Propensity Score Diagnostics

Emily Granger
Jamie C. Sergeant
Mark Lunt
Propensity scores are becoming increasingly popular

Figure 1: Number of propensity score publications in medical research by year
Recent review on the use of propensity score diagnostics in the applied medical literature [Granger et al. 2020]

Inclusion criteria:
- Publication years 2014-2016
- High-impact journals (Impact Factor > 4)

Extracted data on:
- Research area
- Propensity score method used
- Diagnostics used
Review on the use of propensity score diagnostics

Key Findings:

- 894 studies included
- 20.9% did not report use of any diagnostic
- 36.6% used hypothesis tests
Aims of research

Aim 1:
Review and compare the existing propensity score diagnostics.

Aim 2:
Develop guidelines for how to build and assess propensity score models.
Aims of research

Aim 1: Review and compare the existing propensity score diagnostics.

Aim 2: Develop guidelines for how to build and assess propensity score models.
Aims of research

Aim 1: Review and compare the existing propensity score diagnostics.

Aim 2: Develop guidelines for how to build and assess propensity score models.
### Individual diagnostics

#### Mean-based
- Standardised difference (SD)
- t-test statistic ($t$)
- Percent reduction in mean difference (PR)

#### Distribution-based
- Overlapping coefficient (OVL)
- Kolmogorov-Smirnov Statistic (KS)
Notation: exposure indicator for subject $i$: $E_i$, propensity score for subject $i$: $PS_i$, sample size: $n$.

For continuous variable $X$:

- $OCP_X(X_0) = \frac{1}{n} \sum_{i: X_i \leq X_0} E_i$
**Notation:** exposure indicator for subject $i$: $E_i$, propensity score for subject $i$: $PS_i$, sample size: $n$.

For continuous variable $X$:

- $OCP_X(X_0) = \frac{1}{n} \sum_{i: X_i \leq X_0} E_i$
- $ECP_X(X_0) = \frac{1}{n} \sum_{i: X_i \leq X_0} PS_i$
*Notation*: exposure indicator for subject $i$: $E_i$, propensity score for subject $i$: $PS_i$, sample size: $n$.

For continuous variable $X$:

- $OCP_X(X_0) = \frac{1}{n} \sum_{i : x_i \leq x_0} E_i$
- $ECP_X(X_0) = \frac{1}{n} \sum_{i : x_i \leq x_0} PS_i$
- $D_X = | OCP_X - ECP_X |$
Propensity score model:

\[ \logit(PS) = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_7 X_7 + \alpha_8 X_8 \]

Variation between scenarios:

**Correct PS:**

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Equation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1: ( X_8 = 0 )</td>
<td>Linear model</td>
<td></td>
</tr>
</tbody>
</table>
Propensity score model:

- \( \text{logit}(PS) = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_7 X_7 + \alpha_8 X_8 \)

Variation between scenarios:

**Correct PS:**

- **S1:** \( X_8 = 0 \) \quad \text{Linear model}
- **S2:** \( X_8 = 0.4(3.5^{X_1} - 1) \) \quad \text{Nonlinearity added (monotonic)}
Propensity score model:

- \( \text{logit}(PS) = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_7 X_7 + \alpha_8 X_8 \)

Variation between scenarios:

**Correct PS:**

- **S1:** \( X_8 = 0 \)  
  Linear model
- **S2:** \( X_8 = 0.4(3.5^{X_1} - 1) \)  
  Nonlinearity added (monotonic)
- **S3:** \( X_8 = X_4 X_5 \)  
  Binary-binary interaction
- **S4:** \( X_8 = X_4 X_1 \)  
  Binary-continuous interaction
- **S5:** \( X_8 = X_1 X_2 \)  
  Continuous-continuous interaction

Simulated data
Simulated data

Propensity score model:

\[ \text{logit(PS)} = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_7 X_7 + \alpha_8 X_8 \]

Variation between scenarios:

<table>
<thead>
<tr>
<th>Correct PS:</th>
<th>Incorrect PS:</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1: ( X_8 = 0 )</td>
<td>Linear model</td>
</tr>
<tr>
<td>S2: ( X_8 = 0.4(3.5^{X_1} - 1) )</td>
<td>Nonlinearity added (monotonic)</td>
</tr>
<tr>
<td>S3: ( X_8 = X_4X_5 )</td>
<td>Binary-binary interaction</td>
</tr>
<tr>
<td>S4: ( X_8 = X_4X_1 )</td>
<td>Binary-continuous interaction</td>
</tr>
<tr>
<td>S5: ( X_8 = X_1X_2 )</td>
<td>Continuous-continuous interaction</td>
</tr>
</tbody>
</table>
Scenario 1: Omission of a linear term

**Diagram:**
- Sample size 5000

**Legend:**
- Correct PS
- Incorrect PS

**Axes:**
- Standardised balance measure

**Variables:**
- SD: standardised difference
- t: t-test statistic
- PR: percent reduction in mean prevalence
- KS: Kolmogorov-Smirnov statistic
- OVL: overlapping coefficient
- CP: cumulative prevalence
Scenario 1: Omission of a linear term

<table>
<thead>
<tr>
<th>SD: standardised difference</th>
<th>KS: Kolmogorov-Smirnov statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>t: t-test statistic</td>
<td>OVL: overlapping coefficient</td>
</tr>
<tr>
<td>PR: percent reduction in mean prevalence</td>
<td>CP: cumulative prevalence</td>
</tr>
</tbody>
</table>
Simulated data

Decreasing sample size

Decreasing $R^2$

1. 20%, 5000
2. 20%, 2000
3. 20%, 500
4. 10%, 5000
5. 10%, 2000
6. 10%, 500
7. 5%, 5000
8. 5%, 2000
9. 5%, 500

Decreasing sample size
Scenario 2: Misspecification of a non-linear term
Scenarios 3-5: Omission of an interaction term

Figures:
- Scenario 3 (top left) binary-binary
- Scenario 4 (top right) binary-continuous
- Scenario 5 (bottom left) continuous-continuous

20
Conclusions (so far)

• Mean-based diagnostics can fail to identify nonlinear misspecifications in the propensity score

• Distribution-based diagnostics least reliable at identifying omission of interactions terms.

• Cumulative prevalence diagnostics most useful for identifying all types of propensity score misspecification.
What about the outcome?
Aims of research

Aim 1: Review and compare the existing propensity score diagnostics.

Aim 2: Develop guidelines for how to build and assess propensity score models.
Overall diagnostics

Which balance metric?

- Standardised difference (SD)
- Overlapping coefficient (OVL)
- Kolmogorov-Smirnov Statistic (KS)

Which weighting scheme?

Let $w_{ji}$ denote the $j^{th}$ weight for covariate $i$. Then:

- $w_{1i} = \gamma_i Std. Dev(x_i)$ [Caruana et al. 2015]
  - $\gamma_i$ is the coefficient for $x_i$ obtained after regressing outcome on $x_i$.
- $w_{2i} = \delta_i Std. Dev(x_i)$
  - $\delta_i$ is the coefficient for $x_i$ obtained after regressing outcome on all covariates.
Overall diagnostics

Which balance metric?

• Standardised difference (SD)
• Overlapping coefficient (OVL)
• Kolmogorov-Smirnov Statistic (KS)

Which weighting scheme?

Let $w (%)$ denote the $j$th weight for covariate $i$. Then:

- $w_#% = \gamma % SD(x_%)$ [Caruana et al. 2015]
- $\gamma !$ is the coefficient for $x_i$ obtained after regressing outcome on $x_i$.
- $w_.% = \delta % SD(x_%)$
- $\delta !$ is the coefficient for $x_i$ obtained after regressing outcome on all covariates.

Disease risk scores (DRS) defined as predicted outcome under the control condition

- Standardised mean difference in DRS as a propensity score diagnostic [Stuart et al. 2013]
Propensity score model:
• logit(PS) = $\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 +, ..., \alpha_9 X_9$

Outcome model:
• $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 +, ..., \beta_9 X_9 + \beta_{10} X_{10}$

Linear and Non-linear Scenarios:

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Equation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1:</td>
<td>$X_{10} = 0$</td>
<td>Independent baseline covariates</td>
</tr>
<tr>
<td>S2:</td>
<td>$X_{10} = 0$</td>
<td>Correlated baseline covariates</td>
</tr>
<tr>
<td>S3:</td>
<td>$X_{10} = 0.2(6.0^{X_1} - 1)$</td>
<td>Monotonic non-linearity</td>
</tr>
</tbody>
</table>
Propensity score model:
• logit(PS) = $\alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 +, \ldots, \alpha_9 X_9$

Outcome model:
• $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 +, \ldots, \beta_9 X_9 + \beta_{10} X_{10}$

Non-additive Scenarios:

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Interaction Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>S4: $X_{10} = X_1 X_5$</td>
<td>Binary-binary interaction</td>
</tr>
<tr>
<td>S5: $X_{10} = X_1 X_2$</td>
<td>Binary-continuous interaction</td>
</tr>
<tr>
<td>S6: $X_{10} = X_2 X_7$</td>
<td>Continuous-continuous interaction</td>
</tr>
</tbody>
</table>
Scenarios 1 and 2: Linear outcomes

**Table 1: Spearman rank correlation between overall diagnostics and bias**

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Balance Metric</th>
<th>Weights 1</th>
<th>Weights 2</th>
<th>SD(DRS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scenario 1</td>
<td>SD</td>
<td>0.992</td>
<td>0.996</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>KS</td>
<td>0.137</td>
<td>0.134</td>
<td></td>
</tr>
<tr>
<td></td>
<td>OVL</td>
<td>0.012</td>
<td>0.016</td>
<td></td>
</tr>
<tr>
<td>Scenario 2</td>
<td>SD</td>
<td>0.129</td>
<td>0.995</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>KS</td>
<td>0.102</td>
<td>0.142</td>
<td></td>
</tr>
<tr>
<td></td>
<td>OVL</td>
<td>0.031</td>
<td>0.053</td>
<td></td>
</tr>
</tbody>
</table>

*SD: Standardised difference; KS: Kolmogorov-Smirnov statistic; OVL: Overlapping coefficient; DRS: Disease risk score*
### Table 1: Spearman rank correlation between overall diagnostics and bias

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Balance Metric</th>
<th>Weights 1</th>
<th>Weights 2</th>
<th>SD(DRS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scenario 1</td>
<td>SD</td>
<td>0.992</td>
<td>0.996</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>KS</td>
<td>0.137</td>
<td>0.134</td>
<td></td>
</tr>
<tr>
<td></td>
<td>OVL</td>
<td>0.012</td>
<td>0.016</td>
<td></td>
</tr>
<tr>
<td>Scenario 2</td>
<td>SD</td>
<td>0.129</td>
<td>0.995</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>KS</td>
<td>0.102</td>
<td>0.142</td>
<td></td>
</tr>
<tr>
<td></td>
<td>OVL</td>
<td>0.031</td>
<td>0.053</td>
<td></td>
</tr>
</tbody>
</table>

*SD: Standardised difference; KS: Kolmogorov-Smirnov statistic; OVL: Overlapping coefficient; DRS: Disease risk score
### Scenarios 1 and 2: Linear outcomes

*Table 1: Spearman rank correlation between overall diagnostics and bias*

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Balance Metric</th>
<th>Weights 1</th>
<th>Weights 2</th>
<th>SD(DRS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scenario 1</td>
<td>SD</td>
<td>0.992</td>
<td>0.996</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>KS</td>
<td>0.137</td>
<td>0.134</td>
<td></td>
</tr>
<tr>
<td></td>
<td>OVL</td>
<td>0.012</td>
<td>0.016</td>
<td></td>
</tr>
<tr>
<td>Scenario 2</td>
<td>SD</td>
<td>0.129</td>
<td>0.995</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>KS</td>
<td>0.102</td>
<td>0.142</td>
<td></td>
</tr>
<tr>
<td></td>
<td>OVL</td>
<td>0.031</td>
<td>0.053</td>
<td></td>
</tr>
</tbody>
</table>

*SD: Standardised difference; KS: Kolmogorov-Smirnov statistic; OVL: Overlapping coefficient; DRS: Disease risk score*
### Scenarios 1 and 2: Linear outcomes

*SD: Standardised difference; KS: Kolmogorov-Smirnov statistic; OVL: Overlapping coefficient; DRS: Disease risk score

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Balance Metric</th>
<th>Weights 1</th>
<th>Weights 2</th>
<th>SD(DRS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scenario 1</td>
<td>SD</td>
<td>0.992</td>
<td>0.996</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>KS</td>
<td>0.137</td>
<td>0.134</td>
<td></td>
</tr>
<tr>
<td></td>
<td>OVL</td>
<td>0.012</td>
<td>0.016</td>
<td></td>
</tr>
<tr>
<td>Scenario 2</td>
<td>SD</td>
<td>0.129</td>
<td>0.995</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>KS</td>
<td>0.102</td>
<td>0.142</td>
<td></td>
</tr>
<tr>
<td></td>
<td>OVL</td>
<td>0.031</td>
<td>0.053</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Spearman rank correlation between overall diagnostics and bias
Simulated data

Decreasing sample size

1  2% , 5000
2  2% , 2000
3  2% , 500
4  5% , 5000
5  5% , 2000
6  5% , 500
7  10% , 5000
8  10% , 2000
9  10% , 500

Increasing $R^2$
Scenario 2: Non-linear term in outcome model
Scenario 2: Non-linear term in outcome model
Scenarios 3-5: Interaction term in the outcome model

Figures:
- Scenario 3 (top left) binary-binary
- Scenario 4 (top right) binary-continuous
- Scenario 5 (bottom left) continuous-continuous
Main finding: Standardised mean difference in the disease risk score is a promising overall diagnostic

Limitations:
1. Not robust to misspecifications in the outcome model
2. Performance dependent on sample size

Possible solutions:
1. Use of CP diagnostics to check specification
2. Using full sample or historic cohort to estimate DRS
Aims of research

Aim 1: Review and compare the existing propensity score diagnostics.

Aim 2: Develop guidelines for how to build and assess propensity score models.
So, how best to assess propensity scores?

**STEP 1:**
Choose variables

**STEP 2:**
Check individual covariates using **CP diagnostics**

**STEP 3:**
Check overall balance using **DRS**
So, how best to assess propensity scores?

**STEP 1:** Choose variables

**STEP 2:** Check individual covariates using **CP diagnostics**

**STEP 3:** Check overall balance using **DRS**
So, how best to assess propensity scores?

**STEP 1:** Choose variables

**STEP 2:** Check individual covariates using **CP diagnostics**

**STEP 3:** Check overall balance using **DRS**
So, how best to assess propensity scores?

STEP 1: Choose variables

STEP 2: Check individual covariates using **CP diagnostics**

STEP 3: Check overall balance using **DRS**
Thank you for listening
References


Scenario 2: Non-linear (stratification)
Scenarios 3-5: Interaction terms (stratification)

**Figures:**
- Scenario 3 (top left)  
  binary-binary
- Scenario 4 (top right)  
  binary-continuous
- Scenario 5 (bottom left)  
  continuous-continuous
Additional weights: Binary outcome

\[ w_{3i} = 1 + \log(OR_{X_iY}) - \frac{1}{p} \sum_{k=1}^{p} \log(OR_{X_kY}) \]

\[ w_{4i} = 1 + \sqrt{\log(OR_{X_iY})} - \frac{1}{p} \sum_{k=1}^{p} \sqrt{\log(OR_{X_kY})} \]

\[ w_{5i} = 1 + |\log(OR_{X_iY})| - \frac{1}{p} \sum_{k=1}^{p} |\log(OR_{X_kY})| \]

Additional scenario: Binary outcome (matching)
Additional scenario: Binary outcome (stratification)