Connecting Families to Benefits Using Linked Data

Lessons from California's efforts to increase the take-up of anti-poverty credits

APARNA RAMESH, EVAN WHITE, CHARLES DAVIS, SAMANTHA FU, AND JESSE ROTHSTEIN

CALIFORNIA POLICY LAB

MARCH 2022

Introduction

Policymakers rely heavily on the tax system to distribute direct payments to low-income families. Anti-poverty tax credits such as the Earned Income Tax Credit, the advanced Child Tax Credits, and the federal stimulus payments combined to keep millions of Americans out of poverty during the pandemic. Such credits have strong potential to continue to reduce poverty.

These credits only work, of course, if eligible families receive them. To do so, they must file a tax return. But many low-income families who are at or below the federal poverty level are not legally required to file taxes. Policymakers need a better understanding of how many low-income families don't file taxes (and therefore miss out on these valuable credits) in order to address this problem.

While state and federal tax agencies know who files taxes, they have very little information on the families who do not file, especially those below the poverty level with little or no earnings. State and local human-service agencies, however, serve many families below the poverty level, placing them in a unique position to assist eligible families to receive these credits.

To help the State of California understand who may be at risk of not receiving anti-poverty tax credits, the California Policy Lab (CPL) facilitated a linkage of two individual-level datasets held by state agencies: one with safety-net enrollment data and one with state tax filing data. CPL served as a trusted third party by implementing a "hashed linkage" — linking data that was de-identified through "hashing" (a one-way encryption process) by each agency.

By linking this data, we were able to help California measure how many Californians receiving safety-net benefits were at risk of not receiving federal stimulus payments, the state Earned Income Tax Credit, and the advanced Child Tax Credit. We also helped the state learn that the majority of its safety-net beneficiaries were already receiving benefits through tax filing — allowing the state to focus limited resources on those who were not receiving these benefits. This linked data also equipped the California Department of Social Services (CDSS) to conduct targeted outreach to Californians who had not filed state returns (and therefore were missing out on thousands of dollars in credits) in recent years and to direct them towards intensive resources that could help them file a return and claim these credits. This linkage led to millions of dollars in tax credits delivered to Californians who otherwise may not have received them.

The benefits of linking administrative data go beyond the take-up of anti-poverty tax credits. Administrative data can help answer many vexing policy questions faced by policymakers. However, much of the value in administrative data can be obtained only when data can be linked across multiple sources. By linking across systems at the individual level, administrative data, which is often topically narrow, can replicate the cross-domain scope of survey data.

This toolkit is intended to help staff in state governments outside of California who are interested in using administrative data and linking it across agencies to measure the take-up of safety-net benefits. We are also releasing a technical how-to toolkit for those interested in operationalizing a hashed linkage.

Determining the linkage strategy

Step 1: Define your goal.

Why do you want to link data? The policy goals of the linkage will define the strategy departments should pursue, the way the data should be linked, and the permissions around data use. Your goals may include some or all of the following:

Goal 1: Identifying the magnitude of the problem. Linked data provides insight into how many individuals are at risk of not taking up benefits. For example, through linking safety-net and tax-filing data in California, it became clear that the majority of safety-net enrollees in California actually do file a tax return annually — and that any efforts to increase tax filing should focus on the smaller population of non-filers. Linked data gave insight into who is missing out on benefits, and the potential barriers they face. Understanding patterns of take-up by demographic factors can help inform the type of solution departments may pursue. For example, analysis by CPL showed that single adults with no observable earnings were at highest-risk for not receiving credits, which helped CDSS inform its messaging and outreach. As a result, CDSS underscored that during the pandemic, individuals need not have worked to receive stimulus payments and should file their taxes to receive these payments, regardless of their income.

Goal 2: Targeting solutions and outreach. Using a targeted approach to direct more intensive resources to those who need help enrolling for a program can improve the success and return on investment of efforts to increase take-up. Instead of trying to reach out to the whole safety-net caseload, linking data can provide a list of individuals and households to target and tailor interventions.

Most interventions aimed at closing a take-up or enrollment gap will require some type of outreach. But using linked data beyond the purposes of research — for example, to link hashed data back to identified data and conduct actual outreach — will require different data-use permissions than simply using the data to analyze the magnitude of the problem.

If outreach is a core priority, below are three guiding questions to define your linkage strategy:

1. What department or entity is best positioned to conduct outreach?
2. What data is required for this department to conduct outreach?
3. What permissions need to be put in place to allow this department to use the linked data for this purpose?

Goal 3: Evaluating success. Linked data can be used to evaluate the effectiveness of specific efforts. It can also help gauge overall progress towards closing any take-up gaps. If you are relying on a third-party researcher or outside entity to help with these questions, you will need to ensure that data can be used for the purposes of research and evaluation.

Step 2: Identify barriers to linkage.

Can the relevant government agencies exchange data to meet the above goals? Unfortunately, the answer is often "no." But consulting with the legal teams at each agency may yield opportunities that may not be obvious at first. For example, some statutes allow agencies to share data for the benefit of enrollees, or to improve a certain program. Such statutes can sometimes be used to facilitate data linkages.

As a means of protecting sensitive information, data disclosure and confidentiality laws often restrict the sharing of government records. These laws almost always prohibit the disclosure of personally identifiable information, and often prohibit the disclosure of de-identified records as well. But these laws usually have several exceptions, which mostly fall into three categories:

1. **Certain types of data:** These laws allow certain categories of data to be disclosed, for example if the data is de-identified (stripped of all personally identifiable information) or aggregated so as to reveal information about groups, not individuals.
2. **Certain purposes:** These laws allow data to be disclosed if the purpose to which the data will be used fits within a defined category. For example, data may often be disclosed for purposes of verifying programmatic eligibility (e.g., verifying income). Other common purpose exceptions are to improve the program for which the data was collected, or for purposes of conducting research or evaluation.
3. **Certain users:** These laws allow data to be disclosed to certain types of end users, such as contractors, other government agencies, researchers, or specified departments.

Sometimes these types of exceptions are combined: for example, a law might allow de-identified data to be shared for program improvement only with other government agencies.

Understanding the landscape of data-disclosure laws is essential to determining how data can be linked between two departments. For purposes of linking data, it is necessary to have data that can be used to uniquely identify different records. So if the statute prevents the disclosure of personally identifiable data,

which is not uncommon, one solution is to use techniques such as hashing to de-identify the data but keep it in a format that can be linked with other hashed data. Data-disclosure laws will also dictate whether linking data directly with other departments will be easier or harder than using a third party, such as a research organization.

Step 3: Decide on the best pathway to linkage

Though this report discusses the value of hashed linkages, an *identified* linkage is usually advisable if it is possible to do using existing authorities. In some cases it may even be easier to enact new legislative authority allowing an identified linkage rather than pursue a hashed linkage — for example if government agencies wish to conduct the linkage on an ongoing basis. Hashed linkages should be used only if there are substantial obstacles to conducting identified linkages. This is because hashed linkages can be more time-consuming and less accurate than identified linkages.

Pathway 1: Interdepartmental linkage. To the extent possible, government agencies are best served by pursuing an internal solution to linking identified data using existing statutory authority. Interdepartmental data-sharing agreements are becoming more common, especially when executive leadership in both departments share the policy goal that the data linkage would make possible.

Pathway 2: Statutory change. When statutory authority does not exist, or if obstacles to getting inter-agency approval prove substantial, identified linkages can be facilitated via statutory change. For example, states can often add a new exception to a disclosure prohibition that allows the relevant departments to exchange data for the purposes of identifying take-up gaps, conducting outreach to eligible non-recipients, and evaluating success.

In California the Legislature passed a law allowing the state's tax department (the Franchise Tax Board) to analyze the magnitude of the take-up problem related to the state Earned Income Tax Credit using data from the Department of Health Care Services:

> *Notwithstanding any other law, the State Department of Health Care Services shall exchange data with the Franchise Tax Board upon request, including, but not limited to, sufficient identifying information to allow the State Department of Health Care Services and the Franchise Tax Board to assess the extent to which the State Department of Health Care Services and the Franchise Tax Board can identify individuals enrolled in Medi-Cal who may be eligible for the California Earned Income Tax Credit and the federal Earned Income Tax Credit. The data provided pursuant to this section shall remain confidential and shall be used only for the following purposes:*

a. *To analyze and develop a plan to increase the number of eligible claims of the California Earned Income Tax Credit and the federal Earned Income Tax Credit.*

b. *To reduce any barriers to tax filing for nonfilers of tax returns who may be eligible for the California Earned Income Tax Credit and the federal Earned Income Tax Credit.*

c. *To develop an outline of the changes needed to increase collaboration and coordination among state agencies to inform the greatest number of individuals eligible for the California Earned Income Tax Credit or the federal Earned Income Tax Credit of their eligibility.*[1]

However, such statutory change may not always be possible. There may be a lack of political will or a concern about data privacy or security. If, for whatever reason, an identified linkage isn't possible,a hashed linkage can help.

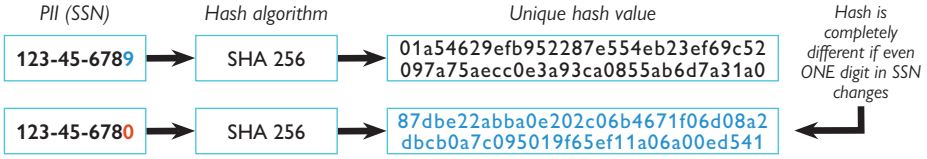
Pathway 3: A hashed linkage. If there's no current pathway to exchanging identified data, one potential option is to employ hashing to de-identify records in a way that maintains their linkability.

Hashing is a method of turning readable data into unique (allowing for analysts conducting the linkage to distinguish between different values) but unreadable (so that analysts will never see underlying sensitive information) data. This strategy can be applied to identifiers such as name, Social Security number, or date of birth to encrypt the values in such a way that they can still be used to link across datasets. When agencies or departments hash data, they apply a one-way encryption that turns identifiers (also known as personally identifiable information, or PII) into indecipherable strings of letters and numbers. Hashing has two main advantages:

1. **Hashing protects privacy** — Hashing turns PII into a string of characters that is no longer identifiable. Changing even one character in the input data will produce a completely different hash string (see box). Because hashing scrambles identifiers into incomprehensible information, it is impossible to recover the PII from the hashed data. By obscuring the PII in this manner, this method can be considered to be legally exempt from laws that prohibit disclosure of PII.

1 Cal. Rev. and Tax. Code § 19551.3.

FIGURE 1: Hashing personally identifiable information



2. **Hashed data can be meaningfully linked** — Hashing the exact same input will always produce the same output. Hashing an individual's PII in two separate datasets will create the same unique hash string. This allows the records from the same person to be linked across different datasets without ever seeing their personal information (assuming identifiers such as an individual's name appear with the same characters in each dataset).

To ensure that a hash cannot be reengineered, the parties that are contributing data ("the data contributors") agree upon and use a "salt". A salt is a string of characters decided upon by the agencies that own the data (for example: "iloveapplepie2!"), that is appended to each piece of original PII. Typically, only the analysts working on hashing the identified data come together to agree upon the salt, and are the only ones who should know the salt. It is mathematically infeasible to determine the initial input that created the hashes unless the salt is shared with the parties who also have access to the hashed data.[2]

Step 4: Decide whether you'll need a third party to perform the linkage.

Depending on the nature of disclosure laws and available resources, a third party may be required to facilitate the hashed linkage.

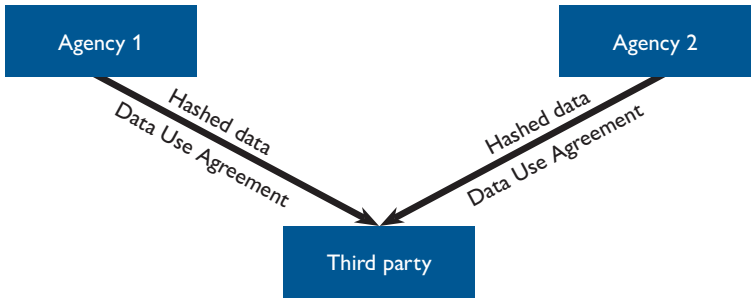
Two-party linkage: Departments may feel comfortable with one of the participating departments linking the hashed data. However, one risk to this approach is that if one department does the linkage, it could re-identify values in the other department's data that match to its own data. To do the linkage in a way that maintains the de-identified nature of the linkage, the department charged with hashing and linking the data should separate the team that hashes the identified data from the team that links the data ("the linkage team"). Said another way, to ensure that the data is not re-identified, the team that links the hashed data should not have access to unencrypted PII or to the salt used to create the hashes.

Though our discussion assumes that only two departments wish to exchange data, sometimes linkages involve more parties. For example, in California CPL facilitated a four-way hashed linkage between higher-education, student-aid, and safety-net agencies. All the above considerations still applied.

2 This prevents so-called brute-force attacks based on known common names. For example, testing many different encryptions of a common name like "Smith" to find the one that produces a string that is known to be common in the encrypted data.

Third-party linkage: A two-party hashed linkage may not be a viable option in many cases. For example, some parties may not be comfortable exchanging data directly due to political or privacy sensitivities. Or some parties may not have the capacity to perform linkages themselves.

FIGURE 2: Third-party linkages



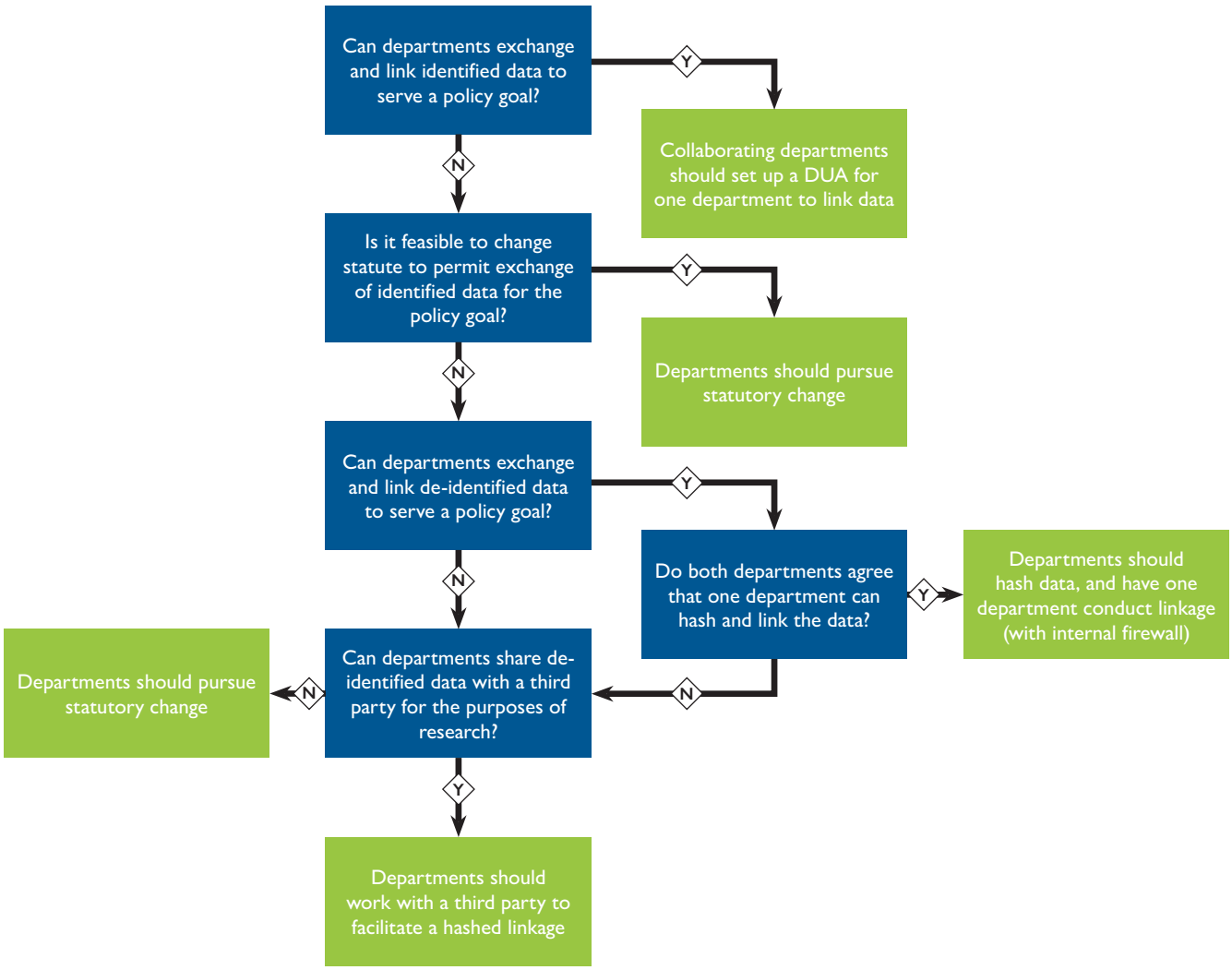
If parties do not want to exchange hashed data directly, a trusted third party could facilitate the linkage (we will also refer to this as "the linkage team"). Using a third party may also be the only legally permissible option: in some cases, disclosure laws prevent the exchange of data with another agency, but do make an exception for the exchange of data with an outside third party for the purposes of research.

Typically, trusted third parties:

- Have a data use agreement in place (or can establish one) with both parties to receive de-identified data
- Have data-storage and data-access infrastructure that meet the privacy requirements of each department
- Have the legal, technical, and research capacity to facilitate the linkage

Third parties tend to be research institutions (either housed at a university or existing as a nonprofit), but could also take the form of government entities or non-governmental partners with strong data capacities. States would be well-served by adopting this function internally, for example through an office of data linkage, which could also play a role in provisioning data to approved parties.

FIGURE 3: What's the best pathway to linkage?



Implementing a hashed linkage

If departments decide to conduct a hashed linkage, implementation requires both legal and technical steps. Below are the basic steps of implementing a hashed linkage. The process can easily take 12 months, start to finish, and is greatly dependent on the speed with which the data-use agreements are executed.

TABLE 1: How long does it take to hash and link data?

STEP	TIME ESTIMATE	PERSONNEL
Draft and sign data-use agreements	6 months	Legal departments at data contributors
Extract data	1–2 months	Data analysts at data contributors
Learn data and refine code	1–2 months	Data analysts from linkage team
Hashing (and troubleshooting)	1–2 weeks	Data analysts from both data contributors and linkage team
Linkage process	1–3 months	Data analysts from linkage team
Total	**9–12 months**	

Step 5: Budget for your team and resources.

To coordinate a hashed linkage, you'll need the following individuals and expertise on your team:

Legal expertise (at least one staff member from each participating entity). All parties need an individual who can help draft language for the legal agreements necessary and shepherd those agreements through an approval process.

- Level of effort required: Work times for drafting the appropriate legal documents are typically low (around 5–20 hours of effort total). However, wait times to get approvals and sign offs on those documents can be quite high — and can lengthen the process significantly.

Data analysts (at least one from each data contributor). Data contributors will need analysts that can perform two functions:

- Extract data from requisite databases (typically someone familiar with the datasource).
    - Level of effort required: Usually spread out over a month of time to understand and help refine the data request, train data analysts from the linkage team on the data, and troubleshoot any issues with the extracted

data (around 15–30 hours of effort for each data extraction, though the effort may vary depending on the complexity of the project, especially if there are multiple data pulls or if the linkage will be refreshed on an ongoing basis).

- Run the hashing code, using a mutually agreed upon programming language.
    - Level of effort required: Usually spread out over a month of time to help implement code, adapt it to the specific dataset, and troubleshoot any issues. This analyst is the one that guides the data contributors in their choice of a salt (around 20–30 hours of effort total).

Linkage team (either a third party or embedded within one of the departments). This team requires staff with the ability to:

- Generate and implement the code needed to both hash and merge the data, and provide technical support to data contributors performing the hash.
- Design the rules of the linkage, with an eye towards how different decisions affect the research or policy goal(s) of the linkage.
- Level of effort required: an estimated 3 months of Full Time Equivalent (FTE) for two analysts to design the rules related to the linkage, provide technical support, and implement the merge. Likely 30–40 hours of effort from an expert who can design the rules of the linkage.

The linked dataset is just a starting point. The linkage team will then need to clean and prepare the linked data and then conduct the analysis. Levels of effort vary widely by project.

Step 6: Set up legal framework and agreements

The legal framework to linking data must be grounded in the goals of the linkage and existing legal barriers to linkage. In our experience, support from executive leadership or program staff is a key ingredient in helping departments find creative solutions to linking data.

Hashing facilitates the exchange of data by de-identifying PII. Many disclosure laws focus on PII, and bar the exchange of PII. However, as long as the correct precautions are taken, hashed PII is not considered PII because it is encrypted in a way that is irreversible by any party that lacks the salt (and even if they had the salt, they would only be able to find individuals for whom they could replicate the input data).

The disclosure of any individual-level data (identified or de-identified) may still be restricted by disclosure laws at a state or federal level. However, third parties, especially those that are research institutions, can sometimes serve as a work-around to exchanging such data. Agencies frequently have statutory exceptions for the disclosure of de-identified data for the purposes of research. This allows a third-party research entity to facilitate a linkage if there are constraints on exchanging data between government agencies.

If data contributors are working with a non-governmental third party, the third party should have a data use agreement (DUA) in place with each participating department. The two departments need not have an agreement with each other. Instead, within each DUA with the third party, departments can agree to allow the third party to link their data with other entities.

Should statute allow for the use of data to be used for outreach, third parties could also help agencies who want to use the linkage to then conduct outreach to individuals or families who may be missing out on benefits. The third party would not have PII, but would instead share a list of IDs from that agency's dataset, which the agency could then link to actual names and contact information.

Step 7: Ensure data security

Whether the hashed linkage occurs at one of the data contributors or at the third-party research entity, data security is paramount, even when all PII is hashed. The linked data could contain sensitive information such as income or disability status. The best approach is to abide by the "Five Safes."

FIGURE 4: The Five Safes

The **project** itself is safe if there is legal authority for linking data and it is ethical to link the data for the specific purpose. In the case of connecting eligible households to valuable benefits, there is a strong rationale.

Those touching the hashed PII should be limited to a small group of safe **people**, which means analysts trained appropriately who know the risks of re-identification. These people should never have access to the hashing salt or the input PII and should make plain to all other parties what procedures they will be using.

The **data** themselves are made safe through the hashing process, and the **setting** they are in is important. The hashed data should be separated from all other data and deleted or walled off from access after the linking procedure is complete. The computing environment in which the linkage occurs should meet appropriate data-security standards, such as NIST 800-53. CPL maintains an environment that blocks all outbound access to outside networks, such as the internet, so as to prevent hashed PII from ever leaving the secure environment. Data are transferred into the environment using a secure file transfer protocol (SFTP), and users are required to use two-factor authentication and connect over encrypted virtual private network (VPN) tunnels. These steps may not be necessary in every case but are good practices.

Depending on the datasets being linked, the compute environment may also need substantial computing resources. Hashes themselves can be long (64 characters) and can take up considerable disk space if there are tens of millions of observations, as there sometimes are.

One can achieve a safe **output** by sharing only the linkage pairs after the merge is complete, and locking away or deleting the hashes.

Important considerations for long-term success: The legal start-up costs for a hashed linkage can be steep and the technical expertise to hash and link data the first time can take time to establish. However, once such infrastructure is set up, routinely receiving data from participating entities and updating the linkage becomes easier. Parties can choose to update the linkage on an annual basis or a more regular basis. As the process for hashing and linking is created, it is important to consider how often the data will be pulled and to ensure that data is consistent across years. Over repeated iterations, the turnaround time for hashing and linking can reduce from months to weeks.

Hashing and linking data through a third party can be a good short-term alternative when parties cannot exchange data directly. A couple rounds of linking, and the use of those linkages to inform policy decisions or inform successful take-up campaigns, can serve as a useful proof of concept that catalyzes codifying the exchange of data directly.

Conclusion

Linking government data is a powerful strategy to connect people to valuable benefits that they might otherwise miss out on. While there are start-up costs, the dividends from this strategy, of better supporting families, injecting federal dollars into state and local economies, and better targeting outreach efforts, are all worth the effort.

Ready to implement a hashed linkage?

Read our guide:
Hashed Linkages for Administrative Datasets: A Technical How-to Guide

Acknowledgments

We are very grateful to the staff at the California Department of Social Services and the California Franchise Tax Board for their thoughtful feedback.

This publication is based on research funded by in part by the Bill & Melinda Gates Foundation. The findings and conclusions contained within are those of the authors and do not necessarily reflect positions or policies of the Bill & Melinda Gates Foundation. All errors should be attributed to the authors.

Any findings reported herein were performed with the permission of the California Department of Social Services and the California Franchise Tax Board. The opinions and conclusions expressed herein are solely those of the authors and should not be considered as representing the policy of the collaborating department, agency, or any department or agency of the California government.

The California Policy Lab builds better lives through data-driven policy. We are an independent, nonpartisan research institute at the University of California with sites at the Berkeley and Los Angeles campuses.