What you need to know about statistics to read a journal article
Graeme L. Hickey

gaeme.hickey@liverpool.ac.uk
www.glhickey.com
@graemeeliehickey
Who am I?

- Statistician (Ph.D. 2011, CStat 2016)
- Former UK National Adult Cardiac Surgery Audit Statistician (2012-14)
- Researcher who has published in cardiothoracic journals
- Assistant Editor (Statistical Consultant) for the EJCTS and ICVTS (2012—present)

900 papers reviewed to-date
Statistics for surgeons

- We use statistical methods we will use to **transform complex raw data into meaningful results**

- We live in a world of **evidence-based medicine**, and statistics is the *lingua franca*

- Choice of statistical methods will depend on several things, including:
  - Clinical question
  - Study design
  - Outcomes
“A mistake in the operating room can threaten the life of one patient; a mistake in statistical analysis or interpretation can lead to hundreds of early deaths. So it is perhaps odd that, while we allow a doctor to conduct surgery only after years of training, we give SPSS® (SPSS, Chicago, IL) to almost anyone.”

What is the study type?

- Clinical Practice Guidelines
- Meta-Analysis
- Systematic Review
- Randomized Controlled Trial
  - Prospective, tests treatment
- Cohort Studies
  - Prospective - exposed cohort is observed for outcome
- Case Control Studies
  - Retrospective: subjects already of interest looking for risk factors
- Case Report or Case Series
  - Narrative Reviews, Expert Opinions, Editorials
- Animal and Laboratory Studies
- Secondary, pre-appraised, or filtered
- Primary Studies
- Observational Studies
- No design
- No humans involved
- Meta-analysis
- RCT design
- Multivariable regression
- Propensity score methods
- Basic summary statistics
- ANOVA

Figure source: https://en.wikipedia.org/wiki/Wikipedia:Identifying_reliable_sources_(medicine)
What are the study outcomes?

- **Continuous**
  - E.g. volume of blood transfused after surgery
- **Dichotomous / binary**
  - E.g. 30-day mortality status (dead versus alive)
- **Time-to-event**
  - E.g. time from surgery to death or re-intervention
- **Ordinal**
  - E.g. MV regurgitation grade at 12-months post-surgery
- **Count**
  - E.g. number of infections in first post-treatment year
Interpretation of clinical trials
Descriptive statistics

• Summarizing a binary outcome: “In-hospital mortality was 3.4% (3 / 87)”
• Summarizing a continuous outcome: “The average length of postoperative stay [PLOS] was…”

• 5 patients [PLOS: 3, 3, 4, 5, 90-days]
• Mean: 21-days
• Median: 4-days
• Skew-distributions are more informatively summarised using quantiles:
  – Median (middle quartile)
  – (Lower (first) quartile, Upper (third) quartile) captures the variability
Relative vs. absolute effects

**INDEPENDENT**

**Vitamin D can prevent asthma, study finds**

Taking oral vitamin D supplements can reduce the risk of severe asthma attacks by 3 per cent, according to research.

Vitamin supplements can help reduce the risk of severe asthma attacks, new research has found.

Taking oral vitamin D tablets can reduce the likelihood of hospital attendance from 6 per cent to 3 per cent, analysts at medical research group Cochrane concluded.

Source: [http://www.independent.co.uk/news/science/vitamin-d-asthma-attacks-prevent-study-cochrane-a7226756.html](http://www.independent.co.uk/news/science/vitamin-d-asthma-attacks-prevent-study-cochrane-a7226756.html)

---

**theguardian**

**Vitamin D supplements could halve risk of serious asthma attacks**

Major research review suggests that people who take vitamin D have fewer attacks requiring hospital treatment.

Vitamin D pills can halve the risk of serious asthma attacks according to a major review into the impact of supplementing on the condition.

People with mild or moderate asthma who took the vitamin with their normal medicine had fewer attacks that required hospital treatment than those who went without, scientists found.

The risk of severe attacks fell from 6% to 3% in patients who had a vitamin D boost for six months to a year. The supplements cut the frequency of attacks too, with cases needing steroid treatment falling from one per person every two or so years, to one every four years.

Source: [https://www.theguardian.com/society/2016/sep/05/vitamin-d-supplements-could-halve-risk-of-serious-asthma-attacks](https://www.theguardian.com/society/2016/sep/05/vitamin-d-supplements-could-halve-risk-of-serious-asthma-attacks)
Example

Randomization
$N = 200$

Treatment
$n = 100$

- Dead at 30-days
  $n = 30$
- Alive at 30-days
  $n = 70$

Control
$n = 100$

- Dead at 30-days
  $n = 40$
- Alive at 30-days
  $n = 60$

Dead at 30-days
Alive at 30-days
### Example

<table>
<thead>
<tr>
<th></th>
<th>Treatment</th>
<th>Control</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Died within 30-days</td>
<td>30</td>
<td>40</td>
<td>70</td>
</tr>
<tr>
<td>Alive at 30-days</td>
<td>70</td>
<td>60</td>
<td>130</td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>100</td>
<td>$N = 200$</td>
</tr>
</tbody>
</table>

A 2x2 contingency table + marginal totals
Example

<table>
<thead>
<tr>
<th></th>
<th>Treatment</th>
<th>Control</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Died within 30-days</td>
<td>a</td>
<td>b</td>
<td>a + b</td>
</tr>
<tr>
<td>Alive at 30-days</td>
<td>c</td>
<td>d</td>
<td>c + d</td>
</tr>
<tr>
<td>Total</td>
<td>a + c</td>
<td>b + d</td>
<td>N = a + b + c + d</td>
</tr>
</tbody>
</table>

A 2x2 contingency table + marginal totals
Example

Absolute risk in treatment group (AR\textsubscript{treat}) = \frac{a}{a + c} = \frac{30}{100} = 0.30

Absolute risk in control group (AR\textsubscript{control}) = \frac{b}{b + d} = \frac{40}{100} = 0.40

Absolute risk reduction (ARR) = AR\textsubscript{control} - AR\textsubscript{treat} = 0.4 - 0.3 = 0.10

Relative risk (RR) = \frac{AR\textsubscript{treat}}{AR\textsubscript{control}} = \frac{0.3}{0.4} = 0.75

Relative risk reduction (RRR) = 1 - RR = 1 - 0.75 = 0.25

Source: http://clinicalevidence.bmj.com/x/set/static/ebm/learn/665075.html
### Results from 3 hypothetical RCTs of the same treatment

<table>
<thead>
<tr>
<th>Risk Level</th>
<th>ARR</th>
<th>RRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>High risk</td>
<td>0.1</td>
<td>0.25</td>
</tr>
<tr>
<td>Intermediate</td>
<td>0.05</td>
<td>0.25</td>
</tr>
<tr>
<td>Low risk</td>
<td>0.01</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Clinical importance depends on underlying prevalence.
Example: ROOBY trial

It is always preferable to report both the absolute and relative effect sizes.

### Table 3. Sensitivity Analysis of 5-Year Follow-up Assessments, According to Treatment Group, among Patients Who Did Not Have Conversion to Other Treatment.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Off-Pump Group (N=967)</th>
<th>On-Pump Group (N=1059)</th>
<th>Absolute Difference (95% CI)</th>
<th>Relative Risk (95% CI)</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Primary outcomes at 5 yr</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Death</td>
<td>131 (13.5)</td>
<td>117 (11.0)</td>
<td>2.5 (-0.4 to 5.4)</td>
<td>1.23 (0.97 to 1.55)</td>
<td>0.09</td>
</tr>
<tr>
<td>Composite MACE outcome with death</td>
<td>281 (29.1)</td>
<td>281 (26.5)</td>
<td>2.5 (-1.4 to 6.4)</td>
<td>1.09 (0.95 to 1.26)</td>
<td>0.21</td>
</tr>
<tr>
<td><strong>Secondary outcomes at 5 yr</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Death from cardiac causes</td>
<td>55 (5.7)</td>
<td>51 (4.8)</td>
<td>0.9 (-1.1 to 2.8)</td>
<td>1.18 (0.82 to 1.71)</td>
<td>0.38</td>
</tr>
<tr>
<td>Acute myocardial infarction</td>
<td>108 (11.2)</td>
<td>99 (9.3)</td>
<td>1.8 (-0.8 to 4.5)</td>
<td>1.19 (0.92 to 1.55)</td>
<td>0.18</td>
</tr>
<tr>
<td>Repeat revascularization procedure</td>
<td>130 (13.4)</td>
<td>127 (12.0)</td>
<td>1.4 (-1.5 to 4.4)</td>
<td>1.12 (0.89 to 1.41)</td>
<td>0.33</td>
</tr>
<tr>
<td>Percutaneous coronary intervention</td>
<td>117 (12.1)</td>
<td>123 (11.6)</td>
<td>0.5 (-2.3 to 3.3)</td>
<td>1.04 (0.82 to 1.32)</td>
<td>0.74</td>
</tr>
<tr>
<td>Repeat CABG</td>
<td>15 (1.6)</td>
<td>5 (0.5)</td>
<td>1.1 (0.2 to 2.0)</td>
<td>3.29 (1.20 to 9.01)</td>
<td>0.014</td>
</tr>
</tbody>
</table>

* The primary 5-year composite MACE outcome was death from any cause, acute myocardial infarction, or any revascularization procedure. The P values are equivalent for both the absolute percentage differences and the relative risks reported. To evaluate for statistical significance, a P value of 0.05 or less was used for the two primary outcomes, and a P value of 0.01 or less was used for the secondary outcomes.

**Source:** Lamy et al. *N Engl J Med* 2016; 375:2359-2368
Odds ratio vs. relative risk

- Often confused with RR
- Exaggerate treatment effect
- Example: $\text{OR} = \frac{ad}{bc} = 0.64$ (recall: $\text{RR} = 0.75$)
- $\text{OR} \approx \text{RR}$ for low baseline risk

- Why do we use them?
  - Logistic regression
  - RRs precluded in some study designs (e.g. case-control)
  - $\text{OR}_{\text{death}} = 1 / \text{OR}_{\text{survival}}$ (not for RRs)

Source: Grant RL. BMJ, 2014; 348(4), f7450.
Time-to-event data

Kaplan-Meier curve

- Hazard: instantaneous rate of occurrence of the event

\[ h(t) = \lim_{\Delta t \to 0} \frac{P[t < T \leq t + \Delta t | T > t]}{\Delta t} \]

- HR = \frac{h_{\text{treat}}(t)}{h_{\text{control}}(t)} \quad [\text{NB: independent of time}]

- HR > 1 \Rightarrow \text{increased hazard}
Time-to-event data

Relative effect:
HR = 0.55

- HR uses all data at each time point
- Not robust to departures from proportionality

Absolute effect:
ARR(12-months) = 20.0%

30.7% in the TAVI group
50.7% in the standard therapy group

## Errors

### Hypothesis test

<table>
<thead>
<tr>
<th>Truth</th>
<th>No evidence of a difference</th>
<th>Evidence of a difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>No difference</td>
<td>True negative</td>
<td>False positive</td>
</tr>
<tr>
<td>Difference</td>
<td>False negative Type II error ((\beta))</td>
<td>True positive</td>
</tr>
</tbody>
</table>
Sample size

• Commonly used values in biomedical research are:
  – $\alpha = 0.05$ (or 5%)
  – $\beta = 0.20$ (corresponding to a power of 0.8, or 80%)
• To estimate sample size needed, we also need the minimum clinically relevant difference (MCRD)
  – Pilot studies
  – Published evidence
  – Clinical knowledge
• Essential that sample size calculation is reported + parameters used
Choosing a statistical test

• Need to know:
  – Continuous, discrete (dichotomous / categorical), or time-to-event data?
  – Independent or paired data?
  – Data satisfy test assumptions?
<table>
<thead>
<tr>
<th>Parametric test</th>
<th>Corresponding nonparametric test</th>
<th>Null hypothesis to be tested</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuous outcome</td>
<td></td>
<td></td>
<td>To compare the average number of sampled lymph nodes in three unpaired/unmatched groups of patients undergoing different resection methods of pancreaticoduodenectomy for pancreatic cancer.</td>
</tr>
<tr>
<td>Two sample (unpaired) t test</td>
<td>Mann-Whitney-U test</td>
<td>Difference between two independent (unpaired) groups</td>
<td>To compare the trajectory of CEA values with respect to race (Caucasian, African American, others) of colorectal cancer patients at several different time points after resection.</td>
</tr>
<tr>
<td>One sample (paired) t test</td>
<td>Wilcoxon matched-pair test</td>
<td>Difference between dependent (paired) groups (eg, before, after) or in matched patients</td>
<td>To evaluate whether age is independently associated with length of hospital stay in breast cancer patient undergoing lumpectomy after adjusting for race, socioeconomic status, comorbidities, tumor stage, and so forth.</td>
</tr>
<tr>
<td>One way analysis of variance (ANOVA)</td>
<td>Kruskal-Wallis test</td>
<td>Difference between three or more independent (unmatched) groups</td>
<td></td>
</tr>
<tr>
<td>Parametric test</td>
<td>Corresponding nonparametric test</td>
<td>Null hypothesis to be tested</td>
<td>Example</td>
</tr>
<tr>
<td>-----------------</td>
<td>----------------------------------</td>
<td>------------------------------</td>
<td>---------</td>
</tr>
<tr>
<td>Dichotomous/categoric outcome</td>
<td></td>
<td></td>
<td>To compare the proportion of successful endoscopic cholangiopancreatography in unmatched samples of elderly versus young patients with choledocholithiasis.</td>
</tr>
<tr>
<td>Chi-square test or Fisher’s exact test†</td>
<td>Difference between proportions of the outcome from two (or more) independent/unmatched samples is 0.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>McNemar’s test</td>
<td>Probability of the outcome is not more likely in one setting versus another (eg, pre or post or with one therapy versus another).</td>
<td></td>
<td>To compare the likelihood of response to proton pump inhibitors versus histamine antagonists for gastroesophageal reflux disease in matched patient samples.</td>
</tr>
<tr>
<td>Multiple logistic regression analyses</td>
<td>There is no association between the predictor variable and the categoric outcome after adjusting for potential confounding factors.</td>
<td></td>
<td>To assess the independent (risk-adjusted) impact of length of postoperative immobility on the occurrence of pulmonary embolisms after adjusting for age, race, socioeconomic status, comorbidities, and so forth.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parametric test</th>
<th>Corresponding nonparametric test</th>
<th>Null hypothesis to be tested</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time to event outcome</td>
<td>Log rank test (based on assumption that the relative hazard does not change over time)</td>
<td>Average hazards in the two groups are equal.</td>
<td>To compare the survival curves of colorectal cancer patients with liver metastases undergoing radiofrequency ablation versus cryotherapy.</td>
</tr>
<tr>
<td></td>
<td>Cox proportional hazard regression analysis (semi-parametric model, based on the assumption that the hazard ratio [hazard rate of group 1 divided by hazard rate of group 2] is constant over time).</td>
<td>The relative hazard of two patient samples is one after adjusting for potential confounding factors.</td>
<td>To assess whether patients with cancer of the parathyroid gland with elevated postoperative parathyroid hormone have a shorter overall survival compared with patients with normal parathyroid hormone after adjusting for potential confounding variables such as age, gender, race, socioeconomic status, tumor size, grading, staging, and so forth.</td>
</tr>
</tbody>
</table>

**P-values**

- **Definition**: a $P$-value is the probability under a specified statistical model (null hypothesis) that a statistical summary of the data would be equal to or more extreme than its observed value.

- Absence of evidence is not evidence of absence.

*Source: https://xkcd.com/1478/*
**P-values**

1. *P*-values can indicate how incompatible the data are with a specified statistical model.
2. *P*-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a *P*-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A *P*-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a *P*-value does not provide a good measure of evidence regarding a model or hypothesis.

**Source:** Wasserstein & Lazar. The American Statistician. 2016; 70(2): 129-133.
One vs. two-tailed $P$-values

• Two-tailed tests most commonly used
  – Allows for *either* treatment to be superior

• One-tailed tests only try to detect effect in one direction of interest
  – Can be abused; e.g. two-tailed $P=0.06$, one-tailed $P=0.03$

• One-tailed tests useful if:
  – treatment effect possible in *only* one direction; and
  – it would not be irresponsible or unethical to miss an effect in the opposite direction
Confidence intervals

• Sample $n$ subjects and construct a 95% CI for the mean outcome
• Imagine that you could then independently sample another $n$ subjects and re-calculate the 95% CI
• Do this lots and lots of times
• 95% of those intervals will contain the true population mean $\mu$

• It does not mean that there is a 95% probability that the population parameter lies within the interval

We can use the CI to gauge plausible estimates and assess if clinically relevant

Figure source: http://www.propharmagroup.com/blog/understanding-statistical-intervals-part-1-confidence-intervals
Clinical vs. statistical significance

- $P$-values become smaller as sample size increase
- Which is more clinically significant?
  - Length of stay recorded for $n$ patients randomized to open or EVAR surgery
  - **Scenario 1**: $n = 16$, difference 1-day (SD = 1-days) $P=0.065$
  - **Scenario 2**: $n = 2000$, difference = 0.1-days (SD = 1-day); $P=0.026$

- Clinical significance ≠ statistical significance
- Interpret the confidence interval rather than the $P$-value
Multiple comparisons & subgroup analyses

• Similar issues
• Each involves testing multiple hypotheses
The probability of obtaining $\geq 1$ significant result (at an $\alpha$-level of 0.05) for testing 20 independent null hypotheses = $(1 - 0.95^{20}) = 64\%$
Subgroup analyses

• **ISIS-2 trial**
  – 17,187 randomized patients with suspected acute MI to intravenous streptokinase, oral aspirin, both, or neither
  – Aspirin produced a highly significant reduction in 5-week vascular mortality relative to placebo
  – **Subgroup analysis**: patients were divided into 12 astrological star sign groups
    – In the Gemini and Libra groups, aspirin had a non-significant adverse effect
• Subgroup analyses should only be considered as *hypothesis generating*, rather than hypothesis testing
• A non-significant effect in a subgroup does not mean no effect is present → studies usually not powered for subgroup analyses
Many other statistical issues

- Trial design
  - Superiority
  - Non-inferiority
- Randomization methods
- Outcome definitions
  - Composite or individual components
- Cross-overs
- Losses after randomization
- Interim analyses
  + many non-statistical issues
Observational studies
Observational studies

Typical scenario: want to investigate the possible effect of a treatment on subjects, where the assignment of subjects into a treated group versus a control group is outside the control of the investigator.

Designs:
- Case-control studies
- Cohort studies
- Cross-sectional studies
Example: MVR

**ORIGINAL ARTICLE**

Mitral valve prosthesis choice for patients aged 65 years and over in the UK. Are the guidelines being followed and does it matter?

Ioannis Dimaraki,1 Stuart W Grant,1,2 Graeme L Hickey,2,3 Ramesh Patel,4 Steve Livesey,5 Neil Moat,6 Frank Wells,7 Ben Bridgewater1,2,3

**ABSTRACT**

Objective  Current guidelines recommend that most patients aged ≥65 years should undergo mitral valve replacement (MVR) using a biological prosthesis. The objectives of this study were to assess whether these guidelines are being followed in UK practice, and to investigate whether the guidelines are appropriate based on in-hospital mortality and mid-term survival.

Methods  Data from the National Institute for Cardiovascular Outcomes Research Adult Cardiac Surgery Audit database from all National Health Service (NHS) hospitals and some private hospitals performing adult cardiac surgery in the UK between April 2001 and March 2011 were analysed. The overall cohort included 3862 patients aged ≥65 years who underwent first-time MVR.

**Source:** Dimarakis et al. *Heart* 2014;100:500–507
Example: MVR

In-hospital mortality:

- **Biological prosthesis group**: 7.8% (152/1945)
- **Mechanical prosthesis group**: 5.5% (106/1917)
- $P = 0.005$ (chi square test)

What is your conclusion (and why)?
Example: kidney stone removal

- \( N = 700 \) patients with kidney stones were non-randomly assigned to either open surgery (Group O; \( n = 350 \)) or percutaneous nephrolithotomy (PN) (Group P; \( n = 350 \))
- Successfully treated:
  - Group O: 273 patients (78%)
  - Group PN: 289 patients (83%)
- **Conclusion**: PN is preferable to O
- What if the patients are separated into those with small and large kidney stones?

<table>
<thead>
<tr>
<th></th>
<th>Group O</th>
<th>Group PN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stones &lt;2cm</td>
<td>93% (81/87)</td>
<td>87% (234/270)</td>
</tr>
<tr>
<td>Stones ≥2cm</td>
<td>73% (192/263)</td>
<td>69% (55/80)</td>
</tr>
<tr>
<td>Total</td>
<td>78% (273/350)</td>
<td>83% (289/350)</td>
</tr>
</tbody>
</table>

- **Confounder**: a variable associated with both exposure and outcome

- 270/357 (76%) patients with small stones were assigned to PN, whereas 263/343 (77%) patients with large stones were assigned to open surgery

- **Simpson’s paradox**: confounding reverses effect of exposure
d (or Δ) = the standardized difference (or bias)

\[ d = \frac{100(\bar{x}_{\text{treatment}} - \bar{x}_{\text{control}})}{\sqrt{(s^2_{\text{treatment}} + s^2_{\text{control}})/2}} \] for continuous variables

\[ d = \frac{100(p_{\text{treatment}} - p_{\text{control}})}{\sqrt[(p_T(1 - p_T) + p_C(1 - p_C)]/2} \] for dichotomous variables

|Δ| > 0.1 (10%) represents meaningful imbalance in a given covariate between treatment groups.

### Table 1: Patient characteristics and operative variables before and after propensity score matching

<table>
<thead>
<tr>
<th></th>
<th>Before matching</th>
<th>After matching</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Overall</td>
<td>Biological</td>
</tr>
<tr>
<td>Dimi</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dim1</td>
<td>23.3</td>
<td>4.0</td>
</tr>
<tr>
<td>dim2</td>
<td>661.1</td>
<td>264.2</td>
</tr>
<tr>
<td>Dimi</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DPT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DPT</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DPT</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Overall</th>
<th>Biological</th>
<th>Mechanical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Δ (%)</td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
</tr>
<tr>
<td>79.5</td>
<td>72.9</td>
<td>4.3</td>
<td>73.0</td>
</tr>
<tr>
<td>-7.4</td>
<td>25.9</td>
<td>4.7</td>
<td>25.9</td>
</tr>
<tr>
<td>-6.4</td>
<td>691.4</td>
<td>264.6</td>
<td>692.5</td>
</tr>
<tr>
<td>Δ (%)</td>
<td>n</td>
<td>%</td>
<td>n</td>
</tr>
<tr>
<td>-7.4</td>
<td>3377</td>
<td>58.3</td>
<td>696</td>
</tr>
<tr>
<td>-6.6</td>
<td>367</td>
<td>15.5</td>
<td>189</td>
</tr>
<tr>
<td>-0.1</td>
<td>332</td>
<td>14.1</td>
<td>162</td>
</tr>
<tr>
<td>-6.8</td>
<td>166</td>
<td>7.0</td>
<td>80</td>
</tr>
<tr>
<td>0.1</td>
<td>59</td>
<td>2.5</td>
<td>27</td>
</tr>
<tr>
<td>-0.2</td>
<td>35</td>
<td>1.5</td>
<td>14</td>
</tr>
<tr>
<td>Dyspnoea (NYHA grade)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>History of hypertension</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.2</td>
<td>213</td>
<td>9.0</td>
<td>105</td>
</tr>
<tr>
<td>3.8</td>
<td>688</td>
<td>29.1</td>
<td>346</td>
</tr>
<tr>
<td>-6.1</td>
<td>1188</td>
<td>50.3</td>
<td>594</td>
</tr>
<tr>
<td>-0.5</td>
<td>273</td>
<td>11.6</td>
<td>136</td>
</tr>
<tr>
<td>-2.7</td>
<td>65</td>
<td>2.8</td>
<td>34</td>
</tr>
<tr>
<td>2.2</td>
<td>290</td>
<td>12.3</td>
<td>149</td>
</tr>
<tr>
<td>-38.1</td>
<td>244</td>
<td>52.7</td>
<td>620</td>
</tr>
<tr>
<td>History of hypertension</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16.9</td>
<td>197</td>
<td>50.7</td>
<td>596</td>
</tr>
</tbody>
</table>
Example: MVR

- Dimarakis et al. undertook 2 separate analyses:
  - Multivariable regression
  - Propensity score matching
Multivariable regression

The investigator seeks to assess the relationship between:

1. the primary predictor (mechanical vs. biological valve)
2. and the outcome(s) under consideration
3. after the potential distortion through covariates has been eliminated
Regression models

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Model*</th>
<th>$\beta$ coefficient (for unit increase)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuous</td>
<td>Multiple linear regression</td>
<td>Expected increase in outcome</td>
</tr>
<tr>
<td>(e.g. aneurysm diameter)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Binary</td>
<td>Multiple logistic regression</td>
<td>Log odds ratio</td>
</tr>
<tr>
<td>(e.g. in-hospital mortality)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time-to-event</td>
<td>Multiple Cox proportional hazards regression</td>
<td>Log hazard ratio</td>
</tr>
<tr>
<td>(e.g. time to all-cause mortality)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Other regression models exist as well
Logistic regression

Effect size is the odds ratio

An OR > 1 confers an increase in the odds of the event (outcome) after adjustment for the other covariates.
Cox regression

Effect size is the **hazard ratio**

A HR > 1 confers an increase in the hazard of the event (outcome) after adjustment for the other covariates
Regression is hard

- **How many covariates can we include?**
  - Depends on the number of events (not the sample size)
  - Rule-of-thumb: 1 covariate per 10 events

- **How do I decide which covariates to include?**
  - Univariable pre-screening
  - Stepwise regression
  - Clinical knowledge

- **How do I model continuous covariates?**
  - E.g. very large BMI is usually associated with increased hospital mortality, but so is very low BMI ⇒ U-shape

- **What model assumptions am I making, and how do I check them?**
  - E.g. Cox regression depends on the assumption of “proportional hazards”

- **How to handle missing data?**

*Picture source: Strauss V. *The Washington Post*. March 27, 2013*
Propensity score analysis

- The **propensity score** (PS) is defined as a subject’s probability of treatment assignment conditional on measured covariates
- Can usually estimate the PS using multiple logistic regression
- Different methods available to estimate the treatment effects

Matching
- Match a treated patient to one (or more) controls

Covariance adjustment
- Include the PS as a covariate along with the treatment variable

Inverse probability treatment weights (IPTW)
- Weight every observation according to the PS

Stratification
- Split the data up in 5 (or more) groups using quantiles of the PS
Propensity score matching

Propensity score matching

Table 1  Patient characteristics and operative variables before and after propensity score matching

<table>
<thead>
<tr>
<th></th>
<th>Before matching</th>
<th></th>
<th>After matching</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Overall  (n=3862)</td>
<td>Biological  (n=1945)</td>
<td>Mechanical  (n=1917)</td>
</tr>
<tr>
<td>Continuous variables</td>
<td>Mean  SD</td>
<td>Mean  SD</td>
<td>Mean  SD</td>
</tr>
<tr>
<td>Age (years)</td>
<td>73.0  4.9</td>
<td>74.8  4.7</td>
<td>71.2  4.4</td>
</tr>
<tr>
<td>BMI</td>
<td>25.9  4.6</td>
<td>25.7  4.7</td>
<td>26.1  4.6</td>
</tr>
<tr>
<td>BMI²</td>
<td>691.1 264.8</td>
<td>682.7 267.7</td>
<td>699.7 261.7</td>
</tr>
<tr>
<td>Categorical and dichotomous variables</td>
<td>n %</td>
<td>n %</td>
<td>n %</td>
</tr>
<tr>
<td>Female</td>
<td>2244 58.1</td>
<td>1095 56.3</td>
<td>1149 59.9</td>
</tr>
<tr>
<td>Angina (CCS grade)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Class 0</td>
<td>2352 60.9</td>
<td>1191 61.2</td>
<td>1161 60.6</td>
</tr>
<tr>
<td>Class I</td>
<td>608 15.7</td>
<td>283 14.6</td>
<td>325 17.0</td>
</tr>
<tr>
<td>Class II</td>
<td>527 13.6</td>
<td>265 13.6</td>
<td>262 13.7</td>
</tr>
<tr>
<td>Class III</td>
<td>278 7.2</td>
<td>157 8.1</td>
<td>121 6.3</td>
</tr>
<tr>
<td>Class IV</td>
<td>97 2.5</td>
<td>49 2.5</td>
<td>48 2.5</td>
</tr>
<tr>
<td>IV nitrates</td>
<td>54 1.4</td>
<td>27 1.4</td>
<td>27 1.4</td>
</tr>
</tbody>
</table>

Propensity score matching

**In-hospital mortality**
In the propensity-matched group, the overall in-hospital mortality was 6.4% (151/2362). In the matched biological prosthesis group, in-hospital mortality was 6.9% (81/1181), and in the matched mechanical prosthesis group it was 5.9% (70/1181). The unadjusted OR in the direction of biological prosthesis was 1.17 (95% CI 0.84 to 1.63, p=0.355).

**Source:** Dimarakis et al. *Heart* 2014;100:500–507.

- Matched 1181 mechanical implant patients with 1181 biological implant patients
- Confirmed that they were well-balanced groups on known confounders
- Compared in-hospital mortality using simple univariable analysis

**Question:** should we account for the paired nature of the data?
- Chi-square test vs. McNemar test?
Propensity score matching is hard too

- Getting a good propensity score model often requires several iterations
  - Interaction terms
  - Higher-order terms
  - What if a known confounder is not measured (cf. frailty for TAVI)
- What if we have missing data?
- N-to-1 matching
- Matching with or without replacement?
- …
Evidence synthesis
Forest plot

Knapp-Hartung random-effects OR and 95% CI for 30-day all-cause mortality stratified by study design. NOTION = Nordic Aortic Valve Intervention; OR = odds ratio; PARTNER = Placement of Aortic Transcatheter Valves; SAVR = surgical aortic valve replacement; STACCATO = A Prospective Randomised Trial of Transapical Transcatheter Aortic Valve Implantation Versus Surgical Aortic Valve Replacement in Operable Elderly Patients With Aortic Stenosis; TAVI = transcatheter aortic valve implantation.

* Percentages do not sum to 18.3% and 81.7% for randomized and matched studies, respectively, because of rounding.

Considerations

1. Publication bias
2. Heterogeneity
3. Randomized and non-randomized studies
Publication bias

Asymmetric funnel plot indicating possible publication bias

Symmetric funnel plot consistent with lower likelihood of publication bias

Heterogeneity

• Differences between study results beyond those attributable to chance
• Can be caused by:
  – clinical differences (e.g. all-comers vs. octogenarians)
  – methodological differences (RCT vs. observational study)

• Usual assessment involves:
  – $I^2$-statistic: the percentage of total variation across studies that is due to heterogeneity rather than chance
  – Cochran's Q-test: significant values ($P < 0.1$) provide evidence against homogeneity
Randomized vs. non-randomized studies (NRSs)

- Fewer RCTs in surgery than medicine
- NRS subject to inherent selection bias
- Present separate meta-analyses; avoid pooling RCTs and NRSs
- When pooling NRSs, consider what effect is being pooled:
  - crude (unadjusted)
  - multivariable regression adjusted
  - propensity score adjusted
  - then ask whether they are sufficiently homogeneous to combine

Reporting
Reporting

• Exists continued need to improve the reliability and value of published health research literature
• To encourage this there are several transparent and accurate reporting guidelines available
• Checklists often required by journals at time of submission

http://www.equator-network.org

Graeme L. Hickey, Joel Dunning, Burkhardt Seifert, Gottfried Sodeck, Matthew J. Carr, Hans Ulrich Burger and Friedhelm Beyersdorf on behalf of the EJCTS and ICVTS Editorial Committees

Received 27 March 2015; accepted 2 April 2015
Thank you for listening
Any questions?

New series of statistical “primers” forthcoming in the EJCTS and ICVTS

Acknowledgements
Dr. Stuart J. Head (L)
Dr. Stuart W. Grant (R)