Presenting Data Visually: Finding the Best Fit for Your Data

Mary Putt, PhD, ScD
Associate Professor of Biostatistics and Epidemiology

Kathryn Smitz, PhD, MPH
Associate Professor of Biostatistics and Epidemiology

Mary Leonard, BA
Biomedical Art & Design

September 27, 2012
A picture is worth 1000 words.....

*but not if it takes 1000 words to explain it*

(not original)
Who is the audience?
Patients expired from disease progression within 5 m of PDT dose-limiting toxicity.

Patients alive-no evidence of disease or progression at > 10m.

Desmoplastic disease, with progression at 6 m.
What’s Wrong Here?

Too much text
  – *Use text sparingly*

Small proportion of graphic involves data
  – *Spend ink on data*

Seriously complex axis labels
  – *Save complexity for the talk or the legend*

Clutter
  – *Think Matisse not Rembrandt*

Negative numbers
  – *Avoid*

Multiple messages
  – *Think headlines*
Hypothesis

- Association between marker and hypoxia
- Differences in marker for responders and non-responders
Who is the Audience?

• NIH grant reviewers
• Translational and basic science reviewers
• Eventually journal reviewers e.g. Clinical Cancer Research
What is the message?
Association between M1 and hypoxia

– *M1 predicts hypoxia*

Distribution of M1 and hypoxia among px with good and poor outcomes

– *Tighter and possibly lower among better outcome patients*
M1 versus Hypoxia

Ratio of M1 Post/Pre Surgery vs. Hypoxia (μM)

Outcome
- Good
- Poor

Graph showing the correlation between the ratio of M1 post/pre surgery and hypoxia levels. The diagram includes data points for both 'Good' and 'Poor' outcomes.
Hypoxia (µM)
What Else Could I Do?

- Put aside and look tomorrow
Color

• Great tool
• Chose one per group and stay consistent throughout presentation/paper
• Avoid pure reds and greens
• Need contrast
Plunger Plots

• Drummond, GB, and Vowler, SI. Show the data, don’t conceal them. Journal of Physiology 2011 589.8 1861-1863
  – Display data in raw form
  – Plungers conceal information
  – Pay attention to sample size
Plunger Plots

- Example data next
- What are your impressions?
- What dominates the graphic?
- Does the `distribution’ of the data appear different?
- Does the `mean’ of the data appear different?
- What aspect of the graphic leads you to your conclusion?
test data with SEM

<table>
<thead>
<tr>
<th>Type</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>control</td>
<td>1.5</td>
</tr>
<tr>
<td>treated</td>
<td>2.0</td>
</tr>
</tbody>
</table>
test data with SD

<table>
<thead>
<tr>
<th>outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
</tr>
<tr>
<td>0.5</td>
</tr>
<tr>
<td>1.0</td>
</tr>
<tr>
<td>1.5</td>
</tr>
<tr>
<td>2.0</td>
</tr>
<tr>
<td>2.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type</th>
<th>control</th>
<th>treated</th>
</tr>
</thead>
</table>

Type
test data with 95% CI

Type
outcome

control

0.0
1.0
1.5
2.0

treated
Boxplots

• Identical data
• How do your impressions of the data change?
• What aspect of the graphic causes the change?
• Let’s add in the original data to make it even more informative
control  treated

outcome
type

control  treated
Now Reduce Sample Size

• How does the plunger plot change?
• How does the boxplot change?
test data with SEM

<table>
<thead>
<tr>
<th>Type</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>control</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
</tr>
<tr>
<td>treated</td>
<td></td>
</tr>
</tbody>
</table>
Examples
Preventing Adolescent Obesity with Sleep

(Unpublished data shared with permission from Project PI, Dr. Janet Audrain-McGovern)

• Population based cohort study
• Followed 1200+ teens from 9<sup>th</sup> to 12<sup>th</sup> grade
• Analysis shows that # of hours of sleep per night predicts obesity prevalence
Goal:
Get the Post-Doc published in the highest tier medical journal possible
## Original Data Presentation

<table>
<thead>
<tr>
<th></th>
<th>Body Mass Index</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10&lt;sup&gt;th&lt;/sup&gt; Percentile</td>
</tr>
<tr>
<td><strong>Intercept</strong></td>
<td>18.1 (17.4, 18.9)</td>
</tr>
<tr>
<td><strong>Wave</strong></td>
<td>0.24 (0.20, 0.27)</td>
</tr>
<tr>
<td><strong>Sleep</strong></td>
<td>-0.04 (-0.11, 0.03)</td>
</tr>
</tbody>
</table>
What’s wrong with this?

• Buries the sexy finding
  – Think headlines

• Too many numbers
  – Think headlines

• Requires a high level of statistical training to interpret the sexy finding
  – Know your audience
Better....

Male, wave 7

7.5 hrs: 32% have BMI 25 or greater
10.0 hrs: 26% have BMI 25 or greater
Setting also matters

• What is appropriate in a
  – Peer reviewed journal
    • Audience matters
  – Presentation
    • Graphs, not tables
  – Popular Press
    • HEADLINES
Sleep affects body weight

26% obesity
10 hours of sleep

32% obesity
7.5 hours of sleep
Deliver <37 weeks

Deliver >37 weeks

Women with PTL

Maternal and cord blood samples

Postnatal evaluation (months; corrected age)

6 12 15 18 21 24 27 36

* * *

Women without PTL

Parent Contact Limited Health Survey

Validated Parental Questionnaire

Extensive Neurobehavioral testing

Original idea from researcher
Revised graphic

Women with PTL

- Delivered <37 weeks
- Delivered >37 weeks

Maternal and cord blood samples

Women without PTL

- Delivered >37 weeks

Postnatal evaluation (months, corrected age)

0 3 6 9 12 15 18 21 24 27 30 33 36

* validated parental questionnaire
* validated parental questionnaire
* extensive neurobehavioral testing

* short survey by phone
**Prospecive Cohort of Pregnant Women**

**Prior PTB**

**Screening**

**Data Collection #1**
- US assessment of cervix*
- Cervicovaginal secretions (Protein)
- Endocervical and ectocervix cell collection (DNA)
- **4.** Urine Sample (Biomarker)
- **5.** Maternal Serum Sample (Biomarker)

**Data Collection #2**
- +/- Use of Progestational agents

**Delivery**

**OUTCOME**

1. Gestational age at delivery
2. Neonatal information

**EXPOSURE:**

Premature Cervical Remodeling
Revised graphic

Prospective Cohort of Pregnant Women

Screening

Data Collection

Outcome

Prior PTB

± Use of Progestational Agents

Nullipara, Multipara

Weeks: 5 10 15 20 25 30 35 40

Data Collection #1

US assessment of cervix*
Cervicovaginal secretions (Protein)
Endocervical and ectocervix cell collection (RNA)
Urine Sample (Biomarker)
Maternal Serum Sample (Biomarker)

Data Collection #2

US assessment of cervix*
Cervicovaginal secretions (Protein)
Endocervical and ectocervix cell collection (RNA)
Urine Sample (Biomarker)
Maternal Serum Sample (Biomarker)

Outcome

Gestational age at delivery
Neonatal information

EXPOSURE: Premature Cervical Remodeling
Graphic revised again

**Screening**
- Prospective Cohort of Pregnant Women
- Singleton pregnancies (nullipara, multipara)

**Outcome**
- Gestational age at delivery
- Neonatal information

**Data Collection**
- Maternal DNA
- Maternal blood
- CVF
- “SACC”
- Urine sample
- Screening for Bacterial Vaginosis
All different scales

Too much reading
Same scale
Tells a story visually
It’s important!

You need to budget your time.

Figures and tables are the take home message.
This is how people will respond to your work.
Software Resources

• **ggplot2 in R**
  – [http://had.co.nz/ggplot/](http://had.co.nz/ggplot/),
  – [www.r-project.org](http://www.r-project.org)

Human Resources Biostatistics Cores

• Center for Translational Science Award, ITMAT
  - Kate Propert, ScD, propert@mail.med.upenn.edu

• Cancer Center
  – Dan Heitjan, PhD, heitjan@mail.med.upenn.edu

• IDDRC
  – Mary Putt, ScD, mputt@upenn.edu
Weljie Data

• Reimer et al. 2012. Satiety hormone and metabolomic response to an intermittent high energy diet differs in rats consuming diets high in protein or pre-biotic fiber. Journal of Proteome Research
• Audience: Journal readers
• Question: The idea here was to be able to correlate dietary intake (high fiber, high protein, or control diets) with specific metabolites and/or proteins.
• Explanation: Each large hexagon (green, peach or brown color) represents the weighted contribution of the underlying metabolite/protein measurements, while the small points are the variables themselves.
e_analysis_Nov_23_classes_norm2_v13_2012_1.M45 (OPLS-DA), fasted final model, proteins as X
Colored according to model terms
Colored according to Obs ID (Diet)

R2X[1] = 0.248  R2X[2] = 0.191