Mastering variation: Variance components and personalised medicine

Stephen Senn
Acknowledgements

Many thanks for the invitation
This work is partly supported by the European Union’s 7th Framework Programme for research, technological development and demonstration under grant agreement no. 602552. “IDEAL”
but not necessarily at University Challenge
Genes, Means and Screens

It will soon be possible for patients in clinical trials to undergo genetic tests to identify those individuals who will respond favourably to the drug candidate, based on their genotype…. This will translate into smaller, more effective clinical trials with corresponding cost savings and ultimately better treatment in general practice. … individual patients will be targeted with specific treatment and personalised dosing regimens to maximise efficacy and minimise pharmacokinetic problems and other side-effects.

Sir Richard Sykes, FRS, 1997

My emphasis
Basic thesis

• Both sides of the regulatory divide are convinced that there is a strong element of personal response to treatment
• The truth is that nobody knows
• This is because we statisticians have failed to teach others about components of variation
• And some of us have failed to learn about components of variation also
Seven key aspects

1. Better communication of the problems by statisticians to their colleagues
2. *Application of decision analysis to determine when personalisation is worth pursuing*
3. Appropriate design for teasing out components of variation
4. Application of random effect methodology for improving estimates
5. Translating from additive to relevant scales
6. Application of Deming’s ideas to understanding the system
7. *Realistic monitoring and feedback*
Zombie statistics 1
Percentage of non-responders

What the FDA says

Paving the way for personalized medicine, FDA Oct2013

Where they got it

Spear, Heath-Chiozzi & Huff, Trends in Molecular Medicine, May 2001

Table 1. Response rates of patients to a major drug for a selected group of therapeutic areas

<table>
<thead>
<tr>
<th>Therapeutic area</th>
<th>Efficacy rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alzheimer’s</td>
<td>30</td>
</tr>
<tr>
<td>Analgesics (Cox-2)</td>
<td>80</td>
</tr>
<tr>
<td>Asthma</td>
<td>60</td>
</tr>
<tr>
<td>Cardiac Arrhythmias</td>
<td>60</td>
</tr>
<tr>
<td>Depression (SSRI)</td>
<td>62</td>
</tr>
<tr>
<td>Diabetes</td>
<td>57</td>
</tr>
<tr>
<td>HCV</td>
<td>47</td>
</tr>
<tr>
<td>Incontinence</td>
<td>40</td>
</tr>
<tr>
<td>Migraine (acute)</td>
<td>52</td>
</tr>
<tr>
<td>Migraine (prophylaxis)</td>
<td>50</td>
</tr>
<tr>
<td>Oncology</td>
<td>25</td>
</tr>
<tr>
<td>Osteoporosis</td>
<td>48</td>
</tr>
<tr>
<td>Rheumatoid arthritis</td>
<td>50</td>
</tr>
<tr>
<td>Schizophrenia</td>
<td>60</td>
</tr>
</tbody>
</table>
Zombie statistics 2

Where they got it

Where those who got it got it

Table 1. Response rates of patients to a major drug for a selected group of therapeutic areas.

<table>
<thead>
<tr>
<th>Therapeutic area</th>
<th>Efficacy rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alzheimer’s</td>
<td>30</td>
</tr>
<tr>
<td>Analgesics (Cox-2)</td>
<td>80</td>
</tr>
<tr>
<td>Asthma</td>
<td>60</td>
</tr>
<tr>
<td>Cardiac Arrythmias</td>
<td>60</td>
</tr>
<tr>
<td>Depression (SSRI)</td>
<td>62</td>
</tr>
<tr>
<td>Diabetes</td>
<td>57</td>
</tr>
<tr>
<td>HCV</td>
<td>47</td>
</tr>
<tr>
<td>Incontinence</td>
<td>40</td>
</tr>
<tr>
<td>Migraine (acute)</td>
<td>52</td>
</tr>
<tr>
<td>Migraine (prophylaxis)</td>
<td>50</td>
</tr>
<tr>
<td>Oncology</td>
<td>25</td>
</tr>
<tr>
<td>Osteoporosis</td>
<td>48</td>
</tr>
<tr>
<td>Rheumatoid arthritis</td>
<td>50</td>
</tr>
<tr>
<td>Schizophrenia</td>
<td>60</td>
</tr>
</tbody>
</table>


(c) Stephen Senn
The Real Truth

- These are zombie statistics
- They refuse to die
- Not only is the FDA’s claim not right, it’s not even wrong
- It’s impossible to establish what it might mean even if it were true
88.2% of all statistics are made up on the spot

Vic Reeves
The Pharmacogenomic Revolution?

• Clinical trials
  – Cleaner signal
  – Non-responders eliminated

• Treatment strategies
  – “Theranostics”

• Markets
  – Lower volume
  – Higher price per patient day
Implicit Assumptions

• Most variability seen in clinical trials is genetic
  – Furthermore it is not revealed in obvious phenotypes
    • Example: height and forced expiratory volume (FEV₁) in one second
    • Height predicts FEV₁ and height is partly genetically determined but you don’t need pharmacogenetics to measure height

• We are going to be able to find it
  – Small number of genes responsible
  – Low (or no) interactive effects (genes act singly)
  – We will know where to look

• We are going to be able to do something about it
  – May require high degree of dose flexibility

• In fact we simply don’t know if most variation in clinical trials is due to individual response let alone genetic variability
## Sources of Variation in Clinical Trials

<table>
<thead>
<tr>
<th>Label</th>
<th>Source</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Between treatments</td>
<td>The difference between treatments averaged over all patients</td>
</tr>
<tr>
<td>B</td>
<td>Between patients</td>
<td>The difference between patients given the same treatment</td>
</tr>
<tr>
<td>C</td>
<td>Patient-by-Treatment Interaction</td>
<td>The extent to which the effect of treatment varies from patient to patient</td>
</tr>
<tr>
<td>D</td>
<td>Within patients</td>
<td>The extent to which the results vary from occasion to occasion for patients given the same treatment</td>
</tr>
</tbody>
</table>

## Identifiability and Clinical Trials

<table>
<thead>
<tr>
<th>Type of Trial</th>
<th>Description</th>
<th>Identifiable Effects</th>
<th>Error Term</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parallel</td>
<td>Each patient is randomised to receive one treatment</td>
<td>A</td>
<td>B+C+D</td>
</tr>
<tr>
<td>Cross-over</td>
<td>Each patient receives each treatment in one period only</td>
<td>A and B</td>
<td>C+D</td>
</tr>
<tr>
<td>Repeated cross-overs</td>
<td>Each patient receives each treatment in at least two periods</td>
<td>A and B and C</td>
<td>D</td>
</tr>
</tbody>
</table>
Counterfactual experiment

Patient

Outcome value

Left-hand panel: what you would see if patients could be treated both ways
Note how difference active-placebo is constant

Placebo
Active
A Thought Experiment

• Imagine a cross-over trial in hypertension
• Patients randomised to receive ACE II inhibitor or placebo in random order
• Then we do it again
• Each patient does the cross-over twice
• We can compare each patient’s response under ACE II to placebo twice

(c) Stephen Senn
## Design

<table>
<thead>
<tr>
<th>Sequence</th>
<th>First Cross-over</th>
<th>Second Cross-over</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Period</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>1 A 2 B</td>
<td>3 A 4 B</td>
</tr>
<tr>
<td>II</td>
<td>1 B 2 A</td>
<td>3 B 4 A</td>
</tr>
<tr>
<td>III</td>
<td>1 A 2 B</td>
<td>3 B 4 A</td>
</tr>
<tr>
<td>IV</td>
<td>1 B 2 A</td>
<td>3 A 4 B</td>
</tr>
</tbody>
</table>
Patients are treated in two cross-over trials, thus permitting two estimates of the difference between active treatment and placebo. The difference on the second occasion is plotted against the first. Blue = response on both occasions, red = non-response on both occasions, orange = response on one occasion but not the other.

The marginal distributions are given as green histograms. LHS response on first occasion predicts response on second. RHS response on first occasion does **not** predict response on second. If you had only carried out one cross-over you would have the picture below. Which case does it apply to?
31 trials of cimetidine in ulcer healing

Data from Moerman, 1983
Significant differences in favor of tiotropium were observed at all time points for the mean absolute change in the SGRQ total score (ranging from 2.3 to 3.3 units, P<0.001), although the differences on average were below what is considered to have clinical significance (Fig. 2D). The overall mean between-group difference in the SGRQ total score at any time point was 2.7 (95% confidence interval [CI], 2.0 to 3.3) in favor of tiotropium (P<0.001). A higher proportion of patients in the tiotropium group than in the placebo group had an improvement of 4 units or more in the SGRQ total scores from baseline at 1 year (49% vs. 41%), 2 years (48% vs. 39%), 3 years (46% vs. 37%), and 4 years (45% vs. 36%) (P<0.001 for all comparisons).

(My emphasis)
Two Normal distributions with the same spread but the Active treatment has a mean 2.7 higher. If this applies every patient under active can be matched to a corresponding patient under placebo who is 2.7 worse off.
A cumulative plot corresponding to the previous diagram.

If 4 is the threshold, placebo response probability is 0.36, active response probability is 0.45.
In summary...this is rather silly

• If there is sufficient measurement error even if the true improvement is identically 2.7, some will show an ‘improvement’ of 4

• The conclusion that there is a higher proportion of true responders by the standard of 4 points under treatment than under placebo is quite unwarranted

• So what is the point of analysing ‘responders’?
Who are the authors?


Personal note. I am proud to have been involved in this important study and have nothing but respect for my collaborators. The fact that, despite the fact that two of us are statisticians, we have ended up publishing something like this shows how deeply ingrained the practice of responder analysis is in medical research. We must do something to change this.
Beware of clinically relevant measures

• Such measures may not have good statistical properties
• There may be variation in terms of these measures that can be removed by transformation
• Analysis is better on the ‘additive’ scale
• Back transformation can be used for prediction
An Example from Research

- Four experimental p38α kinase inhibitors
- Vehicle and marketed product as controls
- Thrombaxane B2 (TXB2) is used as a marker of COX-1 activity
- Six rats per group were treated for a total of 36 rats
- At the end of the study rats are sacrificed and TXB2 is measured.
TXB2 in rats for six treatments
log-TXB2 in rats for six treatments

\[ \text{vehicle} \rightarrow a, b, c, d, \text{marketed} \]
## Analyses of Original and Transformed Data

### Original

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treat</td>
<td>5</td>
<td>184596</td>
<td>36919</td>
<td>6.31</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Error</td>
<td>30</td>
<td>175439</td>
<td>5848</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>35</td>
<td>360035</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Transformed

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treat</td>
<td>5</td>
<td>62.7</td>
<td>12.54</td>
<td>40.1</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Error</td>
<td>30</td>
<td>9.4</td>
<td>0.31</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>35</td>
<td>72.1</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Untransformed

*** Bartlett's Test for homogeneity of variances ***

Chi-square  50.87  on 5 degrees of freedom: probability < 0.001

### Log-transformed

*** Bartlett's Test for homogeneity of variances ***

Chi-square  8.95  on 5 degrees of freedom: probability 0.111
In the Meantime

• There is a massive source of unwanted variation

• Doctors

• Variation in practice is so large that it cannot be justified by variation in patients

• This is the basic idea behind the way that Intermountain Health under the leadership of Brent James has been applying Deming’s principles to health care
Tonsillectomy rate for England by local authority

3 years: 2009-2011, persons aged under 15

Raw estimates and 95% confidence intervals

Source: Unit of Health-Care Epidemiology
University of Oxford, 2013
Raw and shrunk tonsillectomy rate by UK local authority

Mean
Ratio of maximum to minimum = 4.7

Maximum

Mean

Source: Unit of Health-Care Epidemiology
University of Oxford, 2013

Persons aged under 15 per 100,000
Period 01/01/2009 to 31/12/2011
“Guys, it’s more important that you do it the same way than what you think is the right way.”

Brent James, Advice to doctors
Giving this medicine to children:
It is important to know how much your child weighs to make sure you give them the correct amount of medicine. As a guide a child of 9 years of age will weigh about 30 kg (four and a half stone). If in doubt weigh your child, then follow the instructions in the table.

Do not give to children who weigh less than 30 kg.
Do not give to children under 2 years.

<table>
<thead>
<tr>
<th>Age</th>
<th>How many to take</th>
<th>How often to take</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adults and children of 12 years and over</td>
<td>One tablet</td>
<td>Once a day</td>
</tr>
<tr>
<td>Children of 2 to 11 years who weigh more than 30 kg</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Children of 2 to 11 years who weigh less than 30 kg</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Who’s to blame?

• Statisticians
  – (Including me)
• Our life scientist colleagues don’t understand variation
• We do
• We should tell them the truth

(c) Stephen Senn
Advice

• Don’t let the label ‘responder’ infect your brain
• A ‘responder’ is a patient who was observed to get better by some arbitrary standard
• A ‘responder’ is not a patient who was caused to get better by the drug
• Subsequence is not consequence
• To establish who really responds and who does not you need to work very hard
The supply of truth always greatly exceeds its demand

John F Moffitt