Biostat 202: Data visualization: Telling stories with data

Aaron Wolfe Scheffler,
Division of Biostatistics,
Department of Epidemiology and Biostatistics
Outline for today

• Tables versus graphs

• Table principles
• Table practice

• Graph principles
• Graph types
• Graph practice
Study of heart disease patients

- Goal is to predict Major Adverse Cardiac Events (MACE) and mortality.
- Possible predictors: Left Ventricular Mass of the heart (LV mass) estimated during three heart phases and normalized by (divided by) body surface area (BSA) or weight. Also ejection fraction.

Cardiac CT assessment of left ventricular mass in mid-diastasis and its prognostic value

Ran Klein¹*, Emmanuelle S. Ametepe², Yeung Yam³, Girish Dwivedi³, and Benjamin J. Chow³
Table example – what is right/wrong?

### Table 3  ROC analysis results for predictors of MACE and all-cause mortality

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Optimal threshold</th>
<th>ROC AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>F</td>
</tr>
<tr>
<td>LVMI_{BSA} (g/m^2)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MD</td>
<td>79.2^a</td>
<td>63.8^a</td>
</tr>
<tr>
<td>ED</td>
<td>N.A.</td>
<td>72.4</td>
</tr>
<tr>
<td>ES</td>
<td>N.A.</td>
<td>72.3</td>
</tr>
<tr>
<td>LVMI_{Weight} (g/kg)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MD</td>
<td>2.13</td>
<td>1.72</td>
</tr>
<tr>
<td>ED</td>
<td>2.09</td>
<td>1.75</td>
</tr>
<tr>
<td>ES</td>
<td>N.A.</td>
<td>1.76</td>
</tr>
<tr>
<td>LVEF%</td>
<td>62.7</td>
<td>67.4</td>
</tr>
</tbody>
</table>

No significant differences in AUC were detected between phases (P > 0.05 using a Bonferroni adjustment for multiple comparisons). ROC AUC, receiver-operator characteristics—area under curve; N.A., not available due to not well-defined optimal threshold; LV, left ventricle. ^aThresholds from Mao et al. (2013).

European Heart Journal - Cardiovascular Imaging (2017) 18, 95-102
doi:10.1093/ehjci/jev357
Table 3  ROC analysis results for predictors of MACE and all-cause mortality

<table>
<thead>
<tr>
<th>Parameter</th>
<th>ROC AUC</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>F</td>
<td>M + F</td>
</tr>
<tr>
<td>LVMI_{BSA} (g/m^2)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MD</td>
<td>0.572</td>
<td>0.744</td>
<td>0.654</td>
</tr>
<tr>
<td>ED</td>
<td>0.561</td>
<td>0.672</td>
<td>0.626</td>
</tr>
<tr>
<td>ES</td>
<td>0.544</td>
<td>0.732</td>
<td>0.631</td>
</tr>
<tr>
<td>LVMI_{Weight} (g/kg)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MD</td>
<td>0.611</td>
<td>0.754</td>
<td>0.685</td>
</tr>
<tr>
<td>ED</td>
<td>0.616</td>
<td>0.675</td>
<td>0.660</td>
</tr>
<tr>
<td>ES</td>
<td>0.583</td>
<td>0.742</td>
<td>0.662</td>
</tr>
<tr>
<td>LVEF%</td>
<td>0.654</td>
<td>0.670</td>
<td>0.664</td>
</tr>
</tbody>
</table>

No significant differences in AUC were detected between phases (P > 0.05 using a Bonferroni adjustment for multiple comparisons. ROC AUC, receiver-operator characteristics—area under curve; N.A., not available due to not well-defined optimal threshold; LV, left ventricle. ^Thresholds from Mao et al. (2013).
Graph example – what is right/wrong?

Prevalence of CKD by CKD Stage and Year, 1988-1994 to 2015-2016
National Health and Nutrition Examination Survey

## Same data in a table – what is right/wrong? Better/worse?

### Prevalence of CKD by CKD Stage and Year, 1988-1994 to 2015-2016$^{a,b}$

National Health and Nutrition Examination Survey

<table>
<thead>
<tr>
<th>Year</th>
<th>CKD Stage 1</th>
<th>CKD Stage 2</th>
<th>CKD Stage 3</th>
<th>CKD Stage 4</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1988–1994</td>
<td>4.1% (CI 3.6-4.6)</td>
<td>3.0% (CI 2.6-3.4)</td>
<td>4.5% (CI 3.9-5.0)</td>
<td>.2% (CI .2-.3)</td>
<td>11.8% (CI 11.0-12.6)</td>
</tr>
<tr>
<td>1999–2006</td>
<td>5.2% (CI 4.7-5.8)</td>
<td>2.7% (CI 2.3-3.0)</td>
<td>6.3% (CI 5.7-6.7)</td>
<td>.4% (CI .3-.5)</td>
<td>14.5% (CI 13.7-15.3)</td>
</tr>
<tr>
<td>2007–2014</td>
<td>4.2% (CI 3.8-4.5)</td>
<td>3.3% (CI 2.9-3.6)</td>
<td>6.3% (CI 5.8-6.7)</td>
<td>.5% (CI .4-.6)</td>
<td>14.2% (CI 13.5-14.8)</td>
</tr>
<tr>
<td>2015–2016</td>
<td>4.7% (CI 3.9-5.5)</td>
<td>3.4% (CI 2.3-4.6)</td>
<td>5.8% (CI 4.6-7.0)</td>
<td>.4% (CI .3-.5)</td>
<td>14.2% (CI 12.5-16.0)</td>
</tr>
</tbody>
</table>

**Notes:**
95% confidence intervals, when available, are shown in parentheses.
Outline for today

• Tables versus graphs
• Table principles
• Table practice
• Graph principles
• Graph types
• Graph practice
Tables vs graphs

Use **TABLES** when

- The user needs to look up specific values
- You want to enable individual quantitative comparisons
- You need to convey results to high precision
- You want to convey both individual and summary information
- You want to display items that are measured on very different scales. Workarounds: left vs right scales; log scales; multiple panels.
Tables vs graphs

Use **GRAPHS** when
- You want to convey shapes
- You want to convey patterns
- You want to convey trends
- You want to convey quantitative ideas quickly

**Corollary:** You should minimize tables in scientific presentations in favor of graphs!
Chronic Kidney Disease

Which when?

Prevalence of CKD by CKD Stage and Year, 1988-1994 to 2015-2016
National Health and Nutrition Examination Survey

<table>
<thead>
<tr>
<th>Year</th>
<th>CKD Stage 1</th>
<th>CKD Stage 2</th>
<th>CKD Stage 3</th>
<th>CKD Stage 4</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1988–1994</td>
<td>4.1% (CI 3.6-4.6)</td>
<td>3.0% (CI 2.6-3.4)</td>
<td>4.5% (CI 3.9-5.0)</td>
<td>.2% (CI .2-.3)</td>
<td>11.8% (CI 11.0-12.6)</td>
</tr>
<tr>
<td>1999–2006</td>
<td>5.2% (CI 4.7-5.8)</td>
<td>2.7% (CI 2.3-3.0)</td>
<td>6.3% (CI 5.7-6.7)</td>
<td>.4% (CI .3-.5)</td>
<td>14.5% (CI 13.7-15.3)</td>
</tr>
<tr>
<td>2007–2014</td>
<td>4.2% (CI 3.8-4.5)</td>
<td>3.3% (CI 2.9-3.6)</td>
<td>6.3% (CI 5.8-6.7)</td>
<td>.5% (CI .4-.6)</td>
<td>14.2% (CI 13.5-14.8)</td>
</tr>
<tr>
<td>2015–2016</td>
<td>4.7% (CI 3.9-5.5)</td>
<td>3.4% (CI 2.9-3.6)</td>
<td>5.8% (CI 4.6-7.0)</td>
<td>.4% (CI .3-.5)</td>
<td>14.2% (CI 12.5-16.0)</td>
</tr>
</tbody>
</table>

Notes:
95% confidence intervals, when available, are shown in parentheses.
Outline for today

- Tables versus graphs
- Table principles
- Table practice
- Graph principles
- Graph types
- Graph practice
Table principles

- Should be **self-explanatory!!!**
- It is easier to compare numbers in columns.
- Round aggressively, preferably to two effective digits (ie where the numbers differ). **Exception:** you need to convey precise results.
- Order rows and/or columns purposefully to convey information. Compare across columns.
- Lightly use visual elements to ease comparisons.
Outline for today

- Tables versus graphs
- Table principles
- Table practice
- Graph principles
- Graph types
- Graph practice
Table 3  ROC analysis results for predictors of MACE and all-cause mortality

<table>
<thead>
<tr>
<th>Parameter</th>
<th>ROC AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
</tr>
<tr>
<td>LVMI_{BSA} (g/m^2)</td>
<td></td>
</tr>
<tr>
<td>MD</td>
<td>0.572</td>
</tr>
<tr>
<td>ED</td>
<td>0.561</td>
</tr>
<tr>
<td>ES</td>
<td>0.544</td>
</tr>
<tr>
<td>LVMI_{Weight} (g/kg)</td>
<td></td>
</tr>
<tr>
<td>MD</td>
<td>0.611</td>
</tr>
<tr>
<td>ED</td>
<td>0.616</td>
</tr>
<tr>
<td>ES</td>
<td>0.583</td>
</tr>
<tr>
<td>LVEF%</td>
<td>0.654</td>
</tr>
</tbody>
</table>

No significant differences in AUC were detected between phases (P > 0.05 using a Bonferroni adjustment for multiple comparisons. ROC AUC, receiver-operator characteristics—area under curve; N.A., not available due to not well-defined optimal threshold; LV, left ventricle.

*Thresholds from Mao et al. (2013).*
Improving the cardiac prediction table

- Improve table with
  - Explanation
  - Rounded to 2 digits
  - Light visuals
  - Order columns to put ‘M+F’ on the left

- What’s good? has primary comparisons in columns, lightly uses visual elements, and is ordered purposefully.
Table 3: Area under the ROC curve (AUROC) analysis for different predictors of MACE and all-cause mortality by phase of the cardiac cycle for all participants, males, and females. Prediction was better for females than for males and differed minimally by phase. LVMI normalized by weight performed better than normalization by body surface area and was slightly better than LVEF.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>AUROC</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
<td>Males</td>
<td>Females</td>
<td></td>
</tr>
<tr>
<td>LVMI&lt;sub&gt;BSA&lt;/sub&gt; (g/m&lt;sup&gt;2&lt;/sup&gt;)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MD</td>
<td>.65</td>
<td>.57</td>
<td>.74</td>
<td></td>
</tr>
<tr>
<td>ED</td>
<td>.63</td>
<td>.56</td>
<td>.67</td>
<td></td>
</tr>
<tr>
<td>ES</td>
<td>.63</td>
<td>.54</td>
<td>.73</td>
<td></td>
</tr>
<tr>
<td>LVMI&lt;sub&gt;Weight&lt;/sub&gt; (g/kg)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MD</td>
<td>.69</td>
<td>.61</td>
<td>.75</td>
<td></td>
</tr>
<tr>
<td>ED</td>
<td>.66</td>
<td>.62</td>
<td>.68</td>
<td></td>
</tr>
<tr>
<td>ES</td>
<td>.66</td>
<td>.58</td>
<td>.74</td>
<td></td>
</tr>
<tr>
<td>LVEF%</td>
<td>.66</td>
<td>.65</td>
<td>.67</td>
<td></td>
</tr>
</tbody>
</table>
Table Title:

Age-Adjusted Mortality Rates Attributable to Leading Cardiometabolic Underlying Causes of Death in the United States, 1999-2017
# Improving the mortality trend table

## Table. Age-Adjusted Mortality Rates Attributable to Leading Cardiometabolic Underlying Causes of Death in the United States, 1999-2017

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Heart Disease</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall AAMR&lt;sup&gt;a&lt;/sup&gt;</td>
<td>266.5</td>
<td>236.3</td>
<td>196.1</td>
<td>173.7</td>
<td>169.8</td>
<td>168.5</td>
<td>165.0</td>
</tr>
<tr>
<td><strong>Subgroups</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black men&lt;sup&gt;b&lt;/sup&gt;</td>
<td>407.2</td>
<td>372.1</td>
<td>312.4</td>
<td>266.1</td>
<td>262.8</td>
<td>258.6</td>
<td>257.5</td>
</tr>
<tr>
<td>White men&lt;sup&gt;b&lt;/sup&gt;</td>
<td>327.1</td>
<td>288.2</td>
<td>240.3</td>
<td>216.9</td>
<td>213.1</td>
<td>211.2</td>
<td>208.3</td>
</tr>
<tr>
<td>AAMR ratio&lt;sup&gt;c&lt;/sup&gt;</td>
<td>1.24</td>
<td>1.29</td>
<td>1.30</td>
<td>1.23</td>
<td>1.23</td>
<td>1.22</td>
<td>1.24</td>
</tr>
<tr>
<td>Black women&lt;sup&gt;b&lt;/sup&gt;</td>
<td>283.7</td>
<td>258.5</td>
<td>209.8</td>
<td>176.2</td>
<td>172.1</td>
<td>165.7</td>
<td>161.9</td>
</tr>
<tr>
<td>White women&lt;sup&gt;b&lt;/sup&gt;</td>
<td>212.8</td>
<td>188.5</td>
<td>155.4</td>
<td>136.5</td>
<td>132.4</td>
<td>132.4</td>
<td>128.4</td>
</tr>
<tr>
<td>AAMR ratio&lt;sup&gt;c&lt;/sup&gt;</td>
<td>1.33</td>
<td>1.37</td>
<td>1.35</td>
<td>1.29</td>
<td>1.30</td>
<td>1.25</td>
<td>1.26</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Trend period</strong></th>
<th>1999-2010</th>
<th>2010-2017</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Rate of change, β (95% CI)&lt;sup&gt;e&lt;/sup&gt;</strong></td>
<td>-8.3 (-8.8 to -7.8) [P&lt;.001]</td>
<td>-1.8 (-2.5 to -1.0) [P=.001]</td>
</tr>
</tbody>
</table>

Abbreviation: AAMR, age-adjusted mortality rate.

<sup>a</sup> The AAMR indicates rate per 100,000 population, directly standardized to the 2000 US Census population.

<sup>b</sup> Race information of the decedent was reported as standard practice by the funeral director, as provided by an informant (often the surviving next of kin), or in the absence of an informant, on the basis of observation.<sup>2</sup>

<sup>c</sup> AAMR ratio indicates a comparison of AAMRs by race within a sex group (e.g., black men compared with white men).

<sup>d</sup> Calculated using linear regression of overall population AAMR before and after inflection point, identified by Joinpoint analysis of heart disease, stroke, diabetes, or hypertension trend. P < .05 for comparison of linear trend of AAMR after vs before inflection point for all diseases.

<sup>e</sup> P value for statistical significance of first and second linear regression around Joinpoint regression identified inflection point in overall AAMR trend. β coefficient represents change in AAMR per year.

<sup>f</sup> The crude mortality rate was 24.5 per 100,000 in 1999 and 25.7 per 100,000 in 2017, but calculation of the weighted average based on the 2000 US Census population resulted in AAMRs of 25.0 per 100,000 in 1999 and 21.5 per 100,000 in 2017.
Improving the mortality trend table

- Improve table with
  - Rounded to 2 effective digits – how?
  - Light visuals instead of overall shading
  - Main comparison in columns?

- Perhaps better as a graph. Why a table?
  - Trying to get a lot of information in here
Outline for today

• Tables versus graphs

• Table principles
• Table practice

• Graph principles
• Graph types
• Graph practice
Graph principles

- Should be self-explanatory.
- Aim for interocular traumatic impact. Use the ideas below to convey your message clearly.
- Choose a graph type to emphasize your desired message.
- **Maximize the data to ink ratio.**
- Minimize the impact of non-data ink.
- Use colors, marker symbols and line types purposefully (Excel is 😞).
CKD graph: how to improve?

Prevalence of CKD by CKD Stage and Year, 1988-1994 to 2015-2016
National Health and Nutrition Examination Survey

How could the CKD graph be improved?

- Explanation
- Drop total columns
- Change scale
- Change color scheme

*Try your hand at it in the homework!*
Graphs can convey data by

- Points – location in space
- Lines – trends (also points+lines)
- Bars – by height (also points+bars)
- Colors
- Point types
- Point size
- Line types
- Bar or symbol fill (color or shading)
Outline for today

• Tables versus graphs
• Table principles
• Table practice
• Graph principles
• Graph types
• Graph practice
Graph types

- Points – scatter plot for two numeric variables
- Lines – trend for numeric over time
- Points and lines – connected plot, smoothed plot, estimate and error bars
- Bars – bar chart for numeric by categorical; histogram for frequency versus numerical value to display shape.
- Bars and lines – bar chart with error bars; boxplot for distributional shape.
Avoid:
pie charts and stacked bar charts
Avoid: pie charts and stacked bar charts

● Comparing the areas is a difficult visual perception task.

● We are really bad at comparing relative areas that are represented in these pie charts and stacked bar chart.

● Wanna bet? Let’s see.
At all costs avoid pie charts and stacked bar charts. Comparing the areas is a difficulty perception task.
At all costs avoid pie charts and stacked bar charts. Comparing the areas is a difficulty perception task.
At all costs avoid pie charts and stacked bar charts. Comparing the areas is a difficulty perception task.
<table>
<thead>
<tr>
<th>Practice Setting</th>
<th>No. of Order Sessions</th>
<th>Odds Ratio&lt;sup&gt;c&lt;/sup&gt; (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Restricted Group&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Unrestricted Group&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wrong-patient order sessions per 100 000</td>
<td>90.7</td>
<td>88.0</td>
</tr>
<tr>
<td>Wrong-patient order sessions</td>
<td>1980</td>
<td>2026</td>
</tr>
<tr>
<td>Total order sessions</td>
<td>2 183 365</td>
<td>2 303 266</td>
</tr>
<tr>
<td>Emergency department</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wrong-patient order sessions per 100 000</td>
<td>157.8</td>
<td>161.3</td>
</tr>
<tr>
<td>Wrong-patient order sessions</td>
<td>560</td>
<td>576</td>
</tr>
<tr>
<td>Total order sessions</td>
<td>3 548 822</td>
<td>3 570 047</td>
</tr>
<tr>
<td>Inpatient</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wrong-patient order sessions per 100 000</td>
<td>185.6</td>
<td>185.1</td>
</tr>
<tr>
<td>Wrong-patient order sessions</td>
<td>1324</td>
<td>1340</td>
</tr>
<tr>
<td>Total order sessions</td>
<td>7 134 172</td>
<td>7 237 462</td>
</tr>
<tr>
<td>Medical/surgical</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wrong-patient order sessions per 100 000</td>
<td>187.9</td>
<td>187.8</td>
</tr>
<tr>
<td>Wrong-patient order sessions</td>
<td>940</td>
<td>879</td>
</tr>
<tr>
<td>Total order sessions</td>
<td>5 003 385</td>
<td>4 679 941</td>
</tr>
<tr>
<td>Critical care</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wrong-patient order sessions per 100 000</td>
<td>247.8</td>
<td>258.2</td>
</tr>
<tr>
<td>Wrong-patient order sessions</td>
<td>161</td>
<td>216</td>
</tr>
<tr>
<td>Total order sessions</td>
<td>6 497 973</td>
<td>8 366 22</td>
</tr>
<tr>
<td>Pediatrics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wrong-patient order sessions per 100 000</td>
<td>122.5</td>
<td>139.2</td>
</tr>
<tr>
<td>Wrong-patient order sessions</td>
<td>65</td>
<td>111</td>
</tr>
<tr>
<td>Total order sessions</td>
<td>5 307 445</td>
<td>7 972 26</td>
</tr>
<tr>
<td>Obstetrics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wrong-patient order sessions per 100 000</td>
<td>201.1</td>
<td>197.1</td>
</tr>
<tr>
<td>Wrong-patient order sessions</td>
<td>85</td>
<td>78</td>
</tr>
<tr>
<td>Total order sessions</td>
<td>4 227 272</td>
<td>39 575</td>
</tr>
<tr>
<td>Outpatient</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wrong-patient order sessions per 100 000</td>
<td>7.9</td>
<td>8.2</td>
</tr>
<tr>
<td>Wrong-patient order sessions</td>
<td>86</td>
<td>97</td>
</tr>
<tr>
<td>Total order sessions</td>
<td>1 082 855</td>
<td>1 176 344</td>
</tr>
</tbody>
</table>
Having your cake and eating it too

<table>
<thead>
<tr>
<th>Practice Setting</th>
<th>No. of Order Sessions</th>
<th></th>
<th>Odds Ratio (95% CI)</th>
<th>Favors Restricted Group</th>
<th>Favors Unrestricted Group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Restricted Group^a</td>
<td>Unrestricted Group^b</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wrong-patient order sessions per 100,000</td>
<td>90.7</td>
<td>88.0</td>
<td>1.03 (0.90-1.20)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wrong-patient order sessions</td>
<td>1980</td>
<td>2026</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total order sessions</td>
<td>2,183,365</td>
<td>2,303,266</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Emergency department</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wrong-patient order sessions per 100,000</td>
<td>157.8</td>
<td>161.3</td>
<td>1.00 (0.83-1.20)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wrong-patient order sessions</td>
<td>560</td>
<td>576</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total order sessions</td>
<td>3,548,82</td>
<td>3,570,47</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Graph types - Two at a time

- Two numerical variables
  - Scatterplot
- Numerical and Categorical
  - Bar chart
  - Box plot
- Two categorical
  - Table
Outline for today

• Tables versus graphs

• Table principles
• Table practice

• Graph principles
• Graph types
• Graph practice
**Ex: HPV vaccine uptake in Netherlands**

Pot et al (Prev Med, 2017) wanted to understand which variables were associated with intent for mothers to get their 12 year-old daughters vaccinated. Intention is measured on a scale from 1 (definitely not) to 7 (definitely yes). 8,062 responses.

One of those predictors was knowledge about the HPV vaccine. Knowledge was assessed as the sum across 8 questions with -1 for wrong, +1 for correct (so a range of -8 to 8).

What type graph to assess association?
Data consistent with Pot

Scatterplot only

Knowledge

Intent (1-7)
LOWESS smoother and jittered data

bandwidth = .8
Just plot mean values

Mean values only
Mean values and regression

Intent (1-7)

Knowledge

---

Fitted regression

Mean intent
Compare with original

Scatterplot only

Knowledge

Intent (1-7)

Chi-square

Intent (1-7)

Knowledge

Chi-square

Intent (1-7)
Graph practice

OAI data – is the volume of cartilage in the knee different between those who have painful knee arthritis and those that do not?

Preferred graph type?
Mean medial cartilage volume
For pain progressors and non-progressors

Cartilage volume

No pain               Pain

Chi-square
Mean medial cartilage volume
For pain progressors and non-progressors

Cartilage volume

No pain    Pain

Mean medial cartilage volume

Cartilage volume

0  0.5  1  1.5  2
Plot of mean cartilage volume with SE bars

For pain progressors and non-progressors

- No pain
- Pain Progression
- Pain

Plot shows the mean cartilage volume with standard error bars for pain progressors and non-progressors.
Boxplots of medial cartilage volume
For pain progressors and non-progressors
More variables at once (big data)

Adding additional variables to a two-variable graph

- Numeric
  - Symbol size
  - Symbol color shading

- Categorical
  - Color
  - Symbol type or line type
  - Different panels
Graph practice

OAI data – is the volume of cartilage in the knee different between those who have painful knee arthritis and those that do not? And different between males and females?
Mean medial cartilage volume

For pain progressors and non-progressors

Bar chart showing mean medial cartilage volume for males and females with and without pain. The chart indicates higher cartilage volume for pain progressors compared to non-progressors.
Mean medial cartilage volume

For pain progressors and non-progressors

Cartilage volume

- Male/Progressor
- Female/Progressor
- Male/Non-progressor
- Female/Non-progressor
Graph practice

Back to the AUROC predictor data. Can we do better with a graph?

Preferred graph type to display predictor, phase and AUROC?

Predictor: categorical
Phase: categorical
AUROC: numerical
Area under the ROC curve for predictors of mortality for all participants by cardiac phase

- LVMIBSA
- LVMIWeight
- LVEF - not dependent on phase
Area under the ROC curve for predictors of mortality
separately for males and females by cardiac phase

**Graph practice**

- AUROC
- ED
- ES
- MD
- LVMI
- BSA
- Male
- Female
- LVMI
- Weight
- Male
- Female
- LVEF
- Male
- Female
Area under the ROC curve for predictors of mortality for males, females and overall by cardiac phase
Graph practice

OAI data – is the volume of cartilage in the knee related to body mass index (BMI)?

Preferred graph type?
Examples of graphs and how they convey data

Scatterplot.

Plot of cartilage volume versus BMI

Cartilage volume

Body mass index (BMI)
Examples of graphs and how they convey data

Plot of cartilage volume versus BMI

With a linear fit
Examples of graphs and how they convey data

Scatterplot with smoother
Plot of cartilage volume versus BMI

With a linear fit and confidence region

Examples of graphs and how they convey data

Scatterplot with smoother and error bars
Graph practice

NHANES data – is trend in chronic kidney disease prevalence over time?

Does it differ between diabetics and non-diabetics?

Preferred graph type? Hint… trend.
Figure. Adjusted prevalence of stage 3 and 4 chronic kidney disease (estimated glomerular filtration rate of 15 to 59 mL/min/1.73 m² calculated with Chronic Kidney Disease Epidemiology Collaboration equation) in U.S. adults, by age (A), sex (B), race/ethnicity (C), and presence or absence of diabetes (D), NHANES 1988-1994 through 2011-2012.
Interactive plots

- Very cool! Very labor intensive
- Consider if this is worth your effort
- If you need to reach a large audience – your work is important and you should budget resources to assist you here
Other kinds of graphics

• Chronic Kidney disease map and animation

https://nccd.cdc.gov/CKD/Detail.aspx?QNum=Q69&Year=2016&colors=all#refreshPosition
Other kinds of graphics

- Genetic heatmaps
Fig 2. NOJAH genome-wide heatmap (GWH) analysis of RNA-Seq derived gene expression data using TCGA BRCA expression dataset.

https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0204542
Other kinds of graphics

- McCulloch co-authors on UCSF Profiles

https://profiles.ucsf.edu/display/176718/network/coauthors/cluster
Summary

- Use tables for specific values, quantitative comparisons, precision, individual+summary information, differing scales.
- Use graphs to convey information quickly, for shapes, patterns and trends.
- Both should be self-explanatory and use a light touch on non-data
- Tables: compare columns, round aggressively, order purposefully
- Graphs: maximize data to ink, use color, symbols and line types purposefully.