To Thicken Other Proofs That Do Demonstrate Thinly:

A Recent Observational Study Used to Illustrate Some Recent Methodology.

Paul R. Rosenbaum, University of Pennsylvania

With a few glosses, the talk is based on:

•P. R. Rosenbaum, R. N. Ross and J. H. Silber. (2007) Minimum distance matched sampling with fine balance in an observational study of treatment for ovarian cancer. *J. Am. Statist. Assoc.*, 102: 75-83.

•J. H. Silber, P. R. Rosenbaum, D. Polsky, R. N Ross, O. Even-Shoshan, S. Schwartz, K. A. Armstrong, T. C. Randall. (2007) Does ovarian cancer treatment and survival differ by the specialty providing chemotherapy? *Journal of Clinical Oncology* (JCO) 2007; 25: 1169-1175.

•Rosenbaum, Paul R. (2007) Confidence intervals for uncommon but dramatic responses to treatment. *Biometrics*, 2007, 63, 1164–1171.

•**Editorial**: Cannistra, S. A. (2007) Gynecologic oncology or medical oncology: What's in a name? *JCO,* 25: 1157-1159. **5 Letters and 2 rejoinders** from S. Blank, J. Curtin, A. Berchuck, M. Hoffman, U. Iqbal, M. Markham, W. McGuire, J. Silber, P. Rosenbaum, and S. Cannistra. *JCO,* 2007; 25: 1151-1158.

OTHELLO: O monstrous! monstrous!

IAGO: Nay, this was but his dream.

OTHELLO: But this denoted a foregone conclusion:

'Tis a shrewd doubt, though it be but a dream.

IAGO: And this may help to thicken other proofs

That do demonstrate thinly.

William Shakespeare,

Othello, Act III.

1 Two methodological topics

Matching with fine balance: A transparent method of adjusting for observed covariates. Typically used in conjunction with propensity scores, optimal matching, and a distance such as the Mahalanobis distance. Can perfectly balance the marginal distributions of a nominal variable without constraining who is matched to whom.

Uncommon but dramatic treatment effects: When both present and noticed, these may be less sensitive to unobserved biases than are smaller typical effects.

2 A "natural experiment" in health care outcomes research

A natural experiment is a type of observational study in which some stable, perhaps even rational, process for allocating treatments is disrupted in an ostensibly aimless or haphazard way. It is a "wild experiment," not a "wholesome experiment." Haphazard is not random.

Outcomes research examines the delivery and outcomes of health care as it actually occurs.

A key difficulty: most of the variation in treatment and in outcome reflects the health of the patient, not activities of the health care providers.

There is a curious spot where there is variation in treatment that is, to a large extent, not a response to the health of the patient.

3 Two specialties provide chemotherapy for ovarian cancer

Medical oncologists (MO's) are specialists in the provision of chemotherapy, but treat cancers of varied types.

Gynecologic oncologists (GO's) are gynecologists with additional training in oncology. They treat a small group of gynecologic cancers, including ovarian cancer.

As gynecologists, GOs are surgeons, and they typically perform the surgery needed for ovarian cancer. They may also provide chemotherapy.

MOs are, almost invariably, not surgeons. They may provide chemotherapy after a GO, a gynecologist, or a general surgeon has performed surgery.

4 How might the specialties differ?

Chemotherapy often has toxic side effects. The alternative to toxic side effects might be metastatic cancer and death. A delicate balance. Possibly, MOs are more aggressive in the initial use of chemotherapy. If so, is more aggressive treatment of benefit to patients?

When cancer spreads, it may involve organ systems remote from the site of origin. Possibly, MOs are more aggressive in treating metastatic cancer. If so, is more aggressive treatment of benefit to patients?

GOs are compensated for surgery and chemotherapy, whereas MOs do not perform surgery.

5 Data from SEER and Medicare

Data were from a file that linked the Surveillance, Epidemiology and End Results (SEER) program of the National Cancer Institute with claims from Medicare.

In addition to survival, SEER provides clinical information, such clinical stage and tumor grade, while Medicare provides information about treatment, including surgeon type, and comorbid conditions.

SEER sites: some are cities, others are states. Examples: Connecticut and San Francisco are both SEER sites.

Years: Patients diagnosed with ovarian cancer between 1991 and 1999 at SEER sites.

Outcomes: survival, amount of chemotherapy, chemotherapy associated side effects.

6 Matching in the clinical paper

We took all 344 patients treated by a GO and matched each one to a similar patient treated by an MO, creating 344 pairs of two patients.

In the clinical paper, we matched on 36 patient variables, plus some interactions. (In the statistical paper, there were 61 variables.)

In the clinical paper, we briefly described the matching procedure, and showed in detail that the procedure had produced groups with similar distributions of covariates.

Will look at it clinically first, then look under the hood at the details of the matching.

7 Covariate imbalances: part 1

Values are percents. Abbreviations: GO=gynecological oncologist, MO=medical oncologist, Gyn=gynecologist, General=general surgeon. Odds ratio before matching, GO/General with GO/MO is about 8.

		GO $n = 344$	matched-MO $n = 344$	all-MO $n = 2,011$
Surgeon	GO	76	75	33
Type	Gyn	15	16	39
	General	8	8	28
Stage	I	9	9	9
	II	11	9	9
	III	51	53	47
	IV	26	26	31
	Missing	3	2	3
Tumor	1	5	4	4
Grade	2	16	13	17
	3	52	55	47
	4	9	8	11
	Missing	18	20	21

8 Covariate imbalances: part 2

Values are percents, except as noted.

	GO $n = 344$	matched-MO $n = 344$	all-MO $n = 2,011$
White	91	94	94
Black	8	5	3
COPD	15	12	13
Hypertension	48	46	42
Diabetes	11	8	8
CHF	2	2	4
Age, mean	72.2	72.2	72.8
$\widehat{e}(\mathbf{x})$, mean	0.23	0.21	0.14

Propensity score $\widehat{e}(\mathbf{x})$ included: SEER sites; year of diagnosis; stage; grade; race; age; and the comorbidities of anemia, angina, arrhythmia, asthma, chronic obstructive pulmonary disease, coagulation disorder, diabetes, electrolyte abnormality, hepatic dysfunction, hypertension, hyperthyroidism, peripheral vascular disease, and rheumatoid arthritis.

9 Covariate imbalances: part 3

SEER site or year	GO $n = 344$	matched-MO $n = 344$	all-MO $n = 2,011$
Connecticut	18	18	15
Detroit	26	26	12
Iowa	17	17	17
New Mexico	7	7	3
Seattle	9	9	16
Atlanta	9	9	7
Los Angeles	12	12	19
San Francisco	1	1	9
1991	4	4	9
1992	7	7	14
1993	10	9	14
1994	11	11	12
1995	11	13	12
1996	10	9	12
1997	16	15	10
1998	13	15	9
1999	18	17	9

Sites &1991-1992, 1993-1996, 1997-1999 balanced.

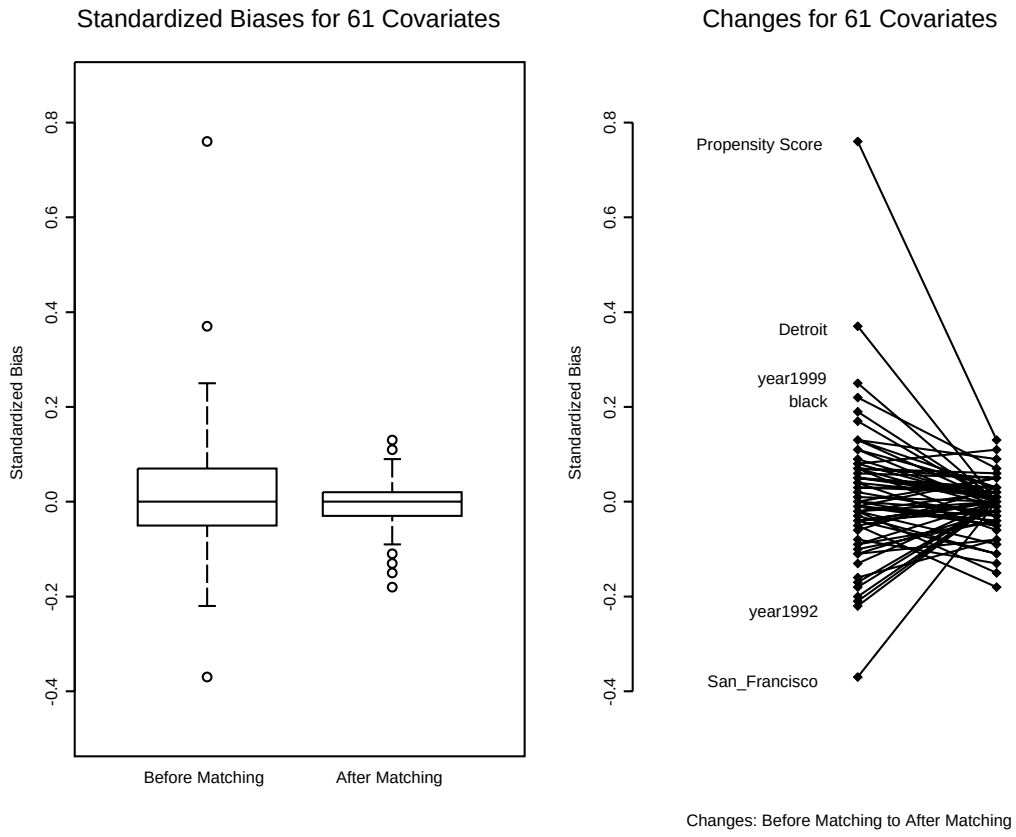


Figure 1: Plots of Imbalance in 61 Covariates, Before and After Matching. Values are differences in covariate means, GO vs MO, before and after matching, divided by the average within group standard deviation before matching. Six covariates with large initial biases are identified by name.

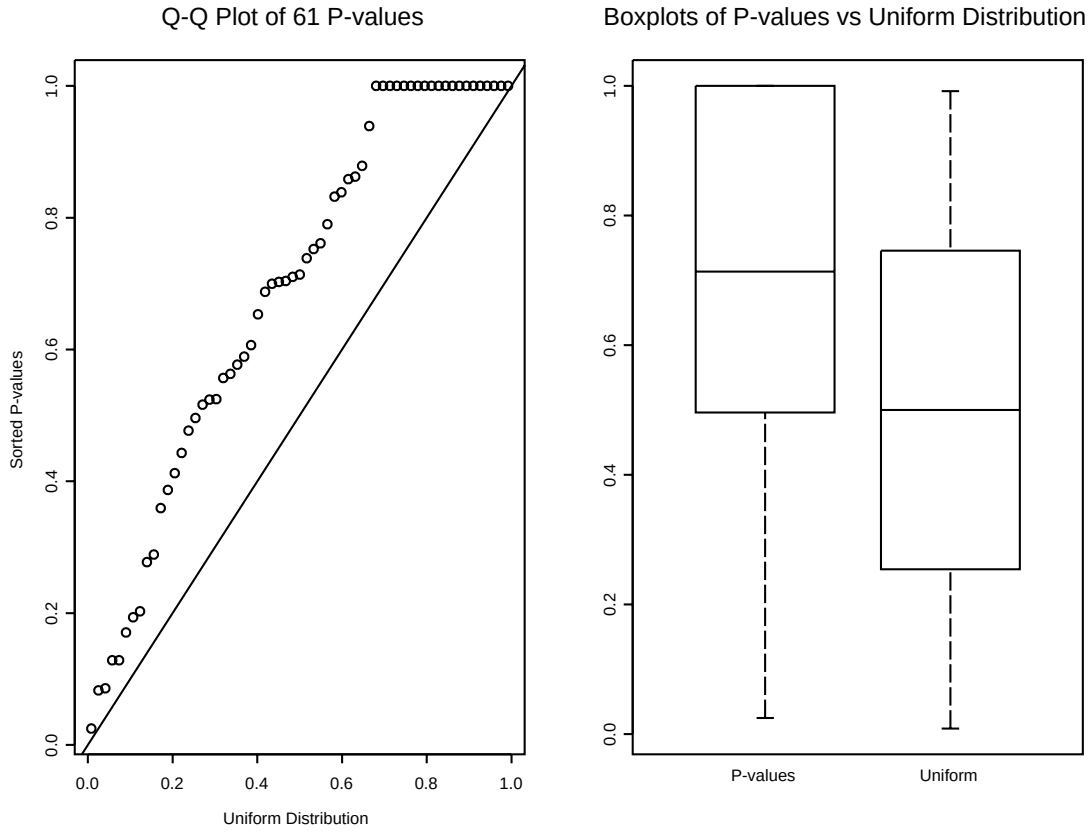


Figure 2: Comparing the Balance on 61 Covariates to the Balance Expected in a Completely Randomized Experiment: 61 P-values From Two-Sample Tests Compared to the Uniform Distribution. The diagonal line is y=x.

10 Matching method

Minimum distance: The optimal assignment algorithm was used with Mahalanobis distances and penalties for mismatches on surgeon type, clinical stage, tumor grade, year of diagnosis, race, congestive heart failure, diabetes, weight loss, and the propensity score.

Within propensity score calipers: Penalty for a large discrepancy on the propensity score based on: SEER sites; year of diagnosis; stage; grade; race; age; and the comorbidities of anemia, angina, arrhythmia, asthma, chronic obstructive pulmonary disease, coagulation disorder, diabetes, electrolyte abnormality, hepatic dysfunction, hypertension, hyperthyroidism, peripheral vascular disease, and rheumatoid arthritis.

With fine balance constraints: The marginal distributions were constrained to agree on the interaction of the 8 SEER sites and the three time intervals $(1991-1992)$, $(1993-1996)$, $(1997-1999)$ or $24 = 8 \times 3$ levels.

11 Older ideas: Propensity scores, minimum distance

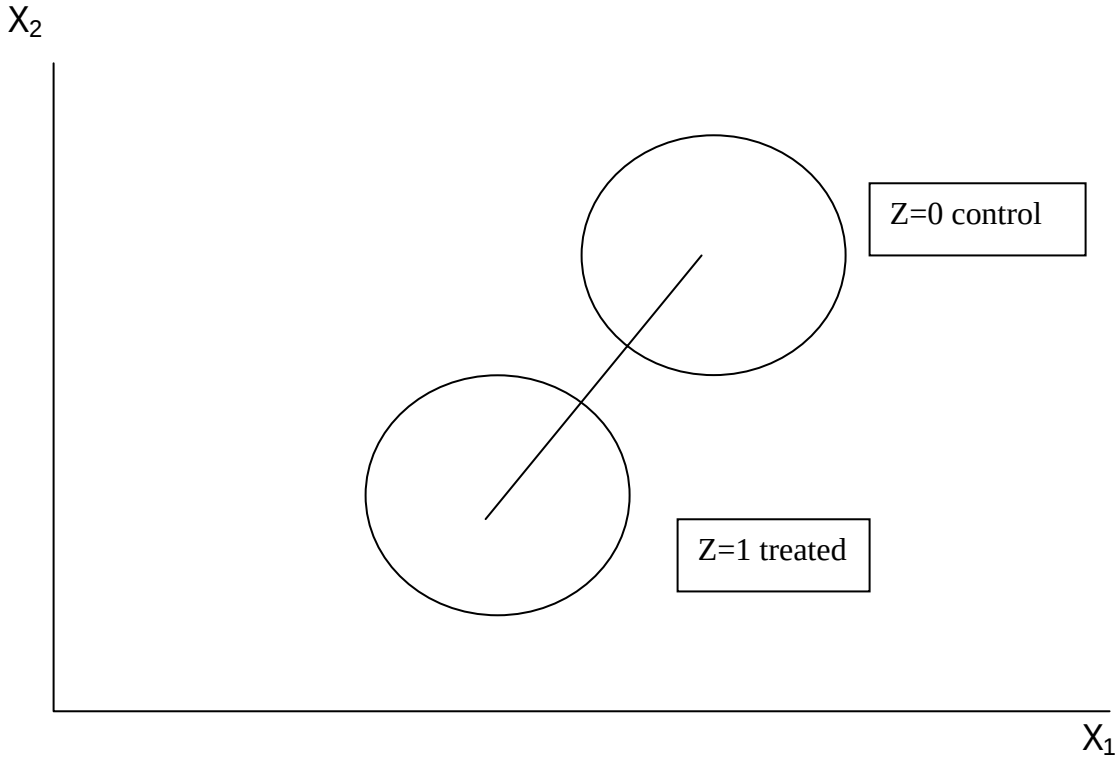
Notation: Covariates $\mathbf{x}_i$ for person i, estimated propensity score $\widehat{e}(\mathbf{x}_i)$, estimated covariance matrix $\widehat{\mathbf{\Sigma}}$.

Propensity score calipers: If $\left|\widehat{e}(\mathbf{x}_i) - \widehat{e}\left(\mathbf{x}_j\right)\right| > \kappa$ do not permit i and j to be matched; set the distance to ∞. A typical value of κ is $0.2 \times st.dev\left\{\widehat{e}\left(\mathbf{x}_j\right)\right\}$. Implement using a penalty, that is, a very large distance, not infinite, distance between i and j.

Other penalties: For example, we penalized mismatch on type of surgeon.

Mahalanobis distances: If i may be matched to j, then the distance between them is the Mahalanobis distance, $\left(\mathbf{x}_i - \mathbf{x}_j\right)^T \widehat{\mathbf{\Sigma}}^{-1} \left(\mathbf{x}_i - \mathbf{x}_j\right)$.

A Matching Picture



--Drawn for two uncorrelated standard Normal variables, but imagine instead K possibly dependent, possibly non-Normal variables.
--If you placed coordinate axis at the K-dimensional median, there would be 2^K quadrants. 2^K is about a billion for K=30 variables.
--If e(**X**) = Pr(Z=1|**X**) then Pr{**X**|Z=1, e(**X**)=a} = Pr{**X**|Z=0, e(**X**)=a}
--For Normal **X**, e(**X**) is a function of the linear discriminant.
1) Failures to match on e(**X**) tend to accumulate, while failures to match on **X** at the same e(**X**) tend to balance out.
2) Once you have a good match on e(**X**), you might as well try to get a close match on the most important coordinates of **X**.
3) Balance from matching on e(**X**) is stochastic. It won't work well if data are thinly spread through many levels of a nominal variable. For this, use "fine balance" to constrain the marginal distributions to agree exactly on the nominal variable.
a) minimum distance matching,
b) within calipers on estimated e(**X**),
c) subject to the fine balance constraint.

12 Older ideas: tips on implementation

Penalties: It is best to use a large finite number for a penalty, and to add $\left(\mathbf{x}_i - \mathbf{x}_j\right)^T \widehat{\mathbf{\Sigma}}^{-1} \left(\mathbf{x}_i - \mathbf{x}_j\right)$ to the penalty. Then, if all the constraints implied by the penalties cannot be respected, the optimal assignment algorithm will not crash. Rather, it will respect as many penalties as possible, and then minimize the total distance subject to respecting that number of penalties.

Varied penalties: With several matching constraints, use penalties of varied sizes to prioritize them.

Distances: The software uses a modified Mahalanobis distance. Uses ranks rather than responses, with average ranks for ties. However, uses the variance associated with untied ranks.

Optimal matching: Pick the matched set to minimize the total distance within pairs. $O\left(n^3\right)$

13 Newer idea: fine balance

What is fine balance? Fine balance imposes a constraint on the previously described match. The constraint is that the marginal distributions of a nominal variable, perhaps with many levels, will be exactly balanced.

How is fine balance different? Fine balance imposes no constraint on who is matched to whom. It is about the marginal distributions only.

Possible uses: (i) A variable with many categories, perhaps growing with the sample size. (ii) A variable formed as the interaction (i.e., direct product) of several key binary variables.

In ovarian study, the nominal variable was SEER site $\times$ year of diagnosis category, 24 levels. (Remember, odds ratio of 19.5 for Detroit vs SF.)

14 Toy Example: the data

In this tiny data set, there are three treated subjects ($Z = 1$), two male, one female, and seven potential controls ($Z = 0$), three male and four female. We want to match closely for X1, X2, X3 and their $X_+ = X1 + X2 + X3$, while balancing gender. In typical practice, the propensity score replaces X_+.

$>$ `data`

Name	Gender	X_1	X_2	X_3	X_+	Z
Harry	M	0	0	1	1	1
David	M	1	0	1	2	1
Susan	F	1	0	1	2	1
Mark	M	0	0	0	0	0
Horatio	M	1	0	0	1	0
Tim	M	1	0	1	2	0
Janet	F	1	1	0	2	0
Diane	F	1	1	1	3	0
Debbie	F	1	1	0	2	0
Sally	F	1	0	1	2	0

15 Toy Example: calipers and distances

A caliper of 1 on X_+ introduces three ∞'s, and elsewhere there is a variant of the Mahalanobis distance. Example: Harry has $X_+ = 1$ Diane has $X_+ = 3$, and $|1 - 3| > 1$.

	Mark	Hor.	Tim	Janet	Diane	Deb.	Sally
Harry	3.3	8.2	3.3	6.3	∞	6.3	3.3
David	∞	3.3	0.0	4.1	3.7	4.1	0.0
Susan	∞	3.3	0.0	4.1	3.7	4.1	0.0

16 Toy example: fine balance

	Mark	Hor.	Tim	Janet	Diane	Deb.	Sally
Harry	3.3	8.2	3.3	6.3	∞	6.3	3.3
David	∞	3.3	0.0	4.1	3.7	4.1	0.0
Susan	∞	3.3	0.0	4.1	3.7	4.1	0.0
α_1	∞	∞	∞	0	0	0	0
α_2	∞	∞	∞	0	0	0	0
α_3	∞	∞	∞	0	0	0	0
α_4	0	0	0	∞	∞	∞	∞

•Not difficult to show that the minimum cost assignment for this array is the optimal balanced match.

16.1 Toy example: fine balance, actual coding & match

	Mark	Hor.	Tim	Janet	Diane	Deb.	Sally
Harry	**3.3**	8.2	3.3	6.3	925.5	6.3	3.3
David	926.2	3.3	0.0	4.1	3.7	4.1	**0.0**
Susan	926.2	3.3	**0.0**	4.1	3.7	4.1	0.0
α_1	10^4	10^4	10^4	**0**	0	0	0
α_2	10^4	10^4	10^4	0	**0**	0	0
α_3	10^4	10^4	10^4	0	0	**0**	0
α_4	0	**0**	0	10^4	10^4	10^4	10^4

•Two male controls, one female, balanced but not matched for gender.

•R function balmatch will do all the work.

17 R software

http://www-stat.wharton.upenn.edu/~rosenbap/index.html

Name	Gender	X_1	X_2	X_3	X_+	Z
Harry	M	0	0	1	1	1
David	M	1	0	1	2	1
Susan	F	1	0	1	2	1
Mark	M	0	0	0	0	0
Horatio	M	1	0	0	1	0
Tim	M	1	0	1	2	0
Janet	F	1	1	0	2	0
Diane	F	1	1	1	3	0
Debbie	F	1	1	0	2	0
Sally	F	1	0	1	2	0

```
> balmatch(gender,Z,X+,X,caliper=1.1)

1 3 2 1 NA 2 NA NA NA 3
```

18 Outcomes

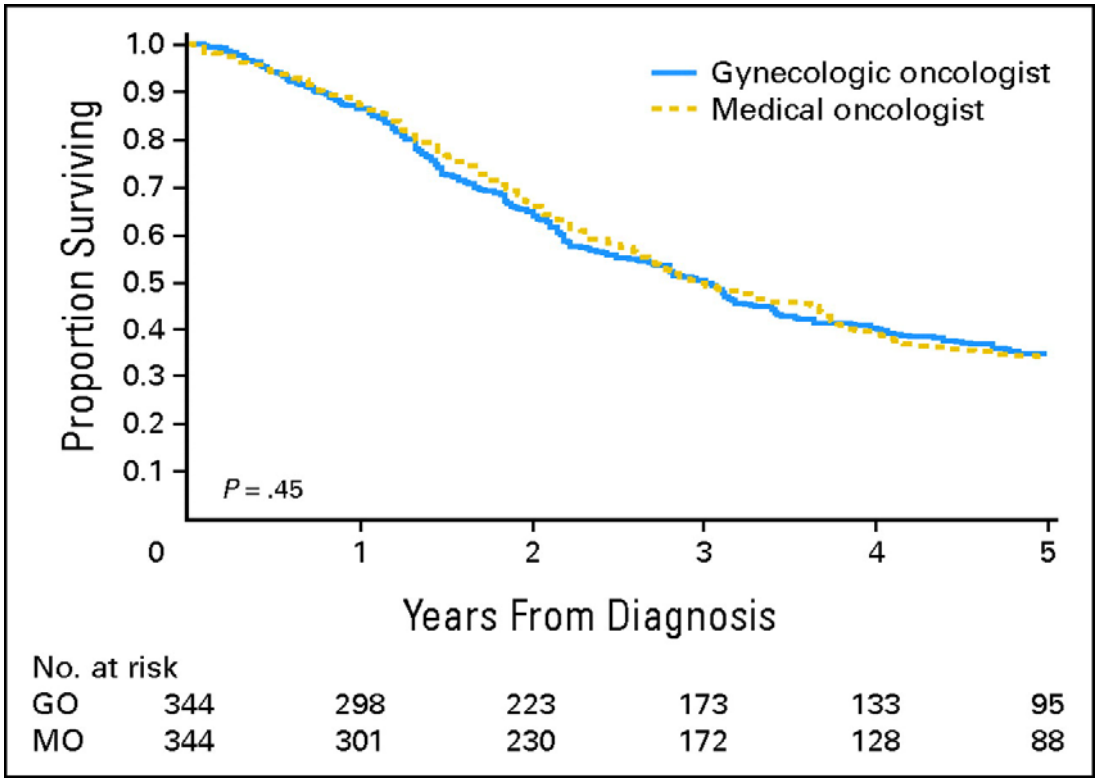
344 matched pairs, GO vs MO.

Outcomes are (i) survival, (ii) intensity of chemotherapy, and (iii) chemotherapy related toxicity.

Weeks of chemotherapy, weeks with toxicity.

Year 1 is initial treatment. Later years are likely to be chemotherapy for recurrence of cancer.

Kaplan-Meier survival plot comparing 344 patients administered postoperative chemotherapy for ovarian cancer by a gynecologic oncologist (GO) and a matched set of 344 patients administered postoperative chemotherapy for ovarian cancer by a medical oncologist (MO). (Figure from JCO 2007)



P-value based on:
O'Brien, P. C. and Fleming, T. R. (1987), "A paired Prentice-Wilcoxon test for censored paired data," Biometrics, 43, 169-180.

Survival (Table from JCO 2007)

Survival, years	GO Group	MO Group	P-value
Median	3.04	2.98	
95% CI	2.50 to 3.40	2.69 to 3.67	
1-year survival, %			.57
Point Estimate	86.6%	87.5%	
95% CI	83.0 to 90.2	84.0 to 90.1	
2-year survival, %			.57
Point Estimate	64.8%	66.9%	
95% CI	59.8 to 69.9	61.9 to 71.8	
5-year survival, %			.81
Point Estimate	35.1	34.2	
95% CI	30.0 to 40.2	29.2 to 39.3	

Abbreviations: GO, gynecologic oncologist; MO, medical oncologist.

Coding Definitions of Chemotherapy and Chemotherapy Associated Adverse Events (from inpatient and outpatient bills) (Table from JCO 2007)

Coding Definitions
Chemotherapy administration
ICD-9 procedure codes
99.25: Injection or infusion of cancer chemotherapeutic substance
HCPCS codes
964.xx: intravenous chemotherapy administration
965.xx: intravenous chemotherapy administration
CPT codes
36640: insertion catheter, artery
36260 insertion of infusion pump
Codes for ovarian cancer drugs
J8999-J9999; Q0163-Q0185
Chemotherapy-associated adverse events: ICD-9 diagnosis codes
Anemia
280.x; 281.x; 283.x; 284.8; 284.9; 285.xx
Neutropenia
288.0
Thrombocytopenia
287.5
Mucositis
528
Dehydration, dehydration, nausea, diarrhea
276.5; 787.01; 787.02; 787.91
Neuropathy (drug associated)
357.6

Abbreviations: ICD, International Classification of Diseases; HCPCS, Healthcare Common Procedure Coding System; CPT, Current Procedural Terminology.

Intensity of Treatment (Table from JCO 2007)

Outcome Measure	GO Group		MO Group		*P*
	Mean	Median	Mean	Median	
Weeks with some chemotherapy					
Over first 5 years	12.1	9.0	16.5	11.0	.0023
For year 1	6.6	6.0	7.7	6.0	.0106
For years 2 to 5	6.3	2.5	10.0	4.0	.0167
Weeks with chemotherapy-associated adverse events*					
Over first 5 years	8.9	5.0	16.2	7.0	.0001
For year 1	3.6	2.0	6.6	3.0	.0001
For years 2 to 5	6.1	2.0	11.0	4.0	.0001

Abbreviations: GO, gynecologic oncologist; MO, medical oncologist.

* Weeks with chemotherapy-associated adverse events was defined as any week that included the following diagnoses occurring as an inpatient or outpatient: anemia, neutropenia, thrombocytopenia, diarrhea, dehydration or mucositis, and neuropathy.

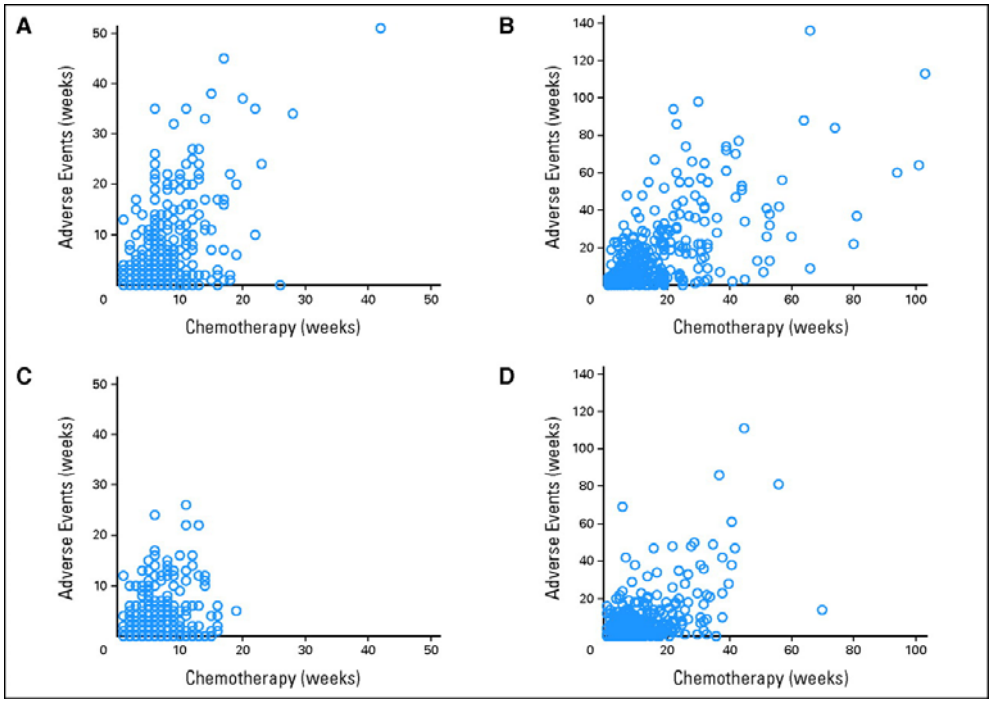
Chemotherapy Intensity and Toxicity for 344 Matched GO and MO patients

A MO patients, year 1.

B MO patients, all years

C GO patients, year 1.

D GO patients, all years



Silber, J. H. et al. J Clin Oncol; 25:3555-3557 2007

19 Summary

Survival: Survival was virtually identical for GO and MO patients.

Intensity: MO's gave more chemotherapy and produced more toxicity, but in the first year and in later years.

Does intensity lengthen survival? No indication that greater intensity of chemotherapy lengthens survival.

20 An editorial, 5 letters, 2 rejoinders

All of the discussion concerned possible biases from unobserved covariates.

A partial success: Balanced matching was transparent. No one doubted that the groups were comparable on variables controlled by matching.

Right orientation: In broad terms, a discussion focused on unobserved biases is, at least, concerned with the correct issue.

Content of discussion: Ranged from what might reasonably be described as speculative to what might charitably be described as speculative. Reasonable $\cong$ a speculative alternative explanation of what was observed. Charitable $\cong$ speculative explanations of different parts that don't form a coherent whole.

Qualitative guesswork: No data, no quantification.

21 Examples

Editorial Stephen Cannistra (an MO) said that a GO might do the surgery, realize that not all the tumor could be removed, and send the patient to an MO for intensive chemotherapy, with the GO keeping the healthier patients. That is, MOs take on the tough patients, prolong their lives with intensive chemotherapy so they live just as long as healthier patients treated by a GO.

First of 5 letters: Stephanie Blank and John Curtin (two GOs in a letter): "Cannistra spins a tale that GOs ... 'attract less proactive patients' [who ...] are not interested in their own health ... [quoting Othello] 'But this denoted a foregone conclusion: Tis a shrewd doubt, though it be but a dream.'"

Cannistra replying to Blank and Curtin: "Drs Bland and Curtin ... have chosen to twist my editorial into

an attack against GOs . . . I offered critical analysis . . . which Drs Blank and Curtin called ‘spinning a tale’.”

22 What makes observational studies insensitive to unobserved biases?

What is a sensitivity analysis? Asks: How much unobserved bias — how large a departure from random treatment assignment — would need to be present to alter the qualitative conclusions of the study? Answer is a number, for instance, the value of a sensitivity parameter, say Γ (to be explained in a moment).

What makes some studies insensitive to biases? Not a subject that has been extensively studied.

Design sensitivity: A tool for investigating this question. Says: imagine the study is actually free of unobserved bias, and there is a treatment effect. In very large samples ($I \longrightarrow \infty$), how sensitive is a particular study design/sampling model? Again, the answer is a number $\widetilde{\Gamma}$ such that, as $I \longrightarrow \infty$, the power of the sensitivity analysis $\longrightarrow 1$ for $\Gamma < \widetilde{\Gamma}$ and $\longrightarrow 0$ for $\Gamma > \widetilde{\Gamma}$.

23 Sensitivity Analysis for Matched Pairs

Covariates: Observed covariate $\mathbf{x}$. Unobserved covariate u.

Matching: I pairs, $i = 1, \ldots, I$, of two subjects, $j = 1, 2$, matched for $\mathbf{x}$, so $\mathbf{x}_{i1} = \mathbf{x}_{i2}$ for each i, but not for u, so typically $u_{i1} \neq u_{i2}$.

Treatment indicator: $Z_{ij} = 1$ if j received treatment, $Z_{ij} = 0$ if j received control, so $Z_{i1} + Z_{i2} = 1$ for $i = 1, \ldots, I$.

Responses: Potential responses, $\left(r_{Tij}, r_{Cij}\right)$, r_{Tij} under treatment, $Z_{ij} = 1$, r_{Cij} under control, $Z_{ij} = 0$, so effect is $r_{Tij} - r_{Cij}$; Neyman (1935) & Rubin (1974).

24 Paired Randomized Experiment

Conditioning: Write

$$\mathcal{F} = \{\left(r_{Tij}, r_{Cij}, \mathbf{x}_{ij}, u_{ij}\right), \\ i = 1, \ldots, I, j = 1, 2\}$$

$$\mathcal{Z} = \{Z_{i1} + Z_{i2} = 1,\ i = 1, \ldots, I\}\,;$$

then $\mathcal{F}$ and $\mathcal{Z}$ are fixed by conditioning in Fisher's theory of randomization inference.

Randomization: $\Pr\left(Z_{i1} = 1 \mid \mathcal{Z}, \mathcal{F}\right) = \frac{1}{2}, i = 1, \ldots, I,$ with independent assignments in distinct pairs.

Observed responses, differences: R_{ij} observed is $R_{ij} = Z_{ij}\, r_{Tij} + \left(1 - Z_{ij}\right)\, r_{Cij}$, and the treated-minus-control difference in responses in pair i is $D_i = (2Z_{i1} - 1)\ (R_{i1} - R_{i2})$.

25 Wilcoxon's Signed Rank Statistic

Fisher's Sharp Null Hypothesis: The hypothesis of no treatment effect is $H_0 : r_{Tij} = r_{Cij}$, $\forall i, j$.

Wilcoxon's Signed Rank Statistic: To test H_0 rank $|D_i|$ from 1 to I; then W, is the sum of the ranks for which $D_i > 0$.

As a randomization test: If H_0 is true, randomization ensures D_i is $r_{Ci1} - r_{Ci2}$ or $r_{Ci2} - r_{Ci1}$, each with probability $\frac{1}{2}$, independently in different pairs. Given $\mathcal{Z}$, $\mathcal{F}$, if H_0 were true in a randomized experiment, then W would be the sum of I independent random variables taking values i or 0 each with probability $\frac{1}{2}$, $i = 1, \ldots, I$.

26 Departures from Random Assignment

1. In the population prior to matching, treatment assignments were independent, with unknown probabilities $\pi_{ij} = \mathsf{Pr}\left(Z_{ij} = 1 \,\middle|\, \mathcal{F}\right)$

2. Two subjects with the same *observed* $\mathbf{x}_{ij}$ may differ in *unobserved* u_{ij} and hence in their odds of receiving treatment by a factor of $\Gamma \geq 1$,

$$\frac{1}{\Gamma} \leq \frac{\pi_{ij}\left(1-\pi_{ik}\right)}{\pi_{ik}\left(1-\pi_{ij}\right)} \leq \Gamma, \;\; \forall\, i, j, k \qquad (1)$$

3. Distribution of treatments within treated/control matched pairs $\mathsf{Pr}\left(Z_{i1} = 1 \,|\, \mathcal{Z}, \mathcal{F}\right)$ is then obtained by conditioning on $Z_{i1} + Z_{i2} = 1$.

27 Departures, continued

$$1/\Gamma \leq \left\{\pi_{ij}\left(1-\pi_{ik}\right)\right\} / \left\{\pi_{ik}\left(1-\pi_{ij}\right)\right\} \leq \Gamma, \;\; \mathbf{x}_{ij} = \mathbf{x}_{ik}$$

No unobserved bias: If $\Gamma = 1$, then $\mathbf{x}_{ij} = \mathbf{x}_{ik}$ ensures $\pi_{ij} = \pi_{ik}$, $i = 1, \ldots, I$, whereupon

$$\Pr\left(Z_{i1} = 1 \mid \mathcal{Z}, \mathcal{F}\right) = \pi_{i1} / \left(\pi_{i1} + \pi_{i2}\right) = \frac{1}{2}.$$

Uncertainty from unobserved bias: If $\Gamma > 1$ in (1), then matching on $\mathbf{x}$ may fail to equalize the π_{ij} in pair i, and $\Pr\left(Z_{i1} = 1 \mid \mathcal{Z}, \mathcal{F}\right)$ is unknown.

Question answered by a sensitivity analysis: Bounds on significance levels, point estimates, confidence intervals for several values of Γ. How large must Γ be before qualitatively different causal interpretations are possible?

28 Sensitivity Analysis Procedure

Two known distributions: For fixed $\Gamma \geq 1$, let $\overline{\overline{W}}$ be the sum of I independent random variables taking value i with probability $\theta = \Gamma/(1+\Gamma)$ and value 0 with probability $1-\theta$, $i = 1, \ldots, I$; and let $\overline{W}$ for the sum of I independent random variables taking value i with probability $1-\theta$ and value 0 with probability θ.

Bounds: If

$$\frac{1}{\Gamma} \leq \frac{\pi_{ij}(1-\pi_{ik})}{\pi_{ik}\left(1-\pi_{ij}\right)} \leq \Gamma, \ \ \forall\, i, j, k$$

and $H_0 : \tau = \tau_0$ are true, then the following bounds are sharp for each $\Gamma \geq 1$:

$$\mathsf{Pr}\left(\overline{W} \geq w\right) \leq \mathsf{Pr}\left(W \geq w \mid \mathcal{Z}, \mathcal{F}\right) \leq \mathsf{Pr}\left(\overline{\overline{W}} \geq w\right)$$

If $\Gamma = 1$, then equality; otherwise the bounds become wider as Γ increases.

29 Sensitivity Analysis Using Wilcoxon Test

Chemo is weeks of chemotherapy in the first year. Adverse is weeks with chemotherapy related toxicity in the first year.

Table gives the (attainable) upper bound on the one-sided significance level for test the null hypothesis of no treatment effect using Wilcoxon's signed rank statistic.

Chemo is sensitive to biases of magnitude $\Gamma > 1.2$ and Adverse is sensitive to biases of magnitude $\Gamma > 1.5$.

Γ	Chemo	Adverse
1	0.0012	6.3×10^{-7}
1.1	0.010	1.9×10^{-5}
1.2	**0.048**	0.00027
1.3	0.14	0.0021
1.5	0.50	**0.036**
1.7	0.83	0.19

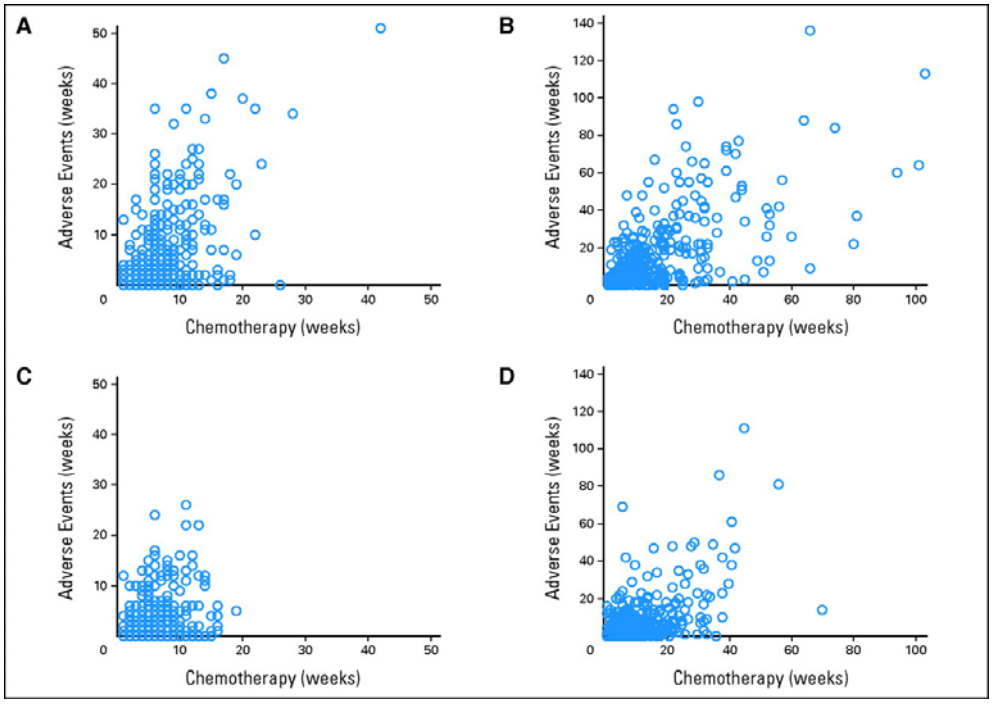
Chemotherapy Intensity and Toxicity for 344 Matched GO and MO patients

A MO patients, year 1.

B MO patients, all years

C GO patients, year 1.

D GO patients, all years



Silber, J. H. et al. J Clin Oncol; 25:3555-3557 2007

30 Uncommon but dramatic responses to treatment 1

Additive effects: $r_{Tij} = r_{Cij} + \tau$. Doesn't look correct here.

Wilcoxon's rank sum: Known to be the locally most powerful rank test in a randomized experiment for an additive effect when r_{Cij} has the logistic distribution and $\tau \to 0$ as $I \to \infty$.

Lehmann's result: Lehmann (1952) showed that the rank sum in a randomized experiment is also locally most powerful for testing

$$r_{Cij} \sim F(\cdot) \text{ and } r_{Tij} \sim (1-p)F(\cdot) + pF^2(\cdot)$$

with $p \to 0$ as $I \to \infty$, where $F^2(\cdot)$ is the distribution of the maximum of two observations from $F(\cdot)$.

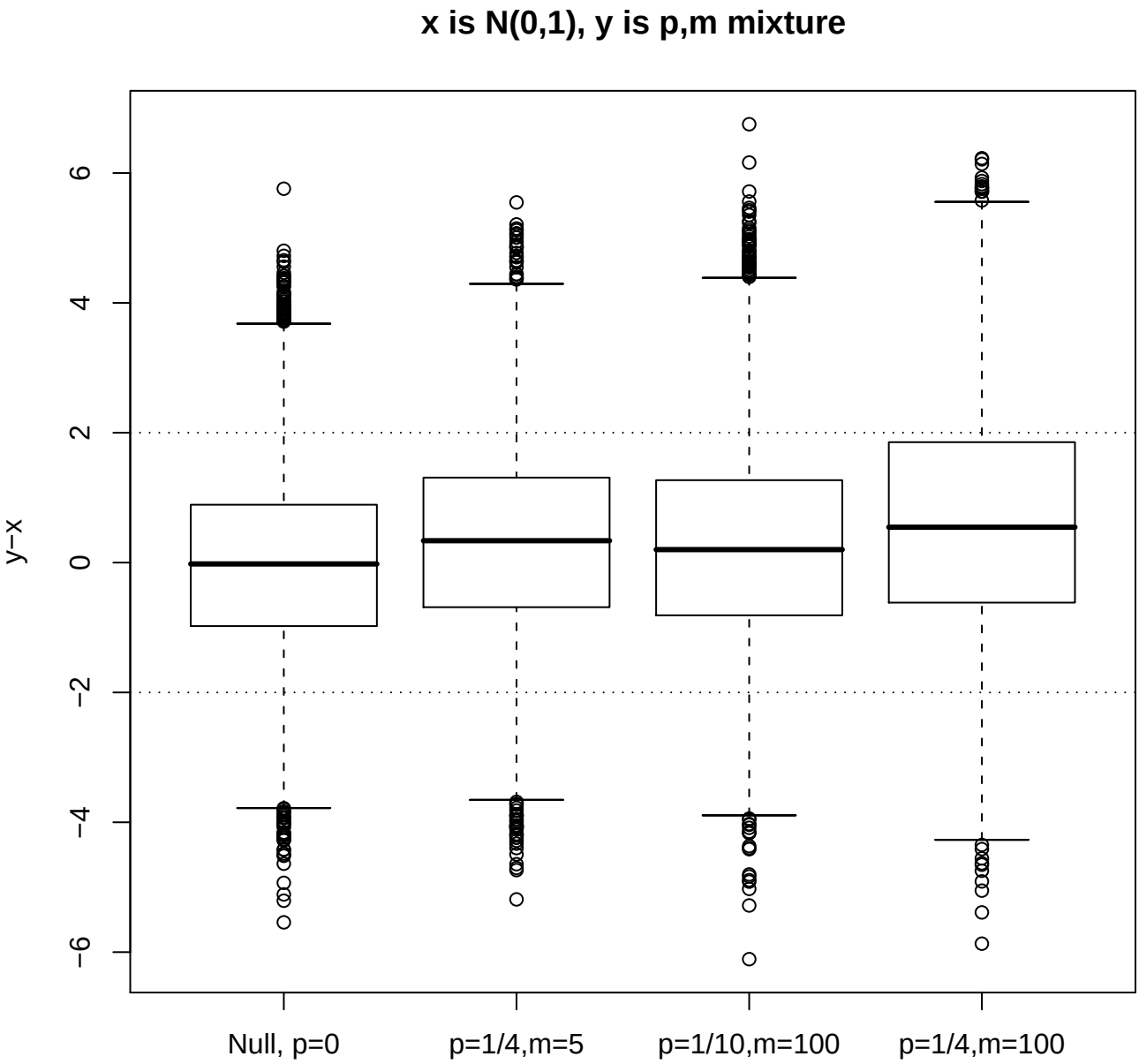
31 Uncommon but dramatic responses to treatment 2

Conover and Salsburg (1988) determined the rank score with is locally most powerful for

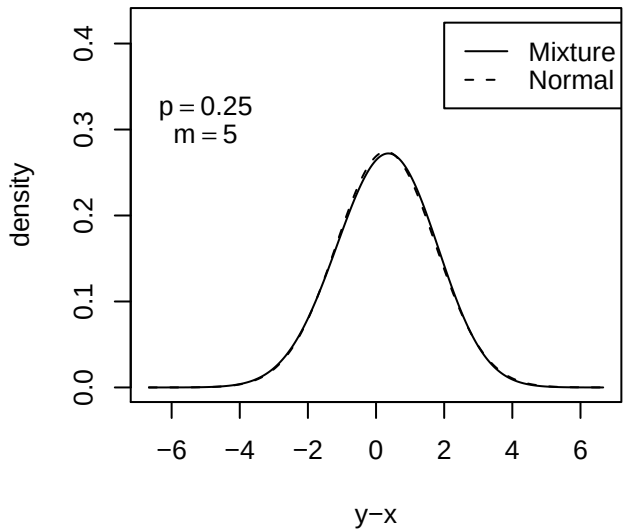
$r_{Cij} \sim F(\cdot)$ and $r_{Tij} \sim (1-p)F(\cdot) + pF^m(\cdot)$

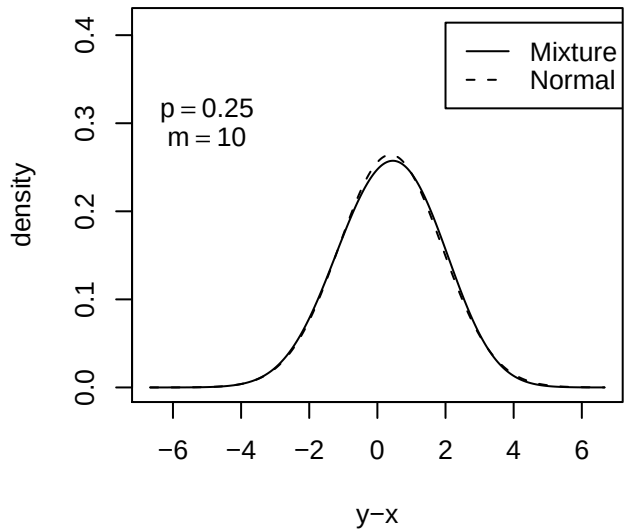
with $p \to 0$ as $I \to \infty$, which are fairly unintuitive in form, but are a polynomial in ordinary ranks whose highest power is $m-1$, consistent with Lehmann's result. Here, $F^m(\cdot)$ is the distribution of the maximum of m independent observations from $F(\cdot)$. They like $m = 5$.

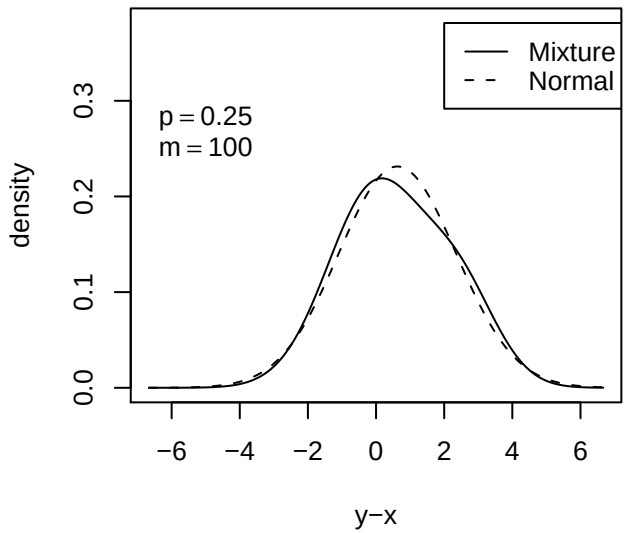
Let's look at 1000 treated- minus-control differences sampled from $r_{Cij} \sim \Phi(\cdot)$ and $r_{Tij} \sim (1-p)\Phi(\cdot) + p\Phi^m(\cdot)$ where $\Phi(\cdot)$ is the standard Normal cumulative distribution.

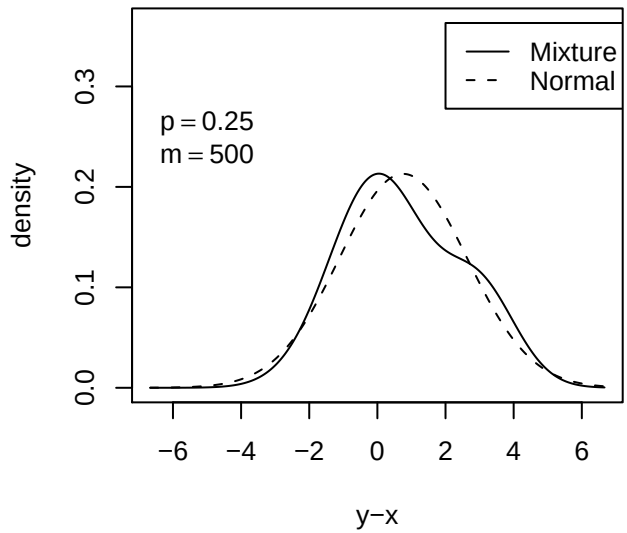

x is N(0,1), y is p,m mixture
y−x
6
4
2
0
−2
−4
−6
Null, p=0
p=1/4,m=5
p=1/10,m=100
p=1/4,m=100

Densities: Now let's look at the densities of treated-minus-control differences sampled from $r_{Cij} \sim \Phi(\cdot)$ and $r_{Tij} \sim (1-p)\,\Phi(\cdot) + p\Phi^m(\cdot)$ compared, not to the standard Normal, but to a Normal with the same expectation and variance.


Mixture
Normal
p = 0.25
m = 5
density
0.0
0.1
0.2
0.3
0.4
−6
−4
−2
0
2
4
6
y−x


Mixture
Normal
p = 0.25
m = 10
density
0.0
0.1
0.2
0.3
0.4
−6
−4
−2
0
2
4
6
y−x


Mixture
Normal
p = 0.25
m = 100
density
0.0
0.1
0.2
0.3
−6
−4
−2
0
2
4
6
y−x


Mixture
Normal
p = 0.25
m = 500
density
0.0
0.1
0.2
0.3
−6
−4
−2
0
2
4
6
y−x

32 Ranks

Wilcoxon's ranks: Ordinary 1, 2, 3, ..., n ranks can be assigned by comparing all $\binom{n}{2}$ pairs of two people and adding 1. 1 wins against no one else and gets a 0+1. 2 wins against 1, and gets a 1+1=2. 3 wins against 1 and 2, and gets a 1+1+1=3, etc.

Stephenson's ranks: Motivated by very different considerations, Stephenson (1981) proposed comparing people not 2 at a time but m at a time. For $m = 3$: both 1 and 2 never win in a group of 3, and get rank 0, but 3 wins in the group (1,2,3) and gets rank 1, while 4 wins in (1,2,4), (1,3,4), (2,3,4) and gets rank 3.

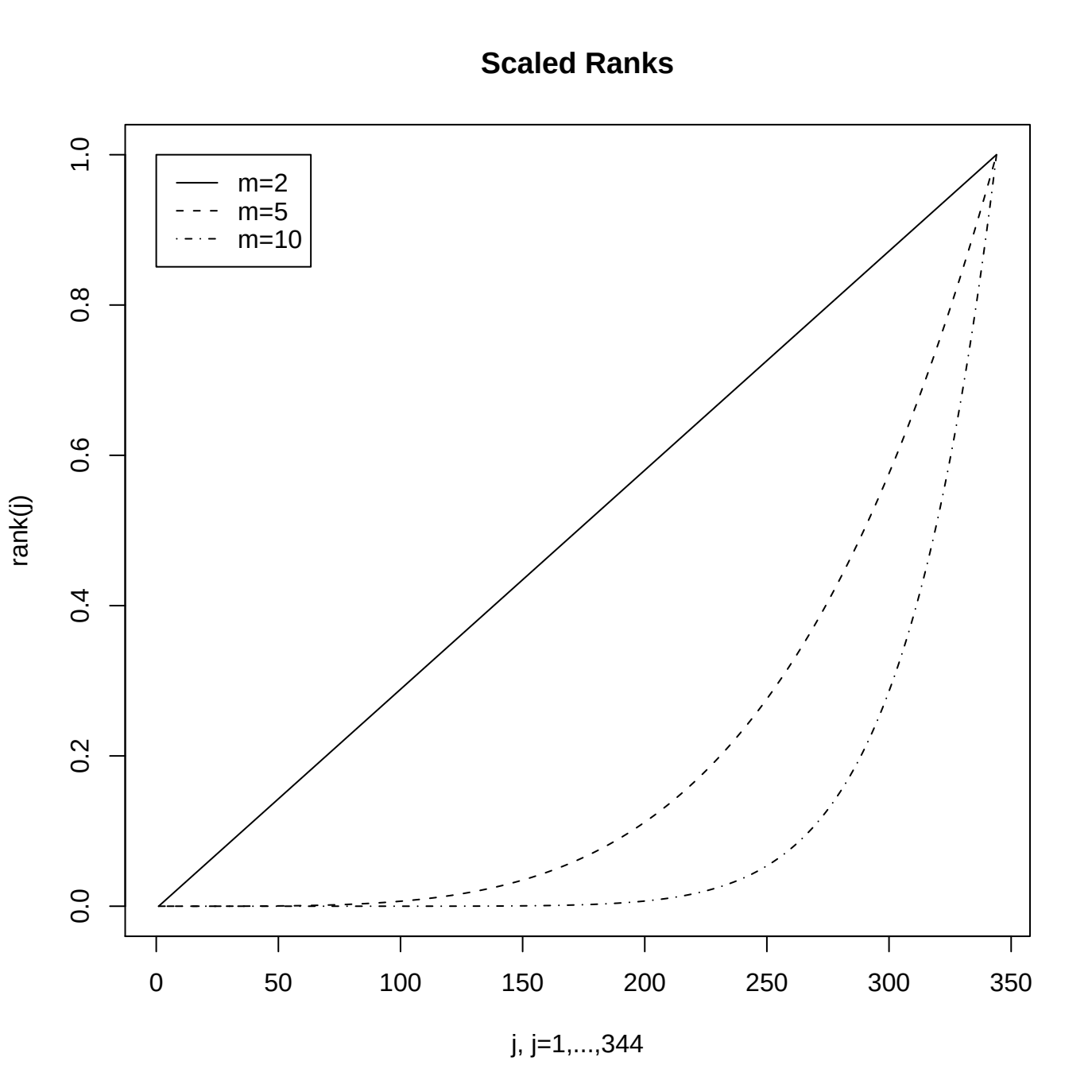
$j = 1, 2, \ldots, n$ are assigned ranks $\binom{j-1}{m-1}$, defined to be zero for $j < m$, because j wins in all $\binom{j-1}{m-1}$ subsets consisting of j and $m-1$ elements picked from $1, \ldots, j-1$.

33 Ranks, continued

Stephenson's ranks: $j = 1, 2, \ldots, n$ are assigned ranks $\binom{j-1}{m-1}$ are also a polynomial in j whose highest power is $m - 1$. For large n, aside from location/scale, they are virtually the same as Conover and Salsburg (1988) locally most powerful ranks.

Relationship to Wilcoxon ranks: For $m = 2$, Stephenson's ranks are $\binom{j-1}{m-1} = \binom{j-1}{2-1} = j - 1$, so they are essentially conventional ranks, again consistent with Lehmann's result.

Confidence intervals: Stephenson's ranks can be used to invert a rank test to obtain confidence statements.


Scaled Ranks
m=2
m=5
m=10
rank(j)
0.0
0.2
0.4
0.6
0.8
1.0
0
50
100
150
200
250
300
350
j, j=1,...,344

Sensitivity Analysis: Chemotherapy Weeks (year 1)

(Chemo1)	Wilcoxon	Stephenson	Stephenson
Γ	$(m = 2)$	$(m = 5)$	$(m = 10)$
1	0.0012	0.00044	0.00080
1.2	**0.048**	0.0086	0.0066
1.4	0.31	**0.055**	0.027
1.5	0.50	0.11	**0.046**

Sensitivity Analysis for Toxicity Weeks (year 1)

(Adv1)	Wilcoxon	Stephenson	Stephenson
Γ	$(m = 2)$	$(m = 5)$	$(m = 10)$
1	6.3×10^{-7}	5.0×10^{-9}	1.8×10^{-8}
1.5	**0.036**	0.00013	2.7×10^{-5}
2	0.62	**0.011**	0.00099
3	1.00	0.34	0.031
3.2	1.00	0.45	**0.046**

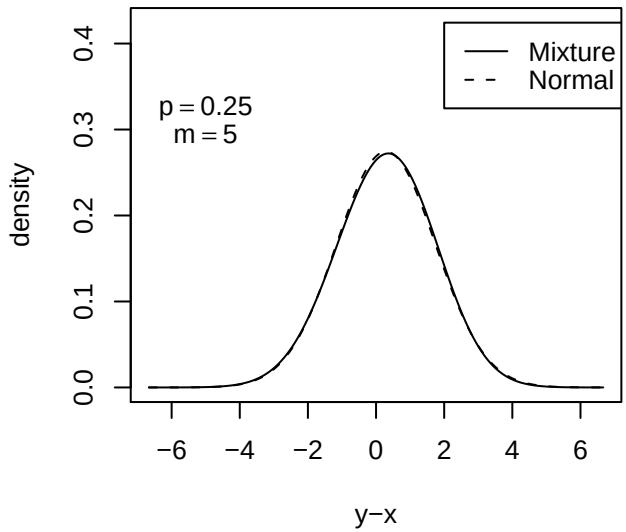
Summary: In this data set, less sensitivity to unobserved bias when the analysis looks for dramatic responses from a subset of treated subjects ($m > 2$).

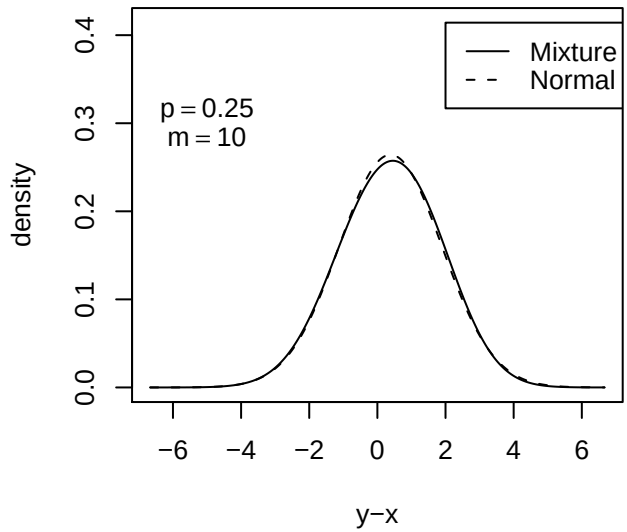
34 Is this pattern true in general?

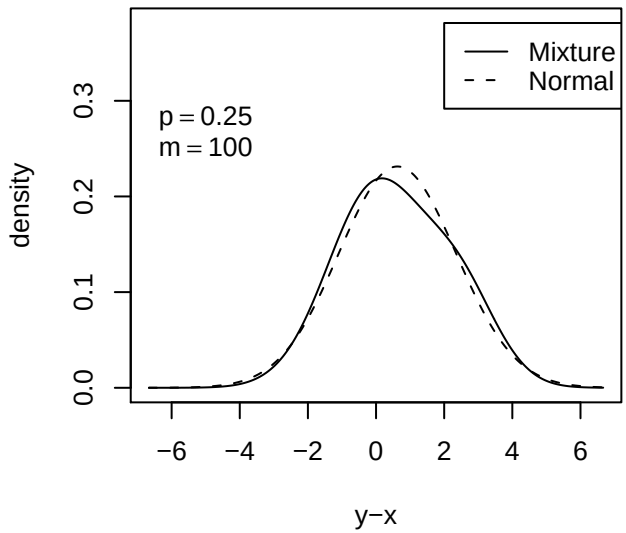
Design sensitivity: Limiting sensitivity to unobserved bias, as $I \rightarrow \infty$, when in fact, unknown to us, the treatment actually had an effect and there is no unobserved bias.

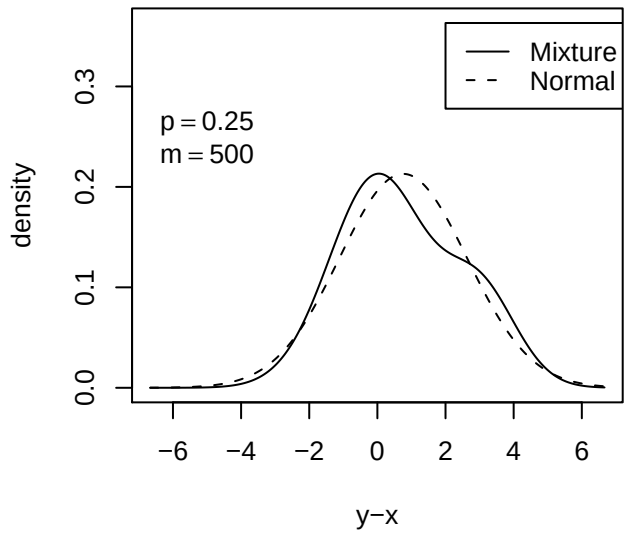
Model: I treated- minus-control differences D_i sampled from $r_{Cij} \sim \Phi(\cdot)$ and $r_{Tij} \sim (1-p)\Phi(\cdot) + p\Phi^{\overline{m}}(\cdot)$ with $p = .25$

Compare: design sensitivity, $\widetilde{\Gamma}$, for Wilcoxon's signed rank test ($m = 2$) and for Stephenson's test with $m = 5$ or $m = 10$. Notice that m may or may not equal $\overline{m}$.


Mixture
Normal
p = 0.25
m = 5
density
0.0
0.1
0.2
0.3
0.4
−6
−4
−2
0
2
4
6
y−x


Mixture
Normal
p = 0.25
m = 10
density
0.0
0.1
0.2
0.3
0.4
−6
−4
−2
0
2
4
6
y−x


Mixture
Normal
p = 0.25
m = 100
density
0.0
0.1
0.2
0.3
−6
−4
−2
0
2
4
6
y−x


Mixture
Normal
p = 0.25
m = 500
density
0.0
0.1
0.2
0.3
−6
−4
−2
0
2
4
6
y−x

35 Table of design sensitivities

Model: I treated- minus-control differences D_i sampled from $r_{Cij} \sim \Phi(\cdot)$ and $r_{Tij} \sim (1-p)\,\Phi(\cdot) + p\Phi^{\overline{m}}(\cdot)$ with $p = .25$

	Wilcoxon ($m = 2$)	Stephenson ($m = 5$)	Stephenson ($m = 10$)
$\overline{m} = 5$	1.6	1.8	2.0
$\overline{m} = 10$	1.8	2.2	2.5
$\overline{m} = 100$	2.2	3.6	5.5
$\overline{m} = 500$	2.4	4.7	8.9

Summary: If only 25% respond to treatment, use of ranks that target responders yields higher values of the design sensitivity, $\widetilde{\Gamma}$.

36 Summary

Re Ovarian Cancer: There was no sign that greater intensity of chemotherapy improved survival, although it did increase serious toxicity.

Re Matching: 'Fine balance' constrains an optimal match to perfectly balance a nominal variable, without constraining who is matched to whom. The nominal variable may have many levels. Easy to do in R.

Re Concerns about this observational study: All of the discussion was about unobserved covariates. Right orientation, but lacking data and quantification.

What makes studies insensitive to unobserved biases? Varied answers, some familiar, others surprising (at least to me). Here, uncommon but dramatic responses to treatment can be quite insensitive to unobserved biases if this pattern targeted in the analysis.

Formula for design sensitivity. Stephenson's ranks are $q_i = \binom{i-1}{m-1}$ with $q_i = 0$ for $i < m$. Also, $\sum_{i=1}^{I} q_i = \sum_{i=m}^{I} \binom{i-1}{m-1} = \binom{I}{m}$. In Stephenson's signed rank statistic, say S, q_i is the 'rank' of the $|D_i|$, and q_i enters the statistic if $D_i > 0$, where $m = 2$ is (essentially) Wilcoxon's signed rank statistic.

In a randomized experiment *under the null hypothesis* of no effect, q_j is added to S with probability $\frac{1}{2}$. With $\pi_{i1}/(\pi_{i1}+\pi_{i2})$ bounded by $\Gamma/(1+\Gamma)$, q_i is added to S with probability at most $\Gamma/(1+\Gamma)$, so the maximum expectation of S is equal to $\Gamma/(1+\Gamma) \times \sum q_i$ or $\Gamma\binom{I}{m}/(1+\Gamma)$.

If no bias from unobserved covariates (randomization, iid differences), with a treatment effect, then S has expectation $\binom{I}{m}p$ where p is the probability that the largest of m absolute differences, $|D_1|, \ldots, |D_m|$ is positive.

Design sensitivity $\widetilde{\Gamma}$ solves $\frac{\Gamma}{1+\Gamma}\binom{I}{m} = \binom{I}{m}p$, so it is $\widetilde{\Gamma} = p/(1-p)$.