Individuals and the group: the challenge of “personalized medicine”

*Why do we need statistics in medicine?*
Because people are *individuals*
(not because they’re all alike).

*It’s all about VARIABILITY*

*On the other hand…*
“I take comfort in thinking of myself as a statistic”
---Nan Laird, Former Chair, Dep’t of Biostats, Harvard School of Public Health

… *What a medicine does to me* teaches *something* about what it does to *you.*
The Lump/Split Dilemma

A new treatment is given to 100 patients. Of them, only 8 respond. But there is a subgroup of 5 in which 3 patients respond, yielding a response rate of 60%!
Should the treatment be recommended for people in the subgroup? (“Personalized”?)

<table>
<thead>
<tr>
<th>Responder (R)</th>
<th>Dark Hair (D)</th>
<th>Light Hair (L)</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonresponder (N)</td>
<td>3</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>TOTAL</td>
<td>5</td>
<td>95</td>
<td>100</td>
</tr>
</tbody>
</table>

Lump? Split?

- Dr. Lump:
  “Of course hair color has nothing to do with it. Response rate = 8/100.
  **Don’t treat** the Dark Hair people!”

- Dr. Split:
  “The latest research says *personalize.* Response rate = 3/5=60%.
  **Treat** the Dark Hair people!”
Some Bayesian analyses

Dr. I. DontKnow: Let the data tell us...
Goldilocks and the three investigators

- Dr. Lump: low variance, high bias
- Dr. Split: high variance, low bias
- Dr. I.Dontknow: JUST…right!
  – Empirical Bayes, hierarchical model

The challenge for the future of medicine:
Let DATA + PRIOR UNDERSTANDING dictate how much to “personalize”.
CLINICAL TRIALS!
Drug development process (idealized)

IDEA

- Molecular
- In vitro
- In vivo
- Epidemiologic
  - “File cabinet”

Phase I

OBJ: “Safety”
ENDPT: Toxicity

Phase II

OBJ: “Efficacy”
ENDPT: Clinical response

Phase III

OBJ: “Effectiveness”
ENDPT: Survival

Phase IV

OBJ: “Outcomes”
ENDPT: Pain, cost, ...

Phase I Clinical Trials: Objectives

OBJECTIVES:
Identification of toxicities to watch out for.
Determination of a “Recommended Phase II Dose”.

DEFINITION:
Maximum tolerated dose (MTD):
“The highest level of a dose that can be tolerated”
(“Tolerated” = an “acceptable” risk of toxicity)

DEFINITION:
Dose-limiting toxicity (DLT):
(1) “an adverse event that is counted against dose escalation”
(2) “a type of adverse event associated with the drug being tested”
Severity grades:

1  Mild
2  Moderate

DLT  3  Severe
DLT  4  Life-threatening
DLT  5  Fatal

The U.S. National Cancer Institute:

Common Toxicity Criteria 1982; CTC Version 2.0 1998;
Common Terminology Criteria for Adverse Events v3.0 (CTCAE)
(an informatics data conversion nightmare)

Phase I: Standard Design (3+3)

3 patients per dose tier
If #DLT = ...

\[\begin{align*}
0/3 & , \text{ then } \textbf{Escalate} \text{ the dose} \\
1/3 & , \text{ then add 3 pts} \\
& \text{if } 1/6, \text{ then } \textbf{Escalate} \\
& \text{if any more, then } \textbf{Stop} \\
2/3 & , \text{ then } \textbf{Stop}
\end{align*}\]

(many better methods, but still the most popular)
Development of a Phase I Dose from a Dose-Ranging Study

<table>
<thead>
<tr>
<th>Steps</th>
<th>Value</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>LD10 in mouse</td>
<td>75 mg/kg</td>
<td>dose which is lethal in 10% of mice</td>
</tr>
<tr>
<td>Divide by 12</td>
<td>6.25 mg/kg</td>
<td>conversion from kg/m² in mice to kg/m² in humans, ( \frac{0.03/0.032}{0.07/0.032} ) = 1/13.3</td>
</tr>
<tr>
<td>Multiple by 70/1.7</td>
<td>250 mg/m²</td>
<td>Convert to m² (people)</td>
</tr>
<tr>
<td>Divide by 10</td>
<td>25 mg/m²</td>
<td>Safety factor</td>
</tr>
</tbody>
</table>

Let’s play Phase I Trial!

You are the Patient.
What’s a “good” toxicity rate?
What’s a bad one?

“too toxic, no matter what!”

“not toxic enough—probably won’t work.”

“just right”

proportion of patients who will get adverse events

Phase I study design worksheet

Answers are now updated. Change numbers to explore other possibilities.

Calculate

Study design

Toxicity model

Operating Characteristics

Conclusion | Pr(Conclusion) |
--- | --- |
Stop 1st level | 0.9 |
MTD=1st level | 0.003 |
MTD=2nd level | 0.264 |
MTD=3rd level | 0.324 |
No MTD found | 0.417 |
How much information do we really have about the “maximum tolerated dose”

- If 0 out of 3 DLT’s,
  - estimated risk of DLT is “zero”
  - 95% confidence interval is 0% to 63%.
- If 1 out of 6 DLT’s,
  - estimated risk of DLT is 17%
  - 95% confidence interval is 4% to 64%.
  ---- not much information!!!
Let’s play Phase I Trial!

You are the Patient.

Phase II: Objectives

Main purpose:

Determine whether there is sufficient evidence of efficacy

Secondary:

Determine (or confirm) safety with greater confidence. (Is the “MTD” or “RP2D” really “tolerable”?)

Usually just one regimen.

Often just one drug or other treatment at a time.
Efficacy vs Effectiveness:
Textbook Definitions

- Efficacy
  - “true biological effect of a treatment”

- Effectiveness
  - “the effect of a treatment when widely used in practice”

What’s a “good” response rate?
What’s a bad one?

0.1    0.2    0.3    0.4     0.5    0.6     0.7    0.8

we love
the new treatment!

indifferent
worse than
what we use now!

p0 (poopee?)
p1 (whoopee?)

proportion of patients who will respond
Some Phase II study designs

“Simon design” - Early stopping for poor response

“Bryant-Day design - Early stopping for poor response

or excess toxicity

Phase II: Ethical Issue

If early results say the treatment is not very good, isn’t it UNETHICAL to continue to accrue patients?

Solution: An early stopping rule

a special case of

ADAPTIVE DESIGN.
Study Design Jargon

“Type I Error” (or “alpha”)

The probability that you say “whoopie” when you shouldn’t.

“Type II Error” (or “beta”)

The probability that you say “poopie” … when you shouldn’t.

“Statistical Power”

The probability that you say “whoopie” ... when you should. So ....

\[
\text{Power} = 1 - \beta = 1 - \text{Type II Error}
\]
This might be a "Type I error".

A class of decision rules:

Reject drug if # responses is...

\[ r_1 \text{ out of } n_1 \text{ (first stage)} \] (SHORT WALL)
or...

\[ r \text{ out of } n \text{ (full trial)}, \] (TALL WALL)

A set of criteria:

\[
\begin{align*}
\alpha &= \text{Type I error} < 0.10, \\
\beta &= \text{Type II error} < 0.10.
\end{align*}
\]

Minimize the average sample size if \( p_0 \) is true: \( E(N \mid p_0) \)

Optimal design:

\[
\begin{align*}
\frac{r_1}{n_1} &= \frac{7}{22}, \\
\frac{r}{n} &= \frac{17}{46}, \\
E(N \mid p_0) &= 29.9
\end{align*}
\]
Let's play Phase II Trial!

You are the Principal Investigator.

Phase III studies

Objective:  Comparative, confirmatory analysis

Endpoint:  Clinically directly meaningful & important
           (Survival; Time to Progression; Symptom relief)

Treatment assignment:  Randomized

Control:  “Standard of care” usually

Early stopping:  Evidence for non-equivalence.
                 - New treatment clearly better.
                 - New treatment clearly worse.
What is a “statistic”?

“Test statistic”

\[ S : \{\text{all possible study outcomes}\} \rightarrow \{\text{numbers}\} \]

measures “surprise if the null hypothesis is true”.

P-value = \( \text{Prob}(S \geq s_{\text{observed}} | \text{null hypothesis}) \)

If P-value \( \leq \alpha \), “whoopie”. ("reject the null")
If not, then “poopie”. ("accept the null")

Strength of evidence: the “P-value”

“P=0.01”:

“In a long series of identical trials, if the null hypothesis is true”, such an unusual result as OUR study would only occur once in a hundred trials (0.01 of the time).”

\[ \{ \text{ordering of possible outcomes} \} \rightarrow \{ \text{what is “unusual”} \} \]
Survival Curves for Arms A and B

Risk Ratio = 1.37

p = 0.005

Example due to Sam Wieand

Let’s play Phase III Trial!

You are a figment in statistician’s computer (a SIMULATION).

Except ONE of you is real.
Arm A is the better regimen!

Arm A = 5-FU and LV (1983 Trial)  
(Advanced Colorectal Cancer)

Arm B = 5-FU and LV (1986 Trial)  
(Advanced Colorectal Cancer)

Historical controls can be misleading…  
This is one reason we randomize!

Stuff we “knew” that ain’t so…  
(animal studies, observational “big data”, …)

**Womens Health Initiative**
The unopposed estrogen trial was halted in February 2004, after an average follow-up period of 6.8 years, on the basis that unopposed estrogen does not appear to affect the risk of heart disease, the primary outcome, which was in contrast to the findings of previous observational studies. On the other hand, there were indications for an increased risk of stroke.

**The Effect of Vitamin E and Beta Carotene on the Incidence of Lung Cancer and Other Cancers in Male Smokers**
Unexpectedly, we observed a higher incidence of lung cancer among the men who received beta carotene than among those who did not.

Randomized trials, hooray!!
Advanced Colorectal Cancer

SYMPTOM STATUS

<table>
<thead>
<tr>
<th></th>
<th>NO</th>
<th>YES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arm A</td>
<td>73 (46%)</td>
<td>86 (54%)</td>
</tr>
<tr>
<td>Arm B</td>
<td>61 (33%)</td>
<td>122 (67%)</td>
</tr>
</tbody>
</table>

Maybe this explains why Arm A did better. Let’s do some statistical magic!

Advanced Colorectal Cancer

Cox Proportional Hazards Model

“Controlling for risk factors”

<table>
<thead>
<tr>
<th>Variable</th>
<th>Risk Ratio</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treat B</td>
<td>1.32</td>
<td>0.02</td>
</tr>
<tr>
<td>Age &gt; 70</td>
<td>1.31</td>
<td>0.06</td>
</tr>
<tr>
<td>PS &gt; 0</td>
<td>1.33</td>
<td>0.04</td>
</tr>
<tr>
<td>Measurable</td>
<td>1.27</td>
<td>0.05</td>
</tr>
<tr>
<td>Grade &gt; 1</td>
<td>1.79</td>
<td>0.02</td>
</tr>
<tr>
<td>Symptoms</td>
<td>1.39</td>
<td>0.01</td>
</tr>
</tbody>
</table>
### Advanced Colorectal Cancer

**Cox Proportional Hazards Model**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Risk Ratio</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment B effect … no “adjustment”</td>
<td>1.37</td>
<td>0.005</td>
</tr>
<tr>
<td>Treatment B effect … “adjusting for covariates”</td>
<td>1.32</td>
<td>0.02</td>
</tr>
</tbody>
</table>

---

### Confounding

**DUKES’ B**

<table>
<thead>
<tr>
<th></th>
<th>Mayo Clinic Treatment A</th>
<th>MD Anderson Treatment B</th>
</tr>
</thead>
<tbody>
<tr>
<td># Patients</td>
<td>60</td>
<td>30</td>
</tr>
<tr>
<td># Deaths</td>
<td>12</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>(20%)</td>
<td>(16.7%)</td>
</tr>
</tbody>
</table>

*Better!*
### Confounding

**DUKES’ C**

<table>
<thead>
<tr>
<th></th>
<th>Mayo Clinic</th>
<th>MD Anderson</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Treatment A</strong></td>
<td># Patients</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td># Deaths</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>(75%)</td>
<td>(63%)</td>
</tr>
</tbody>
</table>

**Treatment B**

|                  | 70          | 44          |
|                  | (63%)       | BETTER!     |

### Confounding

**ALL PATIENTS**

<table>
<thead>
<tr>
<th></th>
<th>Mayo Clinic</th>
<th>MD Anderson</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Treatment A</strong></td>
<td># Patients</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td># Deaths</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BETTER!</td>
</tr>
</tbody>
</table>

|                  | 100         | 49          |

|                  |             |             |
Confounding – “Simpson’s paradox”

<table>
<thead>
<tr>
<th></th>
<th>Treatment A</th>
<th>Treatment B</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dukes’ B</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td># Patients</td>
<td>60</td>
<td>30</td>
</tr>
<tr>
<td># Deaths</td>
<td>12 (20%)</td>
<td>5 (16%)</td>
</tr>
<tr>
<td><strong>Dukes’ C</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td># Patients</td>
<td>40</td>
<td>70</td>
</tr>
<tr>
<td># Deaths</td>
<td>30 (75%)</td>
<td>44 (63%)</td>
</tr>
<tr>
<td><strong>All Patients</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td># Patients</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td># Deaths</td>
<td>42 (42%)</td>
<td>49 (49%)</td>
</tr>
</tbody>
</table>

Afterthoughts

- New ideas in clinical trial design are growing rapidly!
- Clinical trials should become more ethical.
- “Personalized medicine” – a crisis looming.
  - Explosion of “features”.
  - People “like me” – fewer and fewer.
  - Sample sizes smaller, but effect sizes bigger (we hope).