Statistical Issues in Drug Development

Recap of short course from 2011 FDA/Industry Statistics Workshop

Presented by:
Clay Ford, M.S.
8 Nov 2011

The short course

- Presented by Stephen Senn
- "To take some standard topics and look at them afresh"
  - To make the strange familiar
  - To make the familiar strange
- "To provide some suggestions as to how we can present them to medical colleagues"

Warnings

- The lecture was not statistical but meta-statistical
- Stephen Senn's personal view
- You may not agree with his point of view
- He accepts your right to disagree…
  …anybody can make mistakes

Selected Topics

- Causality
- Sources of bias
- Bias-Variance trade-off
- Randomization
- Dichotomizing continuous measurements

Causality

- It would seem commonsense we judge the causal effect of treatment by comparing after with before (i.e., outcome with baseline)
- However, not essential for judging causality
- Example
  - 10 year MS trial
  - subjects are same at end of trial as at beginning
  - Is the effect zero?

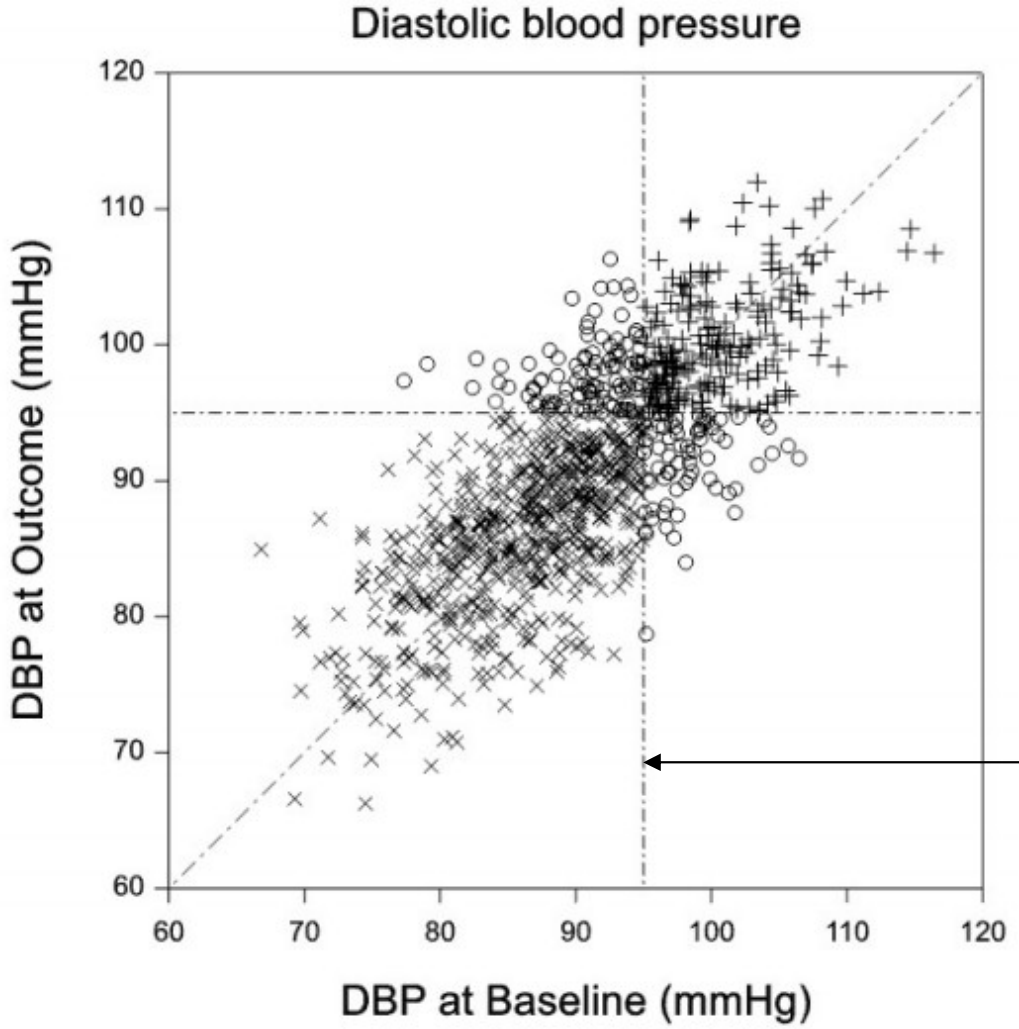
The Counterfactual Approach

- The effect of any treatment for a given patient is the difference between…
  - what happened to the patient as a result of giving the treatment and
  - what *would have happened* had treatment been denied
- That's why we have control groups
- Comparison to baseline *in each group* is a very poor method of comparison

Baseline comparison bias

- Time trends
  - The previous MS example
- Regression to the Mean
  - if subject selected for trial because of extreme measured value…
  - …may see spontaneous movement toward mean and erroneously attribute it to other factors

Regression to the mean



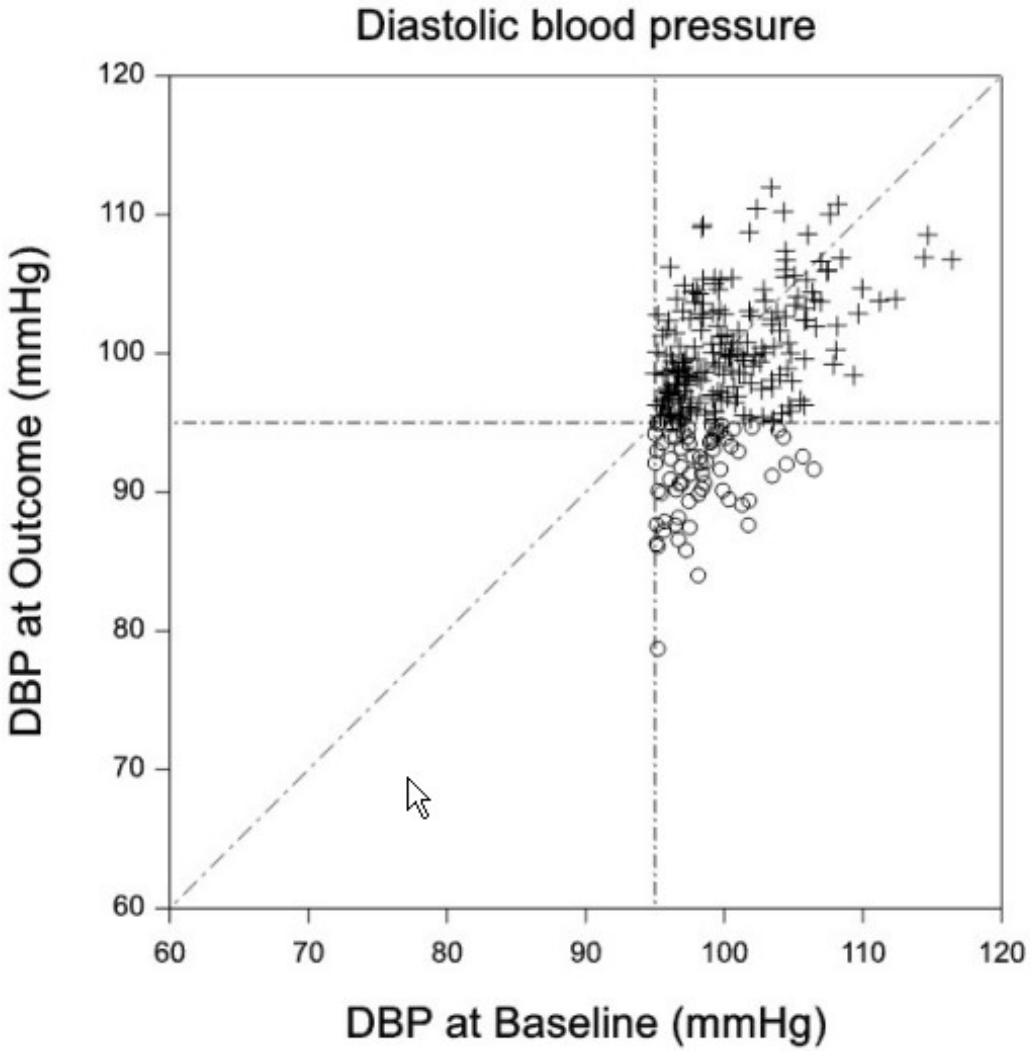
Simulated data for DBP (n = 1000)

Mean is about 90 for both baseline and outcome

Correlation about 0.77

Possible criterion for entry in trial (patients suspected to be hypertensive)

Regression to the mean (2)



Same plot, but with <95 at baseline removed

No restriction that DBP <95 at outcome

Mean of 100 at baseline

Mean of 98 at outcome

The Controlled Clinical Trial

- Addresses "what might have been"
  - study at the same time under similar conditions similar group of patients given control treatment
- Minimize biases
- Allows us to make inferences about the difference between treatment and control
- But controlled clinical trial has implications (that trialists often do not understand)

Between vs. Within Groups

Trialists sometimes focus on mean values *within* groups instead of the difference *between* groups

- Placebo-controlled trials with separate estimates of treatment effect in each group
  - Why use a control group?
  - Why not allocate all patients to active treatment?
- Report standard error of treatment means
  - SE = SD/$\sqrt{n}$ (*assumes simple random sample*)
  - But SRS never occurs in a clinical trial
- The analysis of Clinical Trials should estimate treatment differences

The TARGET Study

- Large osteoarthritis study (18,000 patients)
- Therapeutic Arthritis Research & Gastrointestinal Event Trial (TARGET)
- Purpose: investigate CV and GI tolerability of lumiracoxib
- Randomization took place in two sub-studies of equal size:
  - Lumiracoxib versus ibuprofen
  - Lumiracoxib versus naproxen

The TARGET Study (2)

- Treatment details
  - Lumiracoxib 400 mg once daily
  - Ibuprofen 800 mg three times daily
  - Naproxen 500 mg twice daily
- Complicated blinding scheme if all three in a single study
- To simplify, trial run as two substudies with same protocol
- Randomization within substudy (*not between substudies*)

The TARGET Study (3)

- Distribution of patients at baseline by four binary demographic characteristics
- Excellent balance within substudies
- Poor balance between substudies

Demographic characteristics	Sub-study 1		Sub-study 2	
	Lumiracoxib $n = 4376$	Ibuprofen $n = 4397$	Lumiracoxib $n = 4741$	Naproxen $n = 4730$
Use of low-dose aspirin	975 (22%)	966 (22%)	1195 (25%)	1193 (25%)
History of vascular disease	393 (9%)	340 (8%)	588 (12%)	559 (12%)
Cerebrovascular disease	69 (2%)	65 (1%)	108 (2%)	107 (2%)
Dyslipidaemias	1030 (24%)	1025 (23%)	799 (17%)	809 (17%)

The TARGET Study (4)

- If all treatment arms are considered together this is *not* a randomized study
- In a randomized study we only expect to find chance differences in demographic characteristics between treatment arms
- Check this by
  - analyzing baseline demographic dichotomies using logistic regression (all subjects)
  - comparing models using reduction in deviance

The TARGET Study (5)

- The simple Logistic model with no factors posits independence between demographic characteristic and
  - Substudy
  - Treatment
- Deviance: test that simpler model holds

Effect	Aspirin	Vascular history	Cerebrovascular	Dyslipidaemias
Deviances for the four demographic variables				
Substudy	23.57	70.14	13.538	117.98
Treatment given substudy	0.13	5.23	0.144	0.17
Treatment	13.40	47.41	7.745	54.72
Approximate chi-square probabilities for the four demographic variables				
Substudy	0.0000	0.00000	0.0002	0.0000
Treatment given substudy	0.9365	0.07332	0.9304	0.9194
Treatment	0.0012	0.00000	0.0208	0.0000

The TARGET Study (6)

```
data target;
input subst trt d1 d2 d3 d4 total;
datalines;
1 2 975 393 69 1030 4376
1 0 966 340 65 1025 4397
2 2 1195 588 108 799 4741
2 1 1193 559 107 809 4730
;
run;
```

```
proc genmod data=target;
class trt subst;
model d3/total = subst / dist=bin link=logit;
proc genmod data=target;
model d3/total = / dist=bin link=logit;
run;
```

Deviance = – 2*(LogLH Simple – LogLH Fitted)
= –2*(–1726.4657 – –1719.6966) = 13.5382

P(X>13.5382) = 0.0002
1 degree of freedom

Reject null, demographic characteristic not independent of substudy (or treatment)

Effect	Aspirin	Vascular history	Cerebrovascular	Dyslipidaemias
Deviances for the four demographic variables				
Substudy	23.57	70.14	13.538	117.98
Treatment given substudy	0.13	5.23	0.144	0.17
Treatment	13.40	47.41	7.745	54.72
Approximate chi-square probabilities for the four demographic variables				
Substudy	0.0000	0.00000	0.0002	0.0000
Treatment given substudy	0.9365	0.07332	0.9304	0.9194
Treatment	0.0012	0.00000	0.0208	0.0000

The TARGET Study (7)

- Significant reduction in deviance if treatment or substudy is fitted compared to null model with no factors (i.e., demographic not independent of them)
- We only see chance differences in demographic characteristics between treatment arms *provided we include substudy in model*

Effect	Aspirin	Vascular history	Cerebrovascular	Dyslipidaemias
Deviances for the four demographic variables				
Substudy	23.57	70.14	13.538	117.98
Treatment given substudy	0.13	5.23	0.144	0.17
Treatment	13.40	47.41	7.745	54.72
Approximate chi-square probabilities for the four demographic variables				
Substudy	0.0000	0.00000	0.0002	0.0000
Treatment given substudy	0.9365	0.07332	0.9304	0.9194
Treatment	0.0012	0.00000	0.0208	0.0000

The TARGET Study (8)

```
data target;
input subst trt d1 d2 d3 d4 total;
datalines;
1 2 975 393 69 1030 4376
1 0 966 340 65 1025 4397
2 2 1195 588 108 799 4741
2 1 1193 559 107 809 4730
;
run;
```

```
proc genmod data=target;
class trt subst;
model d3/total = subst trt / dist=bin link=logit;
proc genmod data=target;
model d3/total = subst / dist=bin link=logit;
run;
```

Deviance = – 2*(LogLH Simple – LogLH Fitted)
= –2*(–1719.6966 – –1719.6245) = 0.1442

P(X>0.1442) = 0.9304
2 degrees of freedom

Fail to reject null, treatment effect not significant given substudy (i.e., only chance difference between treatments given substudy)

Effect	Aspirin	Vascular history	Cerebrovascular	Dyslipidaemias
Deviances for the four demographic variables				
Substudy	23.57	70.14	13.538	117.98
Treatment given substudy	0.13	5.23	0.144	0.17
Treatment	13.40	47.41	7.745	54.72
Approximate chi-square probabilities for the four demographic variables				
Substudy	0.0000	0.00000	0.0002	0.0000
Treatment given substudy	0.9365	0.07332	0.9304	0.9194
Treatment	0.0012	0.00000	0.0208	0.0000

The TARGET Study (9)

- Implications
  - Need substudy in the analysis model first with treatment as a subsequently added factor
  - Direct comparison between ibuprofen and naproxen is not possible (potentially biased)
  - The unbiased contrasts are not the efficient contrasts

The TARGET Study (10)

- Unbiased comparison of ibuprofen and naproxen (double contrast):
  - Find difference to lumiracoxib in each study
  - Compare these two differences to each other
- But this double unbiased contrast would have variance twice that of the single (potentially biased) contrast of naproxen and ibuprofen

The TARGET Study (11)

- Suppose there $4n$ patients
  - $2n$ on lumiracoxib (n in each study)
  - n each on ibuprofen and naproxen
- Recall estimated variance of contrast estimator:
  - $\text{MSE}\sum(c_i^2 / n_i)$

Senn calls this the variance multiplier

The TARGET Study (12)

- Est. variance of unbiased double contrast:

$$MSE\left[\frac{(1)^2}{n}+\frac{(-1)^2}{n}+\frac{(1)^2}{n}+\frac{(-1)^2}{n}\right]=MSE\,\frac{4}{n}$$

- Est. variance of potentially biased single contrast (efficient):

$$MSE\left[\frac{(1)^2}{n}+\frac{(-1)^2}{n}\right]=MSE\,\frac{2}{n}$$

Ratio of variances: 2

In an observational study, this would be a dilemma;
In a clinical trial, we use the unbiased estimate.

The TGN1412 Study

- First-in-man study of TeGenero's monoclonal antibody TGN1412
- Carried out by CRO Parexel
- 8 volunteers in first cohort (3:1 ratio)
  - 6 on TGN1412
  - 2 on Placebo
- All 6 given TGN1412 suffered a cytokine storm and admitted to ICU on first day

The TGN1412 Study (2)

- Conventional analysis using Fisher's Exact Test
  - One-sided p-value=0.0357
  - Not significant at 2.5% level
- Alternate analysis using Barnard's Unconditional Test
  - One-sided p-value=0.0111

	Cytokine storm		
Treatment	No	Yes	Total
Placebo	2	0	2
TGN1412	0	6	6
Total	2	6	8

Senn:
"Both P-values are clearly irrelevant."

Digression: that Barnard test

- Recall: Fisher's Exact Test conditions on both row and column margin totals
- Barnard's test does not condition on column totals
- Perhaps better in trial setting where treatment totals are fixed (the row margin) but the response totals are not (the column margin)
- More info: http://www.mcardle.wisc.edu/mstat/barnard/barnard.html
- Barnard GA. Significance Tests for 2x2 Tables. *Biometrika*, Vol 34, No. 1/2 (Jan. 1947), pp. 123-138.

The TGN1412 Study (3)

- The p-values were irrelevant
  - Background risk of cytokine storm is so low that it makes one occurrence, let alone six, indicative of causal relationship
  - All six occurred within a few hours of each other
- "When several individuals suffer severe cytokine storms within a few hours of administering a monoclonal antibody, you do not need a formal statistical analysis to tell you what happened. "
- Lesson: Confining inferences to those permitted by formal controlled analyses is too limiting

The TGN1412 Study (4)

	TGN1412		Placebo
Cohort	Dose [mg/kg bodyweight]	Number of Subjects	Number of Subjects
1	0.1	6	2
2	0.5	6	2
3	2.0	6	2
4	5.0	6	2

Never got to cohorts 2-4

- Why the 3:1 ratio? (Not explained in protocol)
- "This ratio might be viewed as inefficient if the object of the trial is to compare doses with each other or with the placebo."
  - *Report of the Working Party on Statistical Issues in First-in-Man studies.* RSS: London, 2007.

The TGN1412 Study (5)

Design	Number of subjects				Variance of differences between doses, and between placebo and each dose, if a cohort effect is fitted				Variance of differences between doses, and between placebo and each dose, if it is known that there is no cohort effect				
1	Dose	0	1	2	3		1	2	3		1	2	3	Original design
	Cohort 1	2	6	0	0	0	0.67	0.67	0.67	0	0.33	0.33	0.33	
	Cohort 2	2	0	6	0	1		1.33	1.33	1		0.33	0.33	
	Cohort 3	2	0	0	6	2			1.33	2			0.33	
2	Dose	0	1	2	3		1	2	3		1	2	3	Different design
	Cohort 1	4	4	0	0	0	0.50	0.50	0.50	0	0.33	0.33	0.33	
	Cohort 2	4	0	4	0	1		1.00	1.00	1		0.50	0.50	
	Cohort 3	4	0	0	4	2			1.00	2			0.50	

Compare dose 1 to dose 2 (with cohort effect)

Design 1: $1/2 + 1/6 + 1/2 + 1/6 = 8/6 = 1.33\sigma^2$

Design 2: $1/4 + 1/4 + 1/4 + 1/4 = 4/4 = 1.00\sigma^2$

Compare dose 1 to dose 2 (no cohort effect)

Design 1: $1/6 + 1/6 = 2/6 = 0.33\sigma^2$

Design 2: $1/4 + 1/4 = 2/4 = 0.50\sigma^2$

The TGN1412 Study (6)

- If trial had proceeded, it would have had the structure of a *veiled* trial:
  - Patients do not know which treatment they are receiving, but they do know there is a possible treatment they are *not* receiving
  - Different doses allocated to different cohorts, so subjects knew relative dosage level they were receiving
  - This raises an issue…
  - Example: subjects in cohort 4 may believe they are more likely to suffer a problem of tolerability than do subjects in cohort 1. This psychological bias can be eliminated only by comparing subjects given the highest dose to those given placebo at the same time

The TGN1412 Study (7)

- Back to bias-variance trade-off

Design	Number of subjects	Variance of differences between doses, and between placebo and each dose, if a cohort effect is fitted	Variance of differences between doses, and between placebo and each dose, if it is known that there is no cohort effect

1

Dose	0	1	2	3
Cohort 1	2	6	0	0
Cohort 2	2	0	6	0
Cohort 3	2	0	0	6

	1	2	3
0	0.67	0.67	0.67
1		1.33	1.33
2			1.33

Unbiased,
but not efficient

	1	2	3
0	0.33	0.33	0.33
1		0.33	0.33
2			0.33

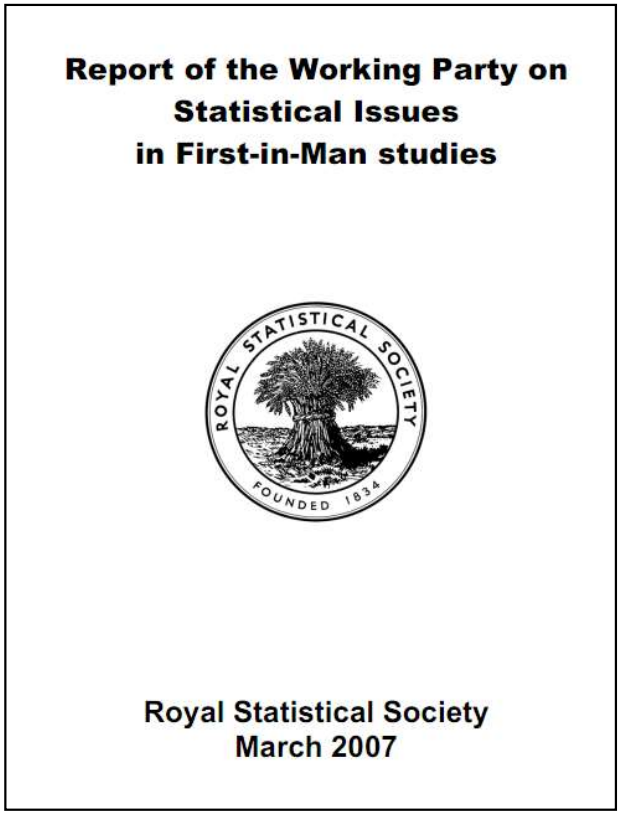
Biased,
but efficient

The biased but efficient design was the plan for TGN1412. The reduction in variance may compensate for the bias, but it was not addressed in protocol.

The TGN1412 Study (8)

- For more info on the TGN1412 study:

http://www.rss.org.uk/uploadedfiles/documentlibrary/736.pdf



Randomization

- What randomization does:
  - Deal validly with effect of unmeasured covariates
- What randomization does not do:
  - Guarantee that patients recruited are representative
  - Guarantee balance
  - Allow us to ignore prognostic factors

Three games with dice

- Object: call the probability of getting 10 in rolling two dice (red die and black die)
- Game 1: call probability beforehand, roll the dice
- Game 2: roll red die, see the result, call probability before rolling black die
- Game 3: call the probability, roll red die, don't look at result, then roll the black die

Three games with dice (2)

	Red Die Score					
Black Die Score	**1**	**2**	**3**	**4**	**5**	**6**
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	(10)
5	6	7	8	9	(10)	11
6	7	8	9	(10)	11	12

Game 1:
36 possible combinations
Three yield 10
Prob = 3/36 = 1/12

Three games with dice (3)

Red Die	Probability Total = 10	Odds
1	0	0
2	0	0
3	0	0
4	1/6	1:5
5	1/6	1:5
6	1/6	1:5

Game 2:
If first roll is 1, 2, or 3, then Prob = 0;
If first roll is 4, 5, or 6, then Prob = 1/6

Game 3: Prob = 1/2*0 + 1/2*1/6 = 1/12
No difference between game 1 and game 3.

The point of the game

- Result of red die: distribution at baseline once a clinical trial has been randomized
- Total score of dice: result observed at end of trial
- Game 1: a trial in which baseline variables are not observed
- Game 2: a trial in which baseline variables are observed
- Game 3: same as game 1

Misunderstandings

- Treating game 1 and game 3 differently
  - Query the validity of a correctly calculated treatment estimate because of unmeasured covariates
- Treating game 2 like game 1
  - Argue that measured covariates can be ignored because one has randomized

Example

- Observe a baseline difference of 3 mmHg in BP between treatment groups and a difference of 8 mmHg at outcome
- Don't ask: "Given no treatment effect, with what probability would randomized trial show a difference of 8 mmHg at outcome"
- Instead ask: "Given no treatment effect, with what probability would randomized trial with a difference of 3 mmHg at baseline show a difference of 8 mmHg at outcome"

The moral

- "Trialists continue to use their randomization as an excuse for ignoring prognostic information, and they continue to worry about the effect of factors they have not measured. Neither practice is logical."
  - Stephen Senn

Dichotomania

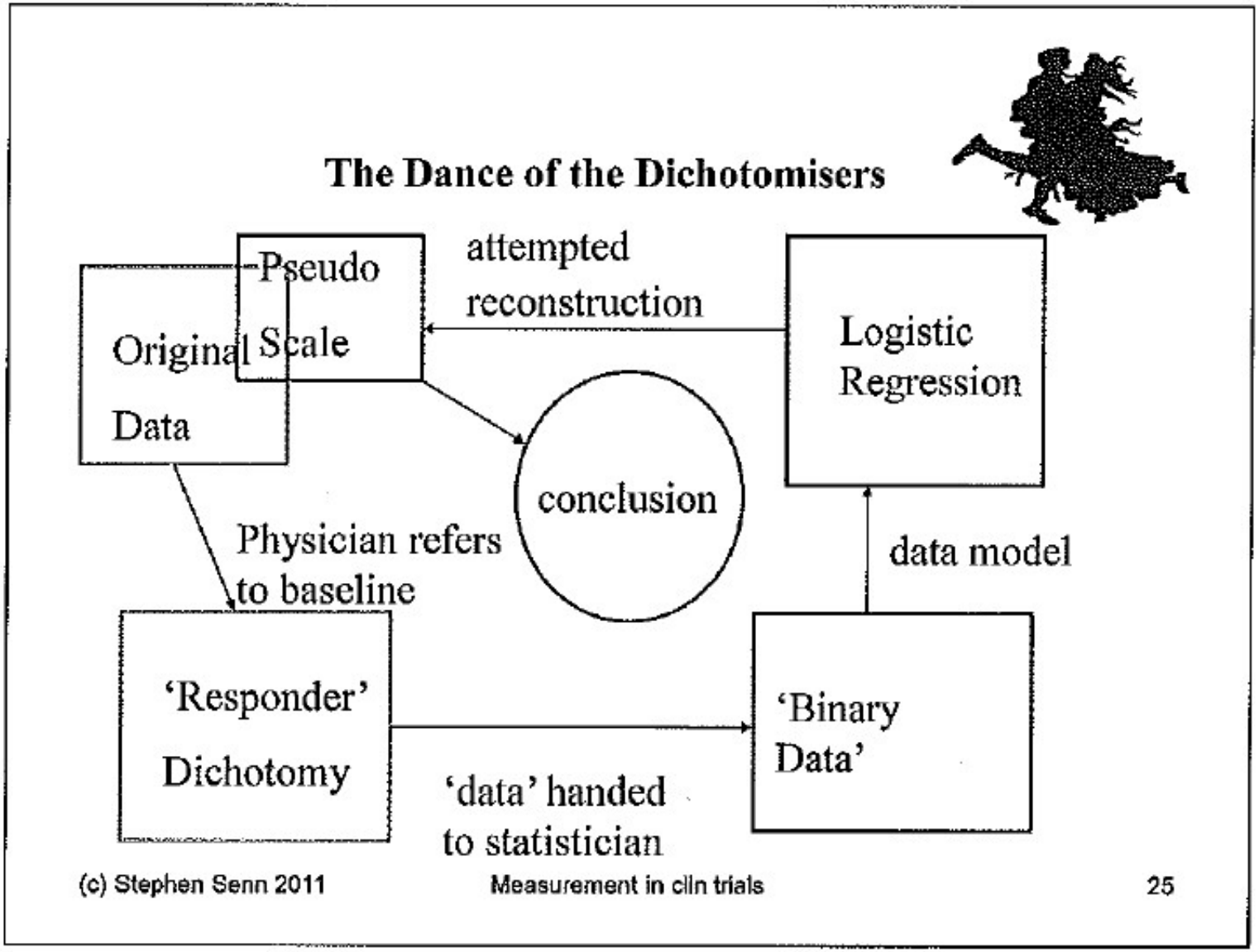
- Division of patients into responders and non-responders by comparing continuous outcome to some cut-point
- Dichotomizing continuous measurements
- Dichotomy: A means by which regulators destroy information, increase the size of clinical trials, delay treatment for patients and cost the economy millions.

  – Senn, *Statistical Issues in Drug Development* (pg. 120)

The Dichotomizer

- The dichotomizer compares continuous baseline measurement to outcome measurement and labels subject as responder (1) or non-responder (0)
- Gives results to statistician, who will have to find a way to model the binary outcome
- Better to use actual and genuine values

The Dance of the Dichotomisers

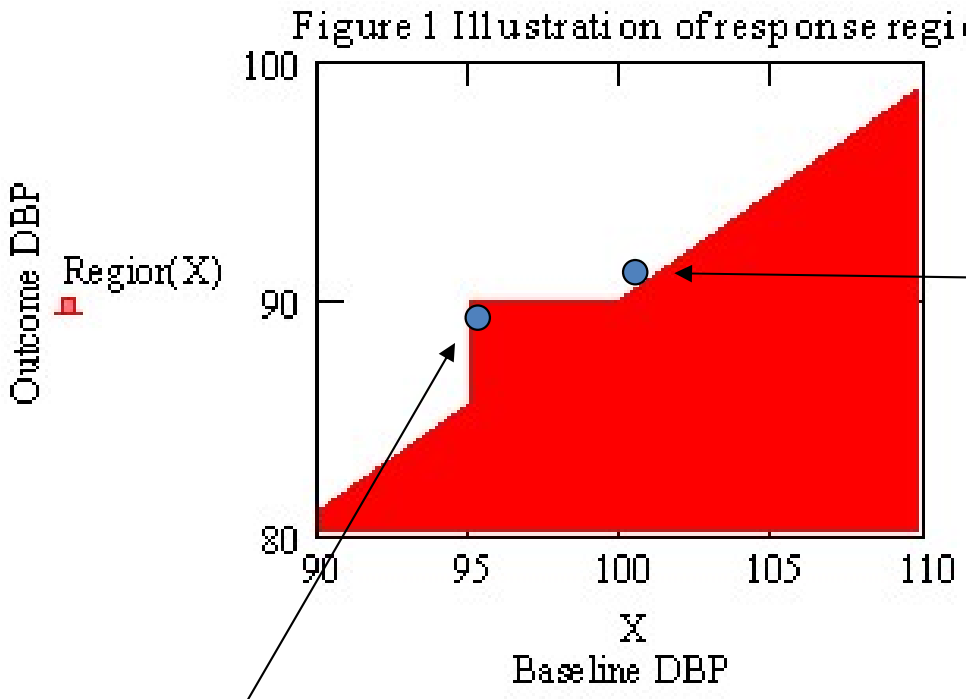


Objectionable Dichotomy*

- Hypertension response: achieved if DBP has reduced either…
  - By 10%
  - Or from ≥95 mmHg to ≤90 mmHg
- Implies a reduction of 10 mmHg and 9.9% from 101 to 91 is not a response
- Implies a reduction of 6 mmHg and 6.3% from 95 to 89 is a response

*see Goetghebeur, et al. (1998) *Statistics in Medicine* 17: 341-355 for example

Hypertension Hill



Red area is region of response.

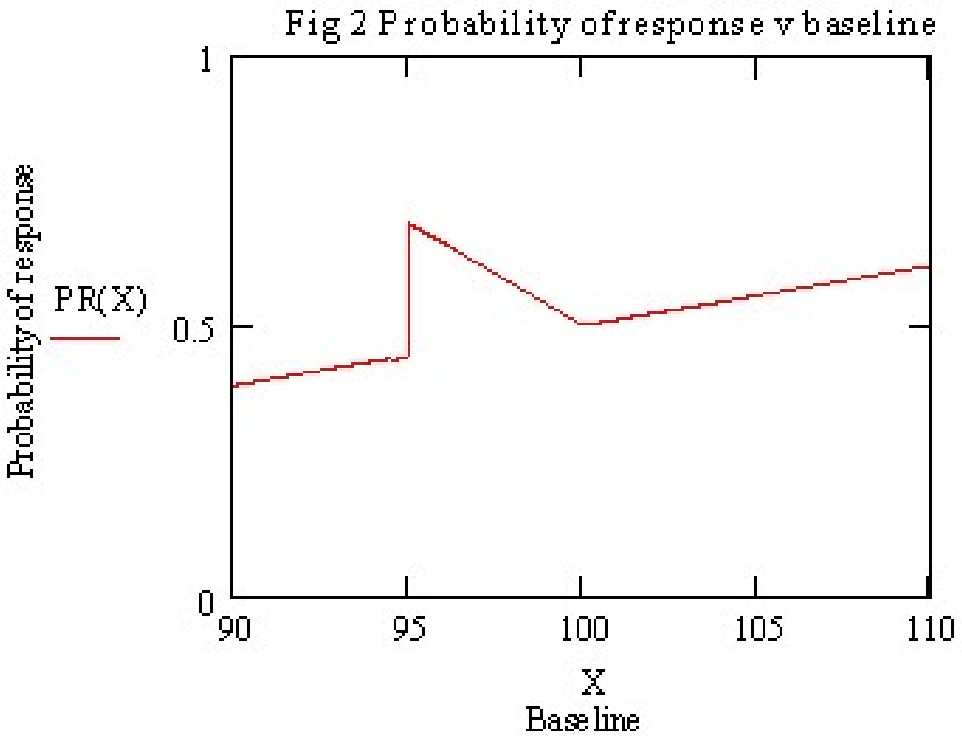
reduction of 10 mmHg and 9.9% from 101 to 91

reduction of 6 mmHg and 6.3% from 95 to 89

Example continued

- Mean DBP at baseline: 100 mmHG
- SD: 10 mmHG at outcome and baseline
- Correlation: 0.7
- True reduction of 10 mmHg for everybody in treatment group
  - Mean DBP at outcome is 90 mmHG
- Check probability of being classified a responder as a function of baseline DBP

Pressure Piste



If baseline is between 95 and 100, you have higher probability of response.

If X ≤ 95 or X ≥ 100 then calculate P(Y < 0.9*X)
Else calculate P(Y < 90)

Recall: responder if…
(1) DBP reduced by 10% or
(2) decrease from ≥95 mmHg to ≤90 mmHg

If X=94: P(Y < 84.6)
If X=96: P(Y < 90) or P(Y < 86.4) → P(Y < 90)
If X = 101: P(Y < 90) or P(Y < 90.9) → P(Y < 90.9)

Pressure Piste details

- Conditional mean and sd of Y|X

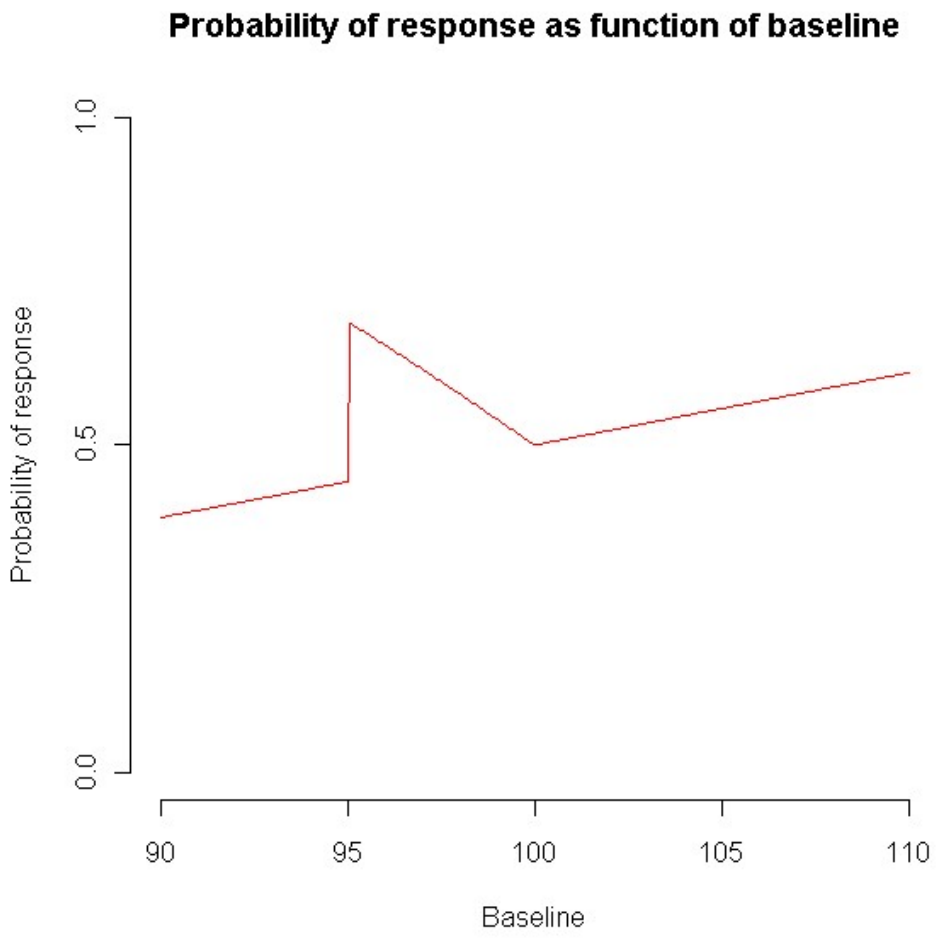
$$\mu_{y|x=x_0} = \mu_y + \rho\sigma_y \frac{x-\mu_x}{\sigma_x}$$

$$\sigma_{y|x=x_0} = \sigma_x\sqrt{1-\rho^2}$$

$$\mu_x = 100$$
$$\mu_y = 90$$
$$\sigma_x = \sigma_y = 10$$
$$\rho = 0.7$$
$$90 \le x \le 110$$

Pressure Piste in R

```
x <- seq(90,110,0.05)
p <- c()
for (i in 1:length(x)){
      if (x[i] <= 95 || x[i] >=100){
            p[i] <- pnorm(0.9*x[i],90+(0.7*10*((x[i]-100)/10)),
                        10*sqrt(1-(0.7^2))) # P(Y > 0.9*X)
      } else {
            p[i] <- pnorm(90,90+(0.7*10*((x[i]-100)/10)),
                        10*sqrt(1-(0.7^2))) # P(Y > 90)
      }
}
plot(x,p,type='l',ylim=c(0,1),col="red",axes=FALSE,
ylab="Probability of response",xlab="Baseline",
main="Probability of response as function of baseline")
axis(2,c(0,0.5,1))
axis(1,c(90,95,100,105,110))
```



More information

- Buy the book:
  - Statistical Issues in Drug Development, 2nd Ed.
- Thank you! Questions or comments?