Working Paper | October 2024

Institutionalising a Culture of Data-Driven Decision Making

By Aparna Krishnan, Tithee Mukhopadhyay, and Gargi Pal

GLOBAL INNOVATION FUND

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

CLEAR South Asia Center

Acknowledgements

We thank all past and current J-PAL South Asia staff who worked on this project: Aishwarya Atluri, Satya Bhushan Chintamaneni, Manvi Govil, Hemanth Guthala, Dona Maria John, Michael Joseph, Samresh Kumar, Apoorva Limaye, Manoj Mitra, Sanchita Ohri, Mustufa Patel, Divya Saravana, Sanya Sareen, Kannan Sobti, Gayathri Sowrirajan, Anuja Venkatachalam, Arpita Wadhwa, as well as the finance/grant team members, A. Narayanee, Arindam Mukherji, and Vrinda Bhandari (legal consultant). Special mention to Nandini Gupta who was instrumental in setting up the project and provided close guidance and leadership as the senior research manager until 2022, and was a core team member of J-PAL South Asia's Innovations in Data for Experiments and Action (IDEA) lab since the start in 2019.

We are grateful to the vision, guidance, and oversight provided by Iqbal Dhaliwal (Global Executive Director, J-PAL) Shawn Cole (co-chair IDEA), Lars Vilhuber (co-chair IDEA), Achyuta Adhvaryu (Scientific Advisor, Indian Scholars Program, 2023), Gautam Rao (Scientific Advisor, Indian Scholars Program, 2023) and Shobhini Mukerji (Executive Director, J-PAL South Asia), as well as the senior management team at J-PAL South Asia (Urvashi Wattal, Megha Pradan, Sharanya Chandran, Shagun Sabarwal (former), Ruben Menon).

We thank all the researchers who were engaged on the use cases: Pascaline Dupas (Scientific Director, J-PAL Africa; Princeton), Radhika Jain (University College London), Girija Vaidyanathan (retired IAS officer of the Tamil Nadu cadre), Sebastian Bauhoff (Harvard), Nikkil Sudarshan (Technical University of Munich), Rashmi Barua (Jawaharlal Nehru University), Deboshmita Brahma (University of Calcutta), Prarthana Agarwal Goel (Guru Gobind Singh Indraprastha University), Remya Tressa Jacob (T A Pai Management Institute), Benston John (St. Stephen's College), Manmeet Kaur (Punjab Agricultural University), E Somanathan (Indian Statistical Institute), Rohini Somanathan (Delhi School of Economics), Shailender Swaminathan (Krea University) and Vijay Victor (Indian Institute of Technology, Roorkee).

We thank all the state government departments we engaged with (Punjab, Haryana, Andhra Pradesh, Tamil Nadu), NHA, and DMEO - NITI Aayog. We are grateful for additional funding support from CLEAR South Asia/GEI for capacity building and advisory activities. Finally, this work wouldn't have been possible without the generous support from the Global Innovation Fund (GIF) and close guidance provided by the team, especially Joseph Ssentongo, Rachna Nag Chowdhuri, Kathi Hanifnia, and Shreya Das.

Table of Contents

1. Executive Summary

Too often, policies and programs in developing countries tend to be informed by a combination of instinct, ideology, or inertia with limited or no use of data for decision-making[1]. With the widespread digitization of data systems and processes over the past decade, especially by national and state governments in India, large volumes of data are now available. This has led to a growing demand from policymakers and researchers alike for data related services and products like access to administrative data for designing new randomised control trials, data catalogues, knowledge products like Data Use Agreements, data talent development programs and resources. At J-PAL South Asia, we continue to sustain the efforts initiated under this engagement through our deep partnerships with governments and a consolidated suite of services combining rigorous research, hands-on capacity building, and timely policy advisory services aimed at integrating data and evidence into each step of the policy decision lifecycle. We summarise insights and lessons from J-PAL South Asia's experience of designing and testing innovative approaches for strengthening the use of administrative data and research in decision-making with state governments across India, particularly with health departments and e-governance/data analytics agencies. This paper attempts to capture and contextualise challenges and innovations that research and policy think tanks/consultants collaborating with governments can use to institutionalise a culture of using data and evidence for policy-making.

The Indian government has increased its IT spending by 4 times over the last decade[2], catalysing digitisation initiatives and setting up digital infrastructure for more accurate policy targeting and implementation monitoring. But despite these efforts, it is observed that such rich administrative data is primarily used for operational purposes rather than being used in advanced ways that can inform decision making. There is immense potential to unlock the value of data and evidence in informing decision making, and many ongoing efforts to enhance technical capacity and strengthen data reliability. Our learnings in this paper are informed by the unprecedented growth of digital data collection and data use in India, the potential role of this growth in improving public service delivery and our deep experience of collaborating with governments, for over a decade.

Advancing data driven decision making in government requires strategic collaborations, targeted capacity-building, technically trained talent and the development of robust data-quality protocols. To support this tall mandate of strengthening the digitalisation and data use ecosystem as well as building collaborative feedback loops, this space requires well designed and responsive funding structures. Through a grant provided by the Global Innovation Fund (GIF) to J-PAL South Asia, we explored nuanced mechanisms to strengthen state capacity for institutionalising the use of administrative data to improve public service delivery. In this paper,

[1] *Poor Economics: A radical rethinking of way to fight global poverty (2011)* by Abhijit Banerjee and Esther Duflo

[2] *Union budget 2024-25 allocates over 550 crores to the IndiaAI mission*. (n.d.). INDIAai. https://indiaai.gov.in/article/union-budget-2024-25-allocates-over-550-crores-to-the-indiaai-mission

we explore the process and insights from adopting a collaborative approach with state governments in India to develop policy research questions (in the health sector) that could be addressed using existing administrative data sources, and engage in complementary efforts through customised capacity building activities to institutionalise data use practices by strengthening data governance and capabilities. This process could take different pathways in different states. For instance, in two states, our work involved identifying relevant policy questions in the delivery of public health insurance programmes, such as unpacking patterns and gaps in utilisation by the intended beneficiaries or improving monitoring and audit of practices and claims processing at hospitals using a combination of existing administrative data and primary data collection in meaningful ways. In other instances, collaborations involved developing metadata catalogues of administrative data, training workshops for staff on data quality and analysis, developing guidelines, templates and processes for secure sharing and use of data for decision making such as data use agreements, protocols and data sharing & use policies.

Could such an approach lead to replicability in other contexts? Does the process lead to greater demand for meaningful and reliable data and in turn strengthen government capacity to generate, leverage and use data effectively? What is required to bring about sustained change in processes, practices and mindset at scale for data use in decision making? In this paper, we reflect on our 3+ years of experience of developing data use cases in health insurance, health systems and climate by exploring the availability of existing administrative data to address policy research questions of interest.

The paper is structured as follows:

- **Vision and Motivation (Sections 2 and 3):** These sections outline the overarching vision to improve the use of administrative data and strengthen state data capacity for better public survey delivery and strategic decision making. GIF's support enabled the demonstration of a use case for the advanced use of administrative data in state health insurance schemes as well as in the broader health and climate change space, through building deep relations with government stakeholders and co-creating scopes of work based on government priority.
- **Lessons and Challenges for Strengthening State Capacity (Section 4):** This section explains the contextual problem statement and the implementation model undertaken. It also details foundational lessons on strengthening state capacity to make effective use of administrative data, setting the stage for subsequent initiatives. Limited use of administrative data is driven by lack of understanding of the data use value proposition, limited technical capacity and lack of data reliability. For demonstrating a data use case, we followed a systematic four-stage approach. Among our key lessons are, the importance of demand for solutions from the government for a successful collaborative engagement, clarity of data ownership for accessing and using data, need for a data strategy to ensure a comprehensive administrative dataset and the need for investment

in capacity building for better data use. In the subsequent sections (5 and 6) we unpack some of the crucial interim steps that served as critical milestones in our overall journey.

- **Building Strategic Partnerships (Section 5):** This section explores the key elements of successful government engagement, including identifying champions, defining boundaries of work, developing flexible talent structures to support data integration in different states, as well as highlights the need for flexible funding outlays that allow responsiveness to policy windows. A double funnel approach bookends the combination of specific impact opportunity and broader policy dialogue and data use case learnings. A formal institutional set-up along with sustained involvement of champions at the government departmental level was crucial for building and sustaining partnerships. Depending on government interests, the co-created scope of work tended to differ which informed the staffing structure and governance mechanisms - generalist policy dialogue vs. specific project discussions driven by researchers. In both new and existing partnerships, senior staff at an organisational level was important to steward partnerships with governments. Further policy windows were discrete and varied, which required flexible and collaborative funding structures.
- **Co-creating Research Questions and Scope (Section 6):** Emphasising the importance of collaborative development, this section introduces two engagement models— sequential (generalist policy dialogues) and researcher-led (specific project discussions driven by researchers) — that were tested to identify state-specific policy questions and priorities that allowed for co-creation of research. The ideal model was dependent on government policy priorities - preference for an explorative approach to co-diagnose avenues for research vs. demand for specific solutions.
- **Execution and Success Drivers (Section 7):** This section identifies the critical factors that influenced project outcomes, including data access, data quality, and continuity in government priorities. It highlights how the absence of any one factor impacts the degree of success achievable. The four-stage model, therefore, achieved different outcomes in different states based on the success of these factors.
- **Integration and Sustainability (Section 8):** This concluding section discusses key takeaways from our experience as well as how the practices and learnings developed through the initiative continue to be embedded and sustained within J-PAL South Asia and government partners' operations to foster a lasting impact on data utilisation and building a culture of data informed decisions. Sustainability is ensured through strategic partnerships and initiatives driven by J-PAL South Asia as well as the work done by researchers associated with us. Further, we revisit the double funnel approach to draw out key lessons including drivers of successful partnerships (multi-level engagement, formal mechanisms and sustained involvement of champions), responsiveness to policy demand (choice of researcher or sequential model and allocation of resources), identifying specific impact opportunity (impact is incremental and has the potential to be a catalyst for further change) and the need for greater investment in building capacity to improve data quality, reliability, strategy and ultimately data use.

2. Vision

Data is increasingly becoming an integral part of our everyday lives. There has been an unprecedented growth in the generation of data due to rapid advancements in digital technology all across the world. Innovative use of data can be transformative in business and in governments alike. The World Development Report of 2021[3] describes how "data gathered by governments, international organisations, research institutions and civil society can improve societal well-being by enhancing service delivery, prioritising scarce resources, holding governments accountable and empowering individuals." We typically refer to such data generated through the course of delivery of public services as administrative data, which are at the core of government efforts in low- and middle-income countries to reduce poverty and improve welfare. Public access to data, especially generated by governments and donor organisations, is often considered valuable in itself since it supports the goals of transparency and accountability[4].

With digitalisation, the volumes, granularity and nature of data have changed drastically, especially with the growth of cell phone and satellite technologies. Therefore, there definitely does emerge an important opportunity to test out innovative approaches to build data infrastructure and institutionalise a culture of data driven decision making. In India, for instance, it is estimated that the digital economy accounts for a tenth of its GDP and is poised to constitute a fifth by 2026[5]. Budgeted allocation for the Ministry of Electronics and Information Technology, Government of India has increased by more than 4 times[6] in the last decade to INR 21,936.9 crore in 2024[7]. There is a growing interest to explore the application of data and latest statistical and analytical methods, including the use of artificial intelligence and machine learning methods for public good. As the scope and volume generated by governments, non-profits and private organisations grows, there is potential for the use of such data in the designing, targeting, monitoring and scaling of innovations, as well as research and evaluations of the impact of such innovations.

However, despite the benefits from digitalisation of the economies and governments in South Asia, the full game-changing potential still remains untapped.[8] In order to effectively use data to

[3] *World Bank. 2021. World Development Report 2021: Data for Better Lives. Washington, DC: World Bank. pg. 54.* https://wdr2021.worldbank.org/

[4] Cole, Shawn, Iqbal Dhaliwal, Anja Sautmann, and Lars Vilhuber. 2022. "Using Administrative Data for Research and Evidence-Based Policy: An Introduction." In: Cole, Dhaliwal, Sautmann, and Vilhuber (eds), *Handbook on Using Administrative Data for Research and Evidence-based Policy*, Version v1.1. Accessed at https://admindatahandbook.mit.edu/book/v1.1/intro.html on 2024-09-30.

[5] *Reserve Bank of India, 2024. Report on Currency and Finance 2023-24, India's Digital Revolution.*

[6] *Outcome Budget 2015-16 | Ministry of Electronics and Information Technology, Government of India*. (n.d.). https://www.meity.gov.in/content/outcome-budget-2015-16

[7] *Union budget 2024-25 allocates over 550 crores to the IndiaAI mission*. (n.d.). INDIAai. https://indiaai.gov.in/article/union-budget-2024-25-allocates-over-550-crores-to-the-indiaai-mission

[8] *World Bank. 2022. South Asia's Digital Opportunity: Accelerating Growth, Transforming Lives. World Bank, Washington, DC., pg 3.*

inform decision making processes within government, there is need for significant investments in data infrastructure, statistical/computational capabilities, data governance, balancing data use purposes with data privacy safeguards and security measures. These investments are interdependent and a failure in any one area could diminish the value of proposition of data for development[9].

We believe that demonstrating the value proposition from using data for decision making with and within governments could help unlock the necessary investments to harness the transformative potential of data for development. Use of existing administrative data combined with survey data and in research studies aimed at addressing policy questions jointly identified with the government, can help generate "data use cases" that could encourage demand for relevant data and the application of insights to policy making processes. An increasing proportion of academic studies undertaken in high income countries now use administrative data.[10] There are several advantages of using administrative data for evidence and data informed policy making[11].

- ➢ Primarily, since such data is collected during the course of routine business, it avoids social desirability and recall biases associated with survey and other forms of self-reported data, therefore can be considered more objective.
- ➢ The coverage of administrative data is much larger and more frequent (routine) relative to periodic, sample surveys. This makes the data more representative and complete (i.e. it is possible to collect longitudinal data with less attrition).
- ➢ Finally, such administrative data also provides a unique opportunity to the researchers to explore novel data driven analyses that are relatively cheaper, recurring and devoid of biases that are often present in primary survey data.

The use of administrative data in undertaking policy research to support evidence-informed decision making could serve twin goals of (a) strengthening the usability of existing and new forms of data for analysis and interpretation and (b) bringing about a change in mindset and culture is the systematic use of data for decision making.

In this paper, we explore the process and insights from adopting a collaborative approach with state governments in India to develop policy research questions (in the health sector) that could be addressed using existing administrative data sources, and engage in complementary efforts through customised capacity building activities to institutionalise data use practices by strengthening data governance and capabilities. Could such an approach lead to replicability in

[9] *World Bank. 2021. World Development Report 2021: Data for Better Lives. Washington, DC: World Bank. pg. 54.* https://wdr2021.worldbank.org/

[10] Chetty, Raj. 2012. "Time Trends in the Use of Administrative Data for Empirical Research." http://www.rajchetty.com/chettyfiles/admin_data_trends.pdf.

[11] Cole, Shawn, Iqbal Dhaliwal, Anja Sautmann, and Lars Vilhuber. 2022. "Using Administrative Data for Research and Evidence-Based Policy: An Introduction." In: Cole, Dhaliwal, Sautmann, and Vilhuber (eds), *Handbook on Using Administrative Data for Research and Evidence-based Policy*, Version v1.1. Accessed at https://admindatahandbook.mit.edu/book/v1.1/intro.html#the-potential-of-administrative-data-for-research-and-policymaking on 2024-10-29.

other contexts? Does the process lead to greater demand for meaningful and reliable data and in turn strengthen government capacity to generate, leverage and use data effectively? What is required to bring about sustained change in processes, practices and mindset at scale for data use in decision making? These are a few guiding questions we hope to explore in the rest of the paper by reflecting on our 3+ years' experience of developing data use cases in health insurance, health systems and climate by exploring the availability of existing administrative data to address policy research questions of interest.

3. Motivation and Context

This working paper reflects on lessons learned from a three-year grant provided by the Global Innovation Fund (GIF) to J-PAL South Asia to explore the mechanisms for strengthening state capacity in order to institutionalise the use of administrative data to improve public service delivery by developing a few health-sector related data use cases. Leveraging the slew of digitisation initiatives within governments in the past decade, and the launch of a J-PAL-wide initiative, "Innovations in Data and Experiments for Action (IDEA)," this project aimed to understand the mechanisms for bringing about sustainable change in government demand, and more importantly, capacity, to leverage large volumes of administrative data generated during programme administration for further use in identifying problems and solutions. This project helped us address and respond to growing demand from our existing and potential government partners on how we could support their intent to institutionalise a more data-driven approach to decision making to improve governance and welfare outcomes.

The lessons from this project derive from experiences of collaborating with health insurance departments of Punjab, Haryana, Tamil Nadu and Andhra Pradesh as well as other departments in Kerala and Assam during 2022-25. Our work in these states also sparked interest in collaborating with researchers on the effective use of administrative data in other sectors such as education in Tamil Nadu, health (routine immunisation) in Haryana as well as work on administrative data best practices through DMEO (NITI Aayog)[12] . Through the course of the grant, we also worked on some public goods which includes data capacity assessment formats, MoU and data sharing/use formats, metadata catalogues as well as codebooks. We also developed guidelines and resources for governments interested in building data use strategies and guidelines for secure sharing and use of data for research and decision making. We have also developed content for and delivered training workshops for government officials on improving quality, usability and analysis of administrative data (through additional funding provided by the CLEAR South Asia, Global Evaluation Initiative). In this paper, we synthesise our overall learnings through rich experiences across multiple states.

[12] DMEO, NITI Aayog, CLEAR at J-PAL SA, Compendium of Case Studies on Using Administrative Data for Evidence-Based Policymaking, link
DMEO, NITI Aayog, CLEAR at J-PAL SA, Administrative Data Toolkit, link

To demonstrate a use case for administrative data use and unpack nuanced learnings, two pathways were undertaken - (i) A major proportion of our work was with the large data systems of the country's health insurance program which had the potential for drawing common lessons across states, (ii) a researcher-based approach on working with administrative data in the health and climate change space.

In the next section, we discuss the challenges and lessons for strengthening state capacity for administrative data use. This will be followed by sections 5 and 6 which unpack two important drivers of our work, namely, building and developing government partnerships and co-creating research questions. Subsequently, we discuss our experience in execution in section 7 and our key learnings and its integration in section 8.

BOX 1: THE CASE FOR LEVERAGING HEALTH INSURANCE TO BUILD STATE CAPACITY FOR INSTITUTIONALISING DATA USE IN DECISION MAKING

Health shocks, particularly critical illnesses that require hospitalisation, have the potential to push poor households back into poverty despite any gains made. Poor households are vulnerable to health shocks due to several reasons: lack of awareness about and coverage of health insurance for critical illnesses; geographical variability to access to quality health care in private and public facilities, and high out-of-pocket expenditure on health.

In 2018, the Government of India introduced a comprehensive, nationwide health insurance program called the Ayushman Bharat - Pradhan Mantri Jan Arogya Yojana (AB-PMJAY). AB-PMJAY is a cashless health insurance cover for secondary and tertiary care hospitalisation across public and private empanelled hospitals for low-income families and is fully financed by the government. While several states had their own health insurance programmes (some with broader coverage than AB-PMJAY), eventually all states joined the national insurance programme with exemptions made to integrate state-specific requirements. This allowed for a common framework across the country. The PMJAY also introduced a common management information system (MIS) to collect data on service delivery, utilisation, and other operational aspects. This allows all state government schemes to collect and report information along common metrics, whether through the new MIS or their legacy systems. Further, different states operationalized the scheme through two standardised operational models, (i) the insurance model, using an insurance company and third-party administrators (TPAs) and (ii) the assurance model, where the administration of the insurance was entirely managed through the government. There were some variations within these models, in whether the scheme is universal or targeted in coverage, or blending of insurance and assurance models for different treatment conditions. However, largely, with the introduction of the PMJAY, all states aligned their programmes to a few common, minimum criteria.

Given this context, administrative data of the government's large-scale health insurance scheme was initially chosen as a specific example for demonstrating a use case for data-driven decision-making. The objective was to demonstrate how administrative data generated through institutional information systems, combined with primary data collected through beneficiary surveys, could be leveraged better to identify gaps, and design and test improvements to the program to ultimately strengthen the health outcomes of citizens. Through these use cases, we hoped to encourage improvements in state capacity for using data meaningfully for strategic planning and course correction, driving optimal resource allocation.

This engagement leveraged J-PAL's global expertise in academic rigour and network across policy areas, strategic government partnerships, and Indian researchers and staff. It was anticipated that insights and experience from this use case could apply to other sectors within these states, and to similar schemes and programmes in other States. In 2023, based on experience and demand, the scope was expanded to more states and sectors of health systems and climate. This paper synthesises lessons and reflects on insights from the experience of executing a range of activities under this grant from January 2021 to May 2024 to ultimately institutionalise effective data use practices within the government. We will explore in detail the approaches and considerations in developing partnerships for the health insurance use case, and other critical policy questions (including the use of a competitive call for ideas). We will do this in the context of the nature of progress made along each step of data access, management, analysis, output, insights, and potential pathways for long-term outcomes. We will also reflect on the various enabling and challenging factors in the course of institutionalising the use of data for decision-making within governments, and share suggestions related to streamlining access to, management and use of administrative data more broadly for policy research and analysis.

4. Challenges and Lessons for Strengthening State Capacity for The Effective Use of Data

As anticipated, understanding and strengthening state capacity to systematically leverage data for decision-making was both complex and multi-faceted. Success with any kind of capacity building initiatives was dependent on building deep relations with departments and identifying champions who could rally and sustain departmental interest in this aspect.

4.1. Reasons Behind Limited Use of Administrative Data

We started with this problem statement: too often, policies and programs in developing countries tend to be informed by a combination of instinct, ideology, or inertia with limited or no use of data for decision-making[13]. With the widespread digitization of data systems and processes over the past decade, especially by national and state governments in India, large volumes of data are now available. However, the use of data for decision-making, beyond the immediate administrative purpose for which it was collected, remains limited for several reasons. This is especially true in the health sector, where governments have launched multiple management information systems (MIS) to capture data on health services delivery, utilisation, hospital records, medicines and vaccine supplies, screening, and tracking of patients, among others. The use of this administrative data, however, is primarily for routine process monitoring and reporting – and is rarely deployed systematically for planning, targeting, course correction and/or innovations to improve the effectiveness of health services. During the Covid-19

[13] *Poor Economics: A radical rethinking of way to fight global poverty (2011)* by Abhijit Banerjee and Esther Duflo

pandemic, we saw that governments that had already invested in robust data systems historically were able to leverage the same for targeted responses[14]. There was also an overall greater movement among other governments to leverage data to inform their plans and responses. However, beyond immediate response during the pandemic, it did not necessarily translate into strengthened systems and processes for the systematic use of administrative data for medium to long-term planning and resource allocation decisions.

The reasons for this can be grouped under:

- *Lack of understanding of the value proposition of using data and evidence proactively*: Given that administrative data is generated as part of the routine process of delivering services and primarily used for reporting upwards, there often isn't a clear understanding or vision of how such data may be leveraged as a tool for strategic planning and course correction. This requires a substantive change in mindset, processes, and systems to realise the potential for data use, and in the absence of sufficient instances, this value proposition may not be obvious and data is only used for the limited purpose of reporting and in the course of administration of services
- *Limited technical capacity to manage and analyse large data*: Digitisation efforts have also resulted in the generation of massive volumes of high-frequency data. However, there is still limited awareness and sensitivity among government officials on how to securely access and use data, while not compromising the privacy of the individuals on whom data is available, especially in absence of legislation and guidelines. Further, their technical capacity for the use of advanced statistical packages and tools to meaningfully analyse and visualise data patterns depending on the questions of interest at different levels remains limited.
- *Lack of readiness and reliability of data:* Digitization from paper to electronic has improved the timeliness of data availability – but not necessarily its usability. This is because data is collected for a specific purpose and without a clear articulation of its intended use. Even in cases where the purpose is articulated, the data is not always made available in formats that allow for statistical analysis. Efforts have to be taken to prepare and clean the data before it can be used for drawing insights. This is further exacerbated by the possible absence of systematic data quality checks at the field level (especially when the data is manually entered by frontline workers) or at the MIS level (logical validations and system design elements), thus rendering the data un/under-utilised for analysis.

4.2. Model for Institutionalising Data Use

Over the past three years, we attempted to develop an innovative model for institutionalising data use for decision-making within the government by collaborating on data use cases and systemic capacity-building efforts in health insurance and related sectors to address the challenges described above. Through this, we also attempted to engage in different capacity settings to observe any differences in the take-up of a data-driven approach among high and

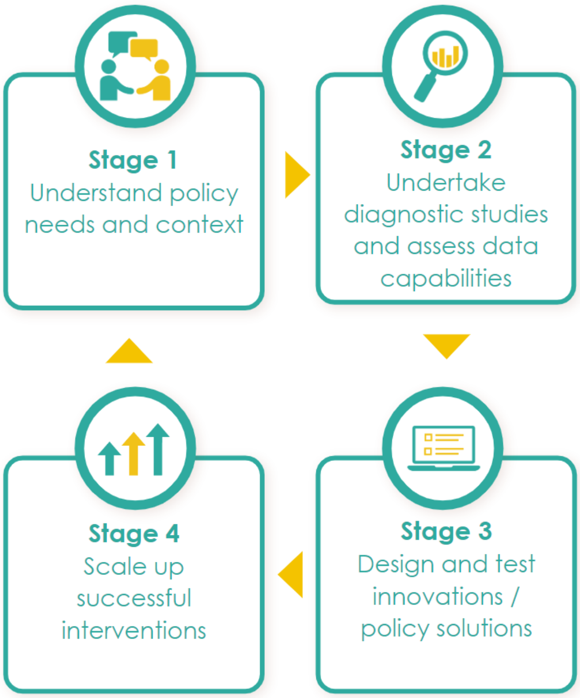
[14] How Kerala's response helped flatten the curve: here and here

low-capacity government departments. Figure 1 depicts these efforts, which is a broad four-pronged strategy -

- Stage 1: Understand policy needs and context
- Stage 2: Undertake diagnostic studies and assess data capabilities
- Stage 3: Design and test innovations/policy solutions
- Stage 4: Scale up successful interventions

One of our key learnings is that we need **multiple parallel strategies and a longer run-up** to bring about changes in outcomes. The links between stages 2 & 3 (i.e. problem identification to solutioning & testing) and between stages 3 & 4 (i.e. insights from design/testing of innovations to scaling up and institutional change) is non-linear and may be iterative and time intensive, depending on several factors. For instance, to take action to develop and adapt solutions to address issues identified from exploratory data analysis (such as gaps in coverage of the target population or inefficiencies in the administration or utilisation of the insurance coverage) a number of factors need to align. These include an appropriate policy window, necessary data infrastructure, required budget allocations, and monitoring and evaluation activities. However, when these factors do align, especially with the vision and leadership of key champions within the government, action tends to be swift and at scale (as we have observed in other instances). Therefore, it is valuable for organisations collaborating with governments on data use mandates to be nimble and responsive to discrete policy windows as they emerge - scaling up and scaling down efforts as needed. A broad-based institutional partnership helps adapt and focus resources where it is most useful.

Figure 1: A collaborative approach to developing data use cases and strengthening government capacity to institutionalise the use of data in decision making


Stage 1
Understand policy needs and context
Stage 2
Undertake diagnostic studies and assess data capabilities
Stage 3
Design and test innovations / policy solutions
Stage 4
Scale up successful interventions

Similarly, in our experience, **considerable preparation and technical assistance may be required to scale up successfully tested innovations, and sustain changes to institutional processes** and systems. For instance, modifications to departmental review processes for integrating data insights, or adoption of new dashboards for data visualisation need persistent championing and integration into ongoing day-to-day operational processes for the changes to be self-sustaining over time. Further, policies and guidelines for data sharing and use, prepared and issued at the state level by nodal departments such as Information Technology or Planning need to be promoted through a series of consultations in order to help implementing line departments to understand and adopt the guidelines to their specific contexts and needs. Moreover, an iterative process that allows stress testing and course correction helps strengthen the applicability and implementation of guidelines across governments, which otherwise remain siloed or as high-level guiding principles. This may need greater engagement initially through nodal officers or committees that review and address issues as they emerge. It may also be useful to phase in the implementation of guidelines, starting with a few departments/ministries, before expanding its applicability government-wide. Understanding and planning for this process of bringing about institutional change within governments, especially innovative ideas or systemic process change, is crucial.

4.3. Lessons

Some specific insights and recommendations for engaging with government partnerships to bring about sustainable and institutional change in culture and systems for effective use of data:

1. Prompt engagement and uptake of proposed solutions on any joint policy-research dialogues from government partners is more likely when there is a **clear and strong demand for solutions**. Our original approach of iterative dialogues during Phase 1 (to identify broad policy challenges and explore available datasets) yielded interesting questions and systematic scopes of work, but in some cases also led to inefficient and cyclic exploration without any concrete outcomes on solutions and next steps. We subsequently added a variation to the model, where, to crowd in more ideas for specific use cases from a larger pool of domain experts based in India, we used a competitive call for proposals for developing short-duration data use cases (6-9 months) on health systems, and climate change in partnership with local and state governments. This partly addressed the demand gap by accelerating a solution-oriented approach to already identified problems. We believe a successful engagement strategy is one which is flexible enough to tap into the key priorities of respective government departments, which might differ from state to state, by proposing specific problem statement-solution proposals for quick exploration. For example, while exploring policy priorities within the health insurance space in a particular state, we uncovered the department's interest to work on routine immunisation as well as the inability of the department to share health insurance administrative data. We were able to accordingly match researchers to take this collaboration forward in the routine immunisation space and explore conversations with

other states pertaining to health insurance. In another state, our exploratory conversations on health insurance bore fruit and led to a successful collaboration to improve hospital monitoring through advanced use of administrative data.

2. Data ownership, and more importantly, **clarity on the roles and responsibilities of data ownership when multiple parties and levels of government are involved is critical** to developing administrative data use cases. Despite having narrowed down on a Scope of Work of mutual interest between the researchers and governments, a lack of clarity on data ownership and rules to secure data access and sharing could stall promising engagements. Each situation and context may be different and many stakeholders could be involved across levels of government. At times, data ownership for an information system may lie with the state government department, and at times with the national ministry. As we saw in the health insurance instance, states with legacy insurance programmes have developed their own data systems and have full access to and control over all aspects. However, newer states that adopt the national MIS have joint ownership of the data with the national Ministry, but access to and control over the full database is restricted depending on where the data servers and management reside (usually at the national level). This adds another level of complexity in securing permissions, especially in the absence of clearly specified rules to the effect and time taken in setting up secure data-sharing arrangements aligned to guidelines. Further, apart from data owners, there could be a range of data intermediaries, such as database/IT managers, Monitoring & Evaluation (M&E) officers, programme officials at the state/district level, consultants etc. who may all have a role to play in providing access or explaining data structures and usability. Mapping out all the stakeholders and their roles/authorities in the data management chain is essential to ensure that complete and secure data access is secured such that it is ready for analysis and use. Finally, data access or transfer is rarely a one-time exchange. It is an iterative and collaborative process that varies by the nature of the data and the associated efforts required to identify, collate, combine, and clean before it can be used. Identifying the stakeholders and their roles helps streamline the process, and reduce time taken in data access.

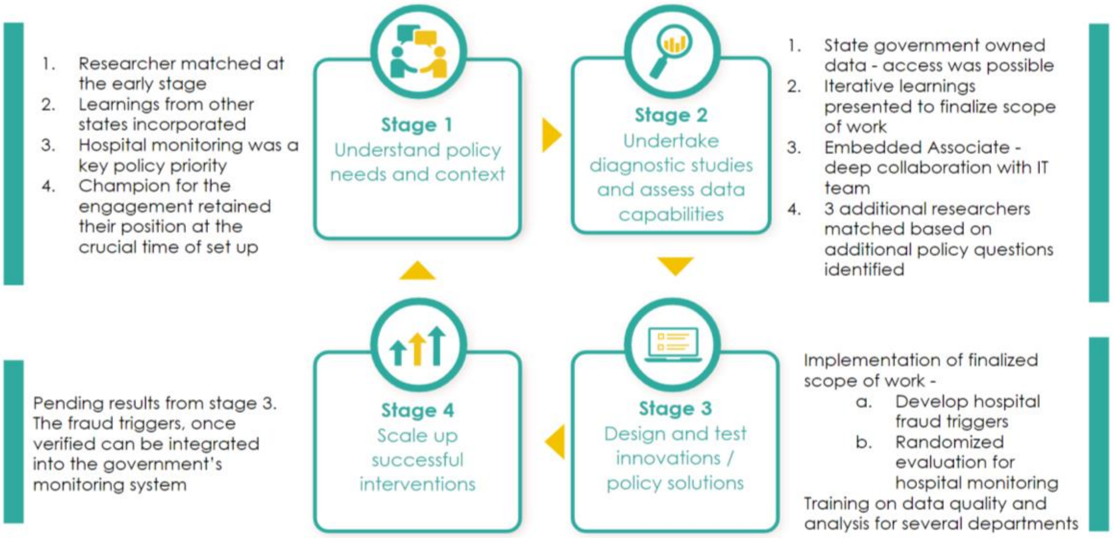
3. **The government, at all levels, needs to define and publicise a comprehensive data strategy** for processes to be streamlined and made sustainable government-wide. This is important not only to establish a clear vision for all stakeholders, but to also allow larger society to identify key areas to contribute through their existing expertise. The development of the Digital Personal Data Protection (DPDP) legislation issued in 2023 is a milestone achievement which we hope will eventually encourage the secure sharing and use of data for public good. However, the legislative mandate needs to be accompanied by policy and institutional systems that prioritise a data and evidence-informed approach to policy-making within the government. The demand for data and clear articulation of its use for planning, strategic resource allocation, and improved governance, among others, would then guide the development of data infrastructure and

policies aligned with that vision. A holistic "data use" strategy by the government may help bring about certain implementation efficiencies. However, currently, the potential for unlocking data use remains limited.

4. **A lot more investment needs to be made by the government to build the capacity of all the officials** operating throughout the life cycle of administrative data, especially those working on the backend IT and MIS infrastructure. This investment in capacity should not just be limited to technical competencies but should sensitise these officials to the importance and potential of using administrative data for policy decisions – to empower them to make critical decisions on data protection, privacy, and sharing. This enables officials across different functions to have the ability to interpret insights from data, combined with domain expertise and local knowledge, and engage with researchers, tech professionals, and others in unique and innovative data use cases that can extend state capacity to solve persistent policy challenges. Moreover, it also ensures that data use is not limited to the senior cadres of the legislative or executive functions of the state or national government. A culture of data use implies that an understanding and application of data insights for functions such as planning, review, and resource allocation are integrated into the work of all officials across levels from field-level functionaries, sub-district offices, district, state-level functions, and ultimately the national level. Training and building capacity of government officials is very effective when combined with data use case instances where they have an opportunity to apply new knowledge to live projects. The unmet potential for data use also lies in the ability to link disparate datasets within the government or with external sources to gain insights that may have not been obvious or intuitive earlier. With the advances in artificial intelligence and machine learning, appropriate and meaningful use of data while ensuring the privacy of its citizens can make a tremendous difference in increasing the efficiency of the delivery of public services, enhancing the welfare of vulnerable populations. This also calls for a cadre of upskilled or newly trained professionals within governments that can handle the collection, storage, management, and sharing of large volumes of high-frequency and granular digital data.

5. Attention needs to be paid to **systematically adopt and institutionalise measures** for data quality and reliability. Typically, there is a lot of investment in tech infrastructure. However, this needs to be accompanied by sufficient training for tech users to populate the systems with reliable data – including frontline community workers, as they are the first point of data generation. There is also a need to institute systemic checks through monitoring, audits, and database-related logical and quality checks to ensure data validity. This is the essential first step to the meaningful use of data, and avoiding unreliable data results that lead to unusable insights or worse – incorrect decisions.

Figure 2: An example of success in developing a data use case



1. Researcher matched at the early stage
2. Learnings from other states incorporated
3. Hospital monitoring was a key policy priority
4. Champion for the engagement retained their position at the crucial time of set up

Stage 1 Understand policy needs and context

Stage 2 Undertake diagnostic studies and assess data capabilities

1. State government owned data - access was possible
2. Iterative learnings presented to finalize scope of work
3. Embedded Associate - deep collaboration with IT team
4. 3 additional researchers matched based on additional policy questions identified

Pending results from stage 3. The fraud triggers, once verified can be integrated into the government's monitoring system

Stage 4 Scale up successful interventions

Stage 3 Design and test innovations / policy solutions

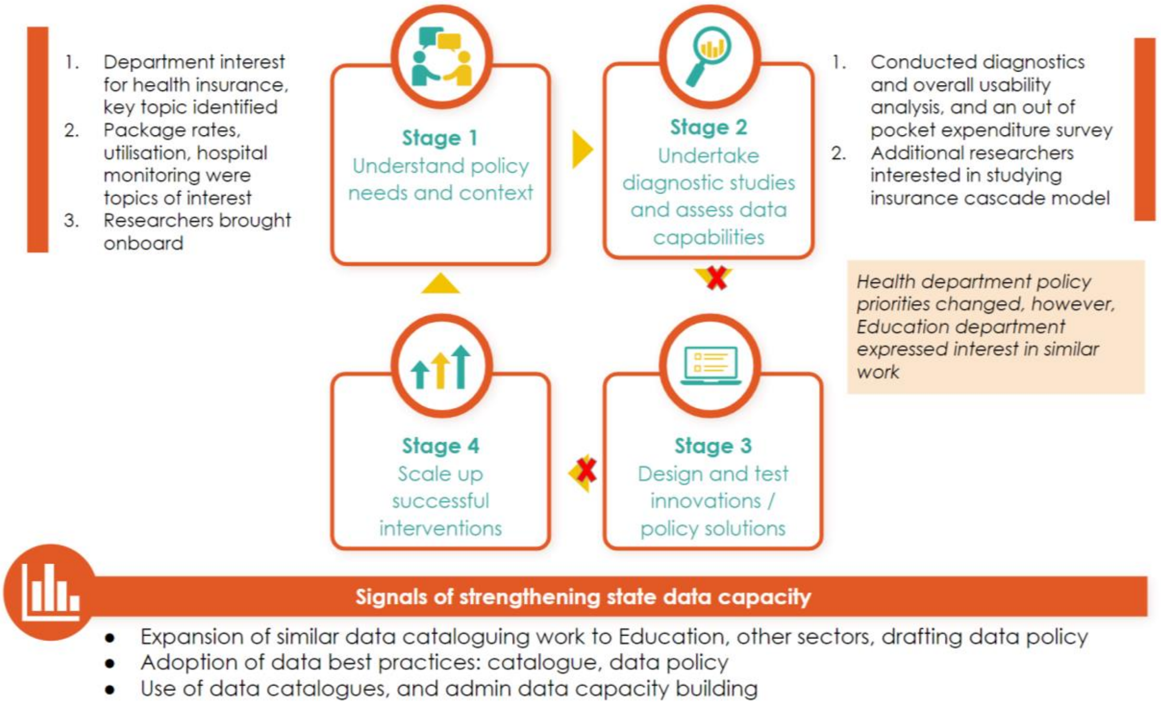
Implementation of finalized scope of work -

a. Develop hospital fraud triggers
b. Randomized evaluation for hospital monitoring

Training on data quality and analysis for several departments

In Figure 2, we showcase an example of success in building a data use case in the health insurance space in a particular state. Whereas, in Figure 3, we showcase an example where there was a break in the sequence of stages, implying specific interventions were not designed or tested in the health insurance space to build a data use case. Instead, there were positive learnings from the process of engagement. These relate to expansion to the education sector (driven by departmental interest), adoption of diagnostic products such as data catalogues as well as contribution to data sharing policy for research through trainings and advisory support to the interested departments and officials.

Figure 3: An example of contribution to state data capacity despite a break in the sequence of stages



1. Department interest for health insurance, key topic identified
2. Package rates, utilisation, hospital monitoring were topics of interest
3. Researchers brought onboard

Stage 1 Understand policy needs and context

Stage 2 Undertake diagnostic studies and assess data capabilities

1. Conducted diagnostics and overall usability analysis, and an out of pocket expenditure survey
2. Additional researchers interested in studying insurance cascade model

Health department policy priorities changed, however, Education department expressed interest in similar work

Stage 4 Scale up successful interventions

Stage 3 Design and test innovations / policy solutions

Signals of strengthening state data capacity

- Expansion of similar data cataloguing work to Education, other sectors, drafting data policy
- Adoption of data best practices: catalogue, data policy
- Use of data catalogues, and admin data capacity building

The following sections share insights on building partnerships, establishing governance and staffing structures, co-creating policy research questions for exploration, using data to execute the defined scope of work and organising capacity building activities.

5. Building and Developing Strategic Partnerships

As described in the preceding section, strengthening government capacity for use of data in decision making requires building a long-term view and a multi-pronged approach, given the complexities in bringing about change within government systems and processes. In order to develop a few data use cases as examples which could help generate demand for data for use in decision making, as well as catalyse investments in data usability and governance within government, it was necessary to adopt a broad approach to creating *enabling conditions* while maintaining the focus through working on a few specific *impact opportunities*. As demonstrated in Figure 4 below, a "double funnel" approach bookends the specific and collaborative data use cases – such that they don't remain isolated efforts, and help translate key lessons to broader, systemic efforts.

At the core of our approach was the need to co-create with the government the space for innovations by using data at all points of the decision lifecycle: from problem identification, to design, targeting and testing of interventions and to the scaling of the most successful and cost-effective interventions. For this, it was necessary to engage at two broad and complementary levels: (the top and bottom two bars of the double funnel):

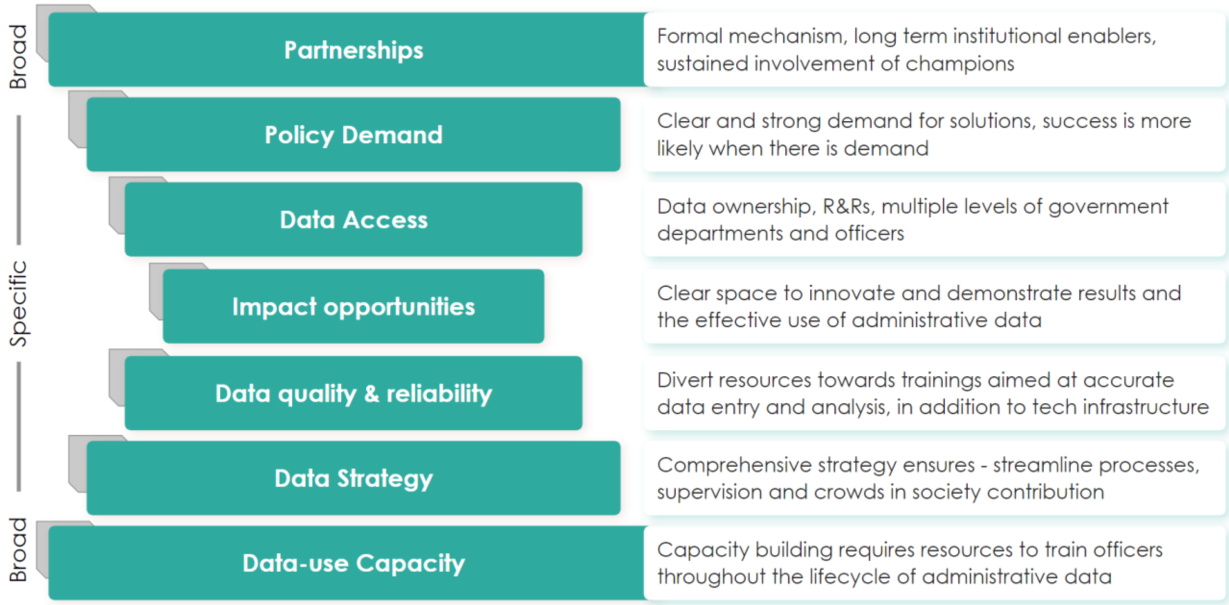
(1) **Formal and institutional partnerships with the relevant Ministry/Department with an explicit and stated commitment to evidence informed policy making:** This provides a clear and institutional mandate for developing data use cases. These formal mechanisms help sustain efforts through political and other transitions over time. Formal mechanisms established through partnerships and MoUs also help streamline formal channels of dialogue between policy makers, researchers and data intermediaries. While individual champions are crucial for bringing about change in existing systems, formal partnerships provide the basis for sustaining good practices beyond individuals to institutional processes. These mechanisms also help elicit demand for policy solutions and help in co-creating a focused scope of work for all participating entities.

In our experience, identifying the policy question(s) to be addressed first, and then assessing the availability and fitment of existing data and the need for additional data sources may be more aligned to encouraging take up of insights and solutions. Our original approach of identifying existing data in relation to broad policy challenges was more useful in identifying potential questions and understanding the context. However, significant effort and time was required to transition to designing and testing solutions to address these challenges. Thus, having a formal partnership with a well-defined scope and governance helps elicit policy demand for data use cases. Further, formal mechanisms also help put in place secure data sharing arrangements through MoUs and data use agreements to support the data use cases, aligning to the purpose of the partnership.

(2) **Assessment of data capacity and strategy**: At the other end of the funnel, attention must be paid to existing data capabilities and the need for strengthening capacity for data generation and integration. Different governments/ministries are often at varying levels in terms of their adoption of digitisation, staff capacity and familiarity with data management and analysis, and the level of integration of data into their review and decision-making processes. Undertaking an assessment of existing capacities of IT/data infrastructure, staff roles and capabilities, and mindset/culture using a structured data capabilities assessment format helped us contextualise the types of data use cases to be developed and the associated capacity building & advisory efforts that may be required. Based on our experience, engaging with government partners to provide inputs on data use potential for evidence-based policy making helps inform the process for developing comprehensive data use strategies. This helps provide perspective on trade-offs faced by the governments in balancing data use for public good with safeguarding data privacy. Ultimately, combining the data use cases with training for government staff on data quality and analysis is quite effective since it provides hands-on training experience on real data use cases. Integrating capacity building with the data use cases also helps embed good practice and process on data use over time, and also help generate more demand for good data and its use in decision making.

Thus, contextualising the core and specific impact opportunities within policy research partnerships and complementing with adjacent data capacity building activities help move the needle on systems change, even while the impact from individual projects are realised over time.

Figure 4: A double funnel approach to strategic partnerships


Broad
Specific
Broad
Partnerships
Formal mechanism, long term institutional enablers, sustained involvement of champions
Policy Demand
Clear and strong demand for solutions, success is more likely when there is demand
Data Access
Data ownership, R&Rs, multiple levels of government departments and officers
Impact opportunities
Clear space to innovate and demonstrate results and the effective use of administrative data
Data quality & reliability
Divert resources towards trainings aimed at accurate data entry and analysis, in addition to tech infrastructure
Data Strategy
Comprehensive strategy ensures - streamline processes, supervision and crowds in society contribution
Data-use Capacity
Capacity building requires resources to train officers throughout the lifecycle of administrative data

In the following subsections, we summarise how partnership development was operationalized through a targeted approach by cultivating champions, embedding teams, and ensuring an optimal governance and staffing structure.

5.1. Process of Developing Partnerships

Dialogues with policy makers aimed at **understanding key policy priorities** formed the first step of our four-pronged[15] approach. For this, we leveraged existing institutional partnerships at J-PAL South Asia and were also able to develop new ones through iterative dialogues aimed at co-creating a scope of work. It is important that these partnerships, while having specific objectives and focus areas, are still broad enough to allow for a range of collaborative activities such as research studies, training, customised workshops, technical support, and policy outreach & advisory engagements. This helps identify the nature of the demand for support from governments, strengthen our partnership and also helps understand baseline state capacity to generate and use data systematically. Specifically, the following aspects are useful in establishing strong and sustainable partnerships.

1. **Broad MoU:** A broad-based MoU anchored in a nodal agency or department of the government, such as Planning or Finance, provides a common framework to engage across many departments and agencies without having to enter into a separate agreement for every project. This reduces the time required for establishing sub-unit partnerships, credibility, and motivations.
2. **Governance structures and financing:** The MoU needs to detail the intent of the collaboration as well as detailed guidelines on governance. Each state government has preferred processes and institutional arrangements to oversee partnerships with external entities. Typically, these consist of committees with senior government officials, and at times political representatives (such as ministers). This helps secure a link to the adoption of insights and relevant decision-making authorities. Regular committee meetings help define which projects/activities are to be prioritised and review progress and outputs for policy action. Depending on the circumstances and resources, MoUs also detail financial arrangements and cost-sharing principles for the projects and activities under the partnership.
3. **Data use agreements:** MoUs for policy-research collaborations anticipate and provide for specific arrangements under which existing and new data would be shared between the government and external research partners. These Data Use Agreements (DUAs) are drafted using institutional review board guidelines from ethics boards of universities, as well as local and national laws on data privacy and security. These are necessary for guiding systematic and streamlined data-sharing arrangements for research activities. These agreements typically define the purposes for which data may need to be shared,

[15] Stage 1: Understand policy needs and context
Stage 2: Undertake diagnostic studies and assess data capabilities
Stage 3: Design and test innovations / policy solutions
Stage 4: Scale up successful interventions

who would have access to it, and the privacy safeguards to be adopted during sharing, analysis, publication, and redressal measures. The Fives Safes Framework provides detailed and helpful guidance on drafting agreements and safeguard measures that can be adopted for secure access to and use of data.

At the time of the start of this grant agreement and project, J-PAL South Asia had several ongoing state-level as well as individual department-level institutional partnerships. After scoping for regions where health insurance was listed among key policy priorities and existing data infrastructure, we were able to explore the potential for establishing a health insurance data use project in six states. Among these there were established state-level partnerships in 3 states (Tamil Nadu, Punjab, and Odisha), new partnership discussions ongoing in 2 states (Andhra Pradesh and Delhi), and a long-running research partnership with the health department in one state (Haryana). This project also helped establish contact and dialogue at the national level with the National Health Authority which oversees the implementation of the PMJAY scheme.

The following considerations further played a role in understanding and developing the value proposition for using data to improve the planning and execution of health insurance programmes, and in strengthening capacity to leverage and use data across the board for decision-making.

5.2. Level of Government and Data Use Champions

The demand for solutions largely drives the uptake of data insights and triggers change within the government. However, the demand can emerge from different levels and units. For a successful collaboration, it is essential to find or develop a champion at the right level of government, who will enable coordination and actively participate in the co-creation process. To develop the health insurance data use cases, there needed to be a certain readiness within the unit that directly oversees the implementation of the health insurance programme (typically a sub-unit under the health department). However, it was also important to assess the relative focus on health insurance compared to other health policy priorities. At the start of the grant period, for instance, most governments had directed a major proportion of resources and efforts towards managing the Covid-19 pandemic and the more urgent challenges emerging out of the pandemic. These included slippages in immunisation schedules, hospitalisation for non-emergency and chronic conditions such as dialysis, etc. Hence, it took some time for appropriate policy windows to emerge – where there was an interest to assess and allocate resources towards health insurance programme design and implementation.

Evidence/Data champions play an important role in formalising the mandate for a Ministry/Department to systematically use data for decision making. Further, continuous engagement and dialogue with these champions goes a long way to co-create and identify relevant policy research questions and better understand the local context and mechanisms. These individuals also play a strong role in institutionalising best practices based on the engagement and learnings, through formalising MoUs and data sharing arrangements, putting in place governance mechanisms for different partnerships etc. In all states, officials led flagship efforts to set up and expand data analytics/evidence units and encourage the use of research and data to inform critical policy questions. This helped anchor the efforts to develop data use cases. Similarly, in some states, efforts to develop State-level data policies helped streamline processes for secure data sharing and governance.

Further, there must be alignment across different decision-making authorities within the government. For instance, the focus of the unit administering the health insurance programme may differ from that of the finance and planning departments or even within the health department, say the public health department. In some states, the policy window to engage arose from the need for renewal of the tender for identifying third-party insurance partners to administer the programme. This allows for renegotiation of the premiums, adjusting package rates, determining inclusion/exclusion criteria etc, and thus analysing the coverage and utilisation data patterns from the previous insurance period was useful. In other states, where an assurance model was being followed, it was important for the government to understand efficiency metrics on turnaround time in claims processing, utilisation of public versus private facilities, conditions for which the insurance was being used etc, since the entire process was managed end-to-end by the health department. Finance departments, especially in states providing near-universal coverage, were keen to understand the coverage and details of utilisation of the insurance programme and its effectiveness in improving health outcomes. There was interest in understanding how insurance coverage and use intersected with other aspects of public health expenditure, for example, free health screenings or other such preventative measures. Across the board, health departments were interested in understanding causes and instances of inefficiencies in implementation - hospital fraud, co-payment by beneficiaries, etc. An indicative list of questions, some across all states and specific to each context are summarised below.

Table 1: DEMAND FOR DATA USE IN HEALTH INSURANCE
Some common questions of interest
Analysis of coverage, enrolment, and eligibility by geographies (districts) and population sub-groups.
Analysis of utilisation across geographies, population sub-groups, types of procedures etc.
Beneficiary experience, quality of care, out-of-pocket expenditure
Identification of Claims Fraud, Evaluating Impact of Increased Hospital Monitoring, and set up Hospital audit mechanisms
Specific questions of interest relative to context, programme design
Choice of care and difference in quality between government and private hospitals

Table 1: DEMAND FOR DATA USE IN HEALTH INSURANCE
Hospital responsiveness to changes in packages and rates: Changes in claim volume in response to changes in treatment packages and their rates.
Benchmarking scheme rates to market rates and identifying over and under-priced treatments.
Analysis of programmes spending and cost effectiveness
Analysis of preventable hospitalizations
Determinants of hospitalisation

Another approach is to complement this exploratory phase of eliciting demand through a competitive call for proposals from innovators and researchers on potential solutions for consideration. Through the course of this project, we followed this approach sequentially, where we introduced the idea of a competitive call for proposals from locally based researchers defining some minimum criteria regarding the use of administrative data in gap identification and solutioning, and contributing to state capacity for leveraging the same. We also expanded the sectors of interest beyond health insurance to leverage opportunities as they arise in discrete policy windows.

A range of ideas was received – from developing a deeper understanding of policy challenges and gaps in service delivery, to innovative solutions that could improve cost-effectiveness. Out of 14 proposals received during the competitive call for ideas that ran for 4 weeks, 4 short-duration proposals were selected. It is valuable to combine both approaches to better match policy questions to quick solutions and encourage the innovative use of data. The broader partnership structure allows for a systematic articulation of policy research questions, terms, and processes of engagement between researchers and government counterparts. The competitive call for ideas to address specific policy questions, similar to a policy hackathon, encourages the identification of new data sources and potential for linkages in non-traditional ways. This may be an efficient process for identifying a suite of solutions that could be tested for feasibility and cost-effectiveness. This helps the process of systematic and meaningful interpretation of insights from data and evidence for making course corrections to existing programme implementation or adoption of new solutions.

5.3. Scope of Engagement: Depth and Breadth

The details of the scope of engagement with each government partner varied considerably based on the nature of demand for solutions and data use, the existing (baseline) technical skills, capacity in the state (including the existence of dedicated data analytics units), overall priority within the political context, data accessibility and quality, and the presence of key champions of data use. To determine the scope of engagement, and ultimately to make progress along the intended outcomes of strengthened and sustained state capabilities for data use, it is very useful to start with a needs/capacity assessment and mapping stakeholders. We followed a process of assessing data capabilities along three main categories: data readiness and usability (focused on data accessibility, documentation of fields, and data quality checks); technical capacity of staff and financial resources available; and mindset for data use (readiness and processes for data sharing, use of visualisations/dashboards, review practices etc.). An indicative format is available here, which could be used for assessing baseline data use capabilities as well as tracking progress and outcomes over time.

This was combined with policy-research dialogues to understand urgent and long-term challenges and goals of the government on health insurance, and the availability of data sources to help determine the specific scope of engagement with the government. This is also, of course, dependent on the duration, financial resources, and interest of the collaborating research organisation, and thus the activities finally selected should be aligned with the goals and comparative advantage of the collaborating institution as well.

While approaching with a focus on health insurance helped with anchoring the discussions to concrete ideas and an actionable process for demonstrating a pathway for a specific use case, the demand for engagement on data use is often much broader within the government. In some states, there was a need to start with capacity-building efforts such as metadata documentation of available datasets (i.e. developing a comprehensive catalogue of what data is being collected/available with different departments, related to say health, education etc., what are the data fields, a high-level assessment of quality, and its accessibility for analysis and use). In other instances, there were requests for training and sensitisation of staff (both functional/ programmatic officers at the state/district level, and M&E, MIS, and other staff with data analysis and data visualisation/reporting responsibilities). A series of workshops and consultations were conducted on data quality, preparation for analysis, and interpretation. These are best done using real datasets in the state – which allows the training participants to apply the skills acquired in practical use cases. In other states, where the data capabilities were more advanced, there was a need to engage with nodal departments like Planning or eGovernance to discuss the contours of designing data use policies and guidelines. These include frameworks that allow for meaningful and relevant data analysis, both within the government and with external collaborators for policy research, while ensuring the protection of individual privacy and data security of the administrative (non-public) datasets. In a couple of states, governments had made efforts to create a government-wide institutional arrangement for encouraging data analysis and use internally. These were typically set up as Data Analytics Units within Planning or Finance departments (or within eGovernance agencies). These units were intended to provide guidelines across the board on standardising data formats, creating platforms for interlinking, developing dashboards and reports, implementing data sharing and use guidelines and policies, and supporting departments in strengthening their data systems and processes. These units also served as important partners for strengthening state capacity for data use through a range of capacity-building efforts - from consultations, embedded staff, and providing advisory inputs.

Table 2: SNAPSHOT OF PHASE-WISE ACTIVITIES UNDERTAKEN

Phase/Activity	Health Insurance theme	Health and climate change theme
Phase 1: Existing broad-based MoU	In 4 states, 1 new MoU established	In 2 departments, 1 NDA signed anew
Phase 1: Commitment to engage on the agreed theme	Outreach to 6 states, 4 commitments received	Commitment received from 3 departments in 2 states
Phase 1: Needs/ data capabilities assessment	Conducted with departments in 4 states	Topic specific discussions conducted with 4 departments in 3 states
Phase 1: Policy dialogues with key officials to identify priorities, challenges	Conducted with departments in 4 states	Conducted with 3 departments in 2 states
Phase 2: Finalised scope of work	Conducted with departments in 4 states	Conducted with 3 departments in 2 states
Phase 2: Data access modalities, access to data obtained	Accomplished with departments in 3 states	Accomplished with 3 departments in 2 states
Phase 2: Capacity building/ advisory inputs	Accomplished with departments in 3 states	Studies ongoing, recommendations to be submitted
Phase 3: Execution of planned scope of work	Accomplished with departments in 3 states	Ongoing with 3 departments in 2 states

5.4. Governance and Staffing Structure

To staff this project, we adopted a hub-and-spoke model. The hub was centrally located and responsible for planning, execution strategy, partnership development, troubleshooting execution-related issues, and supervising and guiding spoke teams. The spoke teams were resources embedded in the respective government departments for day-to-day follow-ups with the respective IT teams to secure access to data, understand the broader data infrastructure, develop tools such as the metadata catalogue, and share interim exhibits with a nodal officer. In addition to this team that was specifically set up for exploring administrative data use cases, support from internal pre-existing teams was leveraged. For example, the policy team at J-PAL South Asia was leveraged to initiate policy dialogues within existing state partnerships and to reach out to the health departments of other states. Selected researchers from the pool of J-PAL affiliated researchers were brought on board as subject matter experts to drive the research agenda that came out of the policy dialogues with respective government departments. Further, the team also received strategic oversight from Scientific Directors, Project Director, Executive and Deputy Executive Directors, as well as the Director and Associate Director of Policy. Therefore, a larger ecosystem of J-PAL South Asia was leveraged to competitively deliver on key engagements with the government.

In our experience, the following considerations may help determine staffing structures and support when working with governments to institutionalise state capacity for data use:

- In both new and existing partnerships, there is an important role to be played in managing and stewarding the partnership by senior staff that can help identify policy windows, unpack priorities and constraints and connect to relevant teams and experts to develop research and capacity-building opportunities. This is best achieved through at least one full-time staff person (with about 8-10 years of experience working with governments and with sufficient domain expertise on location) ideally engaging frequently with all key stakeholders in government.
- Depending on the state capacity, there is a case for embedding data professionals to help streamline and institutionalise processes and practices for data quality, and secure protocols for sharing and use of data. Almost all government ministries and departments can identify and define the roles of IT/MIS professionals, who are more focused on the systems and infrastructure. Data manager roles are relatively new within governments and there is substantial variance across different governments, between national and state, across states, and sometimes even within a state across departments. These newer roles also require clear scopes of work to enable talented professionals to deliver value. Finance or Tax departments typically have clear and stated data use goals, and also have qualified personnel for managing and analysing data. This staffing structure and talent pool is quite limited in social sector departments such as health, education etc. and is now expanding. Through our engagements with government partners on this project, the need for exploring complementary approaches to strengthen the data ecosystem, especially for cultivating data talent came to the forefront. J-PAL South Asia is hosting an innovative engagement with data.org to address this need often expressed by government partners to embed "data fellowships" within government (and social enterprises) for limited periods and for specific goals. Data fellows have been embedded in the researcher model and in some state partnerships on specific projects.

Leveraging and building strategic partnerships anchors an engagement with the right department and data champions committed to a common theme. The next step involves building a detailed scope of work, summarising key aspects of the task involved, dividing responsibilities of each stakeholder (including researchers as well as the government department), and identifying the deliverables to be generated.

6. Co-Creating Questions and A Scope of Engagement

Co-creating policy questions and defining the associated scope of work with government departments offers an opportunity to embed and test policy solutions in real-world settings in collaboration with policymakers. To this effect, we defined a framework, as described in section 2, to give structure to our engagement with government departments which allows for replicability.

Our framework involved a four-stage approach[16], where the latter two stages were largely conceived to occur after the grant period. We expected that the preparatory and exploratory work undertaken through the first two stages would lead to better design of interventions and decisions using insights from data, both in health insurance and in other sectors, along with institutionalising the use of data and evidence in policy-making. We had two strategic approaches in operationalising the framework,

i) a sequential approach of engaging states as per the numbered sequence of stages. This could be done through two types of teams - stages 1 and 2 could be done through generalist policy and data experts, while stages 3 and 4 will require researchers and domain experts,

ii) approaching government departments through stages 2 and 3 in parallel. This entailed identifying domain experts in parallel who have promising policy research ideas and taking these to respective government partners.

In our experience, both approaches had their benefits and challenges. We summarise here, the outcomes that can be achieved through these two approaches.

6.1. The Sequential Approach of Defining Problem Statements

Through a sequential approach, we reached out to four state governments – Punjab, Haryana, Tamil Nadu, and Andhra Pradesh, and were able to execute the activities under each stage to a different extent in each context. Stages 1 and 2 were critical for co-creating research questions. Policy gaps and needs assessments, followed by data diagnostics, helped us understand the perceived gaps as per the government department, as well as the availability, quality, and usability of administrative data. Stage 3 involved engaging researchers, who have expertise in the respective policy areas, to identify potential policy solutions that could be tested. Any intervention that is successful through rigorous testing could then be scaled up. This approach allows us to be thought partners on some very specific aspects of a larger department priority,

[16] Four Stage approach -

- Stage 1: Understand policy needs and context
- Stage 2: Undertake diagnostic studies and assess data capabilities
- Stage 3: Design and test innovations / policy solutions
- Stage 4: Scale up successful interventions

ensuring required government buy-in for the project duration, and as a result greater potential for research to influence policy processes and decisions.

Through our engagement with the state governments, we have observed that research opportunities can be identified on two fronts - (a) the design of the scheme/policy itself (b) policy/ programme implementation monitoring and evaluation involving administrative data.

For the health insurance use case, this meant starting with exploratory and descriptive analysis of the trends and patterns in the main health insurance datasets: eligibility, claims (utilisation), and hospital services. Key questions included identifying variations in health insurance coverage by geographical location or population sub-groups relative to the target population eligible to be covered. Similarly, patterns in utilisation of health insurance were indicative of usage by demographics, across private and public facilities, and by health condition. For questions relating to quality of care and patient experiences, and gaps in service delivery, it was necessary to complement the administrative data with survey data. This involved conducting out-of-pocket expenditure (OOPE) surveys and running diagnostic checks on the data to understand its quality. In one instance, we conducted the first OOPE survey for a small sample of beneficiaries as a gap assessment tool for checking the efficacy of health insurance schemes and identifying instances of payment by beneficiaries in private and public facilities. The survey instrument is easily adaptable and can be used in other instances in other states as well. However, one survey is just one snapshot of the issue. This needs to be complemented with further investigation using internal monitoring records and interviews. Based on the extent of the issue and geographical variation, appropriate corrective measures have to be designed and tested. This process takes time and can be taken up when there is prioritisation, willingness, and resource commitment from the government department to engage in one or more potential solutions.

In other instances, health insurance departments were interested in strengthening their hospital monitoring systems. To describe one such question in detail, with one of the state governments, the aim was to improve fraud detection and deterrence in a cost-effective manner. Through iterative data access and discussions, a scope of work was agreed upon where J-PAL affiliated and invited researchers[17] could offer their expertise. The scope of work involved two broad workstreams -

1. *Analysis of Program Spending and Cost-effectiveness*: This involved identifying areas of excess spending and ways of increasing program effectiveness and efficiency through a detailed analysis of spending and hospital services data.
2. *Identification of Claims Fraud and Evaluating Impact of Increased Hospital Monitoring:* This involved developing a stronger hospital monitoring system that can detect and deter claims fraud in a cost-effective manner. The study consists of two key pieces:

[17] Pascaline Dupas (Scientific Director, J-PAL Africa; Princeton), Radhika Jain (University College London), Shailender Swaminathan (Krea University)

a. Development of a package of fraud triggers based on health insurance claims administrative data. The quality of these triggers will be tested through field surveys and hospital audits.

b. Evaluation of the impact of increased threat of detection and hospital monitoring.

Some key learnings from our experiences in multiple states:

i) willingness to explore solutions for a scheme or policy area in conjunction with researchers varies by state and department;

ii) while government departments have the required IT infrastructure to develop and maintain websites and applications to track the implementation and uptake of a scheme, the capability to use the data effectively for monitoring and decision-making has immense scope for improvement;

iii) Policy research using administrative data often requires in-field validation or data collection for variables that are either not an operational requirement of the administrative dataset, missed in the development of the tracking system, or are unique to the research question.

This space, therefore, can benefit from collaboration with researchers and domain experts who can contribute to developing monitoring metrics and decision-making frameworks based on administrative data as well as primary survey data. Such frameworks can also be tested rigorously before being taken up by the respective government departments. This is a general framework for a strong use case of administrative data. This insight came specifically from our sequential approach.

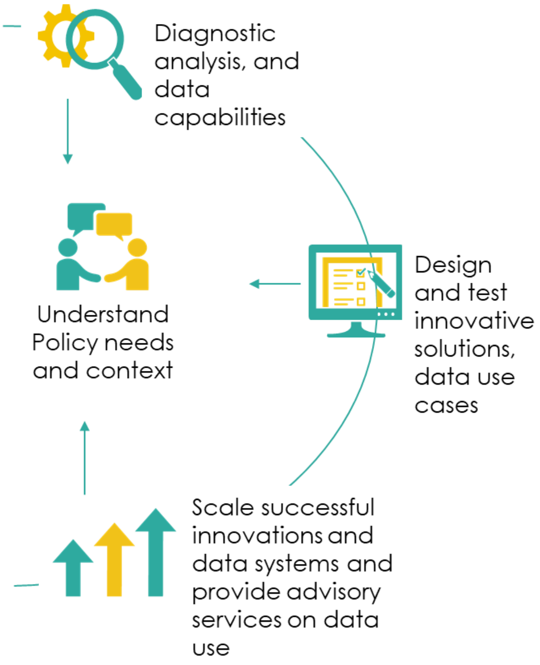
6.2. The Researcher-Led Approach of Developing Solutions

In our journey of working with administrative data, we also explored a different strategy for co-creating policy-research questions and associated scopes of work. This strategy complemented stages 1 and 2 tangentially. It involved approaching respective government departments directly with stage 3. We at J-PAL were able to identify researchers who already had developed a good understanding of policy priorities through secondary research and/or experience in the field or prior engagement with policy stakeholders. Therefore, they had a suite of possible solutions and research ideas that were ready to be tested. We have used this approach often in other thematic areas through the Indian Scholars Program.

- **Why start with stage 3?** With our experience of working on this project we discovered that there are significant value adds when researchers engage with governments on administrative data to co-create policy-research questions. However, the time taken to arrive at this scope of work was longer than we anticipated. This is because our four-stage approach turned out to be iterative and non-sequential in practice as opposed to the assumed sequence of the numbered stages. We observed that stages 1 and 2 particularly were time intensive and often did not linearly translate to stage 3. This was due to many reasons including the need for primary data validation, lack of perfect convergence with government and/or political priorities etc. A valuable learning here is the importance of questioning the assumption of linearity in systems change or process adoption thinking, especially when we are envisioning institutionalising culture in a complex environment.

- **Need for broader thematic areas**: We observed that while there was a larger mandate of investing in data systems and data capacities in some states, the thematic focus did not lie in health insurance. For example, we were not particularly successful in our health insurance work in one of the states despite having steady relations with the government department on other research projects. As a result, the government showed interest in scaling up another study which also involved the assessment of the administrative data use case, though for a separate thematic area – routine immunizations. This served as an unanticipated pathway for developing data use cases and created the learning that having a nimble approach can create more opportunities
- **Leveraging strengths of researchers**: The presence of researchers is crucial in the development of policy-research questions and test different solutions. We engaged with India-based researchers, to develop policy-research ideas that leverage administrative data and engage with departments. In this manner, one can start with stage 3, where stages 1 and 2 are completed through preliminary independent research efforts and then confirmed with the government partners when the relationship is established.

Figure 5: Non-sequential process to institutional change


Diagnostic analysis, and data capabilities
Understand Policy needs and context
Design and test innovative solutions, data use cases
Scale successful innovations and data systems and provide advisory services on data use

BOX 2: IMPROVING AWARENESS OF ARSENIC POLLUTION IN ASSAM

Arsenic pollution affects over 50 million people in India across 20 districts[18], including Jorhat which is home to more than 924,952 people[19]. The researcher-led approach allowed[20] the identification of the prevalence of arsenic pollution, issue of lack of awareness about arsenic pollution in groundwater and its associated adverse health outcomes. The researchers designed and implemented interventions to improve the take up of the Jal Jeevan Mission (JJM)[21] in Assam (Jorhat), where the groundwater is polluted by naturally occurring arsenic.

To increase the demand for JJM among the public, the researchers worked with the Public Health Engineering Department (PHED) to test an information campaign along with other interventions aimed at reducing the transaction cost of applying to this scheme. This turned out to be a key priority of PHED, who also helped streamline the intervention related to reducing transaction costs and supported the research that followed.

Administrative data on tap connections provided under the JJM scheme was used for this study. While the administrative data contained tap connections to track the service delivery against expenditure incurred, the dataset did not have any indicators on utilisation and maintenance of taps as well as the reason for low uptake of the scheme, because such indicators were not relevant from an operational lens. The researchers complemented the administrative dataset by undertaking a household-level survey to understand utilisation, reasons for non-take up as well as long term maternal and child health implications. While the administrative data alone was insufficient to measure changes in outcomes, the study helped highlight the importance and ways of combining and using existing data sources with survey data. This collaboration between researchers and PHED, allowed for exploring different dimensions of utilisation of PHED's data.

Across our engagements, we observe that governments can sometimes be more interested in scaling up a proven study or testing pre-identified solutions rather than taking the longer route of exploratory research for solutions through iterative administrative data analysis. In addition to being responsive to policy demand windows, researcher commitment to policy innovations and solutioning also means bringing together more experts and organisations to solve a broader set of issues, and not being constrained to the resources and mandate of one organisation, thereby creating an ecosystem of experts to complement efforts and solve complex challenges.

In Figure 6, we summarise below the general lessons learnt from both approaches.

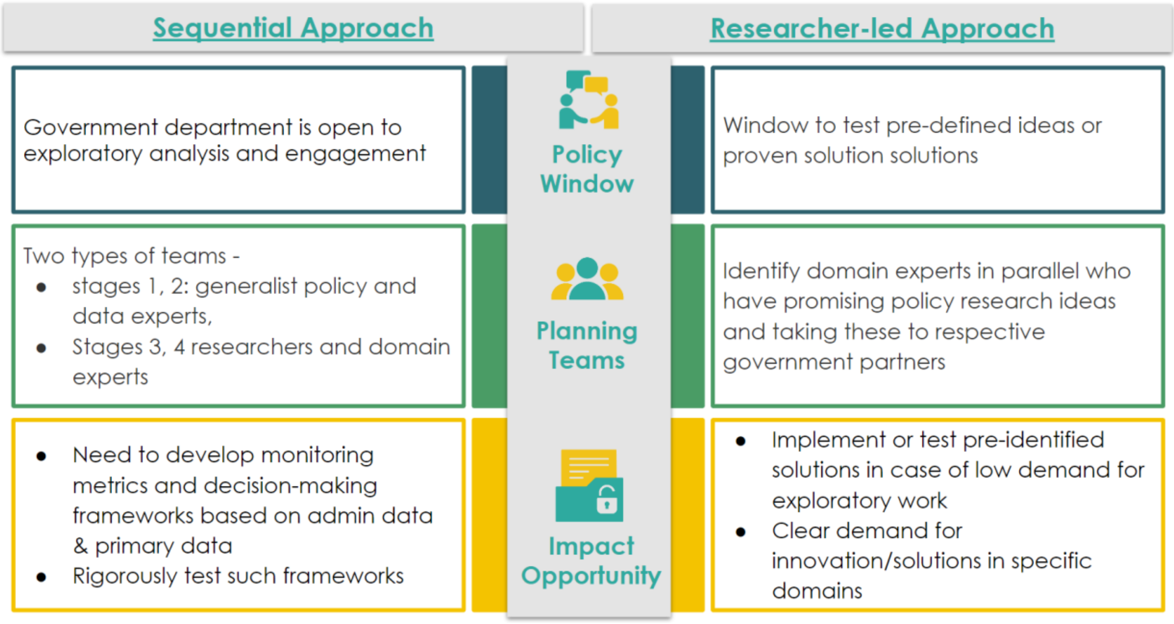
[18] Shaji, E., Santosh, M., Sarath, K., Prakash, P., Deepchand, V., & Divya, B. (2020). Arsenic contamination of groundwater: A global synopsis with focus on the Indian Peninsula. Geoscience Frontiers, 12(3), 101079. https://doi.org/10.1016/j.gsf.2020.08.015

[19] Home | Jorhat District | Government of Assam, India. (2022, April 7). https://jorhat.assam.gov.in/

[20] Dr. Rashmi Barua and Dr. Prarthna Goel

[21] Jal Jeevan Mission (JJM) aims to provide Functional Household Tap Connection (FHTC) to every rural household by 2024, with a minimum water supply service standard of 55 LPCD. JJM subsumed the erstwhile National Rural Drinking Water Programme (NRDWP)

Figure 6: Summary of strategies to approach departments



Sequential Approach		Researcher-led Approach
Government department is open to exploratory analysis and engagement	Policy Window	Window to test pre-defined ideas or proven solution solutions
Two types of teams - • stages 1, 2: generalist policy and data experts, • Stages 3, 4 researchers and domain experts	Planning Teams	Identify domain experts in parallel who have promising policy research ideas and taking these to respective government partners
• Need to develop monitoring metrics and decision-making frameworks based on admin data & primary data • Rigorously test such frameworks	Impact Opportunity	• Implement or test pre-identified solutions in case of low demand for exploratory work • Clear demand for innovation/solutions in specific domains

Administrative data can potentially be a rich source of data for policy research as well as decision-making. However, such data has been historically collected for operational purposes, may require combining different datasets. Many datasets, even today, require extensive cleaning and at times are incomplete for policy-research queries. In Annexure 1 we discuss different stages of understanding, accessing and using administrative data for research.

7. Drivers and Outcomes

Our learnings from this journey are rich and multifaceted. These span from the kind of data infrastructure one is engaged with, the logistics to legalities of data access, the quality and usability of the administrative data, as well as the possibility of integrating our learnings and use cases within the business-as-usual monitoring systems of government departments.

While co-creating questions and executing scopes of work did not guarantee the systematic adoption of insights into decision/action or investments in strengthening state capacity for data use, it did help us identify different pathways, resources, and tools that gained traction within government departments. For instance, metadata catalogues were a key resource that was used by the education department in Tamil Nadu and to an extent in Punjab, whereas, data-driven process improvements gained traction in Andhra Pradesh.

The success of the larger state engagement itself was defined by three key drivers

- Ability to access data;
- Possibility of advanced and aggregate analysis of data along with merger with complementary datasets;
- Continuity of department and state policy priorities to action insights amidst administrative shuffles and other constraints.

Successful execution of the scope of work and improved outcomes relied heavily on the presence of all three drivers. If one was missing, we achieved only partial success in our engagements in the given time. Regardless, the mandate of institutionalising data use has multiple dimensions – suggesting that success, too, has multiple dimensions, and we anticipate policy windows will continue to emerge over time.

A general lesson on using administrative data is that there is a considerable time cost in understanding such data and securing access to the right subset of variables at the required frequency and detail. Our teams often started with a survey conducted with IT officials within departments to gain a comprehensive understanding of the data infrastructure, its usage, and the frequency of data updation, among other things. Following this unstructured survey, the following steps were taken by the embedded teams. This is also explained in detail in Annexure 1.

1. Data quality - assessment of reliability, completeness, representativeness, and interoperability of datasets
2. Data access - ensuring data is accessed in an encrypted format and transferred securely. Internal compliances such as IRB approvals were also secured since many of these datasets contained personally identifiable information (PII)
3. Data analysis, dissemination, and capacity building - Once data is accessed securely, it is cleaned and assessed for fitment for the scope of analysis. Preliminary insights are shared with departments iteratively to keep them involved in the data analysis process and to also understand the reason behind certain trends in a better manner. Administrative data will not have all variables required for the agreed scope of analysis, therefore, complementary primary data collection activities are designed accordingly. Lastly, capacity building workshops are organised for department staff to disseminate our findings and train the staff on improved data monitoring practices relevant to their department.

Figure 7: Key outputs from our engagement with governments on administrative data use

1000+	Datasets catalogued and created respective codebooks *several thematic usability reports were generated as per departmental priority*
100+	Officers trained on the use of administrative data, across 4 states
12	Researchers working in collaboration with state governments to build data use cases and policy-research questions
6	Data use cases/studies contributing to better use of administrative data for decision-making

Our deep engagement has also translated into outcome-level process changes, where we contributed to 2 state data policies, as well as the creation and strengthening of data analytics units in 2 states.

In addition to specific use cases, at J-PAL, we continue to generate evidence and synthesise learnings to inform policy. Through our larger work, we have also been able to inform policy through other embedded teams in different departments, randomised control trials to understand the impact of certain policy ideas, as well as scale-ups of successful studies (such as the scale-up of a gender-based curriculum, ‘Taaron ki Toli’ (TKT) in Odisha and the piloting and expansion of emissions trading in Gujarat). Such cases of adoption of insights from policy are forms of using data/evidence at different stages of the policy lifecycle.

8. Sustaining Learnings for a Stronger Ecosystem

A key theme across all learnings is that the context of impact and collaboration differs by state as well as problem statements under consideration. Therefore, such work requires a ‘plumbing’ approach of iteration-continuously learning and consistent tweaking. For example, we were the most successful in the last state that we worked in because we were able to implement an approach that had undergone several iterations based on learnings from other states. This could only be possible through an innovation-friendly and collaborative funding structure that enabled us to respond to these opportunities and fostered innovative thinking.

8.1. Learnings in Review

Our overall learnings show that building data use capacity forms a double funnel in terms of the scope of an engagement as summarised in section 5. The broader policy priorities are discussed at the partnership building stage. At every subsequent step of success, the scope of work is narrowed to enable appropriate data access and define specific impact opportunities, which forms the data use case. The subsequent steps become broader in terms of learnings, lessons and capacity building as shown in Figure 2.

We summarise key takeaways in each of the above steps for the broader ecosystem working with administrative data. This includes NGOs, think tanks, research institutes, donors as well as government departments.

1. **Developing strategic partnerships**: Establishing effective and lasting partnerships with government departments requires sustained engagement, trust-building, and clear agreements.
    a. **Multi-level engagement**: It is essential for senior leadership to invest time in building trust and understanding government priorities. This can be reinforced by deploying an embedded team that regularly engages with officers across different levels to ensure continuity and responsiveness.
    b. **Formal mechanisms**: MoUs and Data Use Agreements (DUAs) provide a stable framework for collaboration, clarifying roles, responsibilities, and continuity even amid administrative changes. Leveraging long-standing institutional partnerships can also facilitate early conversations with departments, allowing for smoother project initiation.

    c. **Sustained involvement of champions**: Identifying and collaborating with a departmental "champion" who advocates for the project internally is crucial for achieving lasting impact.

2. **Aligning with policy demand**: Policy relevance and demand drive successful data use initiatives. Therefore, ascertaining policy demand implies the need to get more specific about what the department is interested in and is therefore willing to engage and deploy resources in that direction.
    a. **Responsive engagement models**: Depending on the specificity of the policy goals, it may be useful to choose between a sequential or a researcher-first approach. If priorities are highly specific, engaging a domain expert early on can support tailored research questions and more effective relationship building.
    b. **Resource allocation based on policy windows**: Organisations should be prepared to scale resources up or down as policy windows shift, ensuring efficient allocation of time and budget according to evolving priorities.
3. **Securing data access**: Policy demand will determine the direction of engagement and co-creation of a scope of work. At this stage, the topic under consideration becomes even more specific as the government departments will allow access to limited datasets depending on need. At this stage understanding of data ownership is critical, as approvals for access need to be sought from the right owner in the right level of the government.
4. **Identifying impact opportunity**: Once the scope of work is agreed upon, data access is obtained, and the data is deemed usable for aggregate analysis, there is an opportunity to create impact. Researchers can contribute to the government's systems by deploying advanced methods on administrative data as well as conducting field validations, depending on the agreed scope of work. This is the most specific stage in the entire process and the impact is often incremental, yet has the potential to be a catalyst for further change. At this stage the presence of a domain expert is critical to drive the conversation.
5. **Ensuring data quality and reliability**: Ensuring high-quality, reliable data is foundational to effective decision-making. Investing in training for accurate data entry, validation, and regular quality checks helps build a foundation of reliable data. Integrating quality standards and monitoring routines across systems and departments further supports data-driven insights. Lessons from the data use case developed at the impact opportunity stage will have broader insights for data quality and reliability.
6. **Establishing data strategy**: Data use strategy is important in defining the features and functionalities of a portal for a particular program. Aggregate data analysis, along with consultations with experts on optimal monitoring frameworks can ensure that spending on technological infrastructure includes effective program monitoring.
7. **Building data-use capacity**: Building data use capacity for better use of administrative data requires training officers across the lifecycle of the administrative data. The data use case developed will have broader learnings on the need of trainings required depending on the level of data use at a particular department.

8.2. Institutionalising Change at J-PAL South Asia

As we reflect on this journey, we find that we developed deep relations with government partners through our work with administrative data. While the sequential and the researcher-led approaches of engaging with government stakeholders have their respective strengths and disadvantages, both have the potential to foster research and policy action through administrative data. Through the sequential approach, we spent more time on data diagnostics, where we assessed the breadth of data maintained by departments through a metadata catalogue and assessed its depth through data surveys and other secondary diagnostic tools. Through the researcher-led approach, we were able to identify the gaps or potential improvements to data systems or a scheme, thereby demonstrating data use cases by merging datasets of different government departments. We also had the opportunity to evaluate a pilot scheme, where researchers were able to demonstrate key indicators that could be tracked to assess the outcome of the scheme.

All of these efforts were possible since this grant was set up to cover both the development of health insurance use cases and to engage in broader systems-strengthening activities. In some instances, where the timing wasn't relevant for a health insurance use case, other examples from related sectors of health systems and climate change served as the basis for these engagements. There is a need for flexibility in responding to the interests and policy windows of the government counterparts. The same process of engagement can be adopted for a wider range of sectors/policy questions. The funding from GIF was also complemented by additional funding raised for these efforts.

- Notably, CLEAR South Asia is focused on developing Monitoring and Evaluation (M&E) capabilities in the region, especially within governments. This funding helped establish a knowledge-sharing partnership with the NITI Aayog's Development Monitoring and Evaluation Office (DMEO). Under this, two knowledge products[22] on leveraging the potential for administrative data for decision-making have been developed, including insights from the health insurance projects. The CLEAR South Asia support also enabled several customised workshops and consultations on data use.
- A unique initiative was launched in 2022 to address the issue of limited data capacity within governments and social enterprises in the global South. The India Climate and Health Data Capacity Accelerator is focused on training young practitioners in interdisciplinary data sciences for social impact. This is combined with a one-year paid fellowship on data use projects with government partners or non-profit organisations, specifically to solve emerging challenges in climate, health, and their intersection. It is hoped that the fellowship model will help demonstrate the value of data use roles and training within government departments, and provide a pathway to create such roles in the future where they don't exist.

[22] DMEO, NITI Aayog, CLEAR at J-PAL SA, Compendium of Case Studies on Using Administrative Data for Evidence-Based Policymaking, link
DMEO, NITI Aayog, CLEAR at J-PAL SA, Administrative Data Toolkit, link

- Finally, a consolidated initiative focused on building, expanding, and deepening government partnerships at the national and sub-national levels, with a focus on scaling up evidence & data-informed innovations has provided the flexibility and resources to engage in long-term and flexible engagements with governments. This enabled support in a range of exploratory, research, capacity building, technical advisory, and policy support dialogues. The initiative is also set up as an alliance of donors, implementing partners, researchers, and governments to combine efforts and collaborate for greater impact at scale.

At an organisational level, we retain the learnings and practices of the sequential approach through our state partnership model where our engagement with states is on a broader level across sectors and policy priorities. Specifically, for Tamil Nadu, Andhra Pradesh, and Punjab, having an organisation-level state partnership allowed us to remain responsive to the government in the face of potential opportunities and changing government appetite for collaboration. The researcher-led approach on the other hand, is integrated within our organisation through (i) all the work that is independently driven by J-PAL affiliated and invited researchers (ii) researchers who are competitively identified through J-PAL's flagship program scheme - the Indian Scholars Program (ISP).

Considering data as a means of monitoring the status quo to understand and improve public service delivery, the key success of any approach lies in the ability to reach stage 3, which is testing innovations and solutions for the gaps identified, and ultimately scaling and sustaining these solutions to improve welfare outcomes. J-PAL has been at the forefront of evidence-based policy research for more than two decades. Our journey so far has shown us the possibilities of causal inferences with data which was predominantly collected on the field in the early years but is now increasingly leveraging pre-existing large administrative datasets. Many departments within the Indian government have varying quality and quantity of administrative data that could be leveraged for research, even for advanced analysis including AI and machine learning. We continue to explore how such emerging technology can enhance state capacity to generate reliable data and accelerate its use in decision-making while leveraging deep local knowledge of the government officials administering the programme as well as the global domain expertise of researchers to design and test appropriate solutions to improve service delivery, governance, and effectiveness of programmes. We continue to sustain the efforts initiated under this engagement through our deep partnerships with governments and a consolidated suite of services combining rigorous research, hands-on capacity building, and timely policy advisory services aimed at integrating data and evidence into each step of the policy decision lifecycle.

Annexure 1: Guidelines and Resources on Leveraging Administrative Data

In this section, we summarise our experience of working with administrative data. We draw out key lessons which will serve as a guide for researchers and policy professionals interested in including administrative data within their policy-research projects. We summarise our experience under four broad heads - data survey, quality, access, and analysis that we adopted throughout this project.

1. Data Survey

A data survey is a semi-structured survey with the key persons (such as the IT person/team, Project Management Unit) managing the data for the provider (e.g. a government department) to understand the different components of the data, the data collection process and its current use, storage, and access. More like a 'checklist' of things, the data survey is an iterative process and may take place over several discussions, and requires review once the project starts and gets data. If you already have a good sense of the utility of the data for your research purposes, then you can do the survey while waiting for access to save time when the data finally comes.

However, since acquiring access to admin data properly by following the legal processes and protocols is often tricky and time-consuming, it's good to assess whether the data is even useful for your research design up-front. The primary motivation for a data survey is to do a one-time frontloading of this work before you initiate a data access request so that you don't miss anything by just narrowly focusing on doubts that come up after getting access to the data.

Data surveys assist in understanding aspects of the data that are poorly documented or not publicly accessible, which could easily affect analysis and lead to misinformed findings. This is attributed to the very definition of admin data because it is not collected keeping research in mind – only operations. Hence, changes in the process associated with the collection and storage of this data are poorly documented. If some bug or process is fixed or changed going forward, then as long as it works, the historical data often doesn't matter for the sake of operations – while for analysis and research, it does, since it could affect the interpretation of insights.

Broadly, for all data collected or owned by the data provider, a data survey helps to understand the Data universe - who/what is in the data?; Data contents - what information exists about these entities?; Process of data collection; Storage/access; Data use etc.

2. Data Quality

During or after your data survey, a key metric to determine the inclusion of a certain piece of admin data within your policy-research plan is the quality of that data. Data quality has multiple dimensions. A detailed description and discussion of the various dimensions of data quality is provided below:

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, ranging from simple human errors of incorrect data entry or spellings to misaligned reporting incentives for beneficiaries or hospitals to enhance their outcomes from the scheme. For example, during periods of higher volume of responsibilities, frontline workers may tend to focus less and round up certain medical information, such as anthropometric measures or the treatment packages used. In such cases, while information is not entirely incorrect, these variations in the accuracy of reporting can lead to inconsistent results across time.

2.2. Completeness

A dataset is said to be complete if the dataset contains all the required data for the identification and delivery of the scheme.

The first issue within an incomplete dataset is missing information, when 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 registration, but routine updates may be missing (such as changes in phone numbers, claims details, etc.).

Another possibility of this occurring is when a particular scheme has parallel data entry systems where the primary data entry is still done on physical registers despite the existence of a digital infrastructure. For example, we have observed that routine immunisation data is collected physically through registers or forms in session sites, and is later digitised using an application. However, during this digitisation process, it is likely that a great degree of the data actually collected may not be entered into the digitisation system or be lost. This could be due to various reasons including capacity constraints, insufficient monitoring mechanisms, applications not being user-friendly etc. The data survey and recurring presentations to departments can uncover such scenarios.

A similar issue is that of missing data fields, wherein key variables required for decision-making may not be included in the forms/web application. For example, an insurance claims dataset on the usage of insurance coverage for institutions will always provide us with detailed information on the maternal care packages utilised in a particular hospital. However, it is unable to provide us the information on the postnatal condition of the children born through the use of the same packages as it is beyond the purview of the insurance scheme.

2.3. Non-representativeness

A representative dataset is one which contains information regarding all sections of the population without any bias. While working with administrative datasets, it is important to check for representativeness because biases can exist in the dataset due to policy/scheme structures. A stringent eligibility or policy criteria could result in the associated administrative dataset being non-representative of all potential and eligible beneficiaries of a scheme. For example, a strict eligibility cut-off based on income could deprive those just above the cut-off without access to health insurance, while they may need it just as much as those below the cut-off. Logistical challenges associated with enrolment such as accessibility to tribal areas may also leave the most vulnerable sections of the society out of the ambit of the scheme and the dataset. Therefore, in such cases, the denominator used to assess the coverage of the scheme will not elicit the right information.

2.4. Interoperability

Interoperability is the extent to which the data is capable of being integrated with other datasets. Most data is collected and stored in silos within specific implementing departments. The ability to link datasets, such as birth registry to immunisation data to health insurance claims 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., 2022). Similarly, there is potential to link datasets that can help target public health services better. For instance, comparing patterns in the incidence of diabetes and hypertension in any population screenings undertaken at the village level by primary health care (PHC) centres and patterns in insurance claims for cardiac illnesses could help target diagnostic services and early intervention through preventative care by PHCs to reduce complications that require hospitalisation in the future. It is imperative that there are mechanisms to allow the linkages of datasets such as the use of common IDs.

2.5. Suggested practices

For policy research professionals

1. **Navigating reliability and completeness of datasets**: While accessing administrative data, it is important to keep in mind the aspects of data quality issues summarised above. Resources should be set aside for checking the reliability and completeness of datasets at the start, and put in place processes and mechanisms to help governments institutionalise processes for improving data quality. Some of this information could be obtained anecdotally through the data survey. A more systematic approach would involve field visits and verification surveys (data audits). This information will help you understand the drawbacks of the dataset being used, and plan the scope of analysis accordingly. Issues related to reliability and completeness also have the potential to render the dataset unusable for research. This may have consequences for the budgeted expenditures set aside to access the respective information. If the secondary dataset is not usable, researchers will have to acquire this information through primary surveys which will be a lot more expensive. Running analysis on administrative datasets without checking for reliability and completeness runs the risk of eliciting spurious causal inferences or an incorrect understanding of the actual population of analysis or even biases based on data entry or other factors which will not be known to the researcher. Sharing insights from data cleaning and quality checks when using data for exploring a particular question with the appropriate levels within the government can be immensely helpful in bringing about systemic changes in databases or data generation and

monitoring practices. A policy research engagement with routine feedback loops for both insights from the data and also structured recommendations on data quality can be powerful in triggering changes prospectively (even if historical data may be unusable).

2. **Navigating interoperability issues of datasets**: The issue of interoperability is common in administrative datasets, especially if the data is managed by different departments. In such cases, often, the most granular level of merging data is possible only at the geographical level, which also has its challenges when different departments have different administrative boundaries. Therefore, if your data analysis plan involves merging datasets of different departments, it is useful to think of alternative approaches to obtain that data or caveat the possibility of such analysis in your respective engagements.

For professionals engaged in designing MIS systems

In our experience with administrative data, we have observed that data errors are often a function of a lack of simple validation checks. **Built-in validation checks** can improve data quality by designing these checks in the software/app at the data collection stage. For instance, certain critical fields should be made mandatory for data entry so they cannot be left blank. Some other validation checks 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 or phone numbers 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. 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) reduce the scope for spelling errors during entry. Time, date, and GPS 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 upfront, thereby reducing the effort required subsequently to clean and organise the data before use in analysis, making the data more usable. It is always useful to undertake short pilots with quick feedback loops to ensure data collection software and processes are robust.

For government departments managing administrative data

With an increase in digitisation and more sophisticated data entry applications, departments now have access to rich information, summarised through dashboards. These dashboards often only contain outcome indicators, however, indicators related to data quality are often missing. A focus on data quality along with outcome indicators tracked by departments can go a long way in maintaining data quality.

1. **Undertake high-frequency data checks**: Routinely reviewing the admin 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 2022) 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.
2. **Use independent data audits meaningfully:** The extent of reliability of an admin dataset can be gauged through independent data audits, wherein a sample of the admin data 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). Independent data audits may also be taken up by researchers prior to using the respective administrative data.

3. Data Access

Once you have been able to ascertain that the quality of the admin data you intend to integrate into your research design meets your requirements and standards using one or more of the parameters defined above, the next step would be to gain access to that data by navigating the associated data protection and privacy policies of the data-owning body or local jurisdiction. In some cases, these policies are easily found or known publicly. However, in most cases, in the absence of a well-defined over-arching centralised data strategy, this information has to be gathered by having multiple conversations with the respective officials responsible for the data.

Accordingly, an important point to consider at this stage is to determine the granularity and scale of the data you intend to access, because the more identified and sensitive the data, the more challenging it will be to access it. Strict definitions currently don't exist in India, just as in many other jurisdictions, on what constitutes identified data or Personally identifiable information (PII). But it is a spectrum between direct identifiers, indirect identifiers, and de-identified data. Direct identifiers are variables like name, ID number, address etc. Indirect identifiers are things like height, income, occupation etc which, with enough details, you could make a pretty good guess as to who the individual is. Similarly, even if the data are not identified, data providers may not be willing to release data because it is sensitive in a political or corporate sense – whether it be aggregate numbers on certain expenses or other information that can be considered a trade secret.

Given such practical and regulatory uncertainties, a guiding principle is to access as little personally identifiable information (PII) or sensitive data as you can. When developing a data flow strategy, researchers must balance the benefits of access to identified data – which gives them more discretion over the matching process and ensures the ability to match with additional data in the future – with the costs of access to identified data – which include more barriers to access and a more complicated and lengthy data request process.

We will discuss some of these processes below.

3.1. Legal Framework

Depending on the research strategy for administrative data, one may have to undertake some or all of the following processes before being able to access the data: -

1. **Memorandum of Understanding (MOU)** is an overarching institutional agreement between two or more organisations that broadly defines the nature and scope of engagements that the organisations will jointly undertake. It usually outlines the suites of services offered as well as roles and responsibilities for all parties involved. An MoU is not necessary for a data research project, but having an existing institutional partnership with the data owner or service provider may expedite the data acquisition process for the researcher and potentially safeguard the engagement or provide a means of handing over to other officials in the event of administrative shuffles.

2. **Scope of Work/Engagement (SoW/SoE)** is similar to an MoU but has a more defined agenda, objective, and timeline created for each project that is initiated between a researcher/ research organisation and the data-owning partner organisation. SoW/SoE are usually created and effectuated once all parties involved have agreed upon how and why the admin data will be used by the researcher to answer questions of joint interest and how the government plans to action insights from the data analysis and study findings. It also outlines the roles and responsibilities of all parties involved and the types of activities undertaken under the engagement.

3. **Data Use Agreement (DUA)** are agreements that govern access, sharing, and use of data between different entities. DUAs can be created for one-time or multiple engagements between government and external stakeholders (such as civil society, research institutions and universities, think tanks, service providers/vendors, private companies and other entities) for pre-determined and defined purposes where data needs to be exchanged to serve the purpose of the engagement.

   The DUA clauses will most likely be covered in the MoU/project agreement. If the research project doesn't have this and is accessing PII/larger datasets, then it is important to sign a DUA. Some partners may also state this as a requirement for data access.

   Annexure 2 contains an illustrative template for a possible DUA between government and external partners for collaborative social impact and research projects aimed at the use of data for public good. This template[23] pre-supposes an existing agreement between the government and an external partner, agreeing upon the scope of work, for which the data would be used.

4. **Data Request** is submitted to the data provider for accessing closed secondary data, specifying the data that is required by the project and for what purpose, how it needs to be processed (i.e. what should be dropped or encrypted), and options for enabling secure access. Ideally, through prior conversations, the project team should have a good understanding of the specific fields available and the structure of the data. Even if this is not the case, it is advised to make the request as specific as possible, only including the types of variables or information that the project is planning to use.

5. **Non-disclosure agreement (NDA)** is advised in instances when the MoU or DUA might not be sufficient to inspire trust and facilitate data access, especially with government partners. This is valid in cases where the data to be accessed is sensitive such as tax, health, financial, crime records etc.

6. **Acknowledgement Letter/Email of Data Receipt** is to be submitted to the data-owning partner organisation after the researcher (or its team) receives the data. The purpose is to record what data was accessed and protocols followed to ensure data security per the data sharing agreement. The letter should contain the details of the data sets, name and/or designation of the person who accessed the data, the person from whom data access was received/was witness to the process (if applicable), date of access, mode of data access, and protocols maintained to ensure secure access. An associated best practice would be for the data provider to either acknowledge or co-sign this letter.

The next step, after completing the legal requirements, would be to establish the protocols and processes for the safe and secure transfer of data from the data owning partner to you and your team. For more details, please refer to the administrative data toolkit.

3.2. Internal Compliance

Institutional Review Boards (IRB) review is required as the use of administrative data for research always entails the use of an individual's data, irrespective of whether it has identifiers linked or not, beyond what they may have originally shared it for. Moreover, in most cases, the data is collected and owned by governments, which are exempt and do not comply with the IRB mandates. Hence, this data collected, and the associated processes, have never actually been systematically reviewed for ethical research purposes. The data provider may also require IRB approval before signing the DUA. For example, in the IRB approval process, even if you do not come in contact with your study subjects, administrative data may still be considered human

[23] Different DUA templates would have to be developed for other types of engagements, which may be commercial or transactional in nature.

subjects research and may require informed consent. IRB regulations require that researchers consider whether seeking informed consent from individuals to access their records is appropriate. The data provider may also require you to get informed consent from each individual in the study. If it is impossible or impractical to obtain informed consent, you can apply for a waiver from IRB.

Additionally, upon receiving the data from the data owner, ensure that the data stored internally is not only encrypted but can only be accessed by a limited set of team members who are working with the data. Moreover, these team members should be trained and certified to conduct research with Human Subjects (HSC). Another point to note is that the data should not be shared further with any organisation, team, or individual who is not part of your team and not working on the scope of work agreed upon with the data owner. Any such actions should be considered only after duly informing and seeking approval from the data owner. The data owner may require the new party to undergo the same data access protocols that you have had to undergo.

4. Analysis and Preliminary Insights

Administrative datasets, by virtue of being born out of regular and routine monitoring efforts by the government, require additional efforts towards cleaning before being assessed for their fitment towards the RCT study. The following section collates insights from a few instances of the health insurance data use case, and talks through the different stages of cleaning and analysis that need to be done to clean the data, assess its fitment for the study, and leverage its potential for research.

4.1. Data Cleaning

On obtaining the data, we processed it minimally to get to a stage where we could assess its fitment towards a study. This cleaning includes harmonising variable formats, dealing with missing values, duplicates, and other logical checks. While the cleaning may be elementary, each step requires a discretionary step on how to deal with uncleaned/incorrect data. Keeping in mind the study questions, the following checks and cleaning decisions were taken:

a. **Variable Formats:** Variables are either categorical, numeric/dates, or string. Categorical variables include a list of options such as occupation, caste category, or the sex of the patient. Numeric variables are numbers, and strings are texts. A well-cleaned administrative dataset should have harmony among the variable formats across datasets which allows the analyst to check the quality of the data. For example, descriptive information about beneficiaries (such as the occupation of the patients) may be stored as a string variable in the historical datasets, and then as categorical in subsequent datasets. One solution may be to focus the analysis from a specific year onwards and use a suitable variable.

Another common issue is often observed with phone numbers (which are ideally numeric) being incorrectly stored as strings due to data entry issues (for instance, the country code "(+91)" may be attached to some but not all). Such uncleaned entries would need to be manually fixed during the cleaning processes instead of being dropped from the dataset.

b. **Missing Values:** Identifying the extent of missing values would already have been covered during the data access process. However, dealing with missing values is a crucial part of data cleaning. Broadly, missing values are dealt in four ways– a) Dropped from the dataset, b) Replaced with "0", c) Replaced with "NA" / empty cells (depending on the coding software, and d) imputed with expected or average values. The decision of how to deal with these missing values depends entirely on the study and its research questions. For instance, if a variable that denotes the cost of treatment at a hospital has missing values, replacing the missing values with zeroes will massively reduce the average treatment cost; whereas replacing it with empty cells/NAs can prevent doing so. Alternatively, replacing missing values with zeroes in the phone number variable can work.

c. **Duplicates:** Most administrative datasets will have duplicate values – either owing to data entry errors, or (more commonly) because they are pulled in from multiple monitoring systems which may have overlapping information. In general, there are two types of duplicates which need to be dealt with during cleaning.

   i. **Absolute Duplicates:** These are mostly data assimilation/ entry errors which result in absolute duplicates, wherein all values are the same in a few rows. Absolute duplicates can be identified easily from existing commands in R/STATA/Excel and are usually de-duplicated by retaining one observation.

   ii. **Duplicates by Variable:** Even in datasets which do not have absolute duplicates, there can be duplicates at the variable level. For instance, a dataset which logs all claims (and is uniquely identified by the claim ID) has multiple duplicates by the patient variable. This is not necessarily a flag since the scheme allows the same patient to seek treatment multiple times. However, for any patient-level analysis, these duplicates will pose an issue. How to deal with duplicates hinges on the study objectives. There are three common ways: a) to retain one observation (either the latest or the earliest); b) to create one observation by summing up the values in the duplicates; c) to create one observation by averaging the values in the duplicates. In analysing health insurance data, patient-level data is required to understand how much (cost-wise) people utilise the scheme. Hence, it is deemed fit to create one observation per patient by summing up all duplicate values of their subsequent visits.

D. **Logical Checks and Outliers:** The next step is to check if the information in all variables is logically sound. For instance, the discharge date of a patient should not be before their admission date, payment made to the hospital should not be before the claim is filed, or phone numbers cannot be more than 10 digits. Conducting these logical checks on the data can allow assessing its usability at the variable level. Cleaning illogical entries usually involves interacting with the government department to understand why these issues exist. Typically, illogical entries due to data entry/storage errors have to be dropped from the dataset to be used since it is not feasible to rectify it through validation. An extension of looking at logical checks is to assess the soundness of numeric variables. Looking at outliers of numeric variables is one way to assess soundness. For instance, the claim value (cost) variable can have outliers which can be attributed to high-cost treatments being availed. Outliers can have a big impact on any statistical analyses and skew the results – so it is imperative to identify them at the cleaning stage. Depending on the research questions, outliers are either dropped from the study or winsorized at the 90th percentile.

E. Harmony Amongst Datasets: Regular monitoring of large-scale schemes like health insurance also renders multiple datasets for the scheme. Health insurance schemes, for instance, have datasets on all patients treated, all claims filed, all hospitals empanelled, all treatments availed, and all pre-authorizations approved. All of these datasets speak to each other and are generated at different times in the patient treatment cycle. However, depending on the data storing/management systems they may or may not merge entirely with each other. For instance, datasets for all claims approved and all claims rejected may not be mutually exclusive due to errors or inconsistencies. Checks for harmony and consistency across datasets are usually done by merging them on their unique identifiers and checking how much of the datasets merge.

4.2. Assessment for Fitment

Once all datasets are minimally cleaned and processed, the next step is to assess their fitment for the study or analysis purpose. There are three broad ways in which this can be done:

a. Assessing the Quality of Key Variables: Every policy research question will have key variables that the study would want to focus on the most. In the previous section, we identified how clean/logical all variables are. To assess their fitment for the study, it is important to see if the key variables are clean and sound. Research questions where the key variables are mostly missing/have to be imputed may not be successful since a majority of the data is being estimated ex-post. Assessing fitment regarding key variables is an iterative process.

b. Assessing Frequency of Access: Service delivery of health insurance schemes happens very regularly, due to which all administrative data is high-frequency. Data is collected almost every day, ensuring that any changes in service delivery/consumption patterns are adequate and immediately captured. However, the frequency of access to data can be an issue. Depending on the questions for exploration, claims data may need to be analysed every week to track changes in hospital behaviour and also make any immediate tweaks to the package of fraud checks. This would require data sharing/access modalities also to match the use purpose. Assessing the frequency of access and working with the department on setting up regular transfers helps in making the data fit for study use.

c. **Assessing Sufficiency of Data:** Depending on the study question, it is important to also assess whether all the variables in the administrative data suffice the kind of answers that the research wants to seek. For instance, despite the breadth of claims and patient data that the administrative datasets log, aspects of patient quality (such as out-of-pocket expenses and hospital experience) are usually not captured in the datasets (or may not be efficient to do so). There are also other socio-demographic details of patients and hospital locations (such as rural, urban, caste composition at the district level, and distance to the hospital) that the administrative datasets can sometimes lack (or be stored elsewhere). Assessing this (in)sufficiency of data allows the study/analysis to seek supplemental data collection or data use activities that can provide these data points. The most common, and cost-effective route to seek supplementary data is to source them through existing third-party data sources. Datasets like Census, National Family Health Survey (NFHS), and SHRUG provide granular data on socio-economic aspects. These can be merged into the administrative datasets to get information on geolocations (to calculate distance to hospitals); and rural/urban composition.

   Studies could also consider conducting primary data collection efforts to supplement the information being received from administrative datasets. The next section speaks about the utility of primary data collection efforts in admin data-use projects.

4.3. Complementary Primary Data Collection Efforts

Primary data collection refers to targeted efforts by the study to collect the data required directly from the study population. Such efforts serve two primary use cases. The first is to supplement the administrative datasets (which was discussed above). For instance, on assessing that the administrative datasets do not provide variables on out-of-pocket expenses and patient experience, primary data collection of these variables through a phone-based patient experience survey could be undertaken. The data points from these surveys can then be merged with the administrative data of the patients and then analysed for insights.

Another use of such primary data collection efforts is to validate the information being provided in the administrative data. Existing administrative datasets can be supplemented with primary data collection activities to check the accuracy of the admin data. For instance, Child Data Verification Surveys or Zero Dose Vaccination Surveys wherein immunisation information of the children is collected directly from the child's immunisation card through field/household surveys can then be compared to their records in the administrative data to check for accuracy. Similarly, for health insurance, patient experience surveys seek information on the treatment received by the patients so that the information can be validated with the claims filed by the hospital. Conducting primary data collection can thus increase the robustness of the study and provide additional data points to rely on.

4.4. Iterative Interpretation and Learnings

Leveraging administrative datasets for research studies requires an iterative process of cleaning and working with the datasets. It benefits from setting up a feedback loop with the stakeholders wherein data insights from cleaning, fitment checks, and preliminary insights are relayed regularly and can help either validate and/or supplement the findings. In the health insurance data use project, preliminary insights post cleaning and fitment checks revealed were routinely shared with government partners.

In one instance, it was found that the amount paid to the hospitals (as found in the payments data) was not consistent with the amount approved during the claims review process. Setting up feedback loops with the stakeholders helped them learn that approved amounts may be deducted by higher authorities in the claims review process based on the claims filing behaviour of the hospital (not enough evidence/documentation; lack of tests). The research team then adjusted the frequency of data access set-ups to ensure that the most recent data was received and used the payment data to understand how much money was claimed by hospitals instead of relying on the claim-approved amounts. This iterative process thus helps get a better understanding of how data is collected and updated; as well as ensures that all insights accurately reflect the nuances of the scheme.

4.5. Capacity Building

After cleaning and assessing the fitment of the data of the study, it is helpful to relay the insights on data completeness, usability and quality to government stakeholders. This can help facilitate better data-use practices. Under PMJAY, the data checks and the data-driven triggers are being drawn from a variety of resources including automated checks developed by the National Health Authority, India, for fraud control. Research teams can strengthen this process by contextualising local needs while drawing on insights from checks used by insurance programmes in other countries. To be sustainable, a package of checks thus developed would have to be absorbed into the existing insurance MIS' existing dashboard and automated so that it can be run on administrative data by the department regularly to identify hospitals to target for increased oversight.

Annexure 2: A DUA Template

Template for building a 'Data Use Agreement' with an external agency for a collaborative research/public good engagement

Collaborations with academics, research organisations, civil society, and not-for-profit institutions help governments supplement their capacity for data analysis, interpretations, use cases, and use for decision-making. However, these collaborations often require the government to share administrative and personal data of citizens with the partner organisations, albeit for specific purposes under the collaboration (for non-commercial use). Similarly, governments may also often benefit by accessing and using data and information from non-governmental or private sources to validate and complement the administrative data collected to understand complex challenges and design or course-correct the delivery of their schemes and programmes. In this instance, other organizations would be required to share administrative and granular or personal data of clients and beneficiaries with the government to improve public welfare. The benefits of these collaborations and the perceived risks to the privacy and security of data can be balanced and controlled through formalized arrangements and a system of checks and balances. Data use agreements (DUAs) signed between the government agency and the partner organisation are a useful tool to guide such interactions between the data owner or intermediary and the data user such that the collaboration realizes the intended objectives while minimizing risks of loss of privacy and security. A few key points to bear in mind while deciding upon such agreements have been shared below followed by an outline of a generic template that can be modified for use by the government agency.

1. The purpose of the 'Data Use Agreement' is not to limit or discourage data use partnerships with external agencies. Such agreements should serve as evidence that the government is keen to collaborate with defined guidelines that have been put into place to ensure citizen's privacy and security. Excessively limiting agreements hinders productive collaborations, thereby hampering avenues for data-based decision-making and efficient governance.

2. The 'Data Use Agreement' is not a fixed or permanent guideline and may need to be modified based on the nature and purpose of each partnership.

3. The Data Use Agreement should be included in the MOU being signed between the government agency and the partner organisation to provide sufficient context about the scope and purpose of the Data Use Agreement.

4. Each partnership/project and in some cases, every instance of data sharing needs a specific and separate data use agreement between all the parties involved. Efforts to directly apply (without adapting) a previously signed agreement to new assignments and/or collaborations can lead to ambiguity and faults in liabilities and accountabilities.

5. Each agreement should clearly define the data points being shared and its use purpose/s. If a data point or purpose is decided upon after signing the agreement, the agreement needs to be amended or a new agreement signed.

6. The definition and terminology used in the Data Use Agreement should comply with that used in the relevant Ministry's/ State's Data Policy or Data Strategy to minimise future ambiguity in interpretations.

Template for a Data Use Agreement

[This is a suggestive template and should be modified as needed to suit the Government's and the partnership's requirements for a successful and productive collaboration]

Partners to the Agreement:

[Please list all partners (parties) of the agreement here. This should include all direct and indirect stakeholders (institutions and individuals within these institutions) who may gain access to the dataset directly as the recipient or indirectly through the direct recipient for the purposes of the collaboration/ MOU. Broadly, partners to an agreement could be classified as Data owners, data intermediaries (stewards/managers), data users/analysts and any others]

[For the purpose of this template, the government agency signing this agreement has been referred to as GOVT and the Partner organisation as PARTNER]

Purpose and scope:

[Please describe the objectives of the collaboration between the Partners and the scope and means of the engagement: research studies, information portal development, others. The broad methodology and outputs should also be defined]

Definitions

This section should include clear and standardized definitions of all the key terms that are typically included in data use agreements. Most of these are defined in existing legislation on data sharing and privacy. The most recent and standardized definitions may be adopted for the following list of terms.

1. Administrative Data:
   [If possible, the broad list of data points and data sets to be shared under this group should be included under the DUA's Annexure]
2. Data
   [Administrative data is typically a subset of data which includes primary data and other forms of secondary data]
3. Data Owner and Data Intermediary
   [If applicable define data intermediary who will ultimately enable data access - for instance, the technology/MIS vendor, IT department etc., who may be distinct from the data owner]

4. Data Protection Law
   [List all existing national and local legislation applicable to both parties]
5. Data User
6. Personal Data
   [containing Personally Identifiable Information]
7. Sensitive Personal Data or Information

Policies

1. **Sharing and Transfer of Data**
   This section should define:
   - *Specific data to be shared and the level of granularity*
   - *the use purpose for which data will be shared between parties*
   - *types and number of users of the data in the partner agency*
   - *specific processes to be followed by both parties to initiate and execute the process of data sharing (or transfer)*
     - *This should outline institutional processes such as the use of NDAs, data request forms or other documentation, registrations etc. and*
     - *technical processes to be followed such as use of APIs, mirror/remote servers, hard drives etc.*
   - *the time period for which data will be shared*
   - *the privacy safeguards and data security protocols that will be followed by both parties during the process of sharing of data and*
   - *any other relevant aspects during data sharing*

The content for this section would typically draw on principles of "safe projects" and "safe settings" from the Five Safes Framework.

2. **Storage and Security of Administrative Data**
   This section should define:
   - *The format in which data is to be shared, and if containing personal information, whether it would be de-identified or anonymised, and if not, explaining why personal information is required and for what purpose it would be used. Also, fields that cannot be shared (negative list) as per existing legislation and policies*
   - *Safeguards to be followed in storage and analysis of data, including encryption protocols, especially if containing personally identifiable information fields.*
   - *Duration for which the data will be stored by the party that receives access to the data*
   - *Procedures and clauses to be followed in case of any inadvertent disclosures of data by either party*

The content for this section would typically draw on principles of "safe data" and "safe settings" from the Five Safes Framework.

3. **Access and use of Administrative Data**

This section guides post sharing of data regarding:

- *who within the receiving organization is a bonafide user of the data and*
- *the purposes for which it is used (in this instance, for non-commercial or research use only)*
- *protocols for further use of the data for other purposes beyond the originally stated use purpose*

The content for this section would typically draw on principles of "safe people" and "safe projects" from the Five Safes Framework.

4. **Publication**

This section should outline the guidelines for publishing outputs from the analysis of the data and underlying data itself:

- *It should find a balance between academic freedom and safeguards of data. It should define protocols for due review process of the data used to ensure that there is no potential for identifying personal information*
- *It should define a time period within which outputs can be made public and the timeline for the due process to follow for the same.*
- *It should define protocols and safeguards in case of publication of data outputs (Aggregate statistics, tables, visualization etc.)*

The content for this section would typically draw on principles of "safe outputs" and "safe data" from the Five Safes Framework.

5. **New Data Generated/Collected:**

In any instance of use of existing data for a specified purpose, often, additional or new data is collected through primary surveys and interviews or generated using technological devices.

- *The Data Use Agreement should also contain clauses and protocols on the purposes for which new data is required to be collected/ generated, how it is to be accessed, stored and shared, and for what specific purposes for which it will be used.*
- *It should define the rules of engagement between the GOVT and PARTNER on the collection, storage, processing, sharing, ownership and the use of such new data by clearly delineating the responsibilities of all parties engaged in the process.*
- *It should contain all the necessary safeguards for privacy of data, informed consent for sharing and use of data and measures for ensuring data security during storage, analysis and use of data.*

6. **Accessing external/private sources of data**

The GOVT or PARTNER may also access external sources of data from private sources (such as satellite data, cell phone location data, external/independent surveys etc.). If it is anticipated that external data *sources will also be accessed and used with the administrative and primary data already being exchanged between the GOVT and PARTNER, the DUA should also contain clauses on the access, storage, sharing and use of such data, and all the safeguards for data privacy and measures for data security which shall be followed by all parties.*