution control system and quantify its impact

## Emissions trading scheme, Gujarat

### Background

Surat is an industrial city in Gujarat, where textile and dye mills are a major source of particulate matter pollution. The Gujarat Pollution Control Board (GPCB), the State regulatory authority, enforces national pollution laws and regulations. In an effort to curb particulate matter pollution, the Board established an Emissions Trading System (ETS) market in Surat in 2019. As a first step, industries were mandated to install Continuous Emissions Monitoring Systems (CEMS) at the industry site which would transmit information on real time particulate matter emissions to the regulator. The ETS scheme set a cap on emissions equivalent to the amount of pollution that would have been released if all plants had been in compliance and distributed emissions permits to firms. Once permits are allocated by the regulator, industries can then trade among themselves depending on their level of emissions. Firms that found it cheap to reduce pollution, cut back and profited by selling excess permits and hence, the ETS helped in minimising the costs of meeting the emissions target (Greenstone et al. 2023).

In partnership with GPCB, researchers evaluated the impact of the emissions trading program for particulate air pollution on air quality and compliance costs for the industrial plants in the city. The evaluation of ETS leverages an existing innovation of rolling out CEMS devices across the state, which report live readings of particulate emissions. ETS takes advantage of this technology to track industry emissions in a transparent way and better inform emissions regulation.

### Use of administrative data

This study extensively used the administrative data on plant pollution reporting via CEMS, data on plant participation in the market including all permit bids and offers, and records of cleared trades as well as any regulatory penalties to implement the scheme and quantify its impact. In addition to this administrative data, survey data such as plant characteristics covering abatement capital, economic variables like employment and sales, independent measurements of air pollution and abatement costs for all sample plants, was also made use of.

The project evaluating ETS using CEMS is a great example of the innovative use of smart tools and technology for data collection and reporting. The evaluation specifically allows for piloting and assessing the impact of this new regulatory mechanism, so that it can subsequently inform policy options available to the regulator. CEMS devices are installed at the industry premises and the data is reported real time to the GPCB, by the plants. The electronic CEMS devices used, including the probes to measure the readings in plants to estimate PM emissions, are regularly maintained by the industry to ensure that real time data of the mandated quality is transmitted to GPCB. To estimate particulate matter emissions accurately through CEMS, these devices need to be calibrated using air samples, which are manually collected by environmental labs. The calibration exercise entails estimation of particulate matter weight along with around 25-30 estimators required to accurately calibrate the CEMS device as per the Central Pollution Control Board (CPCB) guidelines. A CEMS web portal and mobile application is in place to enable industries to monitor as well as self-regulate their emission trends periodically.

### Key outcomes

The researchers found that ETS reduced emissions by 20-30%, reduced compliance costs by 8-13% for participating industries as well as dramatically improved CEMS data quality. Additionally, they found that the plants traded often and plant permit holdings at the end of each compliance period differed greatly from initial allocations (Greenstone et al. 2023).

The trading platform developed for ETS in Gujarat, to enable industries and State Pollution Control Boards to implement trades, can be adapted to other pollutants, sectors, or geographical clusters. With the success of ETS in Surat, Gujarat has decided to expand the scheme to Ahmedabad and other industrial clusters. The market in Ahmedabad began in September 2023 and is currently ongoing in its fifth trading period. The scheme was also expanded to include 142 Phase-II units in Surat in December 2022.

Another example where accurate data generated in real-time helped drive data-based decision-making is the 'Awaas' app used for the Pradhan Mantri Awaas Yojana by the Department of Rural Development, Government of India (Case Study Box 4). The data from the app was used during monthly review meetings and had reliable information based on which key decisions were taken. Prescriptive analytics were used to understand the resources required to meet set targets.

## Case Study 4:<br>Beneficiary-led data collection for fund releases and monitoring of a universal housing scheme using a mobile application

### Awaas App, Pradhan Mantri Awaas Yojana (Gramin) Scheme, Department of Rural Development

### Background

To achieve the objective of providing "Housing to All", the Government of India rolled out a revamped rural housing scheme, Pradhan Mantri Awaas Yojana-Gramin (PMAY-G) on 20th November 2016 with effect from 1st April 2016. The program envisaged the complete development of 2.95 crore PMAY-G houses with all basic amenities by the year 2022.

Some challenges faced by the Department in the monitoring and evaluation of the scheme included the non-availability of unit/beneficiary level data at the central level to cross-check last mile progress of the scheme, monthly progress reports submitted by states which were only prepared at the district level and prone to human errors and manipulations, a lack of a mechanism to monitor the end mile delivery of scheme benefits to only intended beneficiaries, delays in fund releases, and inefficiencies in scheme implementation.

### Use of administrative data

The Department of Rural Development developed the Awaas App to be used by both program beneficiaries (to report the physical progress of houses under construction) and the designated PMAY house inspectors (to inspect the houses constructed and monitored through AwaasSoft). The mobile application also captured high-quality photographs with time stamps and geo-coordinates at each stage of construction, so that the next instalment of financial assistance can be provided to the beneficiary without any delay – for the automated release of payments after inspections. A Fund Transfer Order (FTO) tracking feature allows the user to track the payment status by directly entering the FTO number of the Beneficiary ID.

The system leverages the use of AI and ML to detect duplication of photos uploaded on the app which helps in automatic verification of data. The app also includes a feature which allows the QR code given in the printed e-sanction order to be scanned which allows the inspector to view details of the household and allows for a hassle-free inspection of the house. The Awaas App also provides geo-referencing features and captures deviation from proposed site coordinates, which helps in on-ground monitoring of the scheme.

With the capability of working in offline mode with dedicated logins, the app proved to be a game-changer in ensuring the accurate monitoring of house construction stages and linked this with the release of instalments to the beneficiary – all while removing inefficiencies in the system. This mobile application not only ensured the systematic recording of assets but also the tracking and release of money to the beneficiary in time.

**Key outcomes**

The mobile application has helped streamline resources for the construction of houses. The adoption of DBT has helped efforts to plug leakages in the system and enabled timely disbursement of funds, enhancing the trust of rural citizens. "This has significantly reduced the average days taken to construct a house from 314 to 114 days, as indicated by a study undertaken by the National Institute of Public Finance & Policy (NIPFP). The citizen-centric architectural approach of the Awaas App has helped the scheme become more citizen-centric and inclusive." (NIPFP Annual Report 2018-19).

Further, during primary data collection, respondents often hesitate to answer sensitive questions related to socio-economic status, income, caste etc. Administrative data can help cut down this issue if there is detailed historical data collected on the population when registering for benefits under any scheme. Such data could reduce the need to seek sensitive information repeatedly through surveys. In these cases, the use of actively or passively collected administrative data rather than data that is reported by respondents or employees in surveys reduces the danger of errors on account of social desirability or enumerator bias (Feeny et al., 2015). This also entails a reduced participant burden of sharing information with researchers. For instance, respondents may be hesitant to report incomes accurately during a survey, while in an income tax form, income may be entered accurately, and can further be validated against amounts reported by their employers.

### 2.1.3 Long-term availability of data at lower costs and greater ease:

In the case of time-intensive programs or interventions, administrative data may be collected systematically across long periods. If designed well, administrative data can be used to track key outcomes of an individual over their lifetime. Surveys tend to be at certain points in time and resources often restrict both the sample size and frequency of data collection. For instance, with the introduction of education management information systems (EMIS), continuous and granular data could become available at the individual student level on attendance and test scores. Such a setup was instrumental in understanding the longer-term outcomes of the age of entry into schools in Chile (Example 4).

### Example 4: Studying the impact of age of entry to school on learning outcomes using student administrative data over 11 years

Researchers in Chile have been able to investigate the impact of age of school entry on academic progression for children using detailed administrative data. They tracked select outcomes by following a cohort of students over 11 years of their school life (Cáceres-Delpiano and Giolito 2019). The data used in this analysis came primarily from public administrative records on educational achievement provided by the Ministry of Education of Chile between 2002 and 2012. These records contained individual information for the whole population of students during the years that a student stays in the system.

They found that a higher age at entry has a positive effect on grade point average and on the likelihood of passing a grade, although this impact tends to wear off over time. They also suggest that children whose school entry is delayed are also more likely to follow an academic track at secondary level. Having access to data for the cohort of students for a long term not only allowed the researchers to understand the "evolution of the impact of age of entry, but also shed light on alternative channels that explain the pattern over time". (Cáceres-Delpiano and Giolito 2019)

Since administrative data is generated or collected at the time of service delivery, it does not incur any additional costs that are typically required by independent survey-based data collection operations. The costs incurred for administrative data may be high upfront in the setting up of information systems, however once set-up, can run for many years with the costs spread out and subsumed as a part of routine 'business operations'.

#### 2.1.4 Timeliness of response and quicker problem identification:

Since data is collected at the point of service delivery and during routine operations, the continuous analysis of such data allows for quicker identification of problems or gaps in service delivery and course correction with appropriate measures.

For instance, the Indian Railways has developed a Real-time Train Information System to facilitate essential train control functions. This system has helped improve the accuracy of reporting train speeds and locations, as well as enabled train movement forecasting based on accurate real-time data. The use of Internet of Things (IoT) devices to inform the system has reduced errors and delays (Case Study Box 5).

# Case Study 5:
# Implementing Real-time Train Information System using Internet of Things (IoT) devices

## Ministry of Railways

### Background

The movement of trains in a given journey in the Indian Railways is managed by plotting a 'time-distance control chart' in the Control Office Application (COA) system. Earlier, information related to the arrival and departure of trains at stations was fed manually by the section controller into the COA system after getting the required information from station masters through hotline communication. This process of manual data collection was very time-consuming, leading to delays in the process and human errors.

Indian Railways has automated the acquisition of train movement data by implementing a Real-time Train Information System (RTIS). The system has been developed by the Center for Railway Information System (CRIS) in collaboration with the Indian Space Research Organization (ISRO). RTIS facilitates efficient train control functions including train movement forecasting based on accurate real-time data.

### Use of administrative data

The RTIS is an Internet of Things (IoT) based system utilising ISRO's SatNav & SatCom services. This ruggedized device has software which helps in determining train movement events such as arrival, departure, and run-through at stations using a pre-defined logic based on spatial coordinates and train speed received from navigation systems at 30-second intervals. Information on these events, along with other location updates, are then communicated to a Central Location Server (CLS) of CRIS, using Mobile Satellite Service (MSS) as well as mobile data service. The CLS processes, hosted in the CRIS data centre, receive data and relay it to COA to automatically plot control charts. As the COA system is integrated with the National Train Enquiry System (NTES), Indian Railways can automatically disseminate this accurate real-time information to passengers.

The RTIS system is a great example of the impact of technological integration in data collection processes, to overcome issues stemming from manual entry, and minimising discrepancies and delays. The RTIS system has also been integrated with FOIS (Freight Operations Information System), ICMS (Integrated Coach Management System) and CMS (Crew Management System). It is also being used for single-touch emergency messaging from loco pilots to control rooms in the event of emergencies.

The data generated under RTIS helps facilitate train control functions in a timely manner and can also contribute to making decisions related to optimum crew booking & reduction in pre-departure detention (PDD). It is being leveraged to monitor the punctuality of trains, real-time tracking of locomotives, and location validation of locomotives in ICMS, FOIS, and COA systems, at the time of their attachments. Moreover, it is being used for drawing trains' speed profile charts as well as for analysis of un-scheduled stoppages of trains using data analytics and machine learning algorithms.

### Key outcomes

Besides improvements in the accuracy of reporting train speeds and locations, the RTIS system has facilitated several intangible benefits in train operations, such as better punctuality monitoring (as accurate train movement is fed into the system without any manual intervention); better and more focused train movement planning by section controllers due to the reduction in their workload of fetching train timings from station masters and manually feeding into the system; better crew planning as accurate train running information is available; and tracking of locomotives by loco sheds for loco maintenance.

However, administrative data may not always contain information on the quality of services and higher-order beneficiary outcomes. There is a need to supplement administrative data with other sources which can provide information on quality, experience and other outcomes of interest directly from beneficiaries, which may not be captured during the provision of services. All government departments have grievance redressal mechanisms to capture complaints and issues. One possibility is to leverage these existing systems to initiate outbound calls as well or visits to a sample of beneficiaries to collect information on beneficiary experience, field-level unanticipated issues, and outcomes. For example, most health insurance schemes now follow a system of calling patients post-hospital discharge to verify services delivered and collate information on the quality of care or expenses incurred. This provides nuanced and disaggregated information on outcomes and helps improve the effectiveness of government programs, as steps can be taken to address specific issues of malpractice or poor quality and timeliness of care.

Administrative data can be combined with different sources to understand the functioning of the programme as well as the impact of the monitoring mechanisms established. In Telangana, the government innovatively used beneficiary phone surveys to ensure timely disbursement of a farmer subsidy program (Case Study Box 6).

## Case Study 6: Leveraging administrative data along with survey data for program monitoring

### Farmer subsidy program, Government of Telangana

**Background**

The Telangana government launched the Rythu Bandhu ("Friend of a Farmer") initiative in May 2018 (Rythu Bandhu Scheme 2018, Telangana). In each of the two crop seasons that year, farmers received payments totalling INR 70 billion (or roughly 7% of the state's yearly budget). These payments, which were made in the form of cheques, were intended to help farmers pay for their purchases of fertiliser and seeds as well as pay off debts before planting season. Every farmer listed as a landowner in the government's digital land registry was the target audience for the campaign. A large-scale experiment was conducted to assess the effect of phone-based monitoring of the programme using high-quality administrative data on the entire universe of 5.7 million prospective programme beneficiaries.

**Use of administrative data**

In collaboration with the government, researchers randomly informed 25% of the state's frontline workers responsible for fund disbursement (122 of the state's 498 Mandal Agricultural Officers(MAO)) of the monitoring system and employed a call centre to reach out to farmers and collect payment information. To monitor the scheme's implementation, the government collected a variety of data points: frontline employees recorded cheque distribution, while banks recorded when farmers received their cheques. Officials from the programme then reconciled check distribution data from MAOs and funds disbursement data from banks. Finally, information from the calls and thorough administrative records were reconciled to establish whether and when farmers received payments and, as a result, frontline worker performance was measured.

“Though the performance reports created using data from the feedback calls themselves were not shared with the frontline officials due to paucity of time, the very announcement of these calls led to a change in their behaviour. The intervention led to a 7.8% reduction in the number of beneficiaries who did not receive their benefits. These effects correspond to an INR 300 million increase in transfers that were delivered on-time, an INR 78 million increase in the amount ever delivered, and 17,771 additional farmers encashing their cheques.” (Muralidharan et al. 2018)

This is an interesting example of combining administrative data from different sources to understand the functioning of the program as well as the impact of the monitoring mechanism established.

**Key outcomes**

Reconciliation of data from different sources enabled the identification of the extent of delays in the disbursement of funds and their potential underlying causes. The study observed that the changes in the performance of MAOs seem to be driven by the knowledge that they were being monitored since the performance reports were shared only after the vast majority of the disbursements were completed. [Muralidharan et al. 2018]

Finally, the data from the phone surveys indicated that in 88.6% of cases, farmers' self-reports on whether they had cashed cheques were consistent with bank records. It helped further validate the accuracy of the data reported on the disbursement and use of the funds transferred.

Such phone-based monitoring systems are in use in other Indian states, and their utility is being studied. First, researchers worked with the Government of Delhi on a high-frequency monitoring system to track the use of mohalla (primary care) clinics in Delhi, and subsequently to monitor COVID-19 relief efforts during the first nationwide lockdown in April-June 2020. Administrative data on patients visiting mohalla clinics was used to identify beneficiaries who are to be followed up with. The system generated a real-time dashboard on food security, economic well-being, and self-reported symptoms that was used to inform policy. The e-Public Distribution System (e-PDS) from Odisha is another example of the innovative use of smart tools and technology to minimise leakages in distribution. Such technology (e-PDS-based selling of ration which generates admin data) supplemented with a phone-based beneficiary monitoring system can effectively ensure timely delivery of food grains.

Compared to more specialised tools for monitoring public service delivery, such as smartphone applications or time clocks, phone call-based monitoring can be inexpensive, quick to deploy, and easy to adapt to changing circumstances. The data from feedback calls can be integrated with the routine performance reports to positively impact the efficiency of service delivery. Thus, a combination of administrative data and survey data can be used as a tool by the government to gain last-mile visibility of programme implementation and monitor front-line worker performance. Investments in high-frequency data collection which leverages mobile phone penetration can be useful in times of crisis to understand and quickly respond to the needs of the poor and vulnerable.

In summary, there are various benefits of using administrative data as it is more exhaustive in coverage, more accurate as it is less prone to sampling or respondent bias, can be collected much more frequently, and also allows for easier availability of temporal data over long periods. The examples and case studies provided above demonstrate how various government agencies have been able to make use of these advantages of administrative data by setting up robust data systems which could be easily used for decision-making as well as research purposes. In the subsequent section, some of the challenges and limitations to the use of administrative data are detailed.

## 2.2 Challenges with the use and interpretation of administrative data:

It is important to assess the shortcomings of the use of administrative data in analysis and policy research. These are briefly described below:

- Administrative data typically does not collect data to "explain" behaviours. Data on factors that influence behaviour change are extremely important to understand the situation, develop well-designed programs, and understand why a program is successful or not. Such data is typically collected through surveys or field studies. There is a need for other forms of data to supplement administrative data. For instance, whether a sanitation programme has been successful in inculcating better hygienic practices such as regular washing of hands is a question of behavioural change, which is not easy to answer through administrative data alone. Such an investigation may require field surveys, focus group discussions with beneficiaries, as well as an understanding of the local context and underlying theory of behaviour change.
- Since administrative data is collected for a particular population, aligning it with the objectives of policy research, as a representative dataset of citizens or an experimental sample, may be challenging in some instances. Very few programmes are universal and are therefore targeted at sub-groups of population or geographies. It is important to assess whether the descriptive data and trends observed for beneficiaries under a specific scheme are representative of a larger population. Statistical techniques can help determine whether the subset is representative of the larger population, and to what extent the observations and trends can be generalizable to the larger universe.

In addition, there are a few other challenges which emerge – not because of innate characteristics of administrative data, but as a result of setting up low-quality administrative data systems that are not fully responsive to the needs of policymakers and researchers. Some of these are explained below:

- Administrative data may be designed to be entered by humans using software on handheld devices, laptops, scanners etc. Human errors in data entry or shortcomings in the software such as lack of validations could affect the quality of data collected. Initial errors can be amplified and persist for a long time as software once developed and deployed can be difficult to recall/revise on a large scale such as in government systems.
- In case administrative data is maintained non-electronically, the manner in which it may be 'digitised' affects its timeliness and possible usage. For instance, if data is collected on paper at the school level and later on fed into digital systems by field workers at the district level, it can result in delays and data entry errors. It also exposes the data to quality issues such as missing information, internal inconsistencies, and invalid information. Further, if field workers aggregate data collected from multiple schools and only submit district-level aggregated data, there is a loss in the granularity of data, as a result of which any unit-level analysis becomes impossible.
- Even when administrative data is well collected, it is often distributed across databases and in silos within organisational structures. There is often a need to integrate multiple administrative data sources to gather useful insights from them. However, different data schemas, the absence of metadata and a lack of clear data exchange standards hinder interoperability. As a result, it becomes difficult to integrate different datasets for purposeful analysis. Often, data on similar indicators is collected in multiple databases, however, due to these constraints, there is no single source of truth.

It is hence evident that there are numerous advantages and use cases of administrative data in public policy. However, to unlock the potential of the use of administrative data for research and policy, it is crucial to design highly granular, near real-time administrative data systems that collect high-quality data with minimum human interference and can be seamlessly connected. The next section focuses on elaborating upon some good practices that can be adopted by government agencies to create such robust administrative data systems.

# Section 3:
# Suggested best practices in setting up administrative data systems to enable data use

In the absence of systematic processes, administrative data is often not well collected and/or is stored in silos in dated file formats and disconnected databases, which do not lend themselves to systematic data analysis. Often, cleaning and preparing data for analysis can take an exorbitant amount of time, which is often a constraint when there is a policy window to inform decision-making. The effective use of administrative data rests on maintaining data quality (Feeny et al., 2015). This entails maintaining accurate records and consistent processes to ensure data is cleaned and cross-checked, independent data audits to ensure high accuracy, and ensuring safeguards in places where subjects have incentives to misreport information by implementing additional validation checks.

A broad set of practices regarding data collection, quality, reliability, and sharing, when implemented well, can prove to be effective in making administrative data usable for effective monitoring, conducting research experiments, and linking with other sources of information (such as survey data).

This section seeks to focus on the best practices relating to such operational aspects around setting up effective administrative data use systems for guiding evidence-based policy. In this section, we aim to provide a set of best practices in technical areas in the form of compelling examples. There are a variety of successful initiatives by state governments, researchers, and other organisations that showcase innovative approaches and effective strategies in working with administrative data. We believe that practical and actionable lessons from these case studies can provide valuable information to other government and non-government institutions, data providers, and researchers on how to collect, use, analyse, and securely share administrative data.

## 3.1 Data Collection methods

The life cycle of data is the order in which it goes through different stages — beginning from its collection/generation, its management and use, to the eventual archiving. The first stage of the life cycle for any data system, including administrative data, is the collection or capture of data. While it is important to adopt standards at every stage of the data lifecycle, it is of particular significance at the data collection stage. Especially from a data quality perspective as issues that could pose a challenge to the effective use of administrative data can be prevented and tackled at this stage.

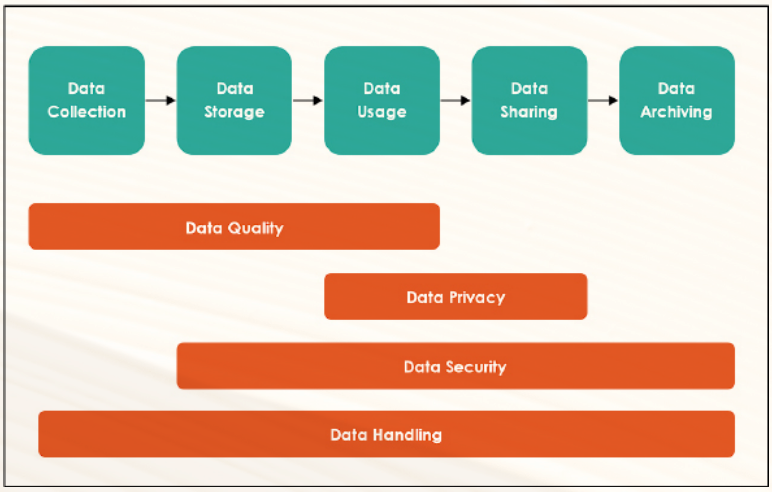

Figure 1: Data standards executed across each of the five phases of the data lifecycle

At the first stage, data collection and storage procedures must be clearly and meticulously documented. Proper oversight of the data collection and flow at the beginning of its life cycle is essential to maximise its usefulness and minimise the potential for errors (Murphy 2020).

### 3.1.1 Types of Data Collection

As we begin to think about the best practices for data collection, there is a need to first understand the underlying type of data collection that is employed in a given data system. Data collection in the context of administrative data can be categorised into two broad categories - active and passive data collection.

#### Active Data Collection

Data collection is termed as "active" when a human manually enters the data as part of the collection process. This typically involves the use of instruments like information forms (both digital and on paper). The data is entered either by the data subject (i.e. beneficiary) directly or by a data collector (i.e. frontline worker or data entry operator).

Below are a few examples of active data collection:

1. Unified District Information System for Education (UDISE) collects information from schools on school profile, physical infrastructure, teachers, enrolments, examination results, etc. through an online Data Collection Form (DCF) which is entered by a data entry operator.
2. Census surveys are an example of active data collection as the process of collecting data involves surveyors who visit households to collect data (usually on paper forms). The collected data is then entered into a computer by a data entry operator. More recently, data is collected using digital devices such as mobile phones or tablets by the surveyors on the field.

#### Passive Data Collection

Data is "passively" collected when the collection of data happens automatically without the involvement of a human actively entering data. This typically involves electronic sensors or other digital tools to capture the data such as the number of hits on a website.

Some examples of passive data collection are :

1. Electricity consumption data is collected digitally through electricity usage metres that collect information on the units of electricity consumed.
2. Transaction data for direct benefit transfers (for example wages under MGNREGA) is automatically generated and captured in the electronic system.

The use of ePoS devices to enable the use of digital transactional data for monitoring the Public Distribution System is another example of how human interference can be minimised at the data generation stage to increase the speed and reliability of data collection processes (Case Study Box 7).

# Case Study 7: Using digital biometric authenticated transactions for real time beneficiary level data generation

## Public Distribution System , Department of Food & Public Distribution

### Background

Dignified access to adequate food throughout the year is a basic necessity since food insecurity can have devastating impacts on human survival, health, and the ability to engage in productive work. To ensure food security, the Government of India runs one of the largest food security programs in the world through the Public Distribution System (PDS). Under this system, the government guarantees food grain access to over 80 crore people in India (almost two-thirds of the country's population as per Census 2011) as a legal right under the National Food Security Act 2013 (NFSA).

However, the delivery of subsidised grains throughout the year to highly vulnerable migrant beneficiaries remained a major challenge in the traditional PDS. As many ration card holders often moved between districts or states in search of better job opportunities, temporary employment, etc., they often lost access to the PDS food grains because their ration cards were tagged to specific Fair Price Shops (FPS) and cardholders were allowed to get the subsidised food grains only from the tagged FPS. Also, due to information gaps, it was often difficult for the migrant beneficiaries to locate a nearby FPS and check the Aadhar seeding status with their ration card. Beneficiaries also faced issues related to a lack of transparency about their transaction history and verification of their entitlement details. Further, when the migrants leave, the unclaimed food grains in the FPS in the migrants' native village or town were also prone to diversion by the FPS dealers, leading to leakages in food subsidies.

### Use of administrative data

The One Nation One Ration Card (ONORC) facility in the PDS addresses the above issues and empowers all beneficiaries under NFSA to seamlessly access their food grains from any FPS in the country by using the same ration card alongside biometric / Aadhar authentication through an electronic Point of Sale (e-PoS) device.

Along with ONORC, greater transparency has been ensured by automating FPS in the country. The e-PoS provides a medium to record and transmit the transactions, authenticate the beneficiary, and ensure that commodities are being issued to the intended beneficiary by biometric verification with the UIDAI server. Fingerprint scanners, IRIS scanners, and printers have been integrated for biometric authentication with UIDAI and printing receipts of sales.

In light of the increasing penetration of mobile phones and internet in the remote areas, the 'Mera Ration' App was also developed to facilitate access to ONORC-related services for National Food Security Act (NFSA) beneficiaries, including migrant workers, FPS dealers, and other key stakeholders, making it a single window information system for the beneficiaries. Using this multilingual app, the beneficiaries can find and locate the nearest FPS. This application is also linked to the Integrated Management of PDS (IMPDS) portal, allowing beneficiaries to view information such as food grain entitlement, transaction history etc. Beneficiaries and FPS Dealers can also submit recommendations or feedback using the application. The real-time feedback system nudges the FPS dealers to improve their service quality to the satisfaction of poor and vulnerable beneficiaries.

**Key outcomes**

The ONORC-based technology-driven reform has enabled migrant beneficiaries to get their entitled quota of food grains in full or part from any e-PoS-enabled FPS in the country. The power to choose any FPS proves extremely helpful to beneficiaries as they do not have to tolerate the FPS with poor service, weighment issues, erratic working timings, etc. This is aimed at infusing a sense of healthy competition among FPS dealers and thereby nudges them to improve service quality to the satisfaction of poor and vulnerable beneficiaries.

The 'Mera Ration' app is convenient for beneficiaries to avail of ration card services like information regarding their transaction history and food grain entitlement, irrespective of their location. More than 20 lakh beneficiaries have downloaded and used the 'Mera Ration' app. Since the inception in August 2019, a cumulative total of around 70 Crore portability transactions have been recorded under the ONORC plan, delivering food grains worth Rs.39,000 Crore as food subsidy, showing a high uptake of ONORC plan by the beneficiaries for accessing their NFSA and PMGKY food security entitlements during the pandemic. Around 3 Crore monthly portability transactions (including inter-state transactions, intrastate transactions & Prime Minister Garib Kalyan Anna Yojna (PMGKY) transactions) are being consistently recorded under the ONORC plan.

The automation of FPS has also led to the generation and collection of Ration Card data, FPS and e-PoS transaction data, and food grain distribution data. This data is maintained by the respective State governments, which then share incremental information with the central repository set up by the Department of Food & Public Distribution. This repository/system also acts as a gateway to check duplicate ration cards/beneficiaries in and across States/UTs through a continuous Aadhaar-based deduplication process. The Department continuously monitors the transactions and distribution of foodgrains through ePoS-based devices on the central Annavitran portal (www.annavitran.nic.in), which fetches intra-state portability transaction data from the servers of the States/UTs using web services. Similarly, the Department monitors inter-state portability transactions and all related aspects of inter-state portability via the IM-PDS portal (www.impds.nic.in). These portals/dashboards provide dynamic information/reports about the intra-state & inter-state portability transactions allowing for continual monitoring of the scheme's performance.

## 3.2 Data Quality

Data quality has multiple dimensions. A detailed description of the various dimensions of data quality is provided below:

### 3.2.1 Reliability

Reliability of data is the overall consistency of the data, i,e, the dataset produces similar results under consistent conditions (such as when a type of analysis is repeated). In many cases, administrative data may not be reliable for various reasons, such as:

a. **Data collection errors:** The procedure involved in data collection is prone to errors of mis-reporting, as human resources manually enter information or response into the system. There could be errors in spellings, incorrect addresses, incomplete information, etc.

b. **Incentives to mis-report:** In case that the performance evaluation of frontline workers is tied to outputs or outcomes, such as school attendance, immunisation completion, thresholds of malnourishment indicators, agriculture output etc, the staff in charge of service delivery may have an incentive to over-report or under-report outcomes based on the scheme benefits.

**Suggested practices to improve data reliability**

a. **Use independent data audits meaningfully:** The extent of reliability can be gauged through independent data audits. An independent data audit is a process by which a sample of the data entered into the database is collected separately by a small team (independent and different from government staff delivering the programme/ service and recording the data). Comparing this subset to the information contained in the larger database would highlight the nature and extent of discrepancies (Gibson 2021). For instance, in a research study conducted with the government of Haryana on improving full immunisation outcomes among children, the researchers used data collected by frontline health workers through a mobile app on the delivery of vaccines (Banerjee et al. 2021). To assess the quality and usability of the data collected, the research team set up an independent system of field audits (known as back checks) to verify the authenticity of the data being collected for a subset of the records. These audits were carried out by an independent group of trained surveyors who visited the households to verify the facts on the delivery of vaccines and the timeline. Comparing the two sets of data allowed the researchers to verify the accuracy of the data on vaccination coverage.

   Further, based on a comparison exercise, one can identify types of errors and most error prone variables. Additional training and monitoring, enhanced validation checks, or "truth-telling" interventions incentivising staff to report the "truth" can be implemented to ensure better data quality.

b. **Built in validation checks** during the design of the software/app used for data collection. The quality of the data can be improved by introducing validation checks at the data collection stage. Data validation checks are described as basic sanity checks on the data, i.e. identifying if there are unrealistic values in the dataset. This could vary from simple typos such as an extra zero in an age field, which makes someone appear 500 years old, or values such as 3213 for the year of birth, or the presence of text letters in a field meant for mobile phone numbers. Other types of errors may be due to manipulation, for instance, pollution levels of particulate matter may be reported by industrial plants as just below the threshold values. It is best to ensure that the data collection software or information system itself has a few data types, logic and validation checks to prevent errors during data entry. Examples include:

   i. **Data type check** ensures that the input data is of the type that is required in a field. For example, the name field should not contain numbers. Similarly, a field that captures net expenditure should not have letters as the input.

   ii. **Logical value check** ensures that the input data is within the logical boundaries of the indicator that is being collected. For example, an age field should not have negative numbers as the input. A latitude value should be between -90 and 90, while a longitude value must be between -180 and 180. Any values out of this range should automatically show an error message by the system as invalid.

   iii. **Automated inputs/ pre-filled data:** Dropdowns for geographical locations (district/ block/village) reduces scope for spelling errors during entry. Time and date stamps of data entry should be automatically recorded to capture any lags in data entry.

Such validations and data checks are especially useful in a digital data collection process as the system makes it impossible to enter invalid values thereby reducing the effort required subsequently to clean and organise the data before use in analysis, and makes the data more usable.

### 3.2.2 Completeness

Data completeness refers to the comprehensiveness of the data. A dataset is said to be complete if all the required data for identification and delivery of schemes and programmes is available in that dataset. Data can be classified as incomplete in three different ways:

a. **Missing beneficiaries:** Data may not capture information from all potential beneficiaries. There could either be an under coverage of data (i.e. the absence of target objects/missing objects in the source), or there could be an over coverage of data (i.e. the presence of non-target objects in the source). These are often described as exclusion or inclusion errors respectively. For example, in order to correctly identify the beneficiaries and address the problem of fake ration cards under the PDS scheme, a new list of people below the poverty line (BPL) was created (NCAER Evaluation Study 2015). The states took the opportunity to make the coverage more inclusive and brought in a larger section of the population (almost three quarters) under the benefit scheme. A few other states have also been experimenting with the use of Aadhaar and registries of other State-level schemes to combine PDS entitlements. These are fully biometric, universal and online systems being used to minimise inclusion and exclusion errors.

b. **Missing information:** Data may not be collected/entered for all the variables/ fields in the forms/web-applications for all beneficiaries. For instance, some basic information may be captured for the purpose of registration, but routine updates may be missing (such as changes in phone numbers, services accessed, etc).

   For example, in the PDS system, in order to address the problem of missing information with respect to the documentation of service delivery at the Fair Price Shops, ePOS machines were introduced. These machines helped create an accurate account of the sales and inventory, as well as increase efficiency of inventory documentation.

c. **Missing data fields:** Key variables required for decision making may not be included in the forms/web application. For data to be used effectively for decision making, there is a need to collect data on a variety of input, process, output, and outcome variables. Often, administrative data is a rich source of information on inputs and outputs, but could lack data on processes or outcomes. For example, the Unified District Information System for Education (UDISE) data collected on school details by the Ministry of Education collects information at the school-level on some input and output indicators pertaining to each school, its students, and teachers – such as demographics, enrollment and retention, and infrastructure. However, data on student learning outcomes are not yet available on the system.

**Suggested practices to improve completeness of data**

a. **Built in validation checks on software** (as described in the previous subsection) could help prevent some of the issues upfront. For instance, certain critical fields should be made mandatory for data entry so it can't be left blank. It is always useful to undertake short pilots with quick feedback loops to ensure data collection software and processes are robust.

b. **Undertake high-frequency data checks -** Routinely reviewing data for missing and incorrect information would help in quicker detection and correction of the issues. This practice is referred to as "high frequency checks" (Gibson 2021) and involves generating summary statistics using statistical analysis - percentage of missing data, plotting frequency distributions, estimating outliers etc. Therefore, it is important that apart from an IT/systems team, data analyst(s) are also present to check the robustness of data.

### 3.2.3 Interoperability

Interoperability is the extent to which data is capable of being integrated with other datasets. The promise of administrative data is amplified when datasets can be linked with one another. Most data is collected and stored in silos within specific implementing departments. The ability to link datasets, such as birth registry to immunisation data, or anganwadi enrolment data to primary school data, can enable policymakers to quickly identify beneficiaries who may get left out through any gaps while transitioning from one stage of life to another (Feeny et al., 2015). It is imperative that there are mechanisms to allow the linkages of datasets such as the use of common IDs. For example, an effort to link the Birth Registry with the mother and child health program data has the potential to reduce leakages (Case Study Box 8).

# Case Study 8:
# Improving comprehensiveness of data to ensure better coverage of potential beneficiaries

## Maternal and child health programs, Tamil Nadu

### Background

The state of Tamil Nadu wanted to streamline its healthcare service delivery and systems through data management and planning. For this purpose, the state pioneered the development of the Pregnancy and Infant Cohort Monitoring and Evaluation (PICME) system, a comprehensive information system to help monitor pregnancies to ensure appropriate healthcare services, provide additional resources and support for mothers, and ensure timely provision of benefits to them.

This system can be used to track and record the information about mother and infant – from the date of prenatal registration of the user, through the first year of their child's life. However, since the PICME data is based on the registration of pregnant women, potential beneficiaries - both women and their children may be missed out if women do not register.  Therefore, the state may not be able to ensure full coverage of its potential beneficiaries.

To overcome this concern, the government of Tamil Nadu linked the PICME system to the Tamil Nadu Civil Registration System (CRS), for the identification of "missed" children from the PICME database and the coverage of child health programs.

### Use of administrative data

Mothers can pre-register their pregnancy on the PICME web or application portal, which is verified through follow-ups by the local village or urban health nurse (VHN/UHN), who generates a Reproductive and Child Health (RCH) ID for the mother. After the preregistration has been approved or initiated by the nurse, and the RCH ID has been generated, the VHN/UHN completes all the required initial medical tests and enters the mother's medical history to register her into the portal. Henceforth, the mother and infant's health indicators are tracked throughout pregnancy and postpartum and inputted on the PICME portal during her visits to health subcenters, primary health clinics, or hospitals/medical colleges.

At the time of delivery, all government hospitals, medical colleges, and certain private hospitals with a high number of annual births are given their own login credentials and they have the ability to upload information about the delivery of the child and the first immunisation doses given to the child. This step is supported by the health nurses who ensure that this information has been uploaded by maintaining a second record in their files of delivery outcomes, and cross-checking this with the data on PICME. If there are missing PICME records, they coordinate with the hospitals and ask them to update the record at the end of the week.

After uploading the delivery information on PICME, the hospitals now have to create a 'birth registration' on the CRS portal with the RCH ID of the mother. The birth information will get auto-populated on this portal if the data tied to that RCH ID has already been added on PICME. The registration of the birth on the RCH ID is necessary for the families to obtain the birth certificate of the child.

**Key outcomes**

The PICME landscape is a good example of a means to ensure the completeness and comprehensiveness of data for overall data quality, as well as the benefits which stem from it. There are also other routine measures which help ensure that the data is reliable and of high quality – such as supervision of data collection and maintenance led by medical officers at primary health centres (PHCs), block review meetings led by block medical officers, and checking and tallying of data on a weekly basis. District officials compare and cross-check the PICME and CRS delivery data weekly to ensure there is no discrepancy in the aggregated numbers. The linking not only ensures the recording and monitoring of births, but also ensures that birth registration formalities, postnatal care, and immunisation is properly provisioned for the mother and child.

Further, the potential for use of the PICME data to track and assist high-risk pregnancies, delivery of scheme benefits, and regular medical checks, points to the promising scope of using data for informed decision-making and ensuring targeted service delivery at the appropriate level.

A health system can hence be transformed by appropriately managing, monitoring, and utilising well-connected health information systems.

## 3.3 Data Storage

After the collection of data, whether passive or active, the next stage in the data lifecycle is the storage of the collected data.

### 3.3.1. Types of storage (databases)

**Transactional databases**

Typically, administrative data is initially stored in transactional databases. Transactional databases are databases that are generated as a result of capturing day-to-day operations via an application. Transactional databases are optimised in a way that allows for multiple transactions and users to use the database at the same time.

Transactional databases[1] are well-suited for querying specific records in the database, for instance, the customer ID of a particular customer. Numerous such queries can be run simultaneously on a transactional database.

An example of a transactional database is the property tax system which provides a digital interface to the government and citizens to make property assessments, pay property tax, generate payment receipts and monitor tax collection. It is used by a variety of users such as citizens, local bodies, service kiosk centres, and field employees to accomplish their specific tasks. Each user can be tracked across multiple tax transactions using a citizen or user ID. The citizens can file for an assessment of the property, search for a property, or transfer ownership. The citizen can also track down the status of their incomplete assessment. The government users can edit the details of the last assessment; generate demand notices, collect payments and receipts and monitor all transactions via dashboard and reports.

**Analytical databases**

On the other hand, analytical databases are optimised to run analytical processes and not just transactional processes. These databases store historical data (i.e. data that is archived) in one centralised inventory. Analytical databases are created by moving the data from transactional databases to a centralised database for long term storage and retrieval.

Analytical databases are well-equipped to compute complex aggregations involving multiple tables. An example is the data system used in the District Development Coordination and Monitoring Committees (DISHAs), formed to ensure better coordination among all the elected

[1] Ministry of Electronics and IT, Government of India. 2022. India Digital Ecosystem Architecture 2.0. Accessed at https://www.meity.gov.in/writereaddata/files/InDEA%202_0%20Report%20Draft%20V6%2024%20Jan%2022_Rev.pdf

representatives in Parliament, State Legislatures, and Local Governments (Panchayati Raj Institutions/Municipal Bodies) for efficient and time-bound development of districts through proper implementation of 43 Central Schemes/ Programmes. The DISHA dashboard, hosted by the Ministry of Rural Development, uses the architecture of an analytical database by integrating historical data across all quarterly review meetings at the district level and storing information on the demands and issues that require follow up during the deliberations at the DISHA meetings. The data and analytics (such as number of households mobilised across years into SHGs under the DAY-NRLM scheme, employment generated under MGNREGA, connections installed under PMUY scheme and so on) are stored and published on the dashboard for read-only consumption by various decision makers.

Since analytical database systems are optimised for analytical processes, multiple users (like data analysts and researchers) can query and fetch data as well as run analysis on the database in a concurrent manner.

Analytical databases store beneficiary databases with a layer of decentralised program-specific monitoring modules. Governments can hence use both recorded cross-cutting data on beneficiaries and programme-specific descriptive analytics on key performance metrics to inform policy. For instance, when it is not possible to link an individual beneficiary in the pension beneficiary database with the Antyodaya or BPL database, the data can be aggregated at the village/block/district level to identify villages where these numbers differ. The reasons for such a mismatch can subsequently be investigated via substantive checks.

### 3.3.2 Metadata Cataloguing

Metadata is the technical information that describes characteristics of the data. It refers to descriptive attributes of a dataset, such as the number and nature of data fields, types of data collected, size etc.

As part of metadata, in addition to “what” data is stored, it is also crucial to document “how” data is stored and any transformations that are applied on the data after its initial storage. These transformations could be processes like

- Converting data of birth to age
- Calculating distance between two geographical locations using the latitude and longitude data
- Calculating the interest amount based on the principal and interest rates data stored in the database

Instituting standards of documenting metadata is therefore crucial for making it easier for administrative data users to access, understand, and use the data. This documentation includes catalogues and inventories of all administrative data collected in a government department or ministry. Metadata documentation for each individual dataset can also make it easier to assess the potential to link datasets.

**Suggested practices to improve metadata documentation**

a. Data cataloguing should be carried out from a data use perspective. This implies that the fields chosen to be documented should include those that will allow users to verify the usability of the existing data including its potential for linkages across datasets.

b. Key fields to document, therefore, include level of observation, frequency of updation of data, eligibility and outcome variables, period of data availability etc. This purposive approach is markedly different from mechanically documenting all aspects of datasets.

c. Metadata standards should ideally be set at source by the people involved in setting up the data system – to adhere to a specific or notified standard for necessary fields to ensure interoperability. In the case of government departments, metadata standards for some indicators are often already defined. For instance, the Ministry of Electronics & Information Technology has already documented metadata standards for several fields as part of e-Governance Standards & Guidelines. When any government department or ministry sets up any data system, they must ensure adherence to these notified standards for necessary fields to ensure interoperability.

d. Systematic documentation of metadata must also ensure that information on transformations of data is also well-documented to reduce errors and inconsistencies.

## 3.4 Data Protection and Privacy

The protection of data and the privacy of individuals are critical prerequisites for the effective storage and use of administrative data (Feeny et al., 2015). High-quality policies and practices that govern data protection and privacy generate trust from all stakeholders and thus enable data use to answer critical stakeholder questions and inform decisions.

Administrative data, because of its predominantly transactional nature, often contains particularly sensitive information, such as names of scheme beneficiaries and their family members, mobile numbers, unique IDs (for example, PAN number, Aadhaar number etc.) and sometimes even financial and biometric details.

While linking disparate datasets using unique IDs can unlock a plethora of opportunities for analysis and research and result in richer and novel insights, there is utmost need to ensure that there is no threat to the privacy of individuals or misuse of their personal information. Hence, there is a need to institute data protection procedures and privacy safeguards when handling and sharing administrative data.

### 3.4.1 Data Protection and Security

Data protection is the practice that includes processes, strategies and controls that physically and technically safeguard data from getting corrupt or lost. Data protection is also important in ensuring the integrity of administrative data.

Data integrity refers to the practice of maintaining the accuracy and completeness of data throughout its life-cycle. This makes sure that the data remains unchanged and accurate from the point of collection to its final stage of data archival.

Data security is the practice of defending data and keeping it secure against internal and external malicious threats from unauthorised users such as virus attacks, cyber attacks, data breaches, etc.

**Suggested practices to ensure data protection and security:**

**During Data Collection:**

In the stage of data collection, it is important to incorporate controls on the device(s) that are being used for collection or data origination. Examples of such controls include passwords, threat scanning for antivirus, etc. Further, a data backup or replication strategy should be set in place (where possible securely) to recover data in the event of a loss of the collection device(s).

**During Data Storage:**

While storing data, it is important to create a restriction on the access of the stored data i.e. who can read and overwrite which type of data. It is imperative to create data backup and store data on cloud servers with good disaster recovery options to avoid data loss. It is also crucial to develop adequate data archival methods for managing historical data. For this purpose, the database should be well equipped to generate audit trails that have detailed information on when was the data collected and by whom, what changes were made to this data at what stages and by which users. It is important to have regular security audits of the data system to ensure that the system is checked for any vulnerabilities and the same are mitigated via necessary corrective measures.

**During Data use and sharing:**

While using data, it is imperative to control access and encrypt it to ensure that different types of users can only accessthe data that they can and need to use. For instance, sensitive information about one department should not be accessible and used by another department. It is also important to have supporting policies and procedures in place to ensure that collected data is used only for its intended purposes with the prior consent of beneficiaries.

### 3.4.2 Data Privacy

**Data privacy refers to the protection of personal data from unintended users or for unintended purposes.**

The personal information contained in administrative data has the potential to be used to identify individuals or groups which could be, for example, used to target them politically, socially, or commercially with malicious intent. It is therefore important to create safe data and ensure it is not misused, and the right to privacy of its participants is not violated.

There are three main "levels" of identifiability as described below (IDEA Handbook, Vilhuber et al. 2020).

- Directly Identifiable Data

In "directly identifiable" data, the identifiers of personal information are present directly in the dataset. What this means is that the data is not coded i.e. personal identifiers are not replaced with any other code.

- De-identified or Pseudonymised Data

Data is termed "de-identified" or "pseudonymised" when personally identifiable information is removed and replaced with a code. The code is linked to the identifying data in a separate document which is known as the key.

- Anonymous Data

Anonymous data is similar to de-identified data in the fact that it also involves the removal of personally identifiable information and replacement with a code. The critical difference between anonymous and de-identified data is that the code is not linked to the personal identifying information in a key document. To completely anonymize a dataset, any fields that can potentially be used in any combination to identify an individual must be either removed, suppressed, or aggregated. This can ensure that the data cannot be used to re-identify the individual.

Instituting data privacy can help identify and address risks associated with the collection, use, and sharing of personal information, and ensure that administrative data is used responsibly and transparently. To begin with, it is imperative to have robust data storage protocols in place, including tools like encryption, and limiting access to the data to a need basis. It is also important to remove all personally identifiable information of individuals from the data before sharing it for analysis.

However, it is not sufficient to only de-identify a dataset to ensure privacy. It is a part of the due diligence of the data collector, researcher or policymaker to ensure that linkages between indicators do not identify vulnerable groups. Many statistical processes and codes can be used appropriately for this purpose. For example, in certain cases, it may be crucial to only report aggregated data trends among certain groups to safeguard individuals' privacy.

Another mechanism to ensure data privacy is to share an artificially synthesised dataset without sharing the actual information of the participants. Additionally, any department handling sensitive information needs to formulate anti-hacking systems in place to ensure that the data is not accessed or tampered with due to the vulnerability of an MIS.

A detailed list of data privacy measures available is described in (IDEA Handbook[2] , Chapter 2 "Physically Protecting Sensitive Data", Schmutte and Vilhuber 2020). The State governments of Punjab and Tamil Nadu, among others, have developed extensive

[2] Handbook on Using Administrative Data for Research and Evidence-based Policy: Aims to provide researchers and data providers with guidance on best practices in legal and technical areas; using a set of compelling examples on a range of topics: drafting data use agreements, cleaning and linking data sets, implementing secure computer systems and managing the data infrastructure, designing an application workflow for granting access to multiple researchers, analysing data for decision-making, and facilitating collaborations between researchers and data providers.

### Example 4: Punjab State Data Policy 2020 and Tamil Nadu Data Policy 2022

Punjab and Tamil Nadu are two states that have made progress in data digitisation and are now working to institutionalise data policy frameworks for data collecting, sharing, storage, and use.

The Punjab State Data Policy 2020 (PSDP) and the Tamil Nadu Data Policy 2022 (TNDP) seek to balance the twin objectives of data protection and facilitating secure data usage for evidence-based decision-making. With the overarching aim of using data for the public good, the policies take cognizance of the principles of openness and interoperability, while trying to balance it against the right to privacy.

The PSDP classifies administrative data into open access, registered access, and non-shareable. Particularly, the policy also details security audits at the state data centre, periodic analyses of system audit logs, and has implemented data back-up and recovery processes to ensure data protection in government systems.

The TNDP has provisioned for the Tamil Nadu e-Governance Agency (TNeGA) to provide data support to line departments and the Planning, Development and Special Initiatives Department of the state to ensure data protection and privacy.

guidelines relating to data protection in their respective data policies (Example 4).

## 3.5 Data Use

To implement a program or make a policy, the government is required to make a variety of decisions. This decision-making process often flows in the form of a cycle. It begins with the primary decision of identifying an area in the current system which requires a change and then moves on to an array of decisions behind designing, implementing and in turn evaluating the impact of the respective program.

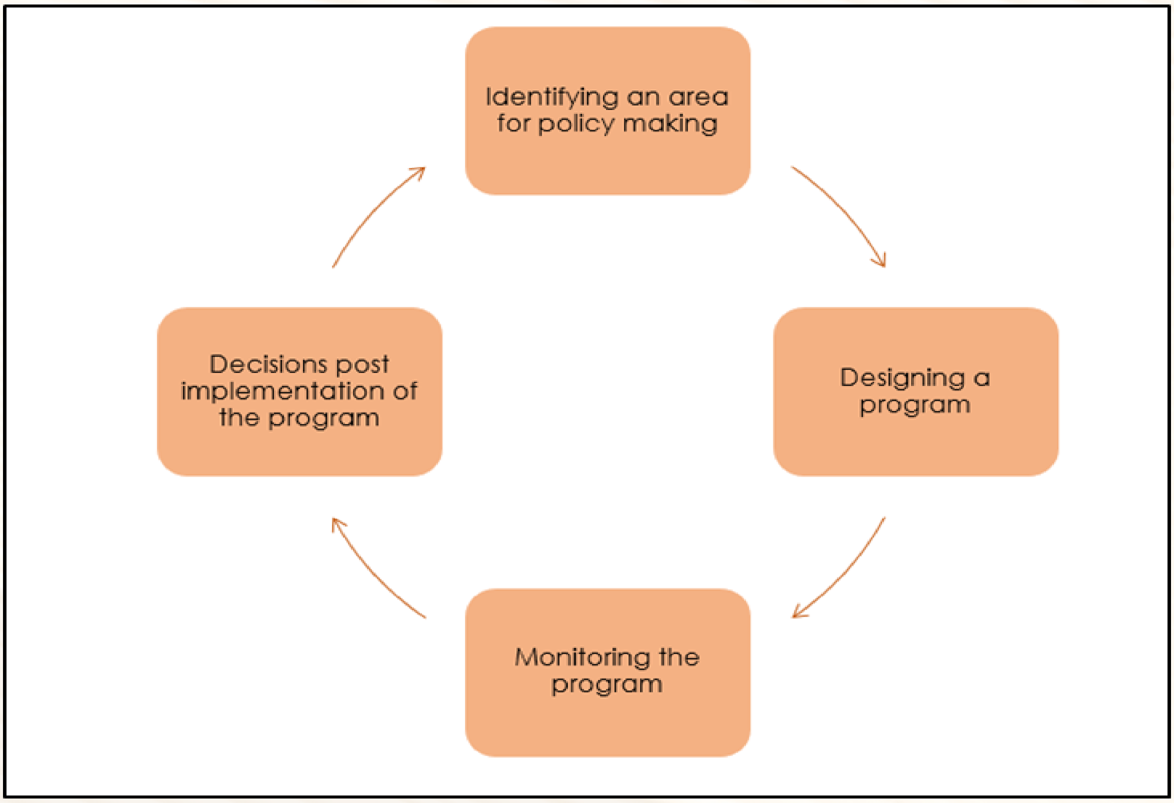
Identifying an area for policy making
Designing a program
Monitoring the program
Decisions post implementation of the program

Figure 2: Policy Lifecycle

The scope and benefits of the use of administrative data in each step of the policy making cycle are explored below:

### 3.5.1 Identifying an area for policy making

A key step in the policymaking process is identifying an area in the existing system which requires improvement or change. The use of administrative data helps substantially in this process as it is a complete representation of the population, unlike sample surveys. For instance, using administrative data from the State health insurance programme of Rajasthan called Bhamashah Swasthya Bima Yojana (BSBY) on the insurance claims on dialysis patients, combined with a sample survey of beneficiaries, found that hospital services were underutilised and mortality rates were higher during Covid-19 associated lockdowns (relative to the previous year). (Jain and Dupas 2022). It was hence identified that patients may need reminders or logistics support to catch up on non-Covid critical care that they may have missed over the past couple of years (other details in Case Study Box 9).

## Case Study 9:

## Using administrative data to identify areas for improvement in social security programmes

**Health Insurance Programme, Rajasthan**

**Background**

The Government of Rajasthan (GoR) launched the Bhamashah Swasthya Bima Yojana (BSBY) program in 2015 to increase equity in healthcare access, utilisation, and outcomes. In accordance with the program, the hospitals are paid by the government to provide free secondary and tertiary care to poor households. Researchers worked alongside the Government of Rajasthan to understand and analyse several aspects of the BSBY implementation over the first three and a half years of the program.

**Use of administrative data**

For the study, trends in the administrative data from BSBY insurance claims, i.e. information/variables notably unique transaction ID, filing date, all packages of care claimed in the visit and the associated package rates, patient name, patient contact information, admission and discharge dates, and hospital name and location were observed. Further, to conduct a spatial analysis of access, locations of hospitals empanelled in BSBY were geocoded using their names and the Google Maps API. Such data helped analyse program access and utilisation over time.

The researchers found that service utilisation was low among children, particularly young girls and that hospitals still charged patients out-of-pocket. This study helps highlight the importance of maintaining and using a repository of data to understand the utilisation patterns, especially in times of need, to identify areas of intervention to improve awareness as well as the take up of the BSBY program.

**Key outcomes**

As per the analysis of the administrative data, it was understood that service utilisation was low among children, particularly young girls and that hospitals still charged patients out-of-pocket. It highlighted the need for better awareness and education among vulnerable groups (such as women in this instance) on benefits they can avail of under the health insurance program. This way, administrative data can help in identifying areas for improvement in existing policies and programmes where there may be a need for mid-course corrections or new interventions.

### 3.5.2 Designing a program

To design a program or an intervention, a critical step is to identify a target population which would benefit from the program. To accurately identify a target population, detailed information is required on the general population, from which a target group can be sectioned out.

A lack of information on the target population could lead to an exclusion of potential beneficiaries and the inclusion of individuals who may not need the programme. A consequence of these two cases would be a diminished effectiveness of the program.

One such example is the Public Distribution System (PDS) dataset. PDS is a food security system, established by the Government of India, to distribute food and non-food items at subsidised rates to citizens below the poverty line. The PDS dataset here serves as an administrative dataset which provides information on a target group.

In the state of Tamil Nadu, the PDS database is linked to the state health insurance scheme. This means that if a citizen is registered in the PDS, they are eligible to apply for health insurance by using their PDS ration card. In this case, the PDS database, linked with the state health insurance scheme, provides a near-universal representation of the vulnerable class. Such use of administrative data can provide a better understanding of the target beneficiaries and reduce exclusion errors, especially when combined with survey data or other sources of socio-economic administrative data.

In addition to minimising exclusion errors, the existence of administrative data and its routine analysis can help generate new ideas for the design of innovative and impactful policies as evidenced by the example provided in Case Study Box 10.

### 3.5.3 Understanding the effectiveness of different options before launch of programs

It is important to conduct continuous impact evaluations concurrently along with collecting continuous and good-quality administrative data at the time of launching a program. This will enable policymakers to scale up or discontinue a program based on how it is delivering its desired impact. The use of administrative data for evaluation is beneficial since the data is available continuously, thus removing the need to conduct large-scale primary data collection surveys to track the progress of the programme, and could be used on a need-basis for specific purposes.

For example,  if governments have to decide which educational or health program works at a scale much larger than what is typically included in a research study; the government can randomise programs and then leverage administrative data collected by frontline workers such as teachers and nurses to assess impact "at scale". For instance, a large-scale immunisation study leveraged administrative data collected by frontline workers was used to measure the impact of interventions aimed at encouraging the take up of full immunisation services (Banerjee et al. 2021) (Case Study Box 11).

# Case Study 10:

# Using administrative data from environmental audits to design ratings system to increase transparency in Environmental Regulation

**Maharashtra Star Ratings Scheme, Government of Maharashtra**

## Background

A large body of evidence shows that particulate matter air pollution is harmful to human health and shortens life spans. According to the World Health Organization, it affects more people than any other pollutant, and there is no threshold below which no harm to health occurs, especially to children and infants. Given this, the requirement for a thorough and continuous evaluation of emissions by industrial plants has become imperative over the years.

With this in mind, the Maharashtra government launched a pollution disclosure and environmental performance ratings scheme for in dustrial plants in 2017, known as the Maharashtra Star Rating Scheme. This scheme sought to understand whether increased transparency about industrial plants' particulate matter emissions could improve the enforcement of environmental laws and reduce pollution in India.

This project represents the first initiative in India in which plants have been mandatorily rated based on legally actionable government tests of pollution, the ratings of which are then publicly disclosed on the program's website. Further, through the process of releasing actionable and understandable information to the public, the scheme aims to create a sense of transparency and improve regulatory accountability. The presence of publicly available information could create pressure on the industry from the customers, the employees, the regulators. as well as the affected citizens, and thus bring about a change in behaviour.

Researchers worked alongside the Maharashtra government to conduct a large-scale pilot involving the design, implementation, and evaluation of the scheme, covering 1000 of the most polluting plants in the state. From this initial population, a randomly chosen subset of "treatment" plants was gradually phased into the ratings scheme. Those plants not selected for initial inclusion in the ratings scheme formed a statistically identical "control group" that could be compared with rated plants in the treatment group.

## Use of administrative data

The administrative dataset used was the Air Sampling Inspection Dataset. The data here was collected as a part of the environmental audits of the industrial plants. These audits were either routine, as a response to an industrial plant applying for consent to operate, or as a follow-up to a violation discovered in a previous inspection.

The process of data collection was standard across every industrial audit. A probe would be inserted inside the chimney of the industrial plant, through which a sample of the particulate matter would be collected. The weight of the probe would be compared before and after the insertion. This data would then be completely digitised for documentation and analysis. The data would then be sent to the laboratory of the Maharashtra Pollution Control Board for analysis, and the star rating calculation would be done by the project team.

This study proves to be a great example of identifying the potential uses of data towards research and policy impact. The star ratings were made publicly available through an interactive website, where users can leave comments, complaints, suggestions, or initiate discussions.

The website was further promoted through outreach activities to spread awareness among the public. The evaluation of the environmental performance through the mode of ratings under the Maharashtra Star Ratings Scheme served as a public good. It was seen that being rated publicly significantly increases the probability that a plant receives a legally actionable and formal communication from the regulator, relating to the pollution they produce.

**Key outcomes**

The scheme has demonstrated the feasibility of utilising administrative data to encourage public disclosure of information on the pollution of industrial plants as a tool for environmental regulation. Based on the positive media coverage received by the Maharashtra pilot, two other states - Odisha and Jharkhand - have launched their own Star Rating Program in a bid to curb air pollution. These initiatives are aimed at informing residents and industries, and strengthening the regulatory efforts of the pollution control boards to reduce pollution.

# Case Study 11:

# Using good quality administrative data collected by frontline workers to improve service delivery

**Improving full immunisation coverage, Government of Haryana**

### Background

Immunisation is a highly cost-effective way of improving child survival. However, over two million children around the world die each year from vaccine-preventable diseases. Governments across low and middle-income countries have made immunizations free and have focused on strengthening vaccine delivery through mobile camps and frequent immunisation drives.

Despite large investments to increase access, as of 2016, only 62% of Indian children were fully immunised. A large fraction of children receive the first vaccine but do not complete the full schedule, reflecting high initial motivation but difficulty in completing later vaccination visits.

With this in mind, researchers collaborated with the government to conduct a randomised evaluation in seven districts of Haryana, with especially low vaccine coverage. The study involved 140 primary health centres (PHC) and 755 subcenters and focused on children under 12 months of age receiving five basic immunizations (BCG, Penta-1, Penta-2, Penta-3, and Measles-1).

The study was conducted from 2016 to 2018, where the researchers evaluated the effect of three policy tools on increasing demand for immunisation: small incentives, targeted reminders, and disseminating information on the importance of vaccinations through local immunisation ambassadors.

### Use of administrative data

The nature of the interventions - incentives in the form of mobile credits, conditional on vaccination, and reminder messages to caregivers - necessitated timely and accurate data on vaccines administered and phone numbers of caregivers. Since only physical documentation was maintained in 2016-17, the researchers decided to deploy a frontline worker data collection application that nurses/ANMs can use at the point of service delivery to record vaccine information as well as caregiver numbers. Such an application was in line with mHealth applications that were being piloted across the country at that time.

The administrative data thus collected on vaccinations for children proved to be quite exhaustive and accurate, and therefore, the researchers used the same data for the final outcome analysis.

Since accurate information on vaccinations and phone numbers was key to the effective implementation of interventions; the research team set up a system of field audits to verify the authenticity of collected information for a subset of records. These audits were carried out by an independent group of trained enumerators who visited households to verify information regarding the child's name, date of birth and most recent vaccination.

### Key outcomes

By using the administrative data collected via the application, it was found that a combination of targeted reminders and information dissemination using hubs was the most cost effective way of improving immunisation rates.

While the study relied on data collected by the ANMs using a mobile application developed by the research team, the Government of India has developed an extensive portal to track services and outcomes related to maternal and child health called the Reproductive and Child Health (RCH) portal. State governments typically have state-level information systems such as the Community Health Integrated Platform (CHIP) in Rajasthan or adopt the ANM Online (ANMOL) application developed by the Government of India to collect granular information on vaccination and caregiver details. Data collected through such systems can help measure the impact of new innovations and interventions on improving health outcomes.

### 3.5.4 Monitoring the program during implementation

The most important type of decisions that need to be made in the process of policymaking is during the implementation of the program: by continually measuring progress on outcomes and also identifying any issues/gaps for course correction.

Citizen services portals are a great example of the use of administrative data to monitor outcomes. To enable the efficient tracking and delivery of citizen services, the Government of Punjab used a two-pronged strategy: a) designing an agile data and monitoring system; and b) creating configurable workflows at the beginning of an application request. Such a system ensures the collection of granular data which can be used to understand the factors influencing service delivery and thereby provide insights on improving implementation (Case Study Box 12).

# Case Study 12:

# Using administrative data  for efficient tracking and delivery of citizen services

**eSewa Kendra Service Delivery, Department Of Governance Reforms, Punjab**

### Background

Punjab e-Sewa is an online portal and mobile application developed by the Government of Punjab to enable the seamless delivery of citizen services. The e-Sewa facilitates various online service requests with multiple state government departments in one place and aims at providing faster processing of public cases/appeals/grievances.

It was created to provide easy and convenient services to the citizens through remote access, primarily through common service centres known as e-Sewa Kendras. The major objectives include the integration of data across multiple departments, optimization of manpower and resources engaged in the service delivery mechanism, and providing an efficient and cost-effective method for service delivery to departments.

As a result, a huge amount of data is generated in the processing of these service requests that can help provide insights into the efficiency of service delivery in the state. The entire e-Sewa data system is managed by the Department of Governance Reforms (DGR) in the state.

### Use of administrative data

The e-Sewa Kendra administrative data is used to monitor the performance of citizen services and to track and deliver each service within a given timeframe[3].  The e-Sewa Kendra database captures transactional information on 351 active services that span 28 government departments. Citizens are required to file applications for their respective services at Sewa Kendras. Each application passes through multiple stages under officials, as mandated by protocols. DGR can track actions taken by officials for a specific application, with action timestamps. Roughly 30% of the services are 'online' services, which implies that there is end-to-end computerization, and the entire process of application approval is paperless with every action taking place digitally.

eSewa team uses an agile methodology for the development and deployment of the e-Services. The team conducts 3-hour long monthly meetings where new services and their associated challenges are discussed and solved. There are designated Scrum masters for every project and each Scrum Master conducts frequent meetings (weekly and daily) on every new issue or new request. These meetings often discuss insights from the data collected, outliers, issues in data validation, new rules to be applied, and top and bottom-performing geographies/personnel based on the workflow data in the system. Data reliability and quality have been expected to have improved over time after using this methodology, according to the official personnel interviewed.

As a major practice, the DGR has created a workflow engine for each service. Each predefined workflow is configurable for every actor involved in the processing of the application, and is mapped to the actions they can perform. The roles and responsibilities of each actor are clearly articulated depending on the type of application filed. The entire trail of acceptance, rejection, redirection, verification, and approval is stored and tracked as a complete workflow. The workflow also has other configurable rules present, such as criteria for enrolment under a scheme or service (targeting), mandatory documentation required, standard duration of application processing time etc.

[3] Department of Governance Reforms and Grievances, Government of Punjab. e-Sewa. https://dgrpg.punjab.gov.in/home/projects/e-sewa/

The DGR has implemented a distributed database architecture, as the ownership of the entire service-level dataset lies with the respective departments. DGR only manages and integrates the information required for service delivery. The classification of the services into three categories, namely completely online, completely offline, and hybrid, has enabled customising the workflows, setting up protocols, and integrating the verification and digital signature-based systems within the department seamlessly, especially if there is a redirection within the same application that uses a hybrid service (both paper and digital).

A rigorous self-paced training module and repository are available for the operators and their managers as a combination, adding to the accountability within the system and ensuring complete absorption and testing of the application before the personnel operate the live system.

**Key outcomes**

There is a potential for integrating services across departments and tracking delays or time taken for each application. The insights from the service data generated for each application and scheme can be used for better planning and smoother service delivery. This can be done, for example, by deploying a certain number of eSewa Kendra windows in a particular geography, analysing and benchmarking the performance of personnel during application processing, and frequent analysis of this connected data to ensure that the targeted beneficiaries are being serviced. Future possibilities include a citizen being able to track the complete lifecycle of an application and get in touch with multiple departments through a single sign-on.

### 3.5.5 Summary

In summary, the use of administrative data in each of these decisions proves to be beneficial in the following ways:

- The presence of near-universal data, provided by the use of administrative data, leads to better quantification of the impact and allows for more types of analysis than sample surveys. For instance, it also allows users to analyse state and district-wise variations in programs which may not be possible in some surveys which are only representative at the national level.
- Administrative data acts as a tool to study historical data over time. It can easily provide information on the same individual or case over long periods. For example, this is particularly beneficial if a study is required to understand change in a recipient's behaviour and/or well-being in response to some program.
- Administrative data also provides greater accessibility of monitoring information for administrators to base implementation decisions on. For instance, evidence suggests that using administrative data for monitoring program delivery has reduced delays in the payments of MGNREGA wage payments in Madhya Pradesh and Jharkhand (Case Study Box 13).

# Case Study 13:

# Administrative data improving accessibility to "actionable" information relating to service delivery

### Reducing delays in MGNREGA payments, Madhya Pradesh and Jharkhand

### Background

The ability of government officials to supervise social protection payments, including accurately identifying the causes of payment delays and establishing who should be held accountable, may be constrained in the absence of readily available information. A study was conducted to determine if easier access to monitoring information at various administrative levels could help social protection programme payments promptly reach their intended recipients.

The administration of MGNREGA is carried out by three levels of officials – local-level officials who verify the work and request wage payments, block-level officials who manage local officials and release wage payments, and district-level officials who have an overarching administrative role. In partnership with the Ministry of Rural Development, researchers conducted a randomised evaluation of PayDash, a new internet- and mobile-based management and monitoring platform for MGNREGA wage payments. It tracks when each step in the payment process occurs and generates real-time information on delayed payments, along with information on employees responsible for each administrative step. While information relevant to payment delays is accessible through the MGNREGA website, PayDash presents this information in a more accessible and actionable format to government officials who are trained to use it. The evaluation took place in 73 districts in two states (Dodge et al. 2018), Madhya Pradesh and Jharkhand from April 2016 early 2017 through to March August 2018.

### Use of administrative data

The project used data from PayDash on the time taken to complete each step in the MGNREGA administrative payments process and the overall payment processing time under MGNREGA. An internal monitoring dashboard to collect real-time data on PayDash usage was also employed, by making use of real-time Google Analytics data for the mobile (Android) and web applications. This usage data showed how many sessions (grouping of individual page views within a specific timeframe) each user had on each date and the duration of each session. Additional data showed how many "cards" related to specific subordinate users viewed on the mobile app on each date and whose card they viewed, when users used the call or direct WhatsApp message functions on the mobile app, the duration of the call, and who they called. The usage data also contained a unique identifier that was used to link with data on each trained official. In addition, researchers collected wage payment data at the block level sourced from the MGNREGA public website and implemented surveys to collect additional information on the officials, including demographics, personality traits, work and management practices, and more (Dodge et al. 2018).

The PayDash infrastructure, the associated dashboard and the tracking application provide an innovation in data collation and visualisation that makes critical actionable information easily available to the administration. In particular, the app is an actionable tool in the hands of officials, helping them track the performance of their subordinates, and fixing accountability at a larger scale.

The application, which also functions offline, serves as an intuitive, user-friendly and easily navigable tool, available in English and Hindi, to make information readily accessible where it is required.

### Key outcomes

Overall, researchers concluded that PayDash significantly reduced MGNREGA wage payment delays in areas with the longest delays prior to the intervention, when provided to either block- or district-level officials. Its use was even higher when senior officials were also provided access. Providing PayDash to block-level officials reduced wage payment delays by 2 days on average, representing a 28% reduction in delays over the comparison group (Dodge et al. 2018). In addition, researchers found that higher use of PayDash among block- and district-level officers was associated with faster completion of the administrative tasks involved in processing wage payments. More generally, the study shows the importance of easing access to monitoring data, especially for the agents who are in the best position to act upon this information. Researchers are continuing their research by examining the long-term impacts of PayDash on payment processing and the performance of MGNREGA, while other states such as Bihar are keen on exploring the impact of this innovation in their systems.

# Section 4:
# Lessons, Challenges, and Potential

There are many ways in which high-quality administrative data can be collected and effectively used for decision-making by governments, as showcased in the examples in the preceding sections.

Consequently, central and state governments in India have often expressed their commitment to a data-driven approach to policymaking, and are undertaking several types of digitisation efforts to this end. Internal Management Information Systems (MIS) have been developed for most government programmes. Many schemes also have interactive dashboards with basic data analytics to make complex information available to decision-makers in the form of simple visualisations and reports. To make government data publicly available for research purposes, several open data portals such as data.gov.in and National Data Analytics Platform (NDAP) have been launched. Intra-government data exchange and integration of data is also being facilitated via various platforms and initiatives.

However, there remains a tremendous unmet potential for the innovative use of data in decision-making due to the persistence of several systemic challenges. The availability of disaggregated data is still a challenge for several indicators where only aggregated data is collected at the state or district level due to weak data-capturingng mechanisms.

Owing to the limited use of new technologies that allow for quicker or transactional data collection, administrative data is also often not updated frequently. Further, government data could often be scattered in organisational silos. The use of different data formats, some of which are non-digital or non-machine readable, makes it difficult to access and process data. The absence of unique identifiers and technical standards also inhibits the analysis of multiple datasets. Even if data is available, it is not customised for use by different types of users who have different data needs. For instance, frontline workers, district administrators, and state-level officers have varied needs, and hence require different types of applications and visualisations to use the same data. Consequently, even though governments collect a lot of data, it is not available as per the requirements of administrators, and hence, not proactively used in decision-making. Consequently, citizens also have limited information on what personal information about them is being collected and how it is being used.

A combination of systemic policy, process, and cultural changes is required to bring about and sustain an institutional culture of using data and research to inform policy design and strengthen the delivery of government services.

Experience from NITI Aayog's DGQI initiative suggests that it is important to routinely collect indicators of data quality, showcase best practices for adoption and provide actionable inputs to Ministries to strengthen their data systems and feedback loops to policy decisions.

CLEAR/J-PAL SA has engaged with state governments and individual departments across many states in long-term and institutional policy-research partnerships to strengthen evidence-informed decision-making. Experience from these partnerships highlights the need for a combination of institutional (system-level) and process changes as well as resources for innovations, led by key champions within the government. These partnerships are governed by detailed institutional MoUs that identify key policy challenges to be addressed through research and evidence and are overseen by high-level steering committees chaired by senior bureaucrats. Channels of continuous dialogue are enabled between government officials and researchers to identify gaps and ideate potential solutions. Financial commitment on both sides then encourages the culture of innovative pilots and testing, which can then be scaled up based on rigorous evidence. Underlying these partnerships is also an important process of knowledge sharing and capacity building which improves the data capabilities and adoption of an evidence-informed approach to decision-making.

International examples of successful collaborations between governments and researchers include the Office of Evaluation Sciences in the United States government and the MineduLab in Peru. The Office for National Statistics (ONS) in the United Kingdom has established the Data Science Campus in 2017 to explore the availability of new data sources (administrative data and big data) and their use for public good, as well as build data science capabilities in the UK and internationally. The Handbook on Using Administrative Data showcases several such successful collaborations between governments and researchers in innovative and effective use of data for decision making.

A few enabling factors to streamline and sustain the use of data for decisions include:

- The systematic digitisation of government services and transactions with a focus on the usability of data for analysis by users, and ease of data capture/entry at the point of collection
- End-to-end digitisation of administrative data ensuring the highest granularity
- Systematic monitoring, review of programmes using data and evidence combined with internal reporting
- Mandate for undertaking rigorous impact evaluations and policy research (in-house and collaborating with external partners) to inform critical government challenges
- Dedicated financial resources for data systems, tools, innovations, and evaluations
- Continuous upskilling of staff on data capabilities to collate, analyse, interpret and use large volumes of data in routine decision-making and design of policies

However, there remain several challenges that constrain the systematic integration of data for decisions. These include:

- Limited focus and tools for ensuring quality and accuracy of data – both at the point of generation and analysis. It is important to ensure that data is reliable and usable before generating insights for interpretation
- Lack of a central policy that guides collation, access, sharing, and use of data within government departments and with external partners
- Limited mechanisms to ensure compliance with quality and data access standards
- In the absence of strong data use mandates and use cases, mere digitisation initiatives are insufficient to foster and institutionalise a culture of data-driven decision-making.

Building on existing efforts and opportunities to address the challenges outlined above would help unlock the true potential of using existing and new forms of administrative data to inform policy decisions that improve the welfare of citizens.

# References

1. Banerjee, Abhijit, Arun G. Chandrasekhar, Suresh Dalpath, Esther Duflo, John Floretta, Matthew O. Jackson, Harini Kannan, et al. 2021. "Selecting the Most Effective Nudge: Evidence from a Large-Scale Experiment on Immunization." NBER Working Paper 28726, National Bureau of Economic Research. https://doi: 10.3386/w28726. Accessed at https://www.nber.org/papers/w28726
2. Banerjee, Abhijit, Esther Duflo, Daniel Keniston, and Nina Singh. 2019. "The Efficient Deployment of Police Resources: Theory and New Evidence from a Randomized Drunk Driving Crackdown in India." NBER Working Paper 26224, National Bureau of Economic Research. https://doi.org/10.3386/w26224. Accessed at https://www.nber.org/papers/w26224
3. Cáceres-Delpiano, Julio, and Eugenio P. Giolito. 2019. "The Impact of Age of Entry on Academic Progression". In Data-Driven Policy Impact Evaluation, edited by Nuno Crato, and Paolo Paruolo, 249-67. Springer, Cham. https://doi.org/10.1007/978-3-319-78461-8_16
4. Cole, Hugh, Kelsey Jack, Derek Strong, and Brendan Maughan-Brown. 2020. "City of Cape Town, South Africa: Aligning Internal Data Capabilities with External Research Partnerships." In Handbook on Using Administrative Data for Research and Evidence-based Policy, edited by Shawn Cole, Iqbal Dhaliwal, Anja Sautmann, and Lars Vilhuber. Accessed at https://admindatahandbook.mit.edu/book/latest/cct.html
5. Cole, Shawn, William Parienté, and Anja Sautmann. 2020. "A revolution in economics? It's just getting started..." World Development 127, 104849. https://doi.org/10.1016/j.worlddev.2019.104849
6. Desai, Tanvi, Felix Ritchie, and Richard Welpton. 2016. "Five Safes: Designing data access for research" University of the West of England. Accessed at https://uwe-repository.worktribe.com/output/914745
7. Dodge, Eric, Yusuf Neggers, Rohini Pande, and Charity Troyer Moore. 2018. "Having it at Hand: How Small Search Frictions Impact Bureaucratic Efficiency." Abdul Latif Jameel Poverty Action Lab. Accessed at https://www.povertyactionlab.org/sites/default/files/research-paper/Have-it-at-hand_
8. Feeny, Laura, Jason Bauman, Julia Chabrier, Geeti Mehra, and Michelle Woodford. 2015. "Using Administrative Data for Randomised Evaluations". Research Resources: Abdul Latif Jameel Poverty Action Lab. Last modified November, 2018. Accessed at https://www.povertyactionlab.org/resource/using-administrative-data-randomized-evaluations
9. García, Jorge Luis, and James J. Heckman. 2020. "Life-cycle Benefits of Early Childhood Programs: Evidence from an Influential Early Childhood Program." Microeconomic Insights. Accessed at https://microeconomicinsights.org/life-cycle-benefits-of-early-childhood-programs-evidence-from-an-influential-early-childhood-program
10. Gertler, Paul, and Simon Boyce. 2001. "An Experiment in Incentive-Based Welfare: The Impact of PROGESA on Health in Mexico." Working Paper. Accessed at https://www.povertyactionlab.org/media/file-research-paper/experiment-incentive-based-welfare-impact-progesa-health-mexico
11. Gibson, Michael. "Data quality checks". Research Resources: Abdul Latif Jameel Poverty Action Lab. Last modified March, 2021. Accessed at https://www.povertyactionlab.org/resource/data-quality-checks
12. Greenstone, Michael, Rohini Pande, Anant Sudarshan, and Nicholas Ryan. 2023. "Can Pollution Markets Work in Developing Countries? Experimental Evidence from India." Abdul Latif Jameel Poverty Action Lab. Accessed at https://www.povertyactionlab.org/media/file-research-paper/can-pollution-markets-work-developing-countries-experimental-evidence

13. Jain, Radhika, and Pascaline Dupas. 2022. "The effects of India's COVID-19 lockdown on critical non-COVID health care and outcomes: Evidence from dialysis patients." Social Science and Medicine 296, 114762. https://doi.org/10.1016/j.socscimed.2022.114762

14. Muralidharan, Karthik, Paul Niehaus, Sandip Sukhtankar, and Jeffrey Weaver. 2018. "Improving last-mile service delivery using phone-based monitoring." NBER Working Paper 25298, National Bureau of Economic Research. https://doi.org/10.3386/w25298. Accessed at https://www.nber.org/papers/w25298

15. Murphy, Kathleen. 2020. "Collaborating with the Institutional Review Board (IRB)." In Handbook on Using Administrative Data for Research and Evidence-based Policy, edited by Shawn Cole, Iqbal Dhaliwal, Anja Sautmann, and Lars Vilhuber. Accessed at https://admindatahandbook.mit.edu/book/v1.0/irb.html

16. National Council of Applied Economic Research. 2015. Evaluation Study of Targeted Public Distribution System in Selected States. Department of Food and Public Distribution, Ministry of Consumer Affairs, Food and Public Distribution, Government of India. Accessed at https://www.ncaer.org/wp-content/uploads/2022/09/1460106533TPDS-140316.pdf

17. National Institute of Public Finance & Policy (NIPFP). 2019. 43rd Annual Report, 2018-2019. Accessed at https://www.nipfp.org.in/media/medialibrary/2019/12/English_Annual_Report_NIPFP_18-19.pdf

18. O'Hara, Amy. 2020. "Model Data Use Agreements: A Practical Guide." In Handbook on Using Administrative Data for Research and Evidence-based Policy, edited by Shawn Cole, Iqbal Dhaliwal, Anja Sautmann, and Lars Vilhuber. Accessed at https://admindatahandbook.mit.edu/book/v1.0/dua.html

19. Schmutte, Ian M., and Lars Vilhuber. 2020. "Balancing Privacy and Data Usability: An Overview of Disclosure Avoidance Methods." In Handbook on Using Administrative Data for Research and Evidence-based Policy, edited by Shawn Cole, Iqbal Dhaliwal, Anja Sautmann, and Lars Vilhuber. Accessed at https://admindatahandbook.mit.edu/book/v1.0/discavoid.html

20. Shen, Jim, and Lars Vilhuber. 2020. "Physically Protecting Sensitive Data." In Handbook on Using Administrative Data for Research and Evidence-based Policy, edited by Shawn Cole, Iqbal Dhaliwal, Anja Sautmann, and Lars Vilhuber. Accessed at https://admindatahandbook.mit.edu/book/v1.0-rc5/security.html