About IDEAL: Introduction & Use Cases

World Bank Team

June 9, 2025

1

Introduction to IDEAL

Alaka Holla

How much evidence is there on using pay-for-performance for general practitioners?

Is childcare a good idea for both women and children?

Is preschool more effective when parents also receive counselling?

Which policies improve foundational learning the most?

Status quo: How do people answer these questions right now?

1. They do a quick search on Google or Google Scholar and focus on the studies that have clearly stated, often positive, results.

2. They check if JPAL, IPA, or the World Bank has a nice brief or they check the introduction to a paper they like on the topic.

3. Maybe they look for existing meta-analyses.

4. They ask Alaka, who then takes a deep breath.

What would we need to answer these questions?

How much evidence is there on using pay-for-performance for general practitioners?

Is childcare a good idea for both women and children?

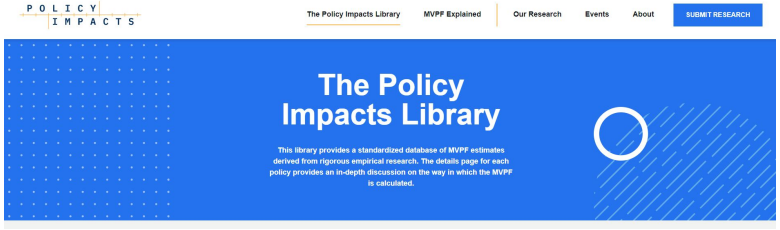
Is preschool more effective when parents also receive counselling?

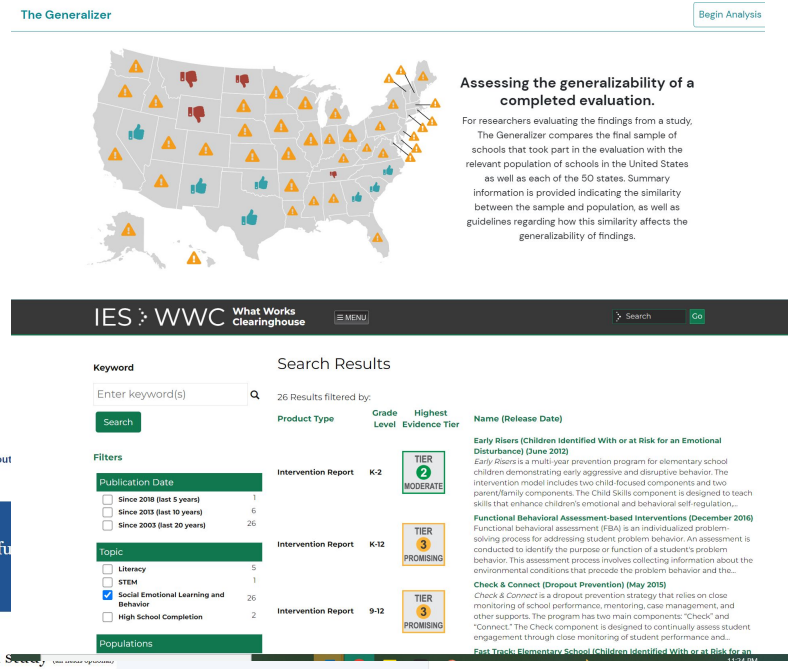
Which policies improve foundational learning the most?

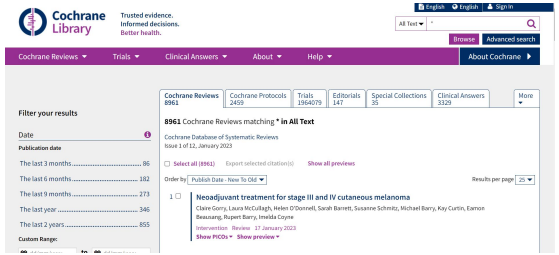
What would we need to answer these questions?

1. A compilation of the **universe** of **estimated treatment effects**, **across disciplines,** that is **up-to-date**.
2. Some way of **standardizing effect sizes** across studies.
3. Some detail on the evaluated **interventions.**
4. Some detail on what is **measured.**
5. A distinction across **contexts**.
6. Some indication of **quality** - i.e. the correct information has been extracted.

There are some existing databases…







The ideal database doesn't exist. Yet.

1. A compilation of the **universe** of **estimated treatment effects**, **across disciplines,** that is **up-to-date**.
2. Some way of **standardizing effect sizes** across studies.
3. Some detail on the evaluated **interventions.**
4. Some detail on what is **measured.**
5. A distinction across **contexts**.
6. Some indication of **quality** - i.e. the correct information has been extracted.

This has led to a situation where:

1. **Evidence aggregation is slow, manual, and limited to researchers with a lot of resources; biased by the questions they pose.**

2. **There are evidence gaps in terms of outcomes, interventions, and country contexts.**

3. **Visibility and publication biases and shortfalls in research transparency distort the evidence base.**

What Really Works to Improve Learning in Developing Countries? An Analysis of Divergent Findings in Systematic Reviews

David K. Evans and Anna Popova

Over the course of just two years, at least six reviews have examined interventions that seek to improve learning outcomes in developing countries. Although the reviews ostensibly have the same objective, they reach sometimes starkly different conclusions. The first objective of this paper is to identify why reviews diverge in their conclusions and how future reviews can be more effective. The second objective is to identify areas of overlap in the recommendations of existing reviews of what works to improve learning. This paper demonstrates that divergence in the recommendations of learning reviews is largely driven by differences in the samples of research incorporated in each review. Of 229 studies with student learning results, the most inclusive review incorporates less than half of the total studies. Across the reviews, two classes of programs are recommended with some consistency. Pedagogical interventions that tailor teaching to student learning levels—either teacher-led or facilitated by adaptive learning software—are effective at improving student test scores, as are individualized, repeated teacher training interventions often associated with a specific task or tool. Future reviews will be most useful if they combine narrative review with meta-analysis, conduct more exhaustive searches, and maintain low aggregation of intervention categories. Education, Impact Evaluation, Human Capital. JEL codes: O15, I21, I28, J13

David K. Evans, Anna Popova, What Really Works to Improve Learning in Developing Countries? An Analysis of Divergent Findings in Systematic Reviews, *The World Bank Research Observer*, Volume 31, Issue 2, August 2016, Pages 242–270

The ideal database doesn't exist. Yet. Enter IDEAL!

1. A compilation of the **universe** of **estimated treatment effects**, **across disciplines,** that is **up-to-date**.
2. Some way of **standardizing effect sizes** across studies.
3. Some detail on the evaluated **interventions.**
4. Some detail on what is **measured.**
5. A distinction across **contexts**.
6. Some indication of **quality** - i.e. the correct information has been extracted.

IDEAL

A **joint library** of RCTs in low & middle-income countries with *accurate*, *reliable*, and *consistent* information on:

- ❑ Estimated impacts and their precision
- ❑ Study design
- ❑ Intervention details
- ❑ Research guidance for users

A **platform** for making data available to inform program and research designs and steer investments.

Collaboration

WORLD BANK GROUP

SIEF Strategic Impact Evaluation Fund

DEVELOPMENT RESEARCH

DIME DRIVING GLOBAL IMPACT

CEGA Center for Effective Global Action

AIDGRADE

Northwestern

ipa Innovations for Poverty Action

3ie International Initiative for Impact Evaluation

IES Institute of Education Sciences

What Works Clearinghouse

2

Use Cases for IDEAL

Anja Sautmann

Use cases in both policy and research

Aggregating Evidence for Policy

Evaluating the Body of Evidence

Econometric Methods

1. Aggregating Evidence for Policy

Most important purpose: evidence aggregation

- Compare (determinants of) effects of the same intervention type on a range of outcomes
- Compare different interventions for effectiveness for the same outcome
- From simple stock-taking to formal meta-analysis

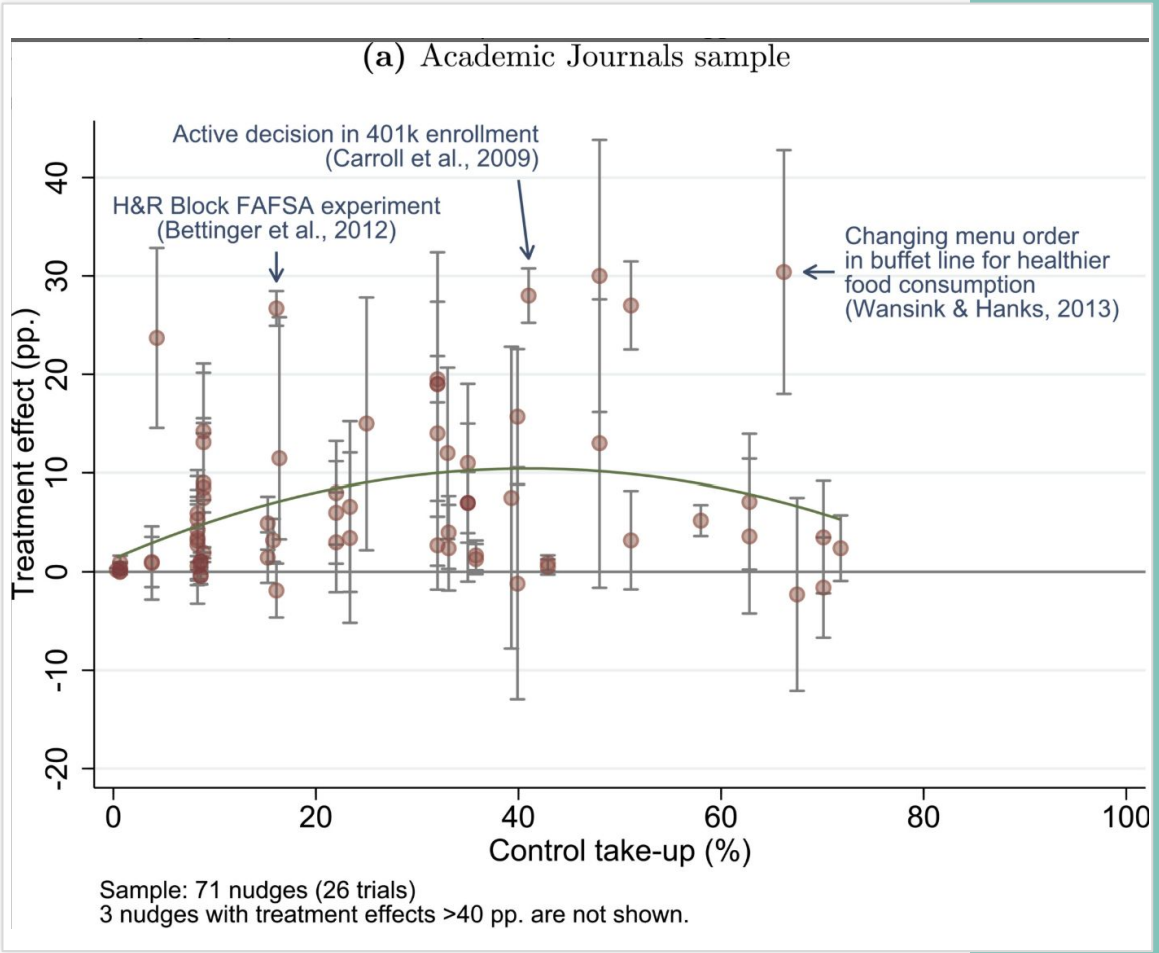
Aggregating Evidence for Policy: Example 1

- Comparing the same intervention across different settings/implementations
  - **“RCTs to Scale: Comprehensive Evidence from Two Nudge Units” (Della Vigna & Linos, 2022, Econometrica)**

- Lots of buzz around nudges: very low cost and effects seemingly large.

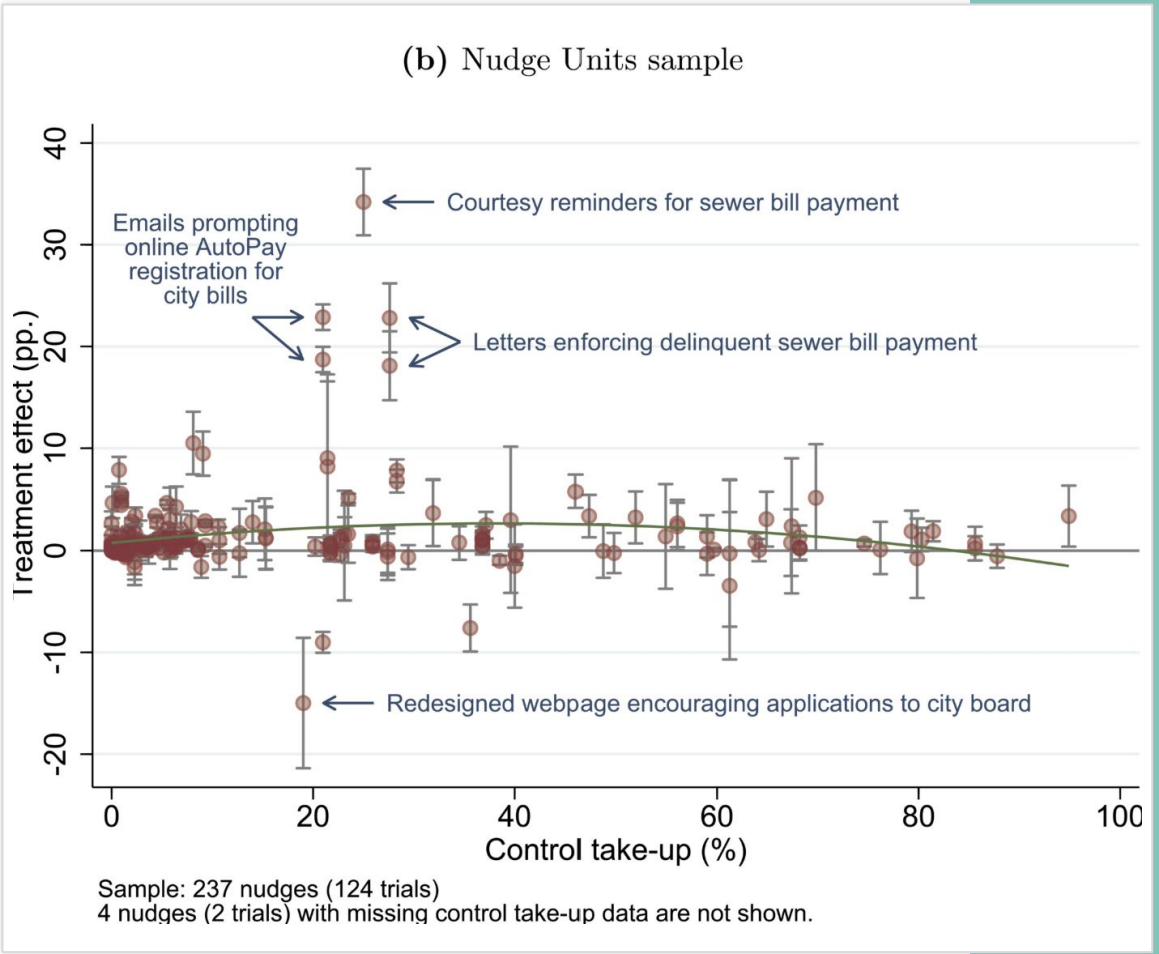
“Academic” nudges

Academic papers on “nudge” interventions: average increase in take up from 26% to nearly 35%.



Nudge unit nudges

In data from 126 trials, 23 million observations: Average increase in take up from 17.3% to 18.7%.



Aggregating Evidence for Policy: Example 2

- Comparing interventions that target the same or related outcomes
    - **Paper: "Cost-effectiveness of 14 Global Fund recommended interventions for HIV/AIDS, malaria, syphilis, and tuberculosis in 128 countries: a meta-regression analysis" (Silke et al., Lancet Global Health, 2024)**
- Incremental cost effectiveness ratio (ICER) estimates from Tufts CEA and GH CEA registries (cost/utility ratios, cost-per-DALY-averted)
- Fill data gaps to provide countries with a "league table" of interventions

2. Evaluating the Body of Evidence

Another important purpose: understand what kind of evidence we currently have (vs. what we would like to have?)

- Simplest case: evidence gaps – on specific outcomes, regions/countries, interventions…
- But: can also ask how studies are done, and how that drives (published) estimates
- Example: Understanding what drives heterogeneity in effect sizes **“How much can we generalize from impact evaluations?”, Eva Vivalt, JEEA 2020**

Evaluating the Body of Evidence: Example

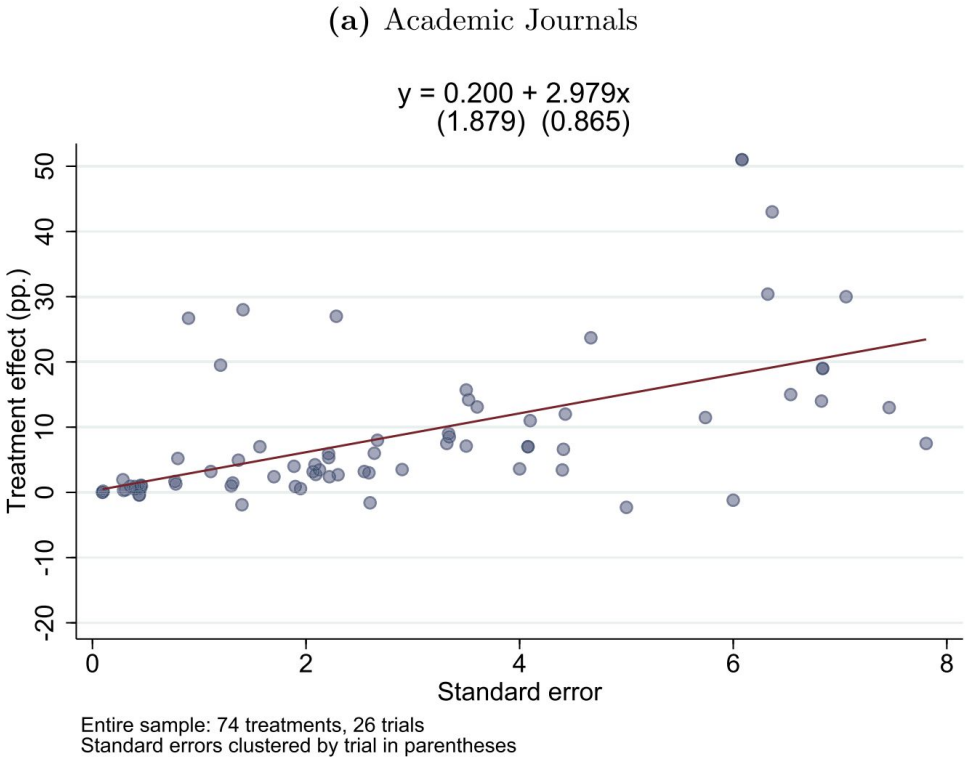
“Identification of and Correction for Publication Bias”, Andrews & Kasy, AER 2019

- Conditional publication probability (as a function of study results): bias if < 1 for effect estimates that are not significant
  - [Note: an estimate can be not significant even if the true effect is not zero]

- With a large enough set of estimates:
  - **Relationship between SE and effect size is a measure of bias**

SE vs. TE size

With publication bias, *published* effect sizes from smaller studies tend to be larger.



Estimated TE as a function of standard error from **academic nudge trials**.

Econometric or other methods

- Many practical and theoretical open questions, and constant innovation, in designing randomized studies and analyzing the resulting data

- Data from existing studies with suitable designs can help **simulate new sampling or estimation methods,** Ex.: “Adaptive treatment assignment for policy choice”, Kasy & Sautmann 2021

- **Body of evidence can…**
  - show how, and how often, specific methods are used in practice
  - **show whether methodological choices affect precision or size of TEs**

Econometric or other methods: Example

Muralidharan et al. "Factorial Designs, Model Selection, and (Incorrect) Inference in Randomized Experiments", ReStat 2025

- From the paper:
  - 27 out of 124 experiments in top five journals over 10 years use factorial designs (cross-randomization)
  - 19 of them do not include interaction effects between cross-randomized treatments in the estimation
  - Inference is only correct if the true interaction effect is zero

- After re-estimating the models including interaction terms:
  - Median absolute change in point estimates is 96%
  - 26% of estimates change sign, 53% (29 out of 55) of TE estimates are not significant at 5% level anymore

3

The IDEAL Pilot 2025

Alaka Holla

Goals of the pilot

- Data extraction from 1,000 RCTs using IDEAL survey fields
- Test the minimum set and build “controlled vocabularies”
- Speed up extraction with machine learning methods
- Present the prototype to key audiences

IDEAL outputs

1	**RCT classifier package**	Open-source tools to automate search for published studies and classification by empirical method and thematic focus.
2	**Metadata schema for minimum fields**	A minimum set of fields needed to standardize, aggregate, and analyze average effects across all RCTs in the social sciences
3	**Survey fields to capture schema fields**	3-stage survey instrument with detailed coding protocol to consistently capture schema fields.
4	**Data entry masks**	Open-source survey forms to extract information on survey fields from published papers
5	**Quality assurance and supervisor protocol**	Procedures, survey forms, and code for quality checks of extracted data that yield calculations of accuracy and inter-rater reliability for the data in library.

IDEAL outputs

6	**Training resources**	Training packages with manuals, slides, videos, and practice papers to guide data extraction following the IDEAL survey instruments.
7	**Papers and extracted data**	Data extracted from 1,000 pilot papers
8	**Data, data documentation and code**	Data and metadata published in MicroData Library and statistical software package to calculate average effect sizes from downloaded IDEAL data.
9	**Partnerships**	Partnerships with other development agencies, research institutes, and universities to test and use IDEAL schema and data.
10	**Library**	A user interface and a public repository with all IDEAL outputs including process documents.

Progress so far

This is YOU!

1 → 2 → 3 → 4

1 Gather and review existing aggregation tools ✔

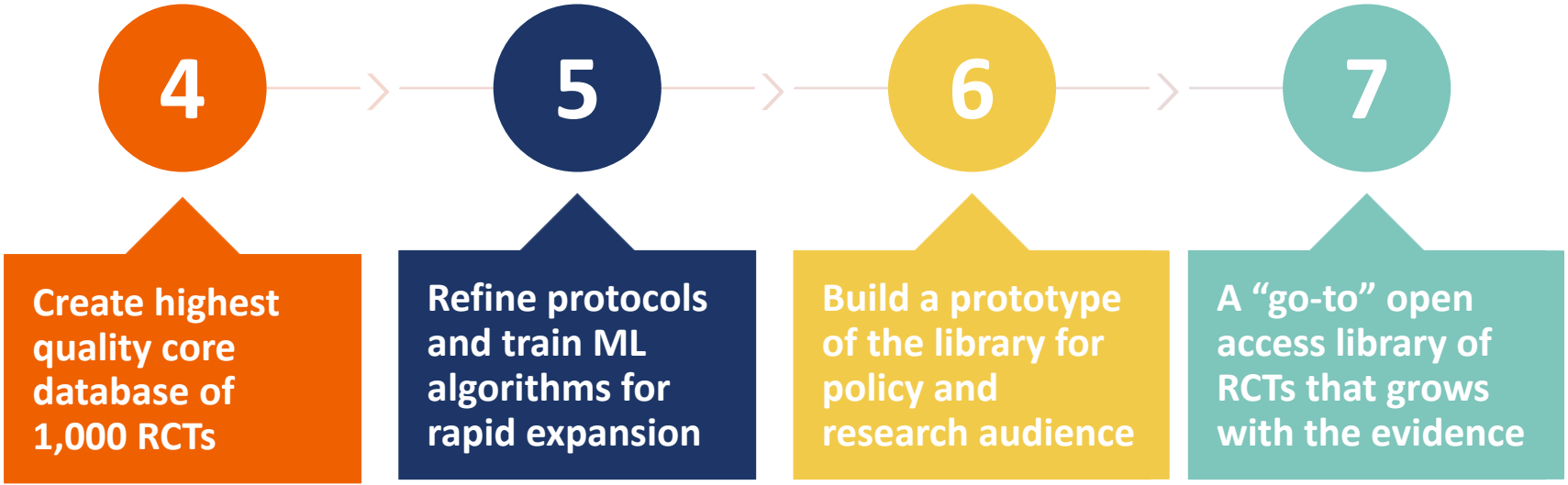
2 Identify the “minimum” set of information that describes an RCT ✔

3 Use international standards and develop coding protocols ✔

4 Pilot minimum set and coding protocols, create core database — In progress

Where we are going

This is YOU!



In progress

Thank you for listening

Alaka Holla
aholla@worldbank.org

Anja Sautmann
asautmann@worldbank.org

IDEAL
Impact Data and Evidence Aggregation Library