11-2009

Analysis of Subgroup Data in Clinical Trials

Kao-Tai Tsai  
Georgia Southern University

Karl E. Peace  
Georgia Southern University, kepeace@georgiasouthern.edu

Follow this and additional works at: https://digitalcommons.georgiasouthern.edu/biostat-facpres

Part of the Biostatistics Commons, and the Public Health Commons

Recommended Citation
https://digitalcommons.georgiasouthern.edu/biostat-facpres/1

This conference proceeding is brought to you for free and open access by the Biostatistics, Department of at Digital Commons@Georgia Southern. It has been accepted for inclusion in Biostatistics Faculty Presentations by an authorized administrator of Digital Commons@Georgia Southern. For more information, please contact digitalcommons@georgiasouthern.edu.
Subgroup Data Analyses of Clinical Trials - *Issues & Methods*

Kao-Tai Tsai\(^1\) & Karl E. Peace\(^2\)

\(^1\)Presented at BASS XVI
Mulberry Inn Suites
Savannah, Georgia
November 10, 2009

\(^1\)tsai0123@yahoo.com
\(^2\)kepeace@georgiasouthern.edu
Analysis of Subgroups: Revisiting an old issue

A.1985: Protocol $H_0: C_{800 \text{mg}} = 0$ vs. $H_a: C_{800 \text{mg}} > 0$ the elderly;
   - Subset the database from a large ($n = 771$) trial of
     $C_{1600}$ vs $C_{800}$ vs $C_{400}$ vs $C_0$; all hs: found # elderly = 101;
     \{95%/90% CI on $(C_{800} - C_0) = (10\% - 66\%) / (14\% - 62\%)$.\}
   - Recommended $N = 0$ for the protocol

B.1986: Developed Protocol with 95% power
   - Randomization within each of two strata
   - Each Stratum: approximately 75% power
   - Analysis plan: Within Strata (Unconditional); if both significant, claim
     reproducibility had been demonstrated, and the two adequate
     and well-controlled requirement was met. If not; analyze as
     one protocol.

C. IND/NDA Rewrite which gave signal that NDAs should do due
diligence regarding inference to demographic subgroups.

   Subpopulations.*
   - Peace KE: *What’s the Question?* Inference within strata? Generalizability?

11/19/2009
Analysis of Subgroups: Revisiting an old issue

E. Various Groups have legitimate interests

- Statisticians
  * Pharmaceutical Industry
  * FDA, other governmental agencies & Academia

- Clinicians
  * Pharmaceutical Industry
  * FDA, other governmental agencies & Academia (Med. Univ.s)

- Practicing Physicians

- Public Health Professionals
Overview of Discussion

1. Randomized Controlled Trials
2. Why or why not perform subgroup analyses?
3. Test of treatment effect in subgroups
   A. False Positives
      - Simulation study (Brookes et al.) & Statistical Remedies
   B. False Negatives
      - Test of interaction & Sample size & power considerations (Brookes et al.)
4. Procedures for Subgroup Analyses
   A. Matching criteria
      - Notation, Propensity Score, Mahalanobis, Genetic, Example
   B. Randomization & Population Model Inference
   C. Randomization test and confidence intervals with Example
5. General guidelines to consider.
6. Summary

Appendix: Outline of Flexible Testing Strategy (Huque/ Alosh)
1. Randomized Controlled Trials (RCTs)

- Why conduct RCTs?
  - RDBCT is the gold standard for evidence
  - Analyses of large randomized controlled trials and systematic reviews (e.g. correct meta-analyses of RCTs) are the most reliable methods for determining the effects of treatments (Rothwell, 2005).

- Rational Criteria for World Health Care Planning
  
  "There is simply no serious scientific alternative to the generation of large-scale randomized evidence. If trials can be vastly simplified, …, and thereby made vastly larger, then they have a central role to play in the development of rational criteria for the planning of health care throughout the world."
  
1. Randomized Controlled Trials (RCTs)

• Purpose of first RCTs.
  – When trials were first developed for use in agriculture, researchers were presumably concerned about the effect of interventions on the overall size and quality of the crop rather than on the well-being of any individual plant.

• Individualized medicine.
  – In modern medical practice, clinicians frequently need to make decisions about individuals, and how best to use results of RCTs and systematic reviews to maximize the well-being of each patient.
2. Why or why not Perform Subgroup analyses?
   - Why Perform Subgroup Analysis?

• Recruitment of a large number of eligible patients from a general population is both a major strength and weakness of large pragmatic trials.
  – Deliberately broad and sometimes generously-defined entry criteria mean that the overall result can be difficult to apply to particular groups.

• Subgroup analyses are necessary if heterogeneity of treatment effect is likely to occur.
2. Why or why not Perform Subgroup analyses?
   - Arguments against Subgroup Analyses

- Statisticians and non-clinical epidemiologists have warned about the dangers of subgroup analyses and other attempts to target treatment:
  - Application of **False positive** subgroup findings may be harmful to patients; such findings may be more common than genuine heterogeneity due to multiplicities or data dredging.
  - Qualitative **heterogeneity** of relative treatment effect (benefit in one subgroup and harm in another) is **rare**. Statistical testing rarely reveals any significant findings.
2. Why or why not Perform Subgroup analyses?

- Arguments for Subgroup Analyses

  • Clinician's warning of the dangers of applying the overall results of large trials to individual patients without consideration of pathophysiology or other determinants of individual response.

  • “. . . It would be unfortunate if the desire for the perfect (i.e., knowledge of exactly who will benefit from treatment) were to become the enemy of the possible (i.e., knowledge of the direction and approximate size of the effects of treatment for wide categories of patients).”

    -- S. Yusuf et al. (1984)
2. Why or why not Perform Subgroup analyses?

- Arguments for Subgroup Analyses

• “The tragedy of excluding cogent pathophysiologic subgroup analyses merely because they happen to be subgroups will occur if statisticians do not know the distinction, and if clinicians who do know it remain mute, inarticulate or intimidated.” A Clinicostatistical Tragedy.
  -- A R Feinstein (1998)

• “Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise.”
  -- J W Tukey (1962)
2. Why or why not Perform Subgroup analyses?

- Arguments for Subgroup Analyses

“Many of the arguments used against subgroup analyses misinterpret their main function. The main potential of subgroup analysis is not in the identification of groups that differ in their response to treatment for reasons of pathophysiology, but is in answering practical questions about how treatments should be used most effectively, such as at what stage of the disease is treatment most effective, how soon after a clinical event is treatment sufficiently safe or most effective, or how are the risks and benefits related to comorbidity? Subgroup analyses related to questions of the practical application of interventions can be vital to effective clinical practice.” -- Rothwell (2005)
2. Why or why not Perform Subgroup analyses?
   - Important Clinical Indications for Subgroup Analyses
     (Rothwell, 2005)

• Potential heterogeneity of treatment effect related to risk
  – Differences in risks of treatment
  – Differences in risk without treatment

• Clinically important questions related to the practical application of treatment
  – Does benefit differ with severity of disease?
  – Does benefit differ with stage in the natural history of disease?
  – Is benefit related to the timing of treatment after a clinical event?
  – Is benefit dependent on comorbidity?
2. Why or why not Perform Subgroup analyses?
   - Important Clinical Indications for Subgroup Analyses
     (Continued: Rothwell, 2005)

- Potential heterogeneity of treatment effect related to pathophysiology
  - Multiple pathologies underlying a clinical syndrome.
  - Differences in the biological response to a single pathology.
  - Genetic variation.

- Underuse of treatment in routine clinical practice due to uncertainty about benefit
  - Underuse of treatment in specific groups of patients, e.g., in the elderly.
  - Confining treatment according to a narrow range of values of relevant physiological variables, e.g., treatment thresholds for cholesterol level or blood pressure.
3. Test of Treatment Effect in Subgroups
   - Categories of Statistical Concerns

   A. False positives – multiplicities of tests
   B. False negatives – interaction tests
      - Findings From a Large Simulation Study

(Brookes et al. (2004): J. Clin. Epidemiology)
3. Test of Treatment Effect in Subgroups
   A. False Positive Concerns (*Brookes et al.*)

Table 5: Summary of FPRs of Tests of No Differential Subgroup Effect

<table>
<thead>
<tr>
<th>True overall Rx effect</th>
<th>Observed overall Rx effect</th>
<th>Both subgroup tests significant</th>
<th>One subgroup test significant</th>
<th>Neither subgroup test significant</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Not significant</td>
<td>0%</td>
<td>&lt; 1%</td>
<td>7%</td>
</tr>
<tr>
<td>Rx effect = 0</td>
<td>Significant</td>
<td>2.5%</td>
<td>0%</td>
<td>64%</td>
</tr>
<tr>
<td>Rx effect ≠ 0</td>
<td>Not significant</td>
<td>0%</td>
<td>&lt; 1%</td>
<td>21%</td>
</tr>
<tr>
<td></td>
<td>Significant</td>
<td>33%</td>
<td>0%</td>
<td>57%</td>
</tr>
</tbody>
</table>
3. Test of Treatment Effect in Subgroups

A. False Positive Concerns:
Statistical Remedies for Multiplicities in Subgroup Analysis

- Multiple test adjustment:
  - Bonferroni, Holm, or bootstrap method (Westfall & Young), etc.
- Use closed or hierarchical testing procedures.
- Spending proportions of $\alpha$-level properly on subgroup analyses within the total Type-I error rate.
  - Huque & Alos (2008): ‘$\alpha$-spending’ and hierarchical? Strongly controls the FWER.
- Improve the precision of the endpoint measurement in the subgroups.
- Choose and justify a different minimum standard of efficacy for subgroups.
3. Test of Treatment Effect in Subgroups

B. False Negative Concerns: – Interaction Tests
Quantitative & Qualitative Interaction (Gail & Simon – 1985)

Remark: $\delta_i = \text{treatment effect in the ith subgroup}$. Off diagonal areas in quadrants 1 and 3 indicate quantitative interaction. Quadrants 2 & 4 indicate qualitative interaction.
3. Test of Treatment Effect in Subgroups

B. Statistical Tests of Interaction

• D. Cox (*International Statistical Review*, 1984)
• M. Silvapulle (*Biometrics*, 2001)
3. Test of treatment effect in subgroups.

B. Test of interaction.

Sample size & power considerations (Brookes et al.)
Frequency of Significant Interaction Tests As a Function of Relative Size

![Graph showing frequency of significant interaction tests as a function of relative size](image-url)
3. Test of treatment effect in subgroups.

B. Test of Interaction

Sample Size Inflation Factor (Brookes et al), when
Power(Test for interaction) = Power(Test for treatment effect)
4. Procedures for Subgroup Analysis
   A. Matching

• Match subjects with similar background to reduce imbalance:
  – Matching criteria:
    • Dividing the values of covariates.
    • Propensity score.
    • Mahalanobis distance, Genetic algorithm.
  – Matching covariates should be based on clinical relevance.

• Divide subjects into a few “homogeneous” strata. Estimate treatment effect within each stratum. Overall result can be derived by aggregation.
4.A.1 Statistical Procedures - Notations

Define treatment effect:

- $Y_{i1}$ = response of active treatment for subject $i$,
- $Y_{i0}$ = response of control treatment for subject $i$.

Define treatment indicator:

- $X_i$ = covariates associated with subject $i$.
- $T_i = 1(0)$ if subject $i$ receives active (control) treatment.

The observed outcome for subject $i$ is

$$Y_i = T_i Y_{i1} + (1 - T_i) Y_{i0}.$$
4.A.2 Propensity Score Matching  
(Rosenbaum & Rubin; 1983)

Definition of Propensity Score:

\[ e(X_i) = P(T_i = 1 | X_i) = E(T_i | X_i). \]

Assume

(i) \(0 < P(T_i | X_i) < 1\) and

(ii) \(P(T_1, \ldots, T_N | X_1, \ldots, X_N) = \prod_{i=1}^{N} e(X_i)^{T_i}(1 - e(X_i))^{(1-T_i)},\)

then the treatment effect, \(\Delta_{(T=1)}\), can be estimated by the following expression:

\[ \Delta_{(T=1)} = E\{E(Y_i | e(X_i), T_i = 1) - E(Y_i | e(X_i), T_i = 0)|T_i = 1\}. \]
4.A.2.1 Remarks on Propensity Score Est.

Let the vector of covariates $X_i = (x_{i1}, x_{i2}, \ldots, x_{ik})$ and $m \leq k$,

- A common method to estimate $e(X_i)$ is via

$$\text{logit}(e(X_i)) = \beta_0 + h_1(\eta_{1i}) + h_2(\eta_{2i}), \quad (1)$$

where $h_1$ and $h_2$ are known functions of smooth functions $f_r$ and $f_q$ with

$$\eta_{1i} = \sum_{r=1}^{m} f_r(x_{ir}) \quad \& \quad \eta_{2i} = \sum_{r,q=1}^{m} f_r(x_{ir})f_q(x_{iq}).$$

- Eq(1) can be estimated using MLE. Goodness-of-fit can be checked via Landwehr, et al (1984) or Tsai (2008) graphically.
4.A.2.1 Remarks on Propensity Score Est.

• According to Rosenbaum & Rubin (1983), it is advantageous to sub-classify or match not only on $e(x)$ but for other functions of $x$ as well;

• In particular, such a refined procedure may be used to obtain estimates of the average treatment effect in subpopulations defined by components of $x$, for example, males and females.
4.A.3 Mahalanobis & Genetic Matching
(Diamond & Sekhon; 2006)

Mahalanobis matching: distance between covariates $X_i$ and $X_j$:

$$md(X_i, X_j) = \{(X_i - X_j)' S^{-1} (X_i - X_j)\}^{1/2}.$$

Genetic matching: distance between covariates $X_i$ and $X_j$:

$$gmd(X_i, X_j) = \{(X_i - X_j)' S^{-1/2} W S^{-1/2} (X_i - X_j)\}^{1/2},$$

where

- $W$ is a diagonal positive definite weight matrix,
- $S^{1/2}$ is the Cholesky decomposition of the covariance matrix of $X$. 

4.A.4 Example for Illustration

- **A rheumatoid arthritis study:**
  - A phase III, multicenter randomized, double blind, placebo controlled comparative study of A or B in combination with methotrexate in controlling disease activity in subjects with rheumatoid arthritis having an inadequate clinical response (DAS_28J) to methotrexate.

- **Sample sizes and baseline comparisons:**
  - A group: 156; B group: 165; Placebo group: 110.
  - All baseline variables were balanced according to t-test.
  - No direct comparison between A and B in design.

- **Objective of this example:** to compare A with B after 6 months of treatment.
4.A.4.1 Balancing Baseline CRP values via Propensity Score Matching
4.A.4.2 Balancing Baseline CRP values via Genetic Matching
4.A.4.3 Balancing Baseline Pain Score via Propensity Score Matching
4.A.4.4 Balancing Baseline Pain Score via Genetic Matching
4.B Randomization & Population Model Inference

• **Randomization model:**
  – The basis for inference in the randomization model is the random assignment of patients to treatments.
  – It is not necessary to have random sample from a population with specific distribution.
  – Strictly speaking, normal theory methods are not appropriate since their distribution theory depends on random sampling.
  – Any inferences from randomization model are limited to the subjects in the study (i.e. a local inference).

• **Population model:**
  – The basis for inference in the population model is the random sample from a population with specific distribution.
  – Any inference from sample can be generalized to the whole population.
4.B.1 Comparison of Pre & Post Matching Estimate of Rx Effect – A Simulation: Response = DAS_28J;
Baseline Covariates = (age, pain, crp, #tend_J, #swol_J, haq)

(1) Combine data from group A and group B (call this sample S).
(2) From S, randomly sample 156 subjects and call this new Group A. The rest is new Group B (to perform permutation test).
(3) Estimate treatment effect (mean difference) Using new Group A and new Group B. No matching here.
(4) Using the same sample as in step 3 above, Genetically match subjects on covariates in new Group A and new Group B. Then estimate treatment effect.
(5) Form difference in estimated Treatment effects from step 3 (no matching) and step 4 (genetic matching).
(6) Repeat steps 2 to 5 for 400 times, producing 400 differences between estimates of treatment effect with matching and no matching. Note: There are \( C_{321}(156) = 321!/156!165! \) possible samples.
(7) Sort the 400 differences from step 6 and plot frequency distribution and generate QQ plot.

11/19/2009
4.B.1 Comparison of Pre & Post Matching Estimate of Rx Effect – A Simulation: Response = DAS_28J;
Baseline Covariates = (age, pain, crp, tend_J, swol_J, haq)
4.B.2 Randomization Test of Pre & Post Matching
Rx Effect: (Pre Est = -0.19 & Post Matching = -0.048)
A Simulation
4.B.3 Randomization Upper Confidence Limit
(Robbins-Monro Stochastic Approx.)

Assume

- the distributions of the two samples are $F(X) = F(Y - \theta)$,
- the hypothesis to be tested is $H_0 : \theta = U$ vs $H_1 : \theta < U$,
- $T(U)$ and $T^*(U)$ are test stat based on the randomized and
  original data, respectively,
- The value of $U$ at the $i$th step, $U_i$, is updated by the following

$$U_{i+1} = \begin{cases} 
  U_i - c\alpha / i & \text{if } T(U_i) > T^*(U_i), \\
  U_i + c(1 - \alpha) / i & \text{otherwise,}
\end{cases}$$

where $c = k(U_i - \hat{\theta})$ and $k = 2 / \{\sqrt{2\pi} \ z_{1-\alpha} \exp(-z_{1-\alpha}^2/2)\}$.
4.B.4 Randomization Lower Confidence Limit
(Robbins-Monro stochastic approx.)

Assume

- the distributions of the two samples are \( F(X) = F(Y - \theta) \),
- the hypothesis to be tested is \( H_0 : \theta = L \) vs \( H_1 : \theta > L \),
- \( T(L) \) and \( T^*(L) \) are test stat based on the randomized and original data, respectively,
- The value of \( L \) at the \( i \)th step, \( L_i \), is updated by the following

\[
L_{i+1} = \begin{cases} 
L_i + c\alpha/i & \text{if } T(L_i) < T^*(L_i), \\
L_i - c(1 - \alpha)/i & \text{otherwise,}
\end{cases}
\]

where \( c = k(\hat{\theta} - L_i) \) and \( k = 2/\{\sqrt{2\pi} Z_{1-\alpha} \exp(-Z_{1-\alpha}^2/2)\} \).
4.B.5 95% CI of Rx Effect Difference

(Based on Randomization Test: 5000 iterations)

Robbins-Monro Estimate of 95% CI = (-0.110, 0.4858)
Example Summary

• Matching is a ‘type of covariate adjustment’
• The adjustment derives from the matching on clinically important covariates rather than from entering covariates in some analysis model
• Inference on treatment effect (-0.048) is from the permutation distribution of the matched data (Pre Est = -0.19); and
• CIs from Robbins-Monro: 95% CI = (-0.110; 0.4858) - which requires one-sided permutation tests in search of LL and UL.
5. General Guidelines for Subgroup Analysis (1)

• **Pre hoc:**
  – Pre-specify the subgroup analyses, the magnitude and direction of the hypothesized subgroup effects; i.e. correctly specify the question
  – Consider subgroups based on clinically important covariates.
  – The number of subgroup analysis should be kept small and well documented
  – If feasible, adequately power planned subgroup analyses.
  – Test for treatment and subgroup interaction to investigate whether the effect is different among subgroups.

• **Post hoc:**
  – Properly match subjects to reduce the potential bias due to the lack of randomization.
  – Post hoc findings should be regarded as exploratory and useful for hypothesis generating.
5. General Guidelines for Subgroup Analysis (2)

- **Interpretation:**
  - The lack of statistical significance only implies the inadequate conviction of effect from the data base.
  - Statistical significant findings of effect should not be overemphasized.
  - Post hoc findings are not automatically invalid, many medical discoveries have been fortuitous; e.g. Penicillin.
  - Interpret the findings in the context of biological plausibility and clinical relevance.

- **Final verdict:**
  - The best measure of the validity of subgroup effects is independent verification in another trial.
6. Summary

• **Interpretation of findings:**
  – The lack of statistical significance only implies inadequate conviction of effect from the data analyzed;
  – ‘Statistically significant findings’ of effect should not be overemphasized;
  – Post hoc findings are not automatically invalid, many medical discoveries have been fortuitous. e.g. Penicillin
  – Interpret findings within the context of biological plausibility and clinical relevance with proper intelligence.

• **A better verdict:**
  – A better measure of the validity of subgroup effects is through independent verification in the proper context if feasible.
6. Summary (cont’d)

- **The crux of scientific exploration:**
  - Positive findings should not be dismissed automatically just because they are counter-intuitive or can’t be interpreted plausibly.
  - Negative findings should not conclude automatically as no effect. According to Dr. Robert Temple of the FDA, about 46% of well-conducted trials of effective antidepressants cannot distinguish drug from placebo [see Leber (1983) Control CTs Soc. Mtg].
  - Sir Isaac Newton once said, “Great discoveries are often achieved off the beaten paths.” Minoxidil, AZT?
6. Summary *(cont’d)*

Do subgroup analysis results have a place in the realm of scientific inference *(Peace; 1994)*?

- Maybe so, maybe not

Suppose proper subgroup analyses lead to statistically significant treatment effects in two subgroups *(Post Hoc)*. Can we infer that treatment effect has been confirmed? And therefore meet the two adequate and well controlled requirement?

If such results are viewed as reproducible, their level of scientific credibility would not be as great as that from results of two prospective CTs *(or two identified subgroups within a CT)* where there was a commitment to the questions and methods.

- Have a place in labeling efforts

One has established that new drug is effective. Explore entire database using interaction tests of treatment and subpopulations *(FDA Demographic Rule)*.

Lack of significant interaction lends support to overall effect being consistent across subpopulations.

For significant interactions perform subgroup analyses to obtain CIs on treatment effect in subpopulations. Qualitative interactions could be very important.
Appendix: Outline of Flexible Testing Strategy
(Huque and Alosh; 2008):

$H_0$: $\delta = 0$ (or $\delta \leq 0$, 1-sided hypothesis) is the null hypothesis for the overall population,

$H_{0s}$: $\delta_s = 0$ (or $\delta_s \leq 0$, 1-sided hypothesis) is the null hypothesis for the targeted subgroup,

$T = \text{test statistic is used for testing } H_0$,

$T_s = \text{test statistic is used for testing } H_s$, and

$(p, p_s) = \text{observed p-values for the testing of } H_0 \text{ and } H_{0s}$, respectively.
Appendix: Outline of Flexible Testing Strategy  
(Haque and Alosh; 2008):

Testing strategy is summarized in the following steps:

Step 1: Reject $H_0$ at level $\alpha_0$ if $T > C_{1;1-\alpha_0}$; where $\alpha_0 < \alpha$.

Step 2:
   (a) If $H_0$ is rejected in Step 1, then test $H_0s$ at the full significance level of $\alpha$; i.e. reject $H_0s$ if $T_0 > C_{2;1-\alpha}$.
   (b) If $H_0$ is not rejected in Step 1, and $\alpha_0 < p < \alpha^*$ ($\alpha^*$ is pre-specified, e.g. 0.10 for 1-sided $p$); then test $H_0s$ at the reduced significance level of $\alpha_s$, where $\alpha - \alpha_0 < \alpha_s \leq \alpha$, i.e. reject $H_0s$ if $T_s > C_{2;1-s}$. If $\alpha^*$ is in the interval $(\alpha_0; \alpha]$, then set $\alpha_s = \alpha$.
   (c) If $p > \alpha^*$, do not test for $H_0s$.

This testing strategy strongly controls the family wise error rate (FWER). Huque and Alosh provide tables for $\alpha_s$ for specified values of $\alpha_0$, $\alpha$ and $\rho$.
Appendix: Flexible Testing Strategy  
(Huque and Alosh; 2008):

When designing a clinical trial in which the interest is in establishing efficacy in one or more subgroups, in addition to that of the overall study population, several factors need to be considered at the design stage:
- pre-specification of hypotheses to be tested for subgroups and the overall population, and
- powering the study overall along with powering for subgroups.

The flexible testing strategy controls Type I error rate and permits testing for the subgroup after taking into account the need for a certain degree of efficacy consistency between the overall study population and that of subgroup. Computation shows that the power of the subgroup is an increasing function of the estimate of the correlation as well as treatment effect in the subgroup relative to that of the total study population. As a consequence of these observations, it is critical at the design stage to use conservative estimate for the correlation and subgroup treatment effect to reduce the chance of under powering the study for the subgroup.
References:


Notes:

- CRP = C-Reactive Protein; a marker of inflammation
- Example: A & B were biologics. One was a Cytokine Inhibitor of TNF (tumor necrosis factor - may regulate immune cells & have a role in inflammation)
- Baseline covariates: AGE, PAIN, CRP, #TEND, #SWOL, HAQ
- Inflation factor (Brookes): is a multiple
- Interpretation of QQ plots: Identical if on y = x; flatter or steeper means more dispersed; arced or S means skewness or one is more heavy tailed.
- Strong Ignorability: P[T| X] = P[T].
- In MHD and Genetic matching, i and j index treatment group. For every subject in A and every subject in B, their proximity in terms of their covariates is calculated according to the formula of GMD. In the end, there will be a big 156-by-165 matrix of proximities. Aggregate the subjects in A and subjects in B with good proximity and estimate the Rx difference. This is similar to Cluster Analysis.
multivariate matching using the Mebane & Sekhon (1998) genetic search algorithm to determine the weight each covariate is given. Balance is determined by examining cumulative pdfs of a variety of standardized statistics. By default, these statistics include t-tests and Kolmogorov-Smirnov tests. Many descriptive statistics based on empirical-QQ plots can be used or any user provided measure of balance.

The statistics are not used to conduct formal hypothesis tests, because no measure of balance is a monotonic function of bias and because balance should be maximized without limit.

The object returned by GenMatch can be supplied to the Match function (via the Weight.matrix option) to obtain causal estimates.

GenMatch uses genoud to perform the genetic search. Using the cluster option, one may use multiple computers, CPUs or cores to perform parallel computations.

Robbins-Monro: Theta-hat is the treatment difference from the 2 samples. Every time a new sample is simulated, theta-hat is updated. Therefore, it affects Ui too. This is not a part of R-GenMatch; was programmed in R.
Results of Six Trials of Exp. Antidepressant (NEW), Imipramine (IMI) and PBO (PBO Data Concealed); HAM_D_S) Mean Scores\(^a\) at last visit\(^b\) Leber: 1983

<table>
<thead>
<tr>
<th>Study</th>
<th>Common baseline</th>
<th>Drug</th>
<th>Sample Size</th>
<th>HDS</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study 1</td>
<td>2.39</td>
<td>NEW IMI</td>
<td>33</td>
<td>13.4</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IMI</td>
<td>33</td>
<td>12.8</td>
<td></td>
</tr>
<tr>
<td>Study 2</td>
<td>26.0</td>
<td>NEW IMI</td>
<td>39</td>
<td>13.0</td>
<td>0.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IMI</td>
<td>30</td>
<td>13.4</td>
<td></td>
</tr>
<tr>
<td>Study 3</td>
<td>28.1</td>
<td>NEW IMI</td>
<td>11</td>
<td>19.4</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IMI</td>
<td>11</td>
<td>20.3</td>
<td></td>
</tr>
<tr>
<td>Study 4</td>
<td>29.6</td>
<td>NEW IMI</td>
<td>7</td>
<td>7.3</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IMI</td>
<td>8</td>
<td>9.5</td>
<td></td>
</tr>
<tr>
<td>Study 5</td>
<td>37.6</td>
<td>NEW IMI</td>
<td>7</td>
<td>21.9</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IMI</td>
<td>8</td>
<td>21.9</td>
<td></td>
</tr>
<tr>
<td>Study 6</td>
<td>26.1</td>
<td>NEW IMI</td>
<td>37</td>
<td>11.2</td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IMI</td>
<td>32</td>
<td>10.8</td>
<td></td>
</tr>
</tbody>
</table>

\(^a\)Scores adjusted by analysis of covariance

\(^b\)Score at 4 weeks or, if none available, last score available prior to 4 weeks
### Results of Six Trials of Exp. Antidepressant (NEW), Imipramine (IMI) and PBO; HAM_D_S) Mean Scores\(^a\) at last visit\(^b\) Leber: 1983

<table>
<thead>
<tr>
<th>Study</th>
<th>Common baseline</th>
<th>Drug</th>
<th>Sample Size</th>
<th>HDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study 1</td>
<td>2.39</td>
<td>NEW</td>
<td>33</td>
<td>13.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IMI</td>
<td>33</td>
<td>12.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PBO</td>
<td>36</td>
<td>14.8</td>
</tr>
<tr>
<td>Study 2</td>
<td>26.0</td>
<td>NEW</td>
<td>39</td>
<td>13.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IMI</td>
<td>30</td>
<td>13.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PBO</td>
<td>36</td>
<td>13.9</td>
</tr>
<tr>
<td>Study 3</td>
<td>28.1</td>
<td>NEW</td>
<td>11</td>
<td>19.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IMI</td>
<td>11</td>
<td>20.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PBO</td>
<td>13</td>
<td>18.9</td>
</tr>
<tr>
<td>Study 4</td>
<td>29.6</td>
<td>NEW</td>
<td>7</td>
<td>7.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IMI</td>
<td>8</td>
<td>9.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PBO</td>
<td>7</td>
<td>23.5(^c)</td>
</tr>
<tr>
<td>Study 5</td>
<td>37.6</td>
<td>NEW</td>
<td>7</td>
<td>21.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IMI</td>
<td>8</td>
<td>21.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PBO</td>
<td>8</td>
<td>22.0</td>
</tr>
<tr>
<td>Study 6</td>
<td>26.1</td>
<td>NEW</td>
<td>37</td>
<td>11.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IMI</td>
<td>32</td>
<td>10.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PBO</td>
<td>36</td>
<td>10.5</td>
</tr>
</tbody>
</table>

\(^a\)Scores adjusted by analysis of covariance  
\(^b\)Score at 4 weeks or, if none available, last score available prior to 4 weeks  
\(^c\)NEW, IMI better than PBO, p<0.001