Optimal Policy Generator: Evidence-Based Policy Recommendations for Jurisdictions

Systematic Generation of Enact/Replace/Repeal/Maintain Recommendations Using Quasi-Experimental Methods and Bradford Hill Criteria

Mike P. Sinn

2025-01-19

Table of contents

> **Working Paper**
>
> This specification describes a framework for generating jurisdiction-specific policy recommendations. It complements the Optimocracy paper's Budget Impact Score (BIS) by extending evidence-based governance from spending allocation to policy adoption and reform.

Abstract

This specification describes the **Optimal Policy Generator (OPG)**, a framework for producing jurisdiction-specific policy recommendations based on quasi-experimental evidence. OPG answers four questions: "What should we add? Change? Remove? Keep?" The framework operates at any jurisdiction level (country, state, county, city) and produces four outputs: (1) **Enact** - new policies the jurisdiction should adopt, (2) **Replace** - existing policies to modify (change level or approach), (3) **Repeal** - harmful policies to remove, and (4) **Maintain** - current policies aligned with evidence. Each recommendation includes expected effect estimates, confidence grades, monetized impact, blocking factors, recommended jurisdictional level (subsidiarity), and tracking guidance for continuous improvement. This specification complements the Optimocracy framework's Budget Impact Score (BIS): where BIS produces jurisdiction-specific spending recommendations ("Texas underspends on education by $X"), OPG produces jurisdiction-specific policy recommendations ("Texas should enact primary seat belt laws, replace tobacco tax: $1.41 → $2.50").

JEL Classification: H10, D72, C54, I18, D61

H10 (Public Finance, Structure and Scope), D72 (Political Economy), C54 (Quantitative Policy Modeling), I18 (Health Policy), D61 (Allocative Efficiency; Cost-Benefit Analysis)

1 System Overview

1.1 What Policymakers See

A jurisdiction-specific dashboard showing which policies to enact, replace, repeal, or maintain, ranked by expected welfare impact:

ℹ Example: Policy Recommendations for Texas

ENACT (New policies Texas should adopt)

Policy	Expected Effect	Evidence Grade	Monetized Impact	Priority
Primary seat belt law	-1.8 deaths/100K	A	+$1.2B/year	High
Motorcycle helmet req.	-1.2 deaths/100K	A	+$680M/year	Medium

REPLACE (Existing policies to modify)

Policy	Current → Optimal	Expected Effect	Evidence Grade	Monetized Impact
Tobacco tax	$1.41 → $2.50/pack	-8.2 pp smoking	A	+$4.8B/year
Speed limit	85 → 70 mph	-0.8 deaths/100K	B	+$350M/year

REPEAL (Harmful policies to remove)

Policy	Current Effect	Evidence Grade	Removing Would Save
None identified	-	-	-

MAINTAIN (Current policies aligned with evidence)

Policy	Current Level	Evidence Grade	Status
DUI threshold	0.08 BAC	A	Continue
Graduated licensing	3-stage system	A	Continue

Expected Total Welfare Gain: +$7.0B/year from adopting all recommendations.
Blocking factors flagged where applicable (constitutional constraints, federal preemption, etc.)

1.2 The Analogy: Spending Gaps vs. Policy Gaps

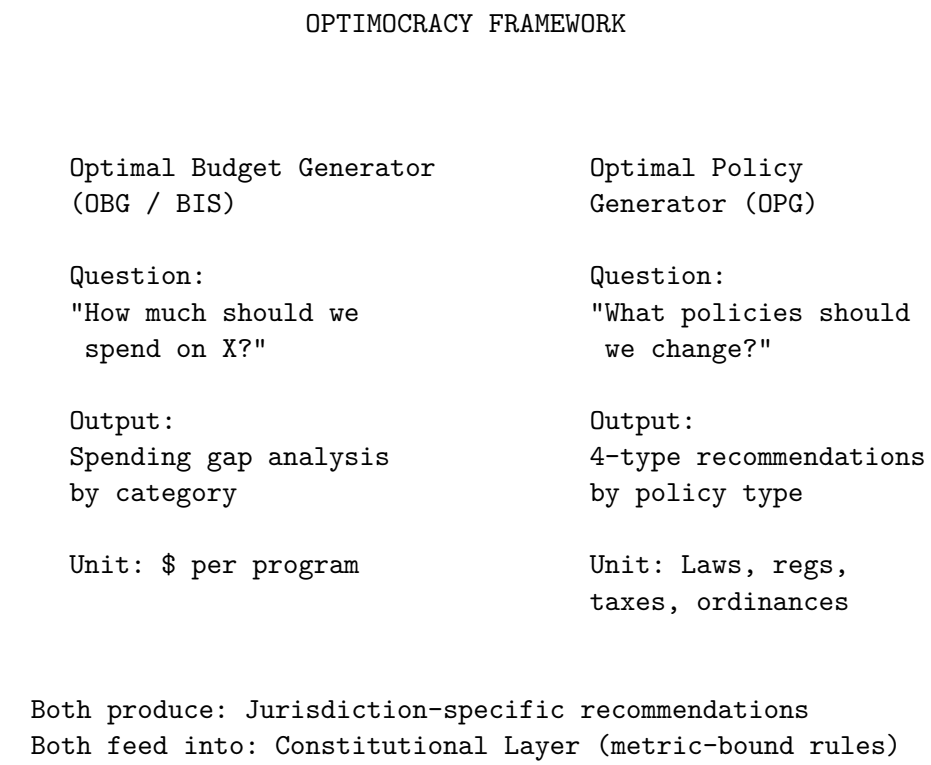
Framework	Question Answered	Primary Output	Example
Optimal Budget Generator (OBG/BIS)	How much should we spend?	Integrated budget recommendations	"Texas underspends on education by $X"
Optimal Policy Generator (OPG)	What policies should we adopt?	Policy gap analysis	"Texas should enact X, repeal Y"

Both frameworks apply the same principle: compare current jurisdiction practice to evidence-optimal practice, quantify the gap, and produce actionable recommendations.

1.3 What Policy Analysts See

- **Effect estimates** with standard errors, confidence intervals, and heterogeneity statistics
- **Policy Impact Scores (PIS)** for each policy-outcome relationship (intermediate metric)
- **Bradford Hill criteria scores** for causality assessment
- **Analysis method** used (synthetic control, DiD, RDD) with quality diagnostics
- **Confounders controlled** and potential threats to validity
- **Natural experiments identified** for validation opportunities
- **Similar policies** via embedding-based similarity for evidence transfer
- **Jurisdiction-specific adjustments** based on demographics, existing policies, and context

1.4 Where This Fits



2 Introduction

2.1 Why Generic Policy Scores Are Not Enough

Previous approaches to evidence-based policy produce generic rankings:

> "Tobacco taxes reduce smoking by 8.2 percentage points on average across 23 states…"

This is useful background but doesn't answer the policymaker's question: **"What should MY jurisdiction do?"**

The Optimal Policy Generator transforms generic evidence into jurisdiction-specific recommendations:

> **Recommendations for Texas**
>
> Texas currently has a tobacco tax of $1.41/pack (below median of $1.91).
>
> **Recommended action**: Increase tobacco tax by $1.00/pack
>
> **Expected effect**:
>
> - -6.5 pp smoking rate (adjusted for Texas demographics)
> - +$1.8B/year health savings
> - +42K QALYs/year
>
> **Evidence grade**: A (based on 8 similar states, synthetic control)
>
> **Blocking factors**: None identified

2.2 Why Policy Ranking Fails Today

Current policy adoption follows a process dominated by political economy dynamics well-documented in the public choice literature[1,2]:

1. **Lobbying intensity**: Policies that benefit concentrated interests (with resources to lobby) are adopted over policies that benefit diffuse majorities[3,4]
2. **Ideological priors**: Policymakers filter evidence through pre-existing beliefs, accepting studies that confirm priors and rejecting those that don't
3. **Anecdote-driven reasoning**: Vivid individual cases drive policy more than systematic evidence ("If it saves one child…")
4. **Status quo bias**: Existing policies persist regardless of evidence because change requires political capital
5. **Salience heuristics**: Policies addressing visible problems (terrorism, rare diseases) receive disproportionate resources relative to invisible problems (air pollution, chronic disease)

The result: trillions of dollars in welfare losses from documented policy failures (see Optimocracy paper, Section 2). Evidence-based policy movements have attempted to address these failures[5,6], but lack the systematic, jurisdiction-specific recommendation generation that OPG provides.

2.3 Scale of Available Evidence

We have access to:

- **202 countries** with political and economic data spanning **1789 to present** (V-Dem)
- **167 countries** with regime and stability data from **1800 to present** (Polity V)
- **36 OECD countries** with detailed policy data from **1960 to present** (CPDS)
- **Thousands of subnational jurisdictions** (US states, EU regions, Indian states) with policy variation
- **Millions of policy changes** documented in legislative databases

Even with imperfect causal inference, systematically analyzing this data and translating it to jurisdiction-specific recommendations will produce outcomes orders of magnitude better than the current system.

2.4 Contributions

This paper makes three primary contributions to the policy evaluation literature:

1. **Methodological**: We develop a systematic framework for translating quasi-experimental evidence into jurisdiction-specific policy recommendations, extending beyond generic evidence ratings to actionable output in four categories (enact/replace/repeal/maintain).

2. **Taxonomic**: We formalize the four recommendation types and introduce the Policy Impact Score (PIS) as an intermediate metric combining effect magnitude, causal confidence (Bradford Hill criteria), and methodological quality. This provides a standardized approach to evidence aggregation.

3. **Applied**: We demonstrate the complete framework with a worked example for Texas traffic safety policy, showing how generic effect estimates are translated into context-adjusted, prioritized recommendations with blocking factors and tracking guidance.

3 Related Work

3.1 Existing Policy Evaluation Frameworks

Regulatory Impact Analysis (RIA): Required by executive order for US federal regulations since 1981[7]. RIA estimates costs and benefits of proposed rules but: (1) applies only to new regulations, not the existing policy stock; (2) lacks systematic cross-jurisdiction evidence aggregation; (3) does not produce jurisdiction-specific recommendations for subnational governments.

What Works Clearinghouse (WWC): The Institute of Education Sciences operates WWC to review education interventions against methodological standards[8]. WWC demonstrates that systematic evidence synthesis is feasible, but: (1) covers only education; (2) provides generic intervention ratings, not jurisdiction-specific recommendations; (3) does not quantify expected welfare gains.

Cochrane and Campbell Collaborations: These systematic review organizations cover healthcare[9] and social policy[10] respectively. They represent the gold standard for evidence synthesis but: (1) produce narrative reviews rather than quantitative recommendations; (2) provide no jurisdiction-specific output; (3) operate on slow update cycles (years between reviews).

Congressional Budget Office (CBO): CBO provides nonpartisan fiscal scoring of proposed legislation. While valuable for budget discipline, CBO: (1) estimates budgetary effects rather than welfare; (2) evaluates what is proposed rather than what should be proposed; (3) is reactive rather than proactive.

Benefit-Cost Analysis Tradition: The broader benefit-cost literature[11,12] provides theoretical foundations for policy evaluation but typically focuses on individual project or regulation assessment rather than systematic cross-jurisdiction recommendation generation.

3.2 This Framework's Contribution

OPG differs from existing approaches by:

1. **Producing jurisdiction-specific recommendations** rather than generic evidence ratings
2. **Covering the full policy stock** (enact/replace/repeal/maintain) not just new proposals
3. **Aggregating quasi-experimental evidence** via meta-analysis with heterogeneity quantification
4. **Applying Bradford Hill criteria** systematically to assess causality confidence
5. **Including subsidiarity guidance** (optimal jurisdictional level) and tracking for continuous improvement

4 Theoretical Framework

4.1 The Policy Optimization Problem

Let $\mathcal{P}$ denote the set of available policies. For jurisdiction j, let $P_j \subseteq \mathcal{P}$ denote the current policy bundle and $W_j(P)$ denote welfare under policy bundle P.

The social planner's problem:

$$P_j^* = \arg\max_{P \subseteq \mathcal{P}} W_j(P) \quad \text{subject to feasibility constraints}$$

Assumption 1 (Additive Separability): For tractability, assume welfare is approximately additively separable across policy domains:

$$W_j(P) \approx \sum_{p \in P} w_j(p) + \varepsilon_{\text{interactions}}$$

where $w_j(p)$ is the marginal welfare contribution of policy p in jurisdiction j, and interaction terms are second-order.

Proposition 1 (Policy Gap Characterization): Under Assumption 1, the welfare-optimal policy set satisfies:

$$P_j^* = \{p \in \mathcal{P} : w_j(p) > 0\}$$

and the policy gap for jurisdiction j is:

$$\Delta_j = (P_j^* \setminus P_j) \cup (P_j \setminus P_j^*)$$

where $(P_j^* \setminus P_j)$ represents beneficial policies the jurisdiction lacks (enact candidates) and $(P_j \setminus P_j^*)$ represents harmful policies the jurisdiction has (repeal candidates).

Proof: Direct consequence of additive separability. Include policy p if and only if $w_j(p) > 0$.

4.2 Evidence Aggregation Properties

Proposition 2 (PIS as Precision-Weighted Evidence): Under random-effects meta-analysis with between-jurisdiction variance τ^2, the pooled effect estimate $\hat{\beta}_{\text{pooled}}$ is:

$$\hat{\beta}_{\text{pooled}} = \frac{\sum_j \frac{1}{\text{SE}_j^2+\tau^2}\hat{\beta}_j}{\sum_j \frac{1}{\text{SE}_j^2+\tau^2}}$$

with variance:

$$\text{Var}(\hat{\beta}_{\text{pooled}}) = \frac{1}{\sum_j \frac{1}{\text{SE}_j^2+\tau^2}}$$

Proof: Standard random-effects meta-analysis derivation (DerSimonian-Laird).

Proposition 3 (Heterogeneity Bounds Transferability): When $I^2 > 75\%$ (high heterogeneity):

$$\text{Var}[\hat{\beta}_j|\hat{\beta}_{\text{pooled}}] > 0.75 \cdot \text{Var}[\hat{\beta}_j]$$

meaning the pooled estimate explains less than 25% of cross-jurisdiction variation. Context-specific estimates are required rather than direct application of the pooled effect.

Proof: By definition, $I^2 = \frac{\tau^2}{\tau^2+\bar{\sigma}^2}$ where $\bar{\sigma}^2$ is typical within-study variance. When $I^2 > 0.75$, between-study variance dominates, and the pooled estimate provides limited information about any individual jurisdiction's true effect.

4.3 Information Value

Proposition 4 (Value of Additional Evidence): The expected value of information from an additional jurisdiction study is:

$$\text{VOI} = E[\max_{a \in \{adopt, reject\}} U(a|\text{new data})] - \max_a E[U(a|\text{current data})]$$

which is maximized when prior uncertainty is high and decision stakes are large.

Proof: Standard Bayesian decision theory[5].

Corollary 1 (Trial Prioritization): Policies with (1) high prior variance in effect estimates, (2) large potential welfare impact, and (3) low trial cost should be prioritized for experimental validation.

5 Core Methodology

5.1 Policy-Outcome Data Structure

The OPG system uses a relational database schema:

5.1.1 Core Tables

```
-- Hierarchical jurisdictions (country > state > county > city)
jurisdictions (
    id, name, jurisdiction_type, -- 'country', 'state', 'county', 'city'
    parent_id, -- FK to parent jurisdiction (e.g., Texas -> USA)
    iso_code, population, gdp_per_capita,
    constitution_type, -- constraints on policy space
    data_quality_score, -- how complete is our policy inventory?
    latitude, longitude, ...
)

-- Policy types (canonical definitions)
policy_types (
    id, name, policy_category_id, policy_type,
    is_continuous, typical_onset_delay_days,
    typical_duration_of_effect_years, canonical_text, ...
)

-- Current policy inventory by jurisdiction
jurisdiction_policies (
    jurisdiction_id, policy_type_id,
    has_policy BOOLEAN,
    policy_strength, -- e.g., tobacco tax amount, not just yes/no
    implementation_date,
    policy_details_json,
    data_source, last_verified
)

-- Outcome variables (welfare metrics)
outcome_variables (
    id, name, category, valence,
    data_source, data_frequency, ...
)

-- Outcome measurements (time series of welfare metrics)
outcome_measurements (
    jurisdiction_id, outcome_variable_id, measurement_date,
    value, confidence_interval_low, confidence_interval_high, ...
)

-- Policy recommendations (generated output)
policy_recommendations (
    jurisdiction_id, policy_type_id,
    recommendation_type, -- 'enact', 'replace', 'repeal', 'maintain'
    -- 'enact': new policy (jurisdiction doesn't have)
    -- 'replace': modify existing policy (change level or approach)
    -- 'repeal': remove policy entirely
```

```
    -- 'maintain': keep current policy (evidence-optimal)
    current_status, -- what they have now (NULL if nothing)
    recommended_target, -- what evidence suggests (for replace/enact with level)
    expected_effect, expected_effect_unit,
    monetized_benefit_annual,
    evidence_grade, priority_score,
    blocking_factors, -- 'constitutional', 'federal_preemption', 'political', etc.
    similar_jurisdictions, -- jurisdictions that adopted this successfully
    -- Jurisdictional level guidance
    minimum_effective_level, -- 'city', 'county', 'state', 'federal'
    recommended_level, -- lowest effective level for max data collection
    -- Tracking for feedback loop
    tracking_metric, -- primary outcome to measure
    tracking_data_source, -- where to get data
    tracking_frequency, -- 'annual', 'quarterly', etc.
    tracking_baseline_method, -- 'pre_implementation_3yr_avg', etc.
    last_generated
)
```

5.1.2 Policy Types

Type	Description	Example	Measurement
`law`	Statutory law passed by legislature	Environmental regulation law	Binary (exists/not)
`regulation`	Administrative rule by agency	Agency emission standards	Continuous (stringency)
`tax_policy`	Tax rate, bracket, credit, deduction	Investment income tax rate	Continuous (rate)
`budget_allocation`	Spending decision	Education spending per pupil	Continuous ($/capita)
`executive_order`	Executive action	Enforcement priority directive	Binary
`court_ruling`	Judicial precedent	Constitutional interpretation	Binary
`treaty`	International agreement	Multilateral cooperation treaty	Binary
`local_ordinance`	Municipal rule	Land use restrictions	Categorical

5.2 Analysis Methods

The OPG system supports multiple quasi-experimental designs, reflecting the "credibility revolution" in applied economics[13]. Each method is appropriate for different data structures[14]:

5.2.1 Synthetic Control Method

Use case: Single treated jurisdiction, good donor pool of similar untreated jurisdictions.

Method: Construct a "synthetic" control as a weighted average of untreated jurisdictions that matches the treated jurisdiction's pre-treatment outcome trajectory. Post-treatment divergence estimates the causal effect.

Quality metrics: - `pre_treatment_rmse`: How well does synthetic control match pre-treatment? (Lower is better) - `placebo_p_value`: Permutation test comparing treated effect to placebo effects (Lower is better)

Example: Effect of a state tobacco tax increase on smoking rates, using similar states without tax changes as donors[15,16]. For comprehensive reviews of the synthetic control method, see[17].

5.2.2 Difference-in-Differences (DiD)

Use case: Multiple treated jurisdictions, staggered adoption timing, parallel trends assumption plausible.

Method: Compare pre-post change in treated jurisdictions to pre-post change in control jurisdictions. Difference of differences estimates treatment effect. For settings with staggered adoption, modern estimators account for heterogeneous treatment effects across cohorts[18].

Quality metrics: - `parallel_trends_test_stat`: Test statistic for pre-treatment trend equality - `parallel_trends_p_value`: P-value for parallel trends test (Higher is better, want to fail to reject)

Example: Effect of occupational licensing reforms across states with different adoption timing.

5.2.3 Regression Discontinuity Design (RDD)

Use case: Sharp eligibility threshold determines treatment assignment.

Method: Compare outcomes just above vs. just below the threshold. If other characteristics are smooth across the threshold, the discontinuity in outcomes estimates the causal effect.

Quality metrics: - Bandwidth selection diagnostics - McCrary density test for manipulation - Covariate balance at threshold

Example: Effect of program eligibility on outcomes at an income or age threshold (e.g., retirement benefits at age 65).

5.2.4 Event Study / Interrupted Time Series

Use case: Need to visualize pre-trends and dynamic treatment effects.

Method: Estimate treatment effects at each time period relative to treatment, including leads (pre-treatment) and lags (post-treatment).

Quality metrics: - Pre-treatment coefficients should be near zero (no anticipation) - Post-treatment coefficients show effect dynamics

Example: Effect of unemployment insurance extensions on job search behavior, showing both anticipation effects (before benefits expire) and persistence of impact (after return to baseline).

5.2.5 Confidence Weighting by Method

Method	Base Confidence Weight	Rationale
Randomized experiment	1.00	Gold standard; rare for policies
Regression discontinuity	0.90	Local randomization at threshold
Synthetic control	0.85	Good pre-treatment fit implies validity
Difference-in-differences	0.80	Requires untestable parallel trends
Event study	0.75	Descriptive of dynamics; less rigorous
Interrupted time series	0.65	Single-unit; history threats
Simple before-after	0.40	No control group; confounding likely
Cross-sectional	0.25	Snapshot; severe confounding

5.3 Bradford Hill Criteria Mapping for Policy

Bradford Hill's criteria for causality[19], originally developed for epidemiology, map to policy evaluation:

Criterion	Definition	DFDA (Drug) Implementation	OPG (Policy) Implementation
Strength	Magnitude of association	Correlation coefficient	Effect estimate magnitude (standardized)
Consistency	Replicated across studies	Across users (N-of-1)	Across jurisdictions (I^2 heterogeneity)
Specificity	Specific exposure → specific outcome	Specific drug-symptom pair	Policy domain → outcome category
Temporality	Exposure precedes outcome	Onset delay optimization	Policy adoption precedes outcome change
Gradient	Dose-response relationship	Dose-outcome curve	For continuous policies (tax rates, spending levels)
Plausibility	Mechanistic explanation	Biological mechanism	Economic mechanism (theory)
Coherence	Consistent with broader knowledge	Medical literature	Economic literature
Experiment	Experimental/quasi-experimental evidence	N-of-1 trial design quality	Quasi-experimental design quality

Criterion	Definition	DFDA (Drug) Implementation	OPG (Policy) Implementation
Analogy	Similar exposures have similar effects	Similar drugs	Similar policies (embedding similarity)

Each criterion is scored 0-1 based on the evidence:

$$\text{Hill}_{\text{criterion}} = f(\text{evidence for criterion})$$

The aggregate causal confidence score combines criteria:

$$\text{CCS} = \frac{\sum_{i=1}^{9} w_i \cdot \text{Hill}_i}{\sum_{i=1}^{9} w_i}$$

Where w_i are criterion weights. Default weights emphasize temporality (must be satisfied), experiment (quasi-experimental design quality), and consistency (replication across jurisdictions).

6 Jurisdiction Policy Inventory

6.1 Tracking Current Policies by Jurisdiction

Before generating recommendations, OPG must know what policies each jurisdiction currently has. The `jurisdiction_policies` table tracks:

Field	Description	Example
`has_policy`	Whether jurisdiction has this policy type	TRUE/FALSE
`policy_strength`	For continuous policies, the current level	$1.41/pack (tobacco tax)
`implementation_date`	When current policy took effect	2009-01-01
`policy_details_json`	Structured details about implementation	{“primary_enforcement”: false}
`data_source`	Where this information came from	“Texas Tax Code §154.021”
`last_verified`	When this was last confirmed accurate	2024-06-15

6.2 Data Sources for Policy Status

Jurisdiction Level	Primary Sources	Update Frequency
Country	WTO, OECD, IMF policy databases	Annual
US State	NCSL, state legislative databases, LexisNexis	Continuous
EU Member	EUR-Lex, national legal databases	Continuous

Jurisdiction Level	Primary Sources	Update Frequency
US City/County	Municipal code databases, Municode	Varies
Other Subnational	National statistics offices, academic datasets	Varies

6.3 Handling Missing Data

Data completeness varies by jurisdiction and policy type:

Data Quality Score	Interpretation	Recommendation Confidence
> 0.9	Comprehensive inventory	Full confidence
0.7 - 0.9	Most major policies tracked	High confidence
0.5 - 0.7	Significant gaps	Medium confidence; flag gaps
< 0.5	Sparse data	Low confidence; prioritize data collection

Recommendations are only generated when policy status is known with reasonable confidence.

7 Policy Gap Analysis

7.1 Comparing Current to Optimal

For each jurisdiction j, the policy gap for policy type p is:

$$\text{Gap}_{jp} = \text{Evidence-Optimal}_{p} - \text{Current}_{jp}$$

Where:

- **Evidence-Optimal**: What the evidence suggests the jurisdiction should have
- **Current**: What the jurisdiction actually has

7.2 Gap Types

Gap Type	Definition	Example
Missing policy	Jurisdiction lacks a policy with strong positive evidence	Texas lacks primary seat belt enforcement
Harmful policy	Jurisdiction has a policy with strong negative evidence	Jurisdiction has policy X shown to increase mortality
Suboptimal strength	Continuous policy set below evidence-optimal level	Tobacco tax at \$1.41 vs. optimal ~\$2.50
Excessive strength	Continuous policy set above evidence-optimal level	Speed limit at 85 mph vs. optimal ~70 mph

7.3 Priority Scoring

Recommendations are ranked by priority score:

$$\text{Priority}_{jp} = |\text{Gap}_{jp}| \times \text{Evidence Grade}_{p} \times \text{Monetized Impact}_{jp}$$

High-priority recommendations have: 1. Large gap between current and optimal 2. Strong evidence (Grade A or B) 3. Large expected welfare impact

7.4 Context Adjustment

Effect estimates are adjusted for jurisdiction characteristics:

Adjustment Factor	Description	Example
Demographics	Age structure, income distribution	Tobacco tax effect varies by income
Existing policies	Interaction with current policy bundle	Effect depends on what else is in place
Institutional capacity	Enforcement capability	Weak institutions → smaller effects
Cultural factors	Compliance norms	Varies by society

$$\text{Expected Effect}_{jp} = \hat{\beta}_{p} \times \text{Context Adjustment}_{j}$$

8 Recommendation Generation

8.1 Recommendation Types

Type	Question	When to Use	Example
Enact	"Add this?"	New policy the jurisdiction doesn't have	"ENACT primary seat belt law"
Replace	"Change this?"	Modify existing policy level or approach	"REPLACE tobacco tax: \$1.41 → \$2.50"
Repeal	"Remove this?"	Remove policy with negative evidence	"REPEAL [harmful policy]"
Maintain	"Keep this?"	Current policy is evidence-optimal	"MAINTAIN DUI threshold at 0.08 BAC"

For continuous policies (taxes, spending levels), **Replace** specifies the change from current to optimal level. **Enact** is reserved for truly new policies that don't exist in the jurisdiction.

8.2 Blocking Factors

Recommendations flag constraints that may impede adoption:

Blocking Factor	Description	Example
`constitutional`	Requires constitutional amendment	2nd Amendment limits on gun regulations
`federal_preemption`	Federal law prevents state/local action	Federal minimum wage floor
`treaty_obligation`	International agreement constrains policy	WTO rules on tariffs
`political_feasibility`	Strong organized opposition	Industry lobbying
`implementation_cost`	High fixed costs to implement	New regulatory agency needed

Important: Blocking factors are flagged but do not filter recommendations. The full evidence-optimal set is always shown; users can filter by feasibility if desired.

8.3 Similar Jurisdictions

For each recommendation, OPG identifies jurisdictions that: 1. Had similar characteristics to the target jurisdiction 2. Adopted the recommended policy 3. Experienced the predicted effects

This provides concrete examples for policymakers: "Vermont (similar demographics, adopted this in 2015, saw -7.1 pp smoking reduction)."

8.4 Recommended Tracking (for OPG Feedback)

Each recommendation includes minimal tracking guidance to enable continuous OPG improvement:

Field	Description	Example
Primary metric	The outcome variable to track	Traffic deaths per 100K
Data source	Where to get it	State vital statistics
Measurement frequency	How often	Annual
Comparison baseline	What to compare against	Pre-implementation 3-year average

This creates a learning loop: OPG recommends → jurisdiction implements → reports outcomes → OPG improves future recommendations.

9 Optimal Jurisdictional Level for Policy Implementation

9.1 The Subsidiarity Principle for Evidence Generation

OPG recommends policies be implemented at the **lowest jurisdictional level where the policy can be effective**, for two reasons:

1. **Maximize experimental data**: 50 states experimenting > 1 federal policy. 3,000+ counties > 50 states. More jurisdictions = more natural experiments = faster evidence accumulation.
2. **Minimize harm from policy failures**: A failed city ordinance affects thousands; a failed federal policy affects hundreds of millions. Lower-level experimentation bounds downside risk.

9.2 When Higher Levels Are Necessary

Some policies require higher jurisdictional levels:

Reason	Example	Recommendation
Externalities	Pollution crosses borders	State or federal
Race-to-bottom risk	Labor standards, tax competition	Federal floor, state variation above
Network effects	Infrastructure standards	Federal coordination
Economies of scale	Defense, diplomacy	National

9.3 Jurisdictional Level in Recommendations

For each policy recommendation, OPG specifies:

Field	Example
Minimum effective level	"City or higher"
Recommended level	"City (maximize data collection)"
Current adoption	"12 states, 47 cities have this"
Level constraints	"Federal preemption prevents city-level"

10 Policy Impact Score (Intermediate Metric)

10.1 Overview

The Policy Impact Score (PIS) is the intermediate metric used to generate recommendations. It quantifies the strength of evidence that a policy affects an outcome.

10.2 Jurisdiction-Level PIS Calculation

For each jurisdiction j, policy p, and outcome o, the jurisdiction-level PIS is:

$$\text{PIS}_{jpo} = |\hat{\beta}_{jpo}| \cdot \text{CCS}_{jpo} \cdot Q_{jpo}$$

Where:

- $|\hat{\beta}_{jpo}|$ = Absolute value of effect estimate (standardized effect size). We take absolute value because PIS measures *strength of evidence*, not direction. Direction is reported separately; negative effects on desirable outcomes are flagged as potential harms.
- CCS_{jpo} = Causal Confidence Score (aggregate Hill criteria)
- Q_{jpo} = Quality adjustment factor based on analysis method and diagnostics

! Reporting Recommendation: Separate Effect Size from Confidence

The multiplicative PIS formula conflates *magnitude* and *certainty*. A policy with large effect but low confidence (=0.8, CCS=0.3 → PIS=0.24) is treated identically to one with small effect but high confidence (=0.3, CCS=0.8 → PIS=0.24).
Best practice: Always report both components separately:

Policy	Effect Size ( )	Confidence (CCS)	PIS	Interpretation
Policy A	0.8 (large)	0.3 (low)	0.24	Promising but uncertain
Policy B	0.3 (small)	0.8 (high)	0.24	Confidently modest

These require different responses: Policy A needs experimental validation; Policy B may not justify further investment despite high confidence.

10.3 Effect Estimate Standardization

Raw effect estimates vary in scale (years of life, dollars of income, crime rates). We standardize to enable comparison:

$$\hat{\beta}_{\text{std}} = \frac{\hat{\beta}_{\text{raw}}}{\sigma_{\text{outcome}}}$$

Where σ_{outcome} is the cross-jurisdictional standard deviation of the outcome variable.

10.4 Quality Adjustment Factor

$$Q = w_{\text{method}} \cdot (1 - \text{violations})$$

Where: - w_{method} = Method confidence weight (see table above) - violations = Proportion of validity checks failed (parallel trends, pre-treatment fit, etc.)

10.5 Confounder Adjustment

For each analysis, we track which confounders were controlled:

```
{
    "confounders_controlled": ["gdp_growth", "unemployment", "population_age_structure"],
    "confounders_not_controlled": ["neighboring_policy_spillovers", "measurement_error"],
    "confounder_sensitivity": 0.85
}
```

The `confounder_sensitivity` field estimates how much the effect estimate might change if uncontrolled confounders were addressed (Oster's delta,[20]).

11 Global (Aggregate) PIS Calculation

Aggregate estimates combine jurisdiction-level analyses via random-effects meta-analysis:

$$\hat{\beta}_{\text{pooled}} = \frac{\sum_j w_j \hat{\beta}_j}{\sum_j w_j}$$

Where weights $w_j = \frac{1}{\text{SE}_j^2 + \tau^2}$ incorporate both within-study variance and between-study heterogeneity (τ^2).

11.1 Heterogeneity Statistics

Following standard meta-analysis conventions[9]:

- **I²**: Percentage of variance due to heterogeneity (vs. sampling error)
  - $I^2 < 25\%$: Low heterogeneity
  - $25\% \leq I^2 < 75\%$: Moderate heterogeneity
  - $I^2 \geq 75\%$: High heterogeneity (effects vary substantially across jurisdictions)
- **²**: Estimated between-study variance
- **Q statistic**: Cochran's test for heterogeneity

High heterogeneity suggests moderators (policy effects vary by context) rather than a single true effect.

11.2 Evidence Grading

Grade	Criteria	Interpretation
A	Multiple high-quality quasi-experiments (synthetic control, RDD) OR RCT; I² < 50%; consistent direction	Strong evidence; ready for implementation
B	Single RCT OR multiple well-designed DiD/event studies; I² < 75%; mostly consistent	Good evidence; consider piloting
C	Well-designed observational studies with confounding control; moderate consistency	Suggestive evidence; needs validation
D	Case studies, weak observational evidence, or high heterogeneity	Weak evidence; exploratory only
F	Expert opinion only OR conflicting high-quality evidence	Insufficient or contradictory evidence

11.3 Context-Specific Confidence

Effects may vary by jurisdiction characteristics. We report confidence separately for:

Context	Description	Example Modifier
High-income countries	OECD members, GDP/capita > $30K	Tax policy effects
Low-income countries	GDP/capita < $5K	Different institutional capacity
Federal systems	Policy set at national level	vs. subnational variation
Subnational	States, provinces, cities	Local policy autonomy

12 Quality Requirements & Validation

12.1 Minimum Thresholds for Inclusion

Criterion	Minimum	Rationale
Pre-treatment periods	4	Need to assess pre-trends
Post-treatment periods	2	Need to observe effect
Outcome observations	20	Statistical power
Control jurisdictions (for DiD)	5	Donor pool size
Pre-treatment RMSE (synthetic control)	< 2 SD	Acceptable pre-treatment fit

12.2 Parallel Trends Testing (DiD)

For difference-in-differences analyses, we test whether treated and control jurisdictions had parallel outcome trends before treatment:

1. Estimate event study with pre-treatment leads
2. Test joint significance of pre-treatment coefficients
3. If $p < 0.10$, flag as potential parallel trends violation
4. Report sensitivity: how different would trends need to be to explain away the effect?

12.3 Pre-Treatment Fit (Synthetic Control)

For synthetic control analyses:

1. Calculate RMSE of synthetic vs. actual treated unit pre-treatment
2. Compare to distribution of placebo RMSEs (treating each donor as "treated")
3. If treated RMSE is in top 10% of placebo RMSEs, flag as poor fit
4. Report ratio of post-treatment effect to pre-treatment RMSE

12.4 Placebo and Robustness Tests

Test	Purpose	Implementation
In-time placebo	Does "treatment" show effect before it happened?	Assign fake treatment date before actual
In-space placebo	Do untreated units show similar effects?	Apply analysis to control jurisdictions

Test	Purpose	Implementation
Leave-one-out	Is result driven by single jurisdiction?	Re-estimate dropping each jurisdiction
Bandwidth sensitivity	(For RDD) Is result robust to bandwidth choice?	Estimate with multiple bandwidths
Covariate adjustment	Does controlling for confounders change result?	Add covariates, compare estimates

13 Interpreting Recommendations

13.1 Priority Tiers

Tier	Criteria	Action
Quick Wins	High impact, low blocking factors, Grade A evidence	Immediate adoption recommended
Major Reforms	High impact, significant blocking factors	Requires political capital; strategic timing
Long-Term	Moderate impact, constitutional or treaty constraints	Requires structural change
Monitor	Moderate impact, Grade C/D evidence	Watch for better evidence

13.2 Political Feasibility Notes

While OPG does not filter by political feasibility, it provides context:

- **Organized opposition**: Industries or groups likely to lobby against
- **Public opinion**: Polling data on similar policies where available
- **Adjacent jurisdictions**: Whether neighbors have adopted (diffusion effects)
- **Historical attempts**: Previous failed attempts and why

13.3 Sequencing Guidance

Some policies are easier to adopt after others:

1. **Quick wins first**: Build political capital with easy, high-impact changes
2. **Complementary bundles**: Some policies work better together
3. **Threshold effects**: Some benefits only appear after critical mass of policies

14 Multi-Unit Reporting

14.1 The Problem with Abstract Scores

Composite scores (like 0-1 PIS values) obscure interpretability. Policymakers and citizens understand concrete outcomes - lives saved, dollars saved, percentage point reductions - not abstract indices. The composite PIS should be a **fallback** when direct interpretation is difficult, not the primary output.

14.2 Reporting at Multiple Levels

Level	Units	Use Case	Example
1. Natural	Domain-specific	Interpretation within domain	"-8.2 pp smoking rate"
2. Monetized	$ equivalent	Cross-domain comparison	"+$2.4B/year health savings"
3. Health	QALYs/DALYs	Health-weighted comparison	"+180K QALYs/year"
4. Composite	0-1 score	Ranking when monetization uncertain	"PIS = 0.85"

Principle: Always report natural units first, then provide monetized equivalents for cross-domain comparison. The composite score is the last resort when monetization is highly uncertain.

14.3 Conversion Factors

Conversion	Value	Source	Notes
Value of Statistical Life (VSL)	~$10M	EPA, DOT	US regulatory standard
Value per QALY	$50K-$150K	ICER, WHO	Context-dependent
QALY → $	$100K/QALY	Mid-range estimate	For cross-domain
Life-year → QALY	~0.8-1.0	Age/health adjusted	Quality weighting
Disability weight	0-1 scale	GBD study	DALY calculation

14.4 Worked Example: Multi-Unit Output for Tobacco Tax

Policy: State tobacco tax increase (+$1/pack)

Unit Level	Value	Interpretation
Natural	-8.2 pp smoking rate	Direct health behavior change
Health impact	+180K QALYs/year	At 23 adopting states scale
Monetized (health)	+$18B/year	At $100K/QALY
Monetized (productivity)	+$4.2B/year	Reduced absenteeism, presenteeism
Total monetized	+$22.2B/year	Health + productivity
Composite (PIS)	1.0	Maximum score (strong evidence)

Interpretation for policymakers: "A $1/pack tobacco tax increase is expected to reduce smoking rates by 8.2 percentage points, generating approximately $22 billion per year in health and productivity benefits. This estimate is based on strong quasi-experimental evidence (Grade A) from 23 US states."

14.5 When to Use Each Level

Situation	Recommended Reporting Level
Single-domain policy (health only)	Natural + Health (QALYs)
Cross-domain comparison	Monetized equivalents
Communication to public	Natural units (most interpretable)
Technical policy analysis	All levels with uncertainty
Highly uncertain effects	Composite score + confidence interval

15 Trial Prioritization

15.1 Value of Information Calculation

The expected value of running a randomized trial on policy p is:

$$\text{VOI}_p = P(\text{adopt}|\text{trial}) \cdot E[\text{benefit}|\text{trial}] - P(\text{adopt}|\text{no trial}) \cdot E[\text{benefit}|\text{no trial}] - \text{Cost}_{\text{trial}}$$

Policies with high VOI have: - **High prior uncertainty**: Current evidence is inconclusive - **High potential impact**: If the policy works, benefits are large - **Low trial cost**: Policy can be randomized in small jurisdictions cheaply - **Decision relevance**: Trial result would change adoption decision

15.2 Natural Experiment Identification

The system automatically identifies potential natural experiments:

Type	Identification Method	Example
Border discontinuity	Adjacent jurisdictions with different policies	Minimum wage differences at state borders
Temporal discontinuity	Abrupt policy change	Court ruling invalidating previous policy
Eligibility threshold	Sharp cutoff for policy application	Income threshold for benefit eligibility
Staggered adoption	Different jurisdictions adopting at different times	ACA Medicaid expansion by state
Lottery	Random assignment (rare)	Charter school lotteries
Court mandate	Externally imposed change	Desegregation orders

Identified natural experiments are stored in `natural_experiments` table for validation.

15.3 Recommended Pilot Jurisdictions

For policies with PIS in the "pilot" range (0.10-0.25), we recommend jurisdictions based on:

1. **Variation feasibility**: Jurisdiction has autonomy to adopt the policy
2. **Data quality**: Good administrative data for outcome measurement
3. **Donor pool**: Similar jurisdictions available as controls
4. **Political openness**: Leadership interested in evidence-based pilots
5. **Scalability**: Results can inform larger-scale adoption

16 Data Sources

16.1 Primary Policy Databases

Database	Coverage	URL	Use Case
V-Dem	202 countries, 1789-present	v-dem.net	Democracy indices, political institutions
Polity V	167 countries, 1800-present	systemicpeace.org	Regime type, political stability
CPDS	36 OECD, 1960-present	cpds-data.org	Economic policy, welfare state
OECD iLibrary	OECD members	oecd-ilibrary.org	Tax, labor, education policy
Congress.gov	US federal, 1973-present	congress.gov	US federal legislation
EUR-Lex	EU, 1951-present	eur-lex.europa.eu	EU legislation and regulations

16.2 Primary Outcome Databases

Database	Coverage	URL	Use Case
World Bank WDI	217 countries, 1960-present	data.worldbank.org	GDP, poverty, education, health
Our World in Data	Global, varies	ourworldindata.org	Curated outcome metrics
WHO GHO	Global	who.int/data/gho	Health outcomes
Penn World Tables	183 countries, 1950-present	ggdc.net/pwt	GDP, productivity, prices
SIPRI	Global, 1949-present	sipri.org	Military spending
IMF	190 countries	imf.org/data	Fiscal, monetary indicators

16.3 Subnational Data

Country	Source	Coverage
United States	Census Bureau, BLS, state agencies	50 states + territories
European Union	Eurostat regional database	~300 NUTS-2 regions
India	CMIE, NSS, state data portals	28 states + territories
China	National Bureau of Statistics	31 provinces
Brazil	IBGE	27 states

16.4 Jurisdiction Policy Inventory Sources

Level	Source	Coverage
US States	NCSL State Legislation Database	All 50 states, continuous updates
US States	State government websites	Primary verification
US Cities	Municode, American Legal Publishing	Major cities
Countries	OECD Government at a Glance	OECD members
Countries	World Bank Doing Business (archived)	190 economies
EU	EUR-Lex	All member states

17 Limitations

17.1 Oracle Capture Risk

As with BIS, the measurement process itself can be captured:

1. **Outcome measurement**: Agencies reporting outcomes have incentives to manipulate
2. **Policy implementation dates**: Recording when policies "really" took effect is subjective
3. **Confounder selection**: Which confounders to control affects estimates

Mitigation: Multiple independent data sources, pre-registered analysis protocols, adversarial audits.

17.2 Confounding Severity

Policy effects face more confounding than drug trials:

Confounder Type	Example	Mitigation
Economic cycles	Recession coincides with policy	Control for GDP growth, unemployment
Secular trends	Improving health over time	Include time trends, compare to controls

Confounder Type	Example	Mitigation
Selection	Jurisdictions adopting policies differ	Matching, synthetic control
Spillovers	Neighboring policies affect outcomes	Spatial controls, SUTVA violations noted
Reverse causality	Outcomes drive policy adoption	Instruments, timing-based identification

17.3 Heterogeneous Effects

Policy effects vary by: - Jurisdiction characteristics (income, institutions, culture) - Implementation fidelity - Complementary policies - Time period

High heterogeneity ($I^2 > 75\%$) suggests context-dependence rather than universal effects.

17.4 Jurisdiction-Specific Caveats

Caveat	Description	Mitigation
Data completeness	Policy inventory may be incomplete	Flag data quality; recommend verification
Context transfer	Effect in State A may not transfer to State B	Adjust for observable differences; widen CIs
Implementation variation	Same policy, different enforcement	Track implementation quality where possible
Interaction effects	Effect depends on other policies in place	Model policy bundles, not just single policies

17.5 Time-Varying Effects

- **Short-run vs. long-run**: Immediate effects may differ from sustained effects
- **Policy drift**: Implementation changes over time (amendment_notes tracking)
- **Adaptation**: Jurisdictions and individuals adapt to policies

The event study design explicitly models dynamic effects; we report both immediate and sustained impact estimates.

17.6 Publication Bias

The policy evaluation literature suffers from systematic publication bias:

1. **Null effects underreported**: Studies finding "no significant effect" are less likely to be published
2. **Positive framing**: Researchers may frame results to emphasize statistically significant findings
3. **File drawer problem**: Failed replications rarely published
4. **Jurisdiction selection**: Jurisdictions with cleaner natural experiments are overrepresented

Mitigation strategies:

- Weight by inverse probability of publication (using funnel plot asymmetry tests)
- Require pre-registration of analysis protocols before data access
- Include unpublished working papers and government reports
- Apply trim-and-fill or PET-PEESE corrections for funnel plot asymmetry
- Report null findings prominently in the database

17.7 Epistemic Limitations

OPG provides *evidence-weighted recommendations*, not causal proof:

What OPG Can Do	What OPG Cannot Do
Rank policies by strength of quasi-experimental evidence	Prove any policy causes an outcome
Generate jurisdiction-specific recommendations	Guarantee effects transfer to new contexts
Identify promising candidates for randomized pilots	Replace randomized policy experiments
Quantify uncertainty and heterogeneity	Eliminate unmeasured confounding
Flag potential harms with moderate confidence	Guarantee a policy is safe
Transfer evidence across similar jurisdictions	Account for all local factors

Important: The quasi-experimental methods used provide evidence *consistent with* causation under assumptions that are often untestable. Synthetic control assumes the donor pool adequately represents the counterfactual; difference-in-differences assumes parallel trends would have continued; regression discontinuity assumes no manipulation around the threshold. These assumptions cannot be verified from data alone.

18 Validation Framework

18.1 The Critical Question

The ultimate test of OPG validity: **Do jurisdictions that adopt high-priority OPG recommendations see better outcomes than those that don't?**

Until this validation is performed, OPG should be treated as a theoretically-motivated heuristic for prioritization, not a validated predictive tool.

18.2 Proposed Validation Study

Design: Retrospective comparison of OPG predictions against subsequent policy outcomes.

Method:

1. Compute OPG recommendations for all jurisdictions using only data available before a cutoff date (e.g., 2015)
2. Identify jurisdictions that adopted high-priority recommendations vs. those that didn't after the cutoff
3. Compare actual outcome changes 2015-2025 in adopting vs. non-adopting jurisdictions

4. Assess whether high-priority recommendations produced larger improvements

Success Metrics:

Metric	Definition	Target
Discrimination (AUC)	Does adopting recommendations predict "welfare improved"?	AUC > 0.65
Calibration	Correlation between predicted effect and actual effect	r > 0.4
Prioritization value	High-priority validation rate vs. low-priority rate	Ratio > 2:1
False positive rate	High-priority recommendations that harmed welfare	< 15%

Expected Outcomes:

- If high-priority recommendations show validation rate of 50%+ and low-priority show rate < 25%, the system has practical utility
- If no discrimination observed, the methodology needs recalibration or fundamental revision

18.3 Known Limitations Requiring Validation

1. **Context adjustment accuracy**: Do jurisdiction-specific adjustments improve prediction?
2. **Blocking factor impact**: Are recommendations with blocking factors less likely to be adopted?
3. **Evidence grade thresholds**: Are the A-F grade cutoffs appropriately calibrated?
4. **Heterogeneity interpretation**: Does high I^2 actually indicate context-dependence vs. measurement noise?

18.4 Continuous Improvement via Adoption Feedback

OPG improves through a learning loop:

1. **OPG generates recommendation** with expected effect ± uncertainty
2. **Jurisdiction adopts** policy at recommended level
3. **Jurisdiction tracks** primary metric per tracking guidance
4. **Jurisdiction reports** outcomes to OPG feedback system
5. **OPG incorporates** new data point into meta-analysis
6. **Future recommendations** reflect updated evidence

This transforms OPG from a static evidence aggregator into a self-improving system where every adoption strengthens the evidence base. The tracking guidance included with each recommendation standardizes what data jurisdictions should collect and report.

19 Future Directions

19.1 Methodological Improvements

1. **Causal discovery algorithms**: Implement PC algorithm, FCI, or GES for policy interaction structure learning

2. **Propensity score integration**: Covariate adjustment for measured jurisdiction characteristics
3. **Bayesian hierarchical models**: More principled cross-jurisdiction pooling with uncertainty quantification
4. **Machine learning for heterogeneity**: Use BART, causal forests to identify which jurisdiction characteristics moderate effects[21]
5. **Text-as-data**: Extract policy features from legislative text for similarity-based evidence transfer
6. **Dynamic treatment regimes**: Model optimal policy sequences, not just single policies

19.2 Validation Priorities

1. **Retrospective validation study** (highest priority): Test OPG predictions against subsequent outcomes
2. **Prospective prediction pre-registration**: Publicly commit to recommendations before policy adoption decisions
3. **Domain expert review**: Have policy experts assess face validity of rankings
4. **Cross-validation**: Hold out jurisdictions, predict their outcomes from others

19.3 Data Infrastructure

1. **Automated policy tracking**: NLP pipeline to detect policy changes from legislative databases
2. **Outcome harmonization**: Standardized outcome definitions across jurisdictions
3. **API access**: Enable researchers to query OPG data programmatically
4. **Version control**: Track how recommendations change as new data arrives

19.4 Integration with Decision-Making

1. **Policy dashboard**: Real-time recommendations for policymakers
2. **Uncertainty communication**: Visualizations that convey confidence appropriately
3. **Scenario modeling**: "What if" analysis for proposed policies based on similar historical policies
4. **Feedback mechanisms**: Track whether recommendations were actually adopted and outcomes realized

20 Conclusion

The Optimal Policy Generator provides a systematic framework for translating policy-outcome evidence into jurisdiction-specific recommendations. By comparing each jurisdiction's current policy inventory to the evidence-optimal set, OPG produces actionable recommendations in four categories (enact/replace/repeal/maintain) ranked by expected welfare impact.

The OPG complements the Optimal Budget Generator (OBG/BIS) in the Optimocracy framework:

- **OBG/BIS** answers: "How much should jurisdiction X spend on each program?"
- **OPG** answers: "What policies should jurisdiction X adopt or repeal?"

Together, they enable evidence-based governance that optimizes both resource allocation and regulatory design, tailored to each jurisdiction's specific context.

Acknowledgments

[To be added: acknowledgments for seminar participants, reviewers, and colleagues who provided feedback.]

21 References

1. Buchanan, J. M. & Tullock, G. *The Calculus of Consent: Logical Foundations of Constitutional Democracy*. (University of Michigan Press, Ann Arbor, 1962).
2. Olson, M. *The Logic of Collective Action: Public Goods and the Theory of Groups*. (Harvard University Press, Cambridge, MA, 1965).
3. Stigler, G. J. The theory of economic regulation. *Stigler* **2**, 3–21 (1971)
   As a rule, regulation is acquired by the industry and is designed and operated primarily for its benefit.
4. Becker, G. S. A theory of competition among pressure groups for political influence. *Becker* **98**, 371–400 (1983)
   Political equilibrium depends on the efficiency of each group in producing pressure, the effect of additional pressure on their influence, the number of persons in different groups, and the deadweight cost of taxes and subsidies.
5. Cartwright, N. & Hardie, J. *Evidence-Based Policy: A Practical Guide to Doing It Better*. (Cartwright, Oxford, 2012).
   The key to evidence-based policy is understanding that evidence of effectiveness elsewhere is not evidence of effectiveness here without support for the claim that the causal mechanism will operate in the new setting.
6. Haskins, R. & Margolis, G. *Show Me the Evidence: Obama's Fight for Rigor and Results in Social Policy*. (Haskins, Washington, DC, 2009).
   The federal government spends hundreds of billions of dollars annually on social programs with little rigorous evidence of effectiveness.
7. Hahn, R. W. Regulatory reform: What do the government's numbers tell us? in 208–253 (2000).
   A review of 48 regulatory impact analyses finds substantial variation in methodology and quality, with many failing to provide adequate justification for regulatory choices.
8. Clearinghouse, W. W. What works clearinghouse standards handbook (version 4.0). (2017)
   The WWC reviews existing research on different programs, products, practices, and policies in education to provide educators with the information they need to make evidence-based decisions.
9. Higgins, J. P. T. & Green, S. *Cochrane Handbook for Systematic Reviews of Interventions*. (Higgins, 2011).
   Heterogeneity in systematic reviews refers to variability among studies. I^2 describes the percentage of variability in effect estimates that is due to heterogeneity rather than sampling error.
10. Petticrew, M. & Roberts, H. *Systematic Reviews in the Social Sciences: A Practical Guide*. (Petticrew, Malden, MA, 2006).
    Systematic reviews can help policymakers by providing a rigorous and transparent method for synthesizing research evidence on the effectiveness of social interventions.

11. Sunstein, C. R. *The Cost-Benefit State: The Future of Regulatory Protection.* (Sunstein, Chicago, 2002).
Cost-benefit analysis, properly understood, is not only a useful tool but also an indispensable safeguard against both excessive and insufficient regulation.

12. Viscusi, W. K. *Pricing Lives: Guideposts for a Safer Society.* (Viscusi, Princeton, NJ, 2018).
The value of a statistical life provides a consistent metric for evaluating the benefits of risk reduction policies across domains.

13. Angrist, J. D. & Pischke, J.-S. The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Angrist & Pischke* **24**, 3–30 (2010)
The primary engine driving improvement has been a focus on the quality of empirical research. designs. Additional sources: https://www.aeaweb.org/articles?id=10.1257/jep.24.2.3

14. Imbens, G. W. & Rubin, D. B. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* (Imbens, Cambridge, 2015).
The potential outcomes framework provides a rigorous foundation for defining causal effects and understanding the assumptions required for their identification.

15. Abadie, A., Diamond, A. & Hainmueller, J. Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program. *Journal of the American Statistical Association* **105**, 493–505 (2010)
The synthetic control method provides a systematic way to choose comparison units in comparative case studies. A combination of comparison units often provides a better comparison for the unit affected by the policy intervention than any single comparison unit alone.

16. Abadie, A., Diamond, A. & Hainmueller, J. Comparative politics and the synthetic control method. *Abadie* **59**, 495–510 (2015)
The synthetic control method provides a systematic way to construct comparison units in comparative case studies, making explicit the weights assigned to each unit.

17. Abadie, A. Using synthetic controls: Feasibility, data requirements, and methodological aspects. *Abadie* **59**, 391–425 (2021)
Synthetic control methods have become one of the most widely used tools for evaluating the effects of policy interventions in comparative case studies.

18. Callaway, B. & Sant'Anna, P. H. C. Difference-in-differences with multiple time periods. *Callaway* **225**, 200–230 (2021)
When treatment timing varies across units, standard two-way fixed effects estimators can be severely biased. We propose alternative estimators that are robust to treatment effect heterogeneity.

19. Hill, A. B. The environment and disease: Association or causation? *Hill* **58**, 295–300 (1965)
None of my nine viewpoints can bring indisputable evidence for or against the cause-and-effect. hypothesis and none can be required as a sine qua non. What they can do, with greater or less strength, is to help us to make up our minds on the fundamental question—is there any other way of explaining the set of facts before us, is there any other answer equally, or more, likely than cause and effect? Additional sources: https://pmc.ncbi.nlm.nih.gov/articles/PMC1898525/

20. Oster, E. Unobservable selection and coefficient stability: Theory and evidence. *Oster* **37**, 187–204 (2019)
A common approach to evaluating robustness to omitted variable bias is to observe coefficient. movements after inclusion of controls. This is informative only if selection on observables is informative about selection on unobservables. Additional sources: https://www.tandfonline.com/doi/abs/10.1080/07350015.2016.1227711

21. Athey, S. & Imbens, G. W. The state of applied econometrics: Causality and policy evaluation. *Athey* **31**, 3–32 (2017)
Machine learning methods offer new tools for causal inference, including methods for estimating heterogeneous treatment effects and for optimal policy learning.

22 Appendix A: Worked Example - Texas Policy Recommendations

22.1 Overview

This worked example demonstrates the complete OPG output for a specific jurisdiction: Texas. It shows how generic policy evidence is translated into jurisdiction-specific recommendations.

22.2 Texas Policy Inventory (Sample)

Policy Type	Has Policy?	Current Level	Evidence-Optimal	Recommendation
Primary seat belt enforcement	No	N/A	Yes	**ENACT**
Motorcycle helmet requirement	No (partial)	Only under 21	All ages	**ENACT**
Tobacco tax	Yes	$1.41/pack	~$2.50/pack	**REPLACE**
Speed limit (rural interstate)	Yes	85 mph	70 mph	**REPLACE**
DUI threshold	Yes	0.08 BAC	0.08 BAC	**MAINTAIN**
Graduated driver licensing	Yes	3-stage system	3-stage system	**MAINTAIN**

22.3 Step 1: Calculate Policy Impact Scores

Example: Primary Seat Belt Law

From meta-analysis of 47 US states (2000-2020):

Parameter	Value
Average effect	-1.8 deaths per 100K
Standard error	0.4
I^2 heterogeneity	28%
Evidence grade	A

Bradford Hill Criteria Scores:

Criterion	Score	Rationale
Strength	0.75	Moderate standardized effect
Consistency	0.82	$I^2 = 28\%$, consistent across states

Criterion	Score	Rationale
Temporality	0.95	Clear temporal ordering
Gradient	0.65	Binary policy, limited dose-response
Plausibility	0.90	Clear mechanism (increased compliance)
Experiment	0.85	Multiple synthetic control studies
...	...	...

CCS = 0.81

PIS = $0.75 \times 0.81 \times 0.85 = 0.52 \rightarrow$ Grade A

22.4 Step 2: Apply Context Adjustment for Texas

Factor	Texas Value	Adjustment
Current seat belt use	91.5%	Effect may be smaller (already high)
Rural driving proportion	High	Effect may be larger (more severe crashes)
Population	29.5M	Scale up total impact

Adjusted expected effect: -1.5 deaths per 100K (slightly smaller due to already-high compliance)

22.5 Step 3: Generate Recommendations

OPG Recommendations for Texas

ENACT (New Policies to Adopt)

1. **Primary Seat Belt Enforcement Law** (Current: None - secondary enforcement only)
   - **Expected effect**: -1.5 traffic deaths per 100K population
   - **Monetized impact**: +$1.2B/year (at $10M VSL × 120 lives saved)
   - **Evidence grade**: A
   - **Priority**: High
   - **Recommended level**: State (maximize data; federal preemption prevents city-level)
   - **Blocking factors**: None identified
   - **Similar jurisdictions**: Florida adopted 2009, saw -1.7 deaths/100K
   - **Tracking**: Traffic deaths/100K (FARS), annual, vs. 2020-2024 baseline
2. **Universal Motorcycle Helmet Requirement** (Current: None for adults - partial coverage under 21 only)
   - **Expected effect**: -1.0 traffic deaths per 100K
   - **Monetized impact**: +$680M/year
   - **Evidence grade**: A
   - **Priority**: Medium
   - **Recommended level**: State (maximize data)
   - **Blocking factors**: Political (strong rider opposition)
   - **Similar jurisdictions**: California (all ages since 1992)

- **Tracking**: Motorcycle fatalities/100K (FARS), annual, vs. 2020-2024 baseline

⚠ REPLACE (Policies to Modify)

3. **Tobacco Tax: $1.41 → $2.50/pack**
   - **Current level**: $1.41/pack (below national median of $1.91)
   - **Recommended level**: $2.50/pack (evidence-optimal range)
   - **Expected effect**: -5.2 pp smoking rate (Texas-adjusted)
   - **Monetized impact**: +$4.8B/year health savings
   - **Evidence grade**: A
   - **Priority**: High
   - **Recommended level**: State (maximize data; city-level preempted)
   - **Blocking factors**: Political (anti-tax sentiment)
   - **Tracking**: Smoking rate (BRFSS), annual, vs. 2020-2024 baseline
4. **Maximum Speed Limit: 85 mph → 70 mph**
   - **Current level**: 85 mph (highest in US)
   - **Recommended level**: 70 mph (evidence-optimal)
   - **Expected effect**: -0.8 deaths/100K
   - **Monetized impact**: +$350M/year
   - **Evidence grade**: B
   - **Priority**: Low (political feasibility concern)
   - **Blocking factors**: Political (driver opposition)
   - **Tracking**: Highway fatalities/100K (FARS), annual, vs. 2020-2024 baseline

🔥 REPEAL (Policies to Remove)

No high-priority repeal recommendations for Texas at this time.
(Example format: If Texas had a policy shown to cause net harm, it would appear here with expected welfare gain from removal.)

ℹ MAINTAIN (No Change Needed)

5. **DUI Threshold at 0.08 BAC**
   - **Current level**: 0.08 BAC (national standard)
   - **Evidence**: Aligned with evidence-optimal level
   - **Status**: Continue current policy
6. **Graduated Driver Licensing Program**
   - **Current level**: Three-stage system with night/passenger restrictions
   - **Evidence**: Consistent with best practices
   - **Status**: Continue current policy

22.6 Step 4: Summary Dashboard

Total Expected Welfare Gain by Recommendation Type

Type	Recommendation	Monetized Annual Impact	Feasibility
ENACT	Primary seat belt law	+$1.2B	High
ENACT	Universal helmet law	+$680M	Medium
REPLACE	Tobacco tax: $1.41 → $2.50	+$4.8B	Medium
REPLACE	Speed limit: 85 → 70 mph	+$350M	Low
MAINTAIN	DUI threshold, GDL	-	N/A
	Total from changes	**+$7.0B/year**	

Note: MAINTAIN items confirm evidence alignment and require no action. REPEAL section empty for Texas - no harmful policies identified with strong evidence.

22.7 Interpretation

This example demonstrates how OPG transforms generic evidence ("seat belt laws reduce deaths by 1.8/100K on average") into actionable, jurisdiction-specific recommendations ("Texas should ENACT primary seat belt enforcement; expected effect -1.5 deaths/100K, +$1.2B/year").

The four recommendation types provide clarity:

- **ENACT**: Policies Texas doesn't have (seat belt, helmet)
- **REPLACE**: Policies Texas has but at wrong levels (tobacco tax, speed limit)
- **REPEAL**: Harmful policies to remove (none identified)
- **MAINTAIN**: Policies already evidence-optimal (DUI threshold, GDL)

The recommendations: 1. Account for Texas's current policy inventory 2. Adjust for Texas-specific context (demographics, existing policies) 3. Flag blocking factors (political, constitutional) 4. Provide concrete comparisons to similar jurisdictions 5. Monetize benefits for cross-domain comparison

23 Appendix B: OPG Analysis Workflow

23.1 Complete OPG Pipeline

```
                OPTIMAL POLICY GENERATOR WORKFLOW


Phase 1: DATA COLLECTION

1. Policy database ingestion
     Parse legislative text (LLM extraction)
     Record implementation dates by jurisdiction
     Classify policy type and category
     Compute policy embeddings for similarity

2. Jurisdiction policy inventory
     Pull current policy status for each jurisdiction
     Record policy strength (for continuous policies)
```

```
      Flag data quality and gaps
      Identify last verification date

3. Outcome data collection
      Pull from primary sources (World Bank, WHO, etc.)
      Harmonize units and definitions
      Identify missing data patterns
      Flag measurement quality issues

4. Confounder data collection
      Economic indicators (GDP, unemployment)
      Demographic variables (age structure, education)
      Political variables (regime type, election cycles)
      Geographic variables (neighbors' policies)

Phase 2: EVIDENCE ANALYSIS (Quasi-Experimental)

5. Policy-outcome pair identification
      Match policies to plausible outcome categories
      Filter by minimum data requirements
      Identify applicable quasi-experimental methods

6. Method selection
      Synthetic control: single treated, good donors
      Difference-in-differences: multiple treated, parallel trends
      Regression discontinuity: sharp threshold exists
      Event study: need dynamic effects
      Interrupted time series: fallback

7. Effect estimation
      Run primary analysis
      Calculate standard errors (clustered)
      Compute confidence intervals
      Store jurisdiction-level results

8. Robustness checks
      In-time placebo tests
      In-space placebo tests
      Leave-one-out sensitivity
      Covariate adjustment sensitivity

Phase 3: AGGREGATION & PIS CALCULATION

9. Meta-analysis
      Pool jurisdiction estimates (random effects)
      Calculate I², ², Q statistics
      Test for publication bias (funnel plot)
      Apply trim-and-fill if needed
```

```
10. Bradford Hill scoring
       Score each criterion (0-1)
       Apply criterion weights
       Compute CCS (causal confidence score)
       Document evidence for each criterion

11. PIS calculation
       Standardize effect estimate
       Calculate quality adjustment
       Compute final PIS
       Assign evidence grade (A-F)

Phase 4: RECOMMENDATION GENERATION

12. Policy gap analysis (per jurisdiction)
       Compare current inventory to evidence-optimal
       Calculate gap magnitude
       Identify gap type (missing, harmful, suboptimal)
       Flag blocking factors

13. Context adjustment
       Adjust effect estimates for jurisdiction characteristics
       Widen confidence intervals for context uncertainty
       Identify similar jurisdictions for comparison
       Note implementation considerations

14. Priority scoring
       Rank by |Gap| × Evidence Grade × Impact
       Assign to priority tiers (Quick Win, Major Reform, etc.)
       Generate enact/replace/repeal/maintain lists
       Calculate total expected welfare gain

Phase 5: OUTPUT GENERATION

15. Recommendation dashboard
       Enact list (new policies to adopt)
       Replace list (existing policies to modify: current → optimal)
       Repeal list (harmful policies to remove)
       Maintain list (policies aligned with evidence)
       Jurisdictional level and tracking guidance for each

16. Multi-unit reporting
       Natural units (domain-specific)
       Monetized equivalents (cross-domain)
       Health units (QALYs/DALYs)
       Composite PIS (when uncertain)
```

```
17. Documentation
      Generate jurisdiction-specific reports
      Create methodology audit trail
      Version control all recommendations
      Publish to API/dashboard
```

23.2 Minimum Data Requirements Checklist

Before generating recommendations, verify:

- ☐ 4 pre-treatment periods in evidence base
- ☐ 2 post-treatment periods in evidence base
- ☐ 20 total outcome observations
- ☐ 5 control jurisdictions (for DiD-based evidence)
- ☐ Policy implementation dates documented
- ☐ Outcome variable has known valence
- ☐ Jurisdiction policy inventory verified within 2 years
- ☐ Data quality score 0.5 for target jurisdiction

24 Appendix C: Glossary

24.1 Core Concepts

- **Optimal Policy Generator (OPG)**: System for producing jurisdiction-specific policy recommendations based on quasi-experimental evidence. Outputs four recommendation types (enact/replace/repeal/maintain) ranked by expected welfare impact, with recommended jurisdictional level (subsidiarity) and tracking guidance for continuous improvement.
- **Policy Impact Score (PIS)**: Intermediate metric quantifying the strength of evidence that a policy affects an outcome. Integrates effect size, causal confidence (Bradford Hill criteria), and analysis quality. Ranges from 0 to 1; higher indicates stronger evidence.
- **Policy Gap**: Difference between a jurisdiction's current policy status and the evidence-optimal policy. Gaps can be: missing (lacks beneficial policy), harmful (has detrimental policy), suboptimal (continuous policy at wrong level).
- **Causal Confidence Score (CCS)**: Weighted average of Bradford Hill criteria scores. Quantifies confidence that observed association reflects true causation rather than confounding, reverse causation, or chance.
- **Evidence Grade**: Letter grade (A-F) summarizing evidence quality. A = strong evidence from multiple high-quality quasi-experiments; F = insufficient or conflicting evidence.
- **Jurisdiction**: Geographic or administrative unit where policies are implemented (country, state, province, city, municipality). OPG operates at any level.
- **Blocking Factor**: Constraint that may impede policy adoption: constitutional (requires amendment), federal preemption (superseded by higher law), political (strong opposition), implementation cost (high fixed costs).

24.2 Quasi-Experimental Methods

- **Synthetic Control Method**: Constructs a weighted combination of untreated jurisdictions that matches the treated jurisdiction's pre-treatment outcome trajectory. Post-treatment divergence estimates causal effect.
- **Difference-in-Differences (DiD)**: Compares pre-to-post change in treated units to pre-to-post change in control units. Valid under parallel trends assumption.
- **Regression Discontinuity Design (RDD)**: Exploits sharp threshold determining policy eligibility. Compares outcomes just above vs. just below threshold.
- **Event Study**: Estimates treatment effects at each time period relative to policy adoption. Visualizes pre-trends and effect dynamics.
- **Interrupted Time Series (ITS)**: Estimates level and slope changes in a single unit's outcome trajectory following policy implementation.

24.3 Statistical Concepts

- **I^2 (I-squared)**: Percentage of variance in effect estimates due to true heterogeneity (vs. sampling error). $I^2 > 75\%$ indicates substantial heterogeneity.
- **^{2} (tau-squared)**: Estimated between-jurisdiction variance in true effects. Used in random-effects meta-analysis.
- **E-value**: Minimum strength of association an unmeasured confounder would need with both policy and outcome to explain away the observed effect. Higher = more robust.
- **Parallel Trends**: Assumption that treated and control jurisdictions would have followed similar outcome trajectories absent treatment. Untestable but can be assessed via pre-treatment data.
- **Pre-treatment RMSE**: Root mean squared error between synthetic control and actual treated unit before policy implementation. Lower indicates better fit.

24.4 Bradford Hill Criteria (Policy Context)

- **Strength**: Magnitude of the effect estimate (standardized).
- **Consistency**: Replication across jurisdictions (inverse of I^2).
- **Specificity**: Whether policy affects specific outcomes vs. everything.
- **Temporality**: Policy adoption precedes outcome change.
- **Gradient**: Dose-response relationship (for continuous policies like tax rates).
- **Plausibility**: Economic or behavioral mechanism exists.
- **Coherence**: Consistent with broader economic theory and evidence.
- **Experiment**: Quality of quasi-experimental design used.
- **Analogy**: Similar policies show similar effects.

24.5 Output Concepts

- **Enact List**: New policies the jurisdiction should adopt (policies that don't exist in the jurisdiction but have strong evidence of benefit).

- **Replace List**: Existing policies the jurisdiction should modify (current policy level or approach differs from evidence-optimal; specifies current → recommended change).
- **Repeal List**: Policies the jurisdiction should remove (has evidence of harm, jurisdiction currently has them).
- **Maintain List**: Policies the jurisdiction should keep unchanged (current policy is aligned with evidence).
- **Priority Tier**: Classification of recommendations by urgency and feasibility: Quick Win (high impact, low barriers), Major Reform (high impact, significant barriers), Long-Term (requires structural change), Monitor (needs more evidence).
- **Subsidiarity Principle**: Recommendation to implement policies at the lowest effective jurisdictional level to maximize experimental data collection and minimize harm from policy failures.
- **Tracking Guidance**: Recommended KPIs and data sources for each recommendation, enabling jurisdictions to report outcomes back to OPG for continuous improvement.
- **Similar Jurisdictions**: Jurisdictions with comparable characteristics that have adopted the recommended policy, used as concrete examples for policymakers.

---

25 Appendix D: Interpreting Effect Sizes

25.1 Standardized Effect Size Benchmarks

$\lvert\hat{\beta}_{\text{std}}\rvert$	Interpretation	Policy Example
< 0.2	Small	Most regulatory changes
0.2 - 0.5	Medium	Typical tax policy effects
0.5 - 0.8	Large	Major reform programs
> 0.8	Very Large	Transformative policies (rare)

25.2 Converting to Interpretable Units

Raw effect estimates should be reported alongside standardized scores:

Outcome	Raw Effect	Standardized	Interpretation
Life expectancy	+0.8 years	+0.32	Medium effect
Unemployment rate	-1.2 pp	-0.45	Medium effect
GDP per capita	+$1,200	+0.18	Small effect
Crime rate	-15 per 100K	-0.61	Large effect

25.3 Confidence Interval Interpretation

CI Width	Interpretation
Narrow (< 0.2 SD)	Precise estimate; high confidence
Moderate (0.2-0.5 SD)	Reasonable precision; moderate confidence
Wide (> 0.5 SD)	Imprecise; low confidence in point estimate

Corresponding Author: Mike P. Sinn, Decentralized Institutes of Health (mike@warondisease.org)

Conflicts of Interest: The author declares no conflicts of interest.

Funding: This work received no external funding.

Data Availability: This specification describes a methodological framework. Policy databases referenced (V-Dem, Polity V, CPDS, World Bank WDI) are publicly available at URLs provided in Data Sources section. A complete replication package including data extraction scripts, analysis code, and recommendation generation algorithms will be deposited in a public repository upon system deployment.

Ethics Statement: This is a methodological specification. No human subjects research was conducted. Policy outcome data used in examples are drawn from published academic studies and government statistical agencies.

Preprint: This working paper has not undergone peer review.