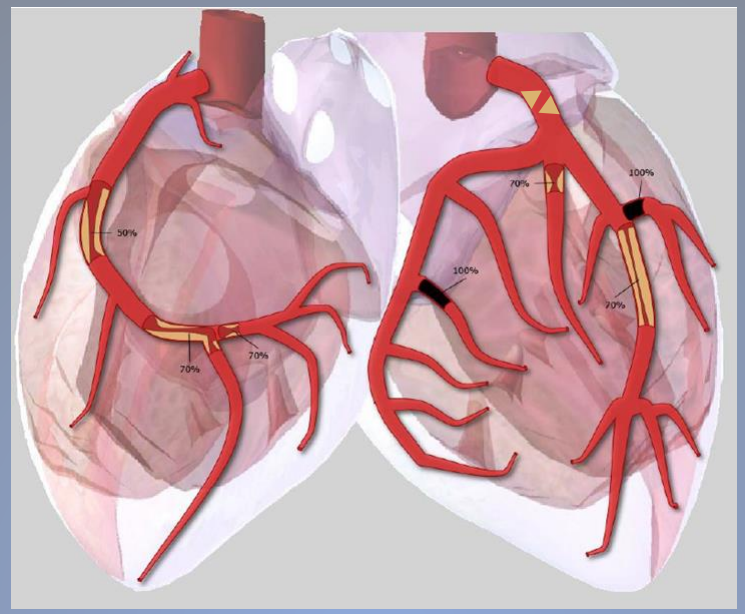
Dealing with Bias in Observational Studies

Porntep Amornritvanich, MD

Bias in observational study

- Systematic error
- Leads to incorrect estimation of treatment effects

- In observational study → treatment is NOT randomly assigned


50%
70%
70%
70%
100%
100%

Type of bias

- **Selection Bias:** Patients receiving treatment differ from those who do not.
- **Confounding Bias:** Other factors influence both treatment assignment and outcomes.
- **Survivor Bias:** Sicker patients might not survive long enough to receive treatment.
- **Information Bias:** Measurement errors in exposure or outcome data.

Limitation of RCT

Comparing to observational study

Methods to Address Bias in Observational Studies

- Common method
  - **Multivariable Risk Adjustment:** Adjust for measured confounders in the model.
  - **Propensity Score (PS) Methods:** Matching, stratification, or inverse probability of treatment weighting (IPTW).
  - **Instrumental Variable (IV) Analysis:** An econometric approach to control for both measured and unmeasured confounders.

PubMed®

propensity score

Search

Advanced Create alert Create RSS User Guide

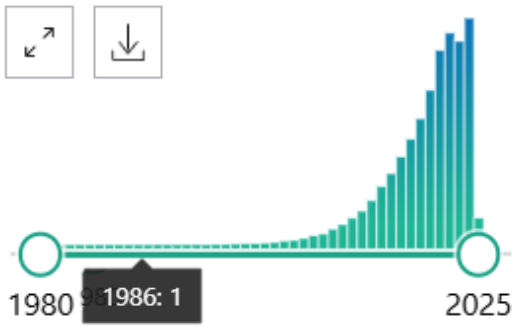
Save Email Send to Sort by: Best match Display options

MY CUSTOM FILTERS

57,340 results

Page 1 of 5,734

RESULTS BY YEAR



PUBLICATION DATE

- 1 year
- 5 years
- 10 years
- Custom Range

1 Cite Share

Propensity Score Matching: A Statistical Method.

Kane LT, Fang T, Galetta MS, Goyal DKC, Nicholson KJ, Kepler CK, Vaccaro AR, Schroeder GD.

Clin Spine Surg. 2020 Apr;33(3):120-122. doi: 10.1097/BSD.0000000000000932.

PMID: 31913173

Propensity score matching (PSM) is a commonly used statistical method in orthopedic surgery research that accomplishes the removal of confounding bias from observational cohorts where the benefit of randomization is not possible. ...PSM is uniquely valuable in its u ...

2 Cite Share

Statistical primer: **propensity score** matching and its alternatives.

Benedetto U, Head SJ, Angelini GD, Blackstone EH.

Eur J Cardiothorac Surg. 2018 Jun 1;53(6):1112-1117. doi: 10.1093/ejcts/ezy167.

PMID: 29684154 Review.

Propensity score (PS) methods offer certain advantages over more traditional regression methods to control for confounding by indication in observational studies. ...

PubMed®

instrumental variable

Search

Advanced Create alert Create RSS User Guide

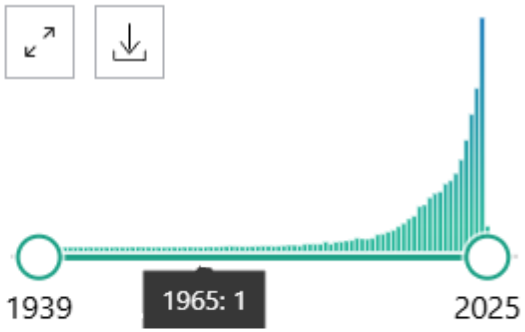
Save Email Send to Sort by: Best match Display options

MY CUSTOM FILTERS

11,536 results

Page 1 of 1,154

RESULTS BY YEAR



PUBLICATION DATE

- 1 year
- 5 years
- 10 years
- Custom Range

1 Cite Share

A review of **instrumental variable** estimators for Mendelian randomization.

Burgess S, Small DS, Thompson SG.

Stat Methods Med Res. 2017 Oct;26(5):2333-2355. doi: 10.1177/0962280215597579. Epub 2015 Aug 17.

PMID: 26282889 **Free PMC article.** Review.

It has gained in popularity over the past decade with the use of genetic variants as **instrumental variables**, known as Mendelian randomization. An **instrumental variable** is associated with the exposure, but not associated with any confounder of the expos ...

2 Cite Share

Meta-analysis and Mendelian randomization: A review.

Bowden J, Holmes MV.

Res Synth Methods. 2019 Dec;10(4):486-496. doi: 10.1002/jrsm.1346. Epub 2019 Apr 23.

PMID: 30861319 **Free PMC article.** Review.

Mendelian randomization (MR) uses genetic variants as **instrumental variables** to infer whether a risk factor causally affects a health outcome. Meta-analysis has been used historically in MR to combine results from separate epidemiological studies, with each study us ...

Propensity score analysis

- A propensity score (PS) is the probability of receiving treatment given observed characteristics.
- It summarizes all covariates into a single score, balancing groups in a way similar to randomization.

$$PS(X)=P(Treatment|X)$$

where X = all observed confounders

Propensity score method

1. Matching
2. Stratification
3. Inverse Probability of Treatment Weighting (IPTW)
4. Regression Adjustment

Propensity score method

1. Matching
2. Stratification
3. Inverse Probability of
4. Regression Adjustme

- Find pairs of treated and untreated patients with similar PS.
- Methods: Nearest-neighbor, caliper, one-to-many matching.
- Example: Matching PCI vs. CABG patients based on age, diabetes, EF.

Propensity score method

1. Matching
2. Stratification
3. Inverse Probability of Tr
4. Regression Adjustment

- Divide patients into quintiles or deciles based on PS.
- Analyze treatment effect within each stratum.

Propensity score method

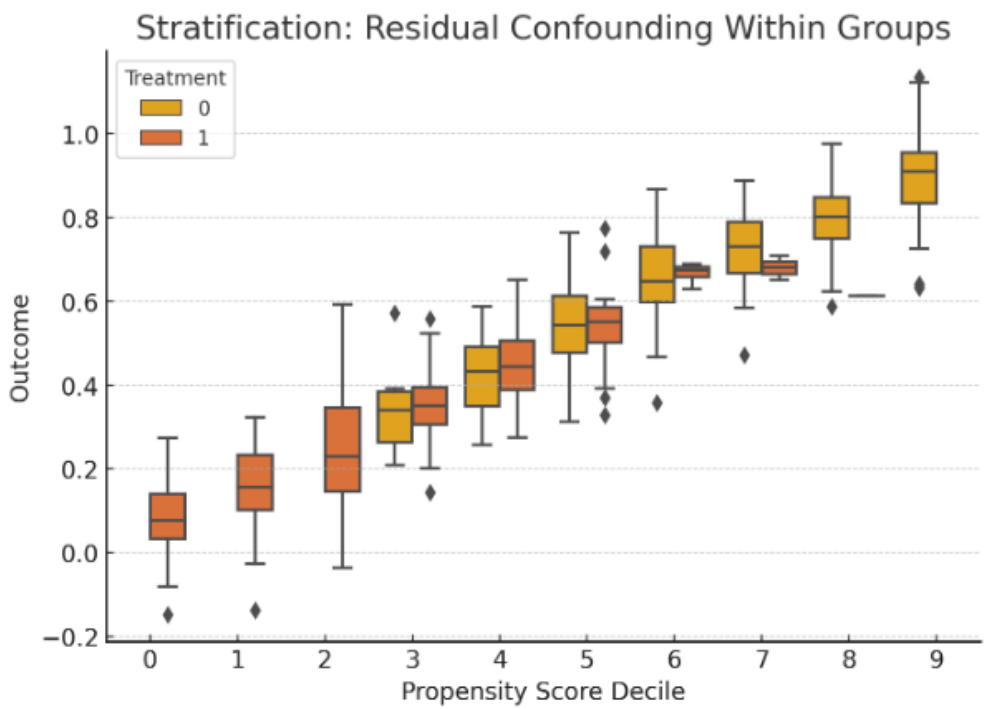
1. Matching
2. Stratification
3. Inverse Probability of Treatment Weighting (IPTW)
4. Regression Adjustment

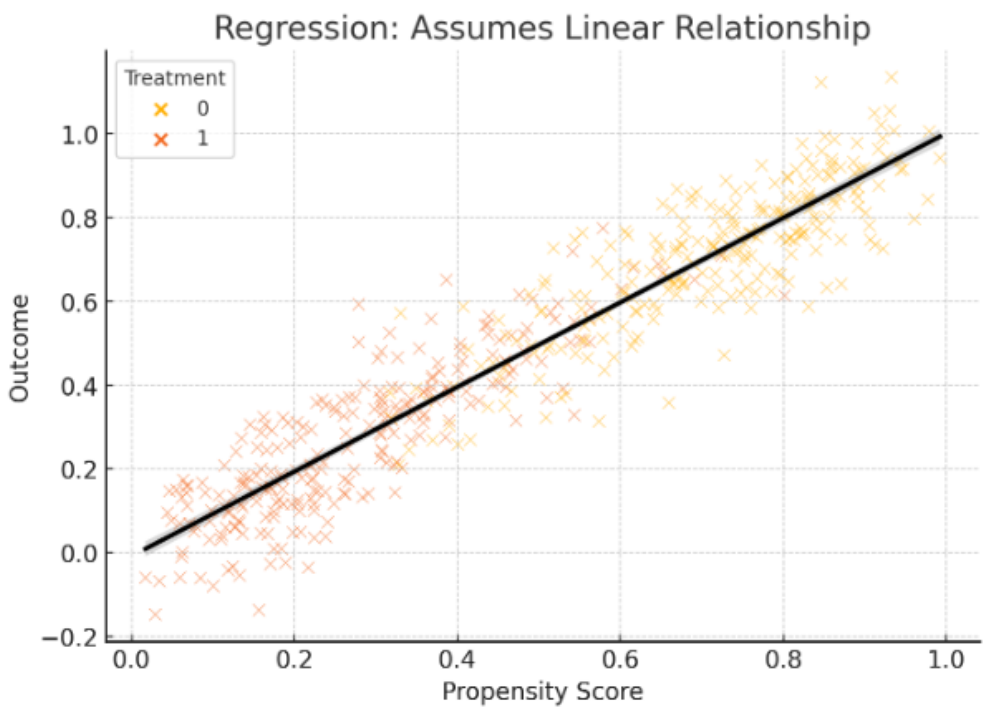
- Patients with low probability of treatment get more weight, balancing the groups.
- Formula:
  - Treated: $w = \frac{1}{PS}$
  - Untreated: $w = \frac{1}{1-PS}$

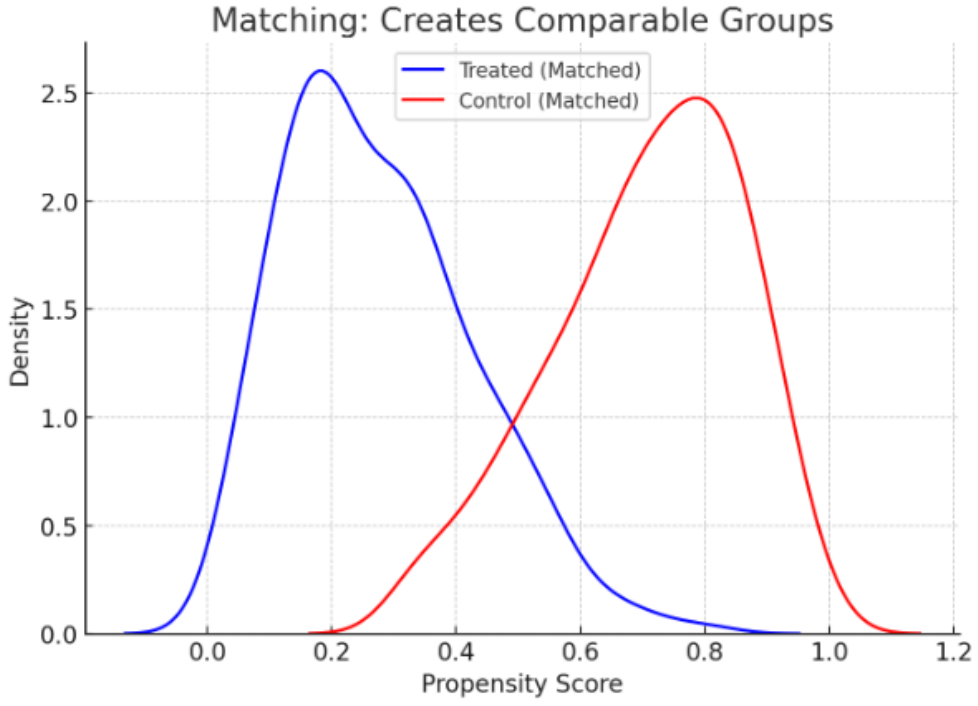
Propensity score method

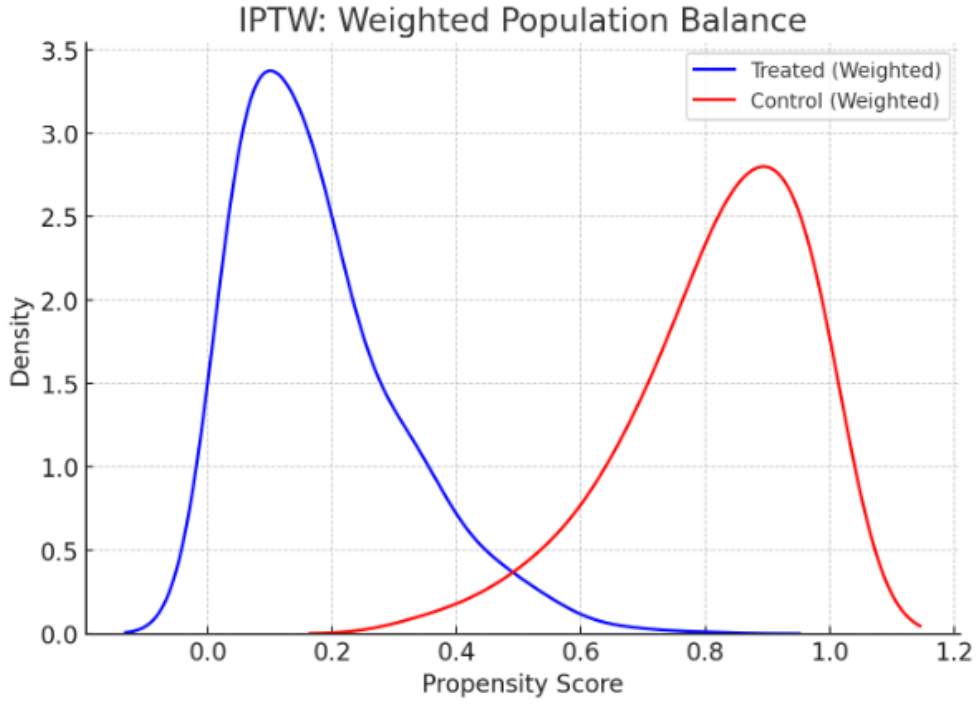
1. Matching
2. Stratification
3. Inverse Probability of Treatment Weighting (IPTW)
4. Regression Adjustment

- Include PS as a covariate in a multivariable model.


Stratification: Residual Confounding Within Groups
Treatment
0
1
Outcome
Propensity Score Decile


Regression: Assumes Linear Relationship
Treatment
0
1
Outcome
Propensity Score


Matching: Creates Comparable Groups
Treated (Matched)
Control (Matched)
Density
Propensity Score


IPTW: Weighted Population Balance
Treated (Weighted)
Control (Weighted)
Density
Propensity Score

Strengths and Limitations of PS

- ☑ **Strengths**
  - Controls for measured confounders.
  - Mimics RCT-like balance in observational data. (emulated RCT; IPW)
  - Matching and IPTW allow for good causal inference.
- ❌ **Limitations**
  - Does NOT account for unmeasured confounders.
  - Requires large sample size for effective matching.
  - Matching may discard a large proportion of patients.
  - IPTW can lead to extreme weights → unstable results.

Propensity scores only work if all important confounders are measured!

Comparison of each method

Method	Handles Confounding?	Emulated RCT?	Best for Causal Inference?
PS Matching	☑ Yes (balances covariates)	☑ Yes (creates matched pairs)	☑ Strong
IPW	☑ Yes (reweights sample)	☑ Yes (creates pseudo-randomization)	☑ Strong
Regression (RA)	⚠ Partially (model-dependent)	❌ No (relies on assumptions)	❌ Weaker
Stratification	⚠ Limited (only categorical confounders)	❌ No (small strata cause issues)	❌ Weaker

From the paper: PS

- The propensity score was computed using logistic regression, where:
  - **Dependent variable:** Receipt of cardiac catheterization.
  - **Independent variables:** 65 patient, hospital, and ZIP code variables.

From the paper: PS matching

- **Matching Method:** 1:1 of patients receiving cardiac catheterization were matched to the closest control (non-catheterized patient).
- **Caliper Width:** ±0.10 of the propensity score.
- **Age Restriction:** Matched within 5 years of age to ensure comparability.
- **Resulting Matches:** 31,193 matched pairs with standardized differences <10%, meaning that key patient characteristics were well balanced.

From the paper: PS risk adjustment

- Patients were grouped into deciles of PS
- Cox proportional hazards models were then used to estimate mortality rates, adjusting for the PS decile.
- Key Assumption:
  - Within each decile, covariates should be balanced between treated and untreated patients.
  - This method removes over 90% of overt bias due to measured covariates

Instrumental Variable (IV) Analysis

- IV Analysis is an econometric method that removes the effect of unmeasured confounders.
- It mimics randomization by using a variable (instrument) that affects treatment but not the outcome directly.

  (Distance, Year, Physician or Hospital preference, Weather, Weekday vs Weekend admission)

  - Example from the paper:

    - Regional cardiac catheterization rate is used as an instrument for receiving PCI.

IV Formula (Two-Stage Least Squares - 2SLS):

1. First stage:

$$Treatment = \alpha + \beta(Instrument) + \epsilon$$

- Predicts treatment based on IV.

2. Second stage:

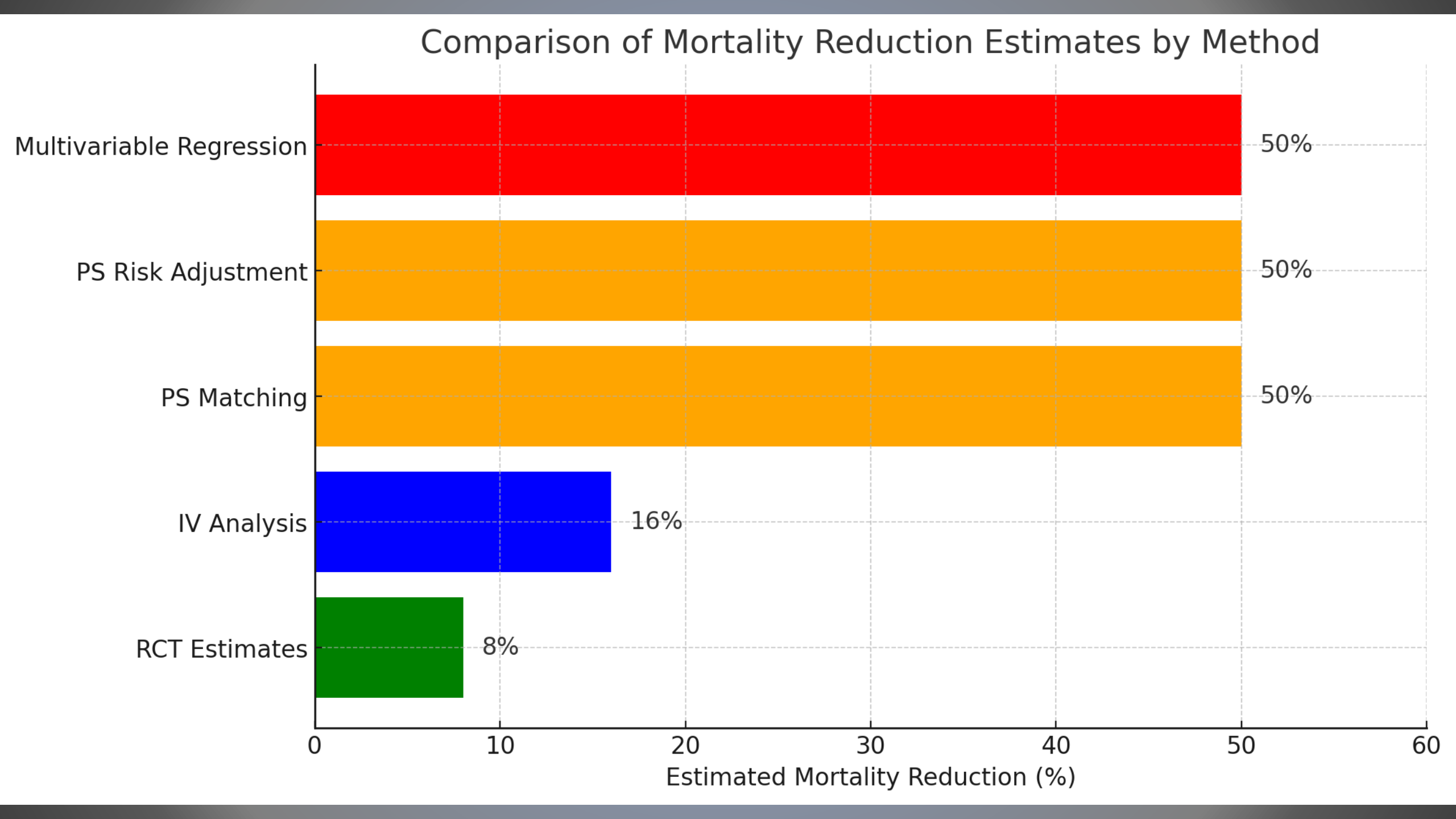
$$Outcome = \gamma + \delta(PredictedTreatment) + \eta$$

- Uses predicted treatment instead of actual treatment.

From the paper: IV analysis

- Instrumental variable used was regional cardiac catheterization rate (i.e., the likelihood of receiving PCI in a given geographic region)

- IV analysis assumes that:
  - Regional variation in treatment affects treatment choice but not survival directly.
  - Patient characteristics are similar across regions.


Comparison of Mortality Reduction Estimates by Method
Multivariable Regression
PS Risk Adjustment
PS Matching
IV Analysis
RCT Estimates
50%
50%
50%
16%
8%
0
10
20
30
40
50
60
Estimated Mortality Reduction (%)

Original Article

Instrumental Variable Analysis to Compare Effectiveness of Stents in the Extremely Elderly

Robert W. Yeh, MD, MSc; Samip Vasaiwala, MD, MSc; Daniel E. Forman, MD; Treacy S. Silbaugh, BSc; Katya Zelevinski, BA; Ann Lovett, RN, MA; Sharon-Lise T. Normand, PhD; Laura Mauri, MD, MSc

Background—Evaluating novel therapies is challenging in the extremely elderly. Instrumental variable methods identify variables associated with treatment allocation to perform adjusted comparisons that may overcome limitations of more traditional approaches.

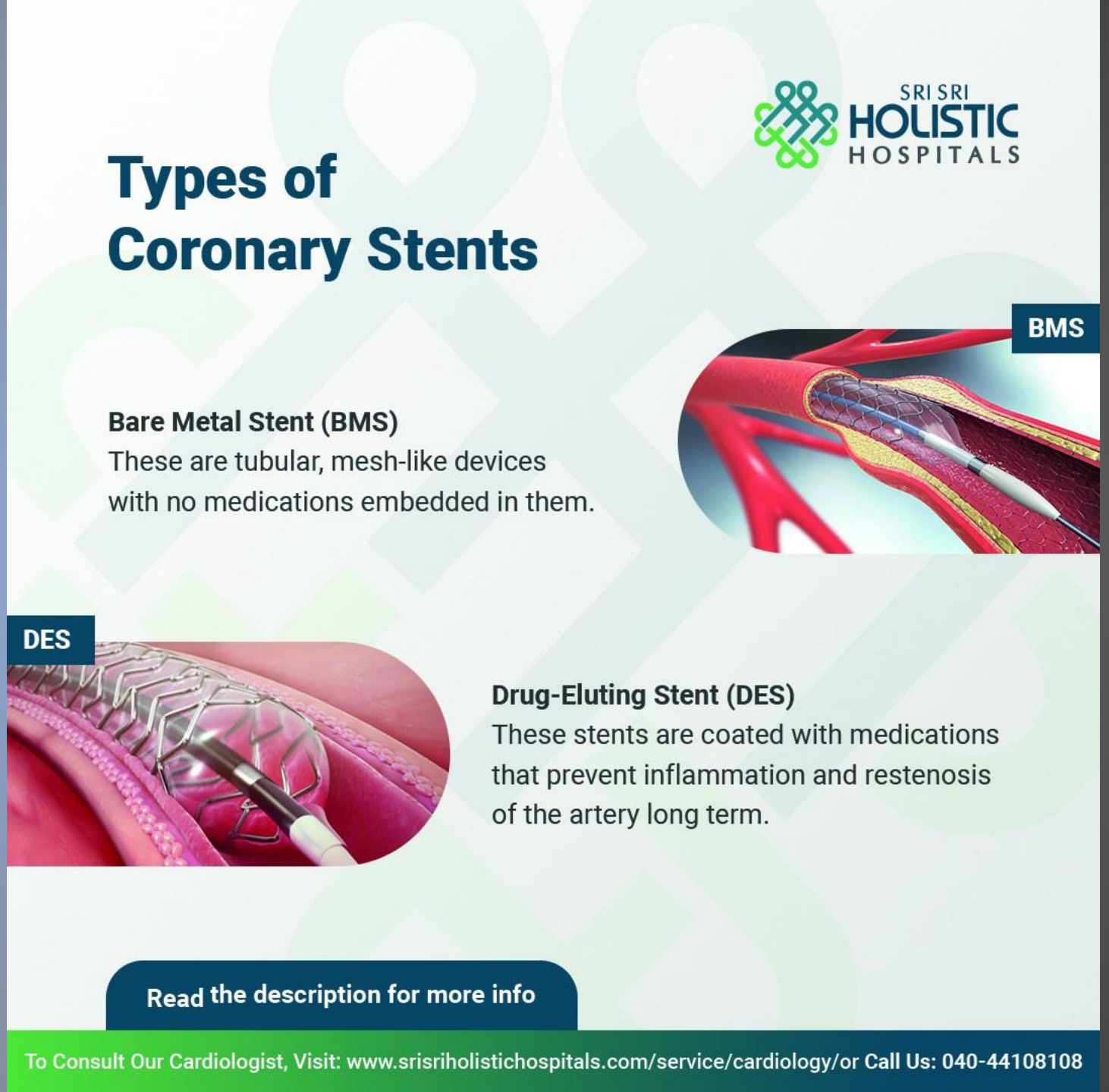
Methods and Results—Among all patients aged ≥85 years undergoing percutaneous coronary intervention in nonfederal hospitals in Massachusetts between 2003 and 2009 (n=2690), we identified quarterly drug-eluting stent (DES) use rates as an instrumental variable. We estimated risk-adjusted differences in outcomes for DES versus bare metal stents using a 2-stage least squares instrumental variable analysis method. Quarterly DES use ranged from 15% to 88%. Unadjusted 1-year mortality rates were 14.5% for DES versus 23.0% for bare metal stents (risk difference, −8.5%; $P<0.001$), an implausible finding compared with randomized trial results. Using instrumental variable analysis, DES were associated with no difference in 1-year mortality (risk difference, −0.8%; $P=0.76$) or bleeding (risk difference, 2.3%; $P=0.33$) and with significant reduction in target vessel revascularization (risk difference, −8.3%; $P<0.0001$).

Conclusions—Using an instrumental variable analysis, DES were associated with similar mortality and bleeding and a significant reduction in target vessel revascularization compared with bare metal stents in the extremely elderly. Variation in use rates may be useful as an instrumental variable to facilitate comparative effectiveness in groups underrepresented in randomized trials.

Circ Cardiovasc Qual Outcomes. 2014;7:118-124.

Result

- Unadjusted results
  - DES had lower mortality than BMS.



IV applied (Two-Stage Least Squares - 2SLS)

1. First Stage: Predicting Treatment

$$Treatment = \alpha + \beta(\text{Quarterly DES Use Rate}) + \epsilon$$

- The probability of receiving DES vs. BMS is modeled based on the overall DES usage rate in that quarter.

2. Second Stage: Estimating the Effect of DES on Outcomes

$$Outcome = \gamma + \delta(\text{Predicted Treatment}) + \eta$$

- Instead of using actual treatment, they used the predicted treatment from the first stage to estimate its effect on mortality, revascularization, and bleeding.

IV analysis result

- DES had no significant difference in mortality or bleeding compared to BMS but significantly reduced target vessel revascularization

Another example

A systematic review of instrumental variable analyses using geographic region as an instrument

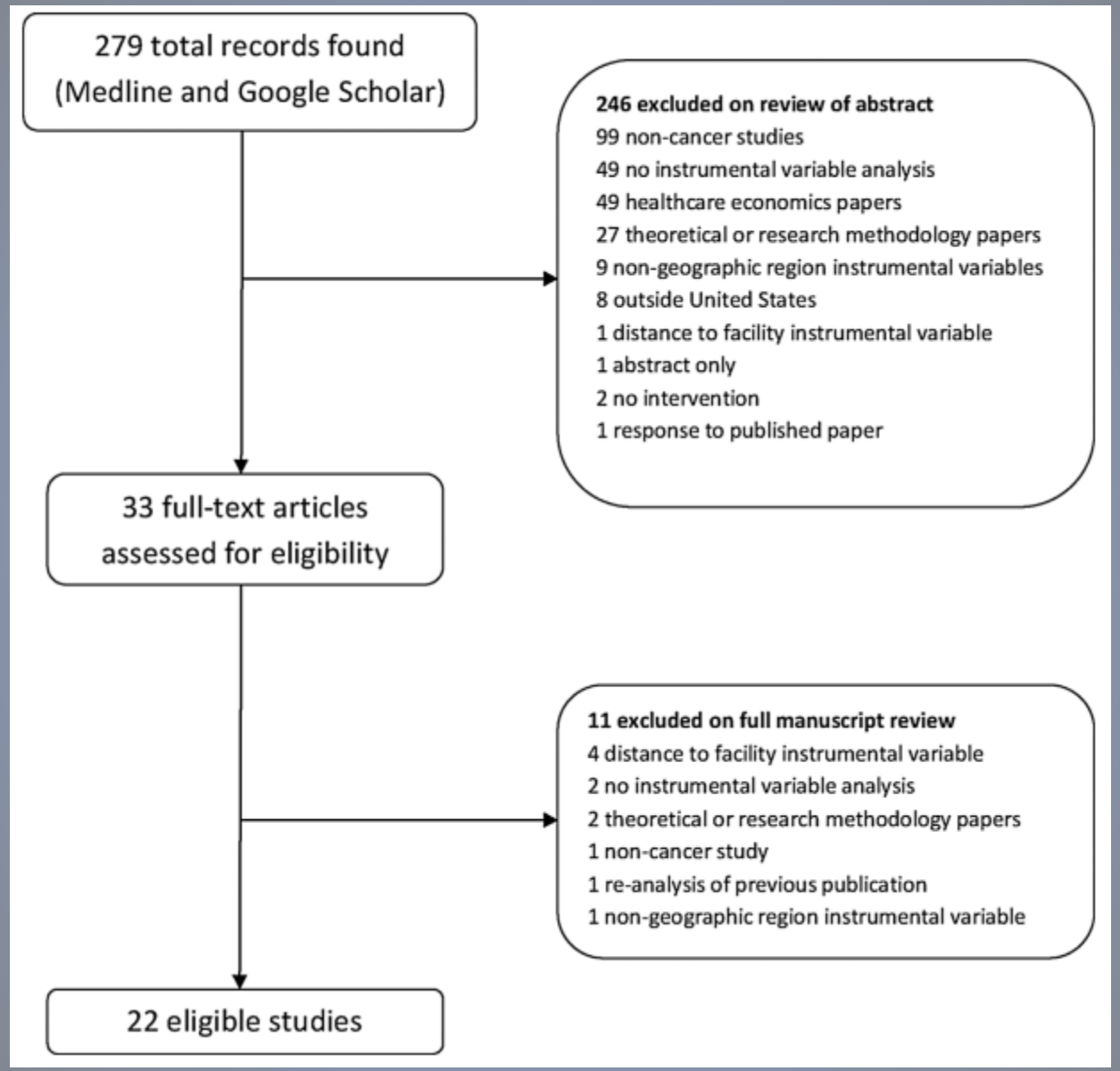
Emily A. Vertosick[1], Melissa Assel[1], Andrew J. Vickers*

Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, 485 Lexington Avenue, 2nd Floor, New York, NY 10017, United States

- Evaluates the use of geographic region as IV in observational cancer treatment studies

- Inclusion criteria
  - Observational study
  - Used IV analysis
  - Compare with Non-IV method: (Multivariable models & Propensity score methods.)

Cancer Epidemiology 51 (2017) 49-55


279 total records found
(Medline and Google Scholar)
246 excluded on review of abstract
99 non-cancer studies
49 no instrumental variable analysis
49 healthcare economics papers
27 theoretical or research methodology papers
9 non-geographic region instrumental variables
8 outside United States
1 distance to facility instrumental variable
1 abstract only
2 no intervention
1 response to published paper
33 full-text articles
assessed for eligibility
11 excluded on full manuscript review
4 distance to facility instrumental variable
2 no instrumental variable analysis
2 theoretical or research methodology papers
1 non-cancer study
1 re-analysis of previous publication
1 non-geographic region instrumental variable
22 eligible studies

Key Finding

- Weakness of Geographic Region as an IV
  - Treatment rates did not vary greatly across regions.
  - In most studies, treatment rate differences between low and high-use areas were only 5-20%, which suggests weak IV strength.
  - F-statistics were high (ranging from 8.57 to 7792), but high F-values in large datasets may not indicate strong instrument validity.
- Lack of Covariate Balance
  - 16 out of 22 studies examined covariate balance.
  - 13 studies reported that at least one covariate was unbalanced across regions (race, socioeconomic status, comorbidities).
  - This violates the exclusion restriction assumption (IV should only affect the outcome via treatment).

Key Finding

- Comparison of IV vs. Traditional Methods
  - 8 out of 11 studies found statistically significant results with multivariable/PS analysis but not IV analysis.
  - 9 out of 11 IV estimates were closer to the null compared to non-IV methods.
  - IV standard errors were much larger, making findings less precise.
- IV Estimates Were Often Misinterpreted
  - Some studies incorrectly cited prior flawed IV research (e.g., Hadley et al., 2010), which used an erroneous reference group in their RCT comparison.
  - This led to continued use of geographic region as an IV despite its weaknesses.

Hadley et al

- Studied the effectiveness of prostate cancer treatment
  - Radical prostatectomy vs. conservative management
  - In early-stage prostate cancer
- Compared difference statistical method
  - Multivariable regression
  - PS adjustment
  - IV analysis
    - Instrumental: geographic region

Hadley et al.

Method	Prostate Cancer–Specific Mortality (HR, 95% CI)	All-Cause Mortality (HR, 95% CI)
Multivariable Survival Analysis	1.59 (1.27–2.00)	1.47 (1.35–1.59)
Propensity Score Adjustment	Similar to multivariable models	Similar to multivariable models
Instrumental Variable (IV) Analysis	0.73 (0.08–6.73)	1.09 (0.46–2.59)
RCT Subset for Elderly Patients	Similar to IV estimates	Similar to IV estimates

Hadley et al

- Problem
  - They use the subgroup of elderly patient from RCT as reference group
  - IV estimate was closer to the RCT subgroup than the PS estimates, they assumed IV was more valid

  - RCT typically measures an Intention-to-Treat (ITT) effect
  - IV estimates Local Average Treatment Effects (LATE)

⃠ Later studies blindly cited Hadley et al. as proof that geographic IVs work.

Key Assumptions for a Valid Instrument

1. Relevance (Strong Association with Treatment)
   - The instrument must strongly predict treatment assignment.
   - **Example:** If a hospital performs more PCI, patients there are more likely to get PCI.
2. Exclusion Restriction (No Direct Effect on Outcome)
   - The instrument should not affect the outcome directly (only through treatment).
   - **Example:** Regional PCI rate should not affect survival directly, except through the effect of PCI.
3. Independence (No Correlation with Unmeasured Confounders)
   - The instrument should not be associated with unmeasured patient factors affecting the outcome.
   - **Example:** If high PCI hospitals also have better medical management, IV estimates may be biased.

Strengths and Limitations of IV analysis

☑ Strengths

- Removes both measured and unmeasured confounding.
- Better for causal inference when hidden biases exist.
- Often used in policy and economic studies.

✗ Limitations

- Finding a valid instrument is difficult.
- If the instrument is weak, results are biased.
- Estimates apply only to the "marginal" population (patients who received treatment due to the instrument).

Comparing PS and IV analysis

Feature	Propensity Score (PS)	Instrumental Variable (IV)
Bias Controlled	Measured confounders	Measured & unmeasured confounders
Key Assumption	No unmeasured confounders	Valid instrument exists
Data Requirement	Large dataset for matching	Requires strong IV
Common Usage	Clinical effectiveness studies	Policy and economic research
Interpretation	Average Treatment Effect (ATE)	Local Average Treatment Effect (LATE)

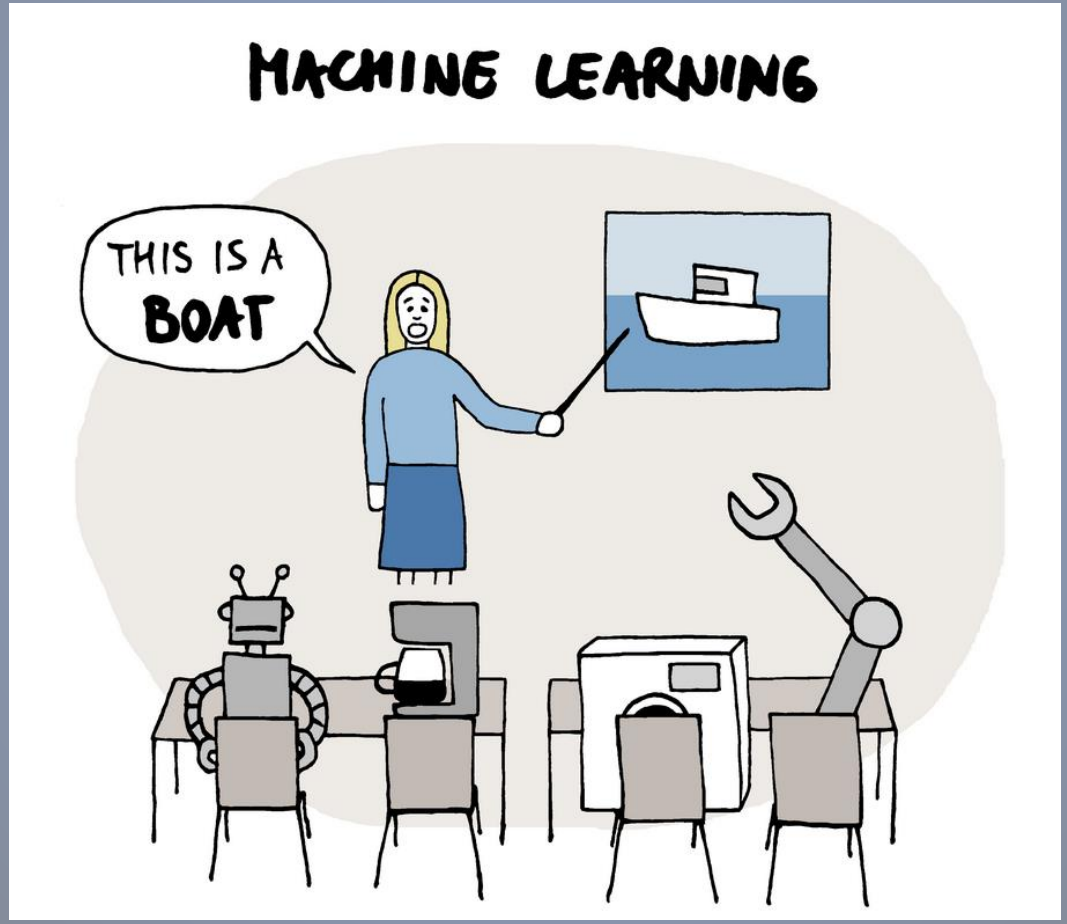
Summary:

- Use PS when confounders are well-measured.
- Use IV when hidden biases are a concern.
- Best practice: Use both methods to validate results.

Conclusion

1. Selection bias is a major challenge in observational studies.
2. PS methods help balance groups but do not remove hidden confounding.
3. IV analysis can remove both measured and unmeasured confounders if a strong IV exists.
4. Each method has strengths and weaknesses – they should be chosen based on the research question and available data.
5. Best practice: Apply both methods to check for consistency.

Can Machine Learning Help?



Can Machine Learning Help?

- Traditional methods (IV, PS) have limitations → strong assumptions, difficulty handling high-dimensional data.
- Machine learning is transforming many fields → Can it also improve causal inference?
- New methods integrate ML with traditional causal frameworks to improve precision and reduce bias.

Learning Causal Effects From Observational Data in Healthcare: A Review and Summary

*Jingpu Shi and Beau Norgeot**

Anthem, Inc., Point of Care AI, Palo Alto, CA, United States

frontiers | Frontiers in Medicine

- Machine Learning is Underutilized in Causal Inference for Healthcare:
  - Despite advances in causal forests, Bayesian methods, and deep learning
  - Traditional methods like Propensity Scores (PS) are still dominant in medical research.
  - PS methods mimic RCT-like conditions and are easier for clinicians to interpret.
- Trade-off Between Bias and Variance:
  - PS-based methods have lower variance but higher bias
  - ML-based causal inference methods (like causal forests, Bayesian approaches, and doubly robust estimators) have higher flexibility but are not widely adopted.

Shi & Norgeot, Front Med, 2022

How ML Enhanced Causal Inference in This Study

✓ **Improved Feature Selection:** Used ML to select the most relevant confounders, reducing bias.

✓ **Better Handling of High-Dimensional Data:** Neural networks and causal forests effectively managed large-scale EHR data.

✓ **Refined Heterogeneous Treatment Effects (HTE) Estimation:** Identified which subgroups benefit most from treatments.

✓ **Automated Instrument Selection for IV Analysis:** Reduced weak instrument bias by applying ML to find stronger IVs.

Proposed idea

- Best causal inference methods combine machine learning with traditional statistical methods

- Algorithm Selection Flowchart for Causal Inference:
  - Choose between causal inference methods based on dataset size, outcome complexity, and the presence of confounders.

  - **For smaller datasets:** PS Matching is recommended.
  - **For large datasets with high-dimensional data:** ML methods like Causal Forests and Targeted Maximum Likelihood Estimation (TMLE) are preferable.

Shi & Norgeot, Front Med, 2022

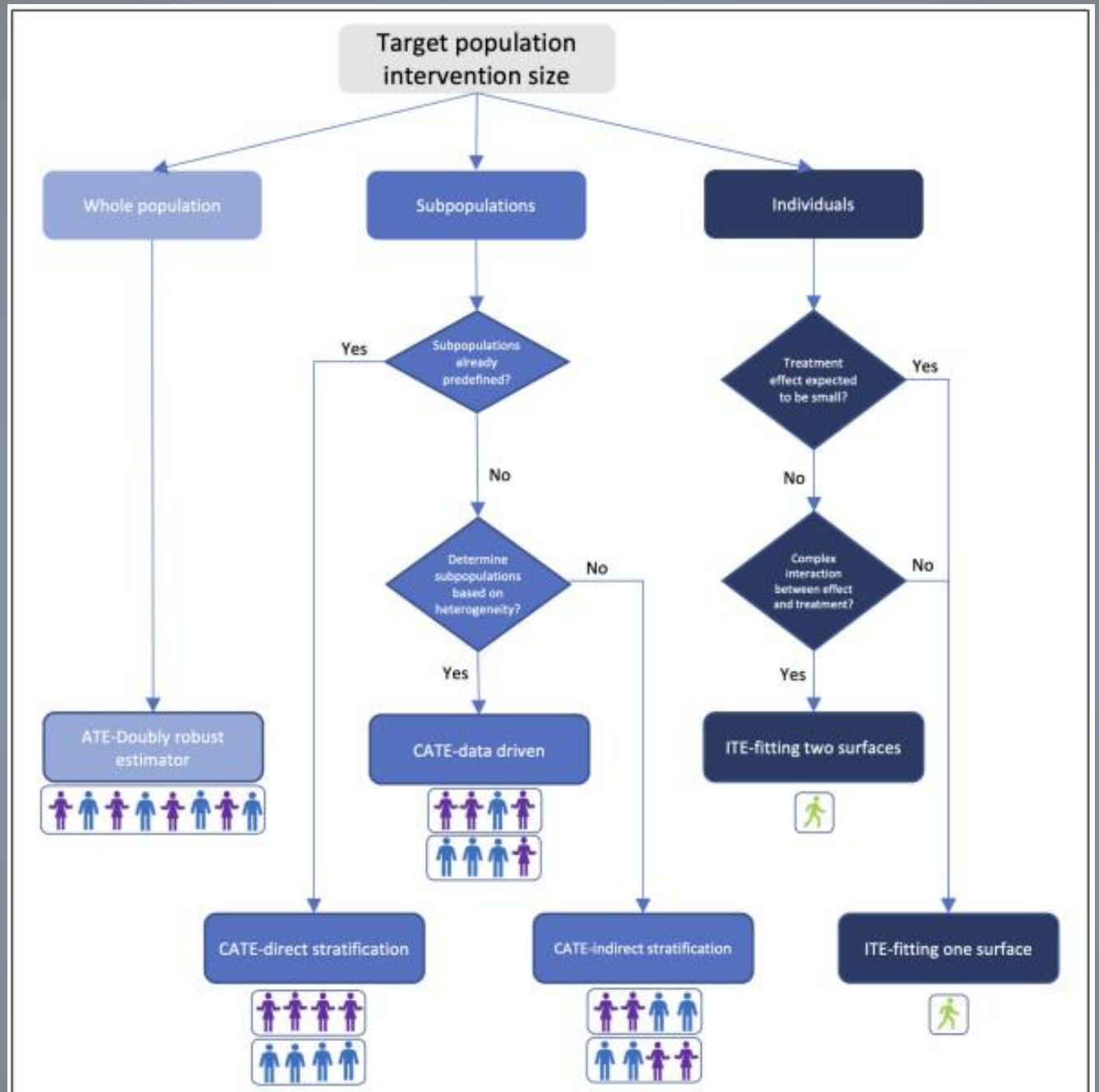


FIGURE 1 | Treatment effect estimator selection guide based on target-population intervention size and prior knowledge. Colors in the figure indicate bias-variance tradeoff. Light blue: high bias and low variance; blue: medium bias and variance; dark blue: low bias and high variance. Person icons under each estimator illustrate the composition of the targeted population.

Shi & Norgeot, Front Med, 2022

Challenge in ML for Causal Inference

- Interpretability
- Computational Complexity
- Bias in High-Dimensional Data
- Limited Adoption in Clinical Research

Challenge in ML for Causal Inference

- Interpretability
- Computational Complexity
- Bias in High-Dimensional Data
- Limited Adoption in Clinical Research

- Clinicians prefer easy-to-understand models like PS matching, but ML models (e.g., deep learning) are often complex and hard to interpret.
- Challenge: How do we ensure ML-based causal inference is trustworthy and explainable?

Challenge in ML for Causal Inference

- Interpretability
- Computational Complexity
- Bias in High-Dimensional Data
- Limited Adoption in Clinical Research

- ML methods require large datasets and high processing power.
- Challenge: Many medical studies have small sample sizes.

Challenge in ML for Causal Inference

- Interpretability
- Computational Complexity
- Bias in High-Dimensional Data
- Limited Adoption in Clinical Research

- ML models can overfit or detect correlations instead of causal effects.
- Challenge: How do we ensure ML learns causality, not just associations?

Challenge in ML for Causal Inference

- Interpretability
- Computational Complexity
- Bias in High-Dimensional Data
- Limited Adoption in Clinical Research

- Traditional methods like PS and IV are widely used in medicine, while ML-based causal inference is still new.
- Challenge: How do we increase the adoption of ML in real-world clinical studies?

Future direction

- Making ML More Explainable for Causal Inference
  - Need hybrid models that combine traditional causal inference (PS, IV) with ML to improve interpretability.
- Causal Representation Learning
  - Using deep learning to automatically find meaningful features from messy data (EHRs, imaging).
  - Helps reduce bias in high-dimensional data without manually selecting confounders.

Future direction

- Better Estimation of Heterogeneous Treatment Effects (HTE)
  - Advancing causal forests, Bayesian nonparametric models to move towards precision medicine.
  - Goal: Not just “Does treatment work?” but “For whom does it work best?”
- Smarter Instrument Selection for IV Analysis 🎯
  - ML can identify stronger, more reliable IVs, reducing weak instrument bias.
  - Example: Instead of using just regional variation, ML finds hidden patterns in physician preference, genetic markers, or policy shifts

Conclusion

- ML enhances causal inference but comes with challenges related to interpretability, computational demands, and adoption.

- Future research should focus on hybrid methods, explainability, and better integration into clinical practice.


THANK
YOU