Über die Umsetzung von Matching-Verfahren mit Hilfe des SAS-Systems

Hans-Peter Altenburg
68219 Mannheim
E-mail: hpa-ma@gmx.de

Zusammenfassung

Im Beitrag sollen verschiedene Varianten dargestellt werden wie mit Hilfe des SAS-Systems Matching-Verfahren realisiert werden können. Neben Standardmethoden wie 1:1- oder 1:n-Matching werden auch besondere Aspekte klinischer Studien oder das Fall-Kontroll-Matching mit Hilfe eines Propensity-Scores vorgestellt.

Schlüsselworte: Matching, Confounding, Bias, SQL, Propensity Score, Prozedur LOGISTIC.

1 Einleitung

In der Epidemiologie und Statistik stellen Matching-Methoden eine wichtige Gruppe von Verfahren dar um den sogenannten Confounding Bias in retro-spektiven Vergleichen zu minimieren. Confounder sind Faktoren, die mit der Exposition assoziiert sind, aber nicht auf kausalem Weg, oder mit der Erkrankung assoziierte Faktoren, unabhängig von der Exposition. Confounder führen zur Verzerrung der Risikoschätzer! Als Confounder gelten für viele Zielgrößen z.B. Alter und Geschlecht (steigendes Risiko für Darmkrebs bei älteren Männern). Die Vermeidung von Verzerrungen kann einmal durch Matching (Matching nach dem Confounder) oder entsprechende Methoden bei der Datenanalyse geschehen (vgl. [1] oder [3]).
Matching nach einer oder mehreren Variablen (wie Alter, Geschlecht,...) ist eine Technik, die sicherstellt, dass Fälle und Kontrollen eine vergleichbare Verteilung besitzen. Sie kann zum einen als Gruppen-Matching (engl. frequency matching) oder als individuelles Matching (engl. matched pairs oder pairing) durchgeführt werden. Der Einfachheit halber sprechen wir im folgenden bei Fällen von Erkrankten und bei Kontrollen von Nichterkrankten.
Beim Gruppen-Matching werden erkrankte Teilnehmer mit einigen bestimmten Merkmalen (z.B. Alter, Größe, Gewicht, Essgewohnheiten, etc.) ausgewählt und dazu die gleiche Anzahl (oder der gleiche Anteil) an Nichterkrankten mit diesen Merkmalen gesucht, d.h. gleiche Verteilung auf einem Gruppenlevel. Die Realisierung dieser

Methode wird immer schwieriger, je mehr Merkmale betrachtet werden sollen. Beim individuellen Matching wird jedem erkrankten Teilnehmer ein Proband aus der Gruppe der Fälle gegenübergestellt, der die gleichen Merkmale besitzt.
Matching erhöht die Power. Jedoch ist es für die Datenanalyse u.U. problematisch, dass Merkmale, die für das Matching genutzt wurden, nicht mehr weiter untersucht, insbes. der Effekt dieser Matching-Variablen nicht mehr analysiert werden kann. Ansonsten können gematchte Studien wie „normale“ Studien analysiert werden, solange nach den Matching-Variablen adjustiert oder stratifiziert wird. Alternativ kann auch mit speziellen Verfahren, wie etwa bedingte logistische Regression, analysiert werden. Matching nach dem Confounder im Studiendesign kann zu einer Verzerrung führen, da die Kontrollen nicht mehr repräsentativ für die Quellpopulation sind. Abhilfe kann hier nur eine stratifizierte Analyse nach den Matching-Kriterien bringen. Unter Umständen kann es bei der Stichprobenerhebung auch zu einem Overmatching kommen, was u.U. das OR hinsichtlich der Exposition in Richtung Eins verzerrt. In diesem Fall erhält man zusätzliche unbeabsichtigte Matchings, z.B. indem durch die Bester-Freund-Methode der Stichprobenerhebung weitere gleiche Merkmale entstanden sind, bedingt durch den gleichen sozio-ökonomischen Status der betrachteten Paare. Vermeidung bzw. Kontrolle des Overmatching kann nur erfolgen, wenn Matching nur nach den wichtigsten Risikofaktoren, die sich als Confounder manifestieren, geschieht und nach den anderen in der Analyse adjustiert wird.
Die Aussagekraft des Matching ist stark von den gewählten Merkmalen abhängig. In den folgenden Abschnitten soll an Hand von Beispielen gezeigt werden, wie mit Hilfe des SAS-Systems Matching durchgeführt werden kann. Die Beispiele gehen über das reine Fall-Kontroll-Matching hinaus.

2 Selektion von Kontrollen (1:n-Matching)

Aufgabe: Fälle und Kontrollen einer Studie sollen nach zwei Variablen, Alter und Geschlecht, gematcht werden.

Hierzu definieren wir zunächst zwei Makro-Variablen, welche die für das Alter maximal zulässige Spannweite (im Beispiel fünf) und die Anzahl von Kontrollen (z.B. vier) definieren:

```
%LET agerange = 5 ;
%LET n_ctrl   = 4 ;
```

```
DATA case control ;
SET rohdaten ;
IF caco = 1    THEN OUTPUT case ;
               ELSE OUTPUT control ;
RUN ;
```

Anschließend erzeugen wir in der Gruppe der Fälle eine eindeutige Liste aller Kombinationen der Ausprägungen der beiden Matching-Variablen und speichern diese in einer SAS-Datei mit dem Namen `kombout` ab:

```
PROC FREQ NOPRINT DATA=case ;
TABLES age * sex / OUT=kombout ;
RUN ;
```

`v-count` speichert, wie oft eine Kombination vorkommt. Zunächst wird eine Untermenge der Kontrollpopulation `subset1` erzeugt und jedem Element eine Zufallszahl zugeordnet. Die SAS-Tabelle `subset2` wählt die Kontrollen aus dieser randomisierten Menge aus. Die Überwachungsvariable `tag` dient lediglich zur Dokumentation, ob die Anzahl gematchter Kontrollen erreicht wurde. Die hieraus resultierende Menge wird an die SAS-Datei `matches` angehängt. Gematchte Kontrollen werden abschliessend aus der Menge der Kontrollen entfernt.

```
%MACRO caseslct(v_age,v_sex,v_count) ;
DATA subset1 ;
 SET control ;
WHERE (&v_age-(&agerange) <= age <=&V_age + (&agerange) )
          AND
     (sex = "&v_sex") ;
case_age=&v_age;
case_sex="&v_sex";
SEED=RANUNI(0);
RUN ;
PROC SORT DATA=subset1; BY SEED;    RUN ;
DATA subset2 ;
SET subset1 NOBS=TOTOBS ;
IF _N_ <= (&v_count)*(&n_ctrl) ;
IF (&v_count)*(&n_ctrl) <= TOTOBS  THEN TAG = 'YES';
                                   ELSE TAG = 'NO ';
PROC APPEND BASE=MATCHES DATA=subset2 ;

PROC SORT DATA=subset2 OUT=TEMP1 (KEEP=id_var);
                                        BY id_var;
PROC SORT DATA=CONTROL OUT=TEMP2 ; BY id_var;
```

```
DATA CONTROL;
MERGE TEMP1(IN=IN1) TEMP2(IN=IN2);
BY id_var;
                IF IN2 AND NOT IN1;
RUN ;
%MEND caseslct;
```

Das Makro wird bedingt ausgeführt, gegeben die Kombinationen der Matching-variablen sind auch in der Menge der Fälle vorhanden:

```
DATA _NULL_; SET kombout ;
CALL EXECUTE
('%caseslct('||Age||','||Sex||','||Count||')');
RUN;
```

Statements zur Kontrolle, ob alles richtig verlaufen ist, können sich anschließen. Die Statements können leicht auf mehrere Matchingkriterien erweitert werden.

3 1:1-Matching mit Hilfe von SQL

Fragestellung: In einer epidemiologischen Studie soll die Frage untersucht werden, ob eine antiretrovirale Therapie die Lebensdauer von HIV-Patienten verlängert. Solche Studien sind selten randomisiert! Der Status des Immunsystems ist ein potentieller Confounder (er ist sowohl mit der Exposition als auch mit dem Ausgang korreliert). Dieses Problem wird reduziert durch „Paarung“ von Fällen und Kontrollen mit ähnlichem Immunsystem-Status, so dass Unterschiede im Status vor der Behandlung reduziert werden.
Situation:
Es liegen zwei Dateien vor: Fälle (SAS-Datei: `Case`) und Kontrollen (`Control`) mit den Variablen:

ID,
Datum Therapiebeginn: `startdt` (bei den Fällen),
Datum, wann T-Zellen gemessen wurden: `dt` (nur Kontrollen),
Maß für den Status des Immunsystems: `tcelln`.

Die Aufgabe besteht darin, eine Datei mit Fällen und Kontrollen zu erzeugen und alle möglichen Matching-Paare zu finden, unter der Restriktion:

Differenz der T-Zellenzahl ist maximal 50,
Differenz im Datum ist maximal 60 Tage.

SAS-Programm:

```
PROC SQL;
CREATE TABLE possmtch
     AS SELECT      case.id AS caseid,
                    control.id AS contrlid,
                    ABS(startdt - dt) AS dtdiff,
                    ABS(case.tcelln - control.tcelln) AS tdiff
     FROM case, control
WHERE ABS(case.tcelln - control.tcelln) < 50
     AND
ABS(startdt - DT) <= 60
ORDER BY caseid, contrlid, dtdiff;
DATA possmtch;
SET possmtch;
BY caseid contrlid;
IF FIRST.contrlid THEN OUTPUT;
```

Es wird eine SAS-Data-Set `possmtch` erzeugt mit den Variablen:

`caseid, contrlid, dtdiff, tdiff`

Die SAS-Data-Set kann mehrere Datums-Objekte für Fälle und Kontrollen enthalten, deshalb der sich anschließende Data Step: Dies liefert den Datensatz mit der kleinsten Datums-Differenz! 1:1-Matching, d.h. eine Kontrolle für jede Fall-ID geht am leichtesten über ein SAS-Makro:

SAS-Makro:

```
%MACRO MATCH121(ds_input, ds_final, case_id, ctrl_id, diff);
%LOCAL i j;
%LET i = 0;
%DO %UNTIL (&SQLOBS = 0);
%LET i = %EVAL(&i + 1);
PROC SORT DATA = &ds_input; BY &case_id &diff ;
DATA bestmtch ;
SET &ds_input ;
BY &case_id;
IF FIRST.&case_id THEN OUTPUT ;
PROC SORT DATA = bestmtch; BY &ctrl_id &diff;
DATA match&i ;
SET bestmtch ;
BY &ctrl_id;
IF FIRST.&ctrl_id THEN OUTPUT;
PROC SQL;
CREATE TABLE &ds_input AS
SELECT &ds_input..*
FROM &ds_input
WHERE &case_id NOT IN
(SELECT &case_id FROM match&i)
AND &ctrl_id NOT IN (SELECT &ctrl_id FROM match&i);
%END;
DATA &ds_final ;
SET     %DO j = 1 %TO &i; match&j %END;
;
PROC DATASETS;
DELETE bestmtch
%DO j = 1 %TO &i ; match&j %END;
;
%MEND MATCH121;
```

Für die Eingabe-Data-Set `ds_input` ist die obige SAS Datei `possmtch` zu verwenden. Die Schleife im Makro (mit der automatischen Makrovariablen `SQLOBS`) wird nur solange durchgeführt wie geeignete Matching-Paare vorhanden sind. Für jede `case_id` werden die Matchingkandidaten nach der Differenzvariablen aufsteigend geordnet, so dass der mit dem kleinsten Wert (d.h. der beste Kandidat) immer am Anfang steht. In der Hilfsdatei `bestmtch` stehen dann jeweils pro `case_id` die besten Matchingkandidaten. Im SQL-Step werden Schritt für Schritt alle bereits gefundenen Matches eliminiert.

4 Klinische Studie – Matching von Informationen

Im Reporting klinischer Studien müssen oft Behandlungsinformationen, wie etwa die verabreichte Dosis mit Safety Parametern wie Vitalitätsangaben, Adverse Events, ECG oder Laborwerten kombiniert werden. Eine wichtige Aufgabe der Dokumentation besteht darin, Dosis-relevante Safety-Angaben zu finden. So ist es etwa wichtig herauszufinden, welche Adverse Events einer bestimmten Dosis zugeordnet werden können bzw. in Folge der Verabreichung einer bestimmten Dosis auftreten. Die meisten solcher Angaben treten nur unregelmäßig auf oder fehlen vollständig. Im folgenden Programmbeispiel soll gezeigt werden, wie solche Daten mit Hilfe von SAS kombiniert (d.h. „gematcht“) werden können.
Wir nehmen an, die Dosis-Angaben liegen in Form einer SAS-Tabelle (`doseadmin`) vor, wie man sie etwa in einem Datenstep der folgenden Form eingelesen hätte:

```
DATA doseadmin ;
INPUT Obs PatNo DoseDtTM DoseAdmin ;
DATALINES ;
 1 0001 30JAN1999:08:00:01 10
 2 0001 06FEB1999:08:00:00 30
 3 0001 20FEB1999:08:00:00 90
 4 0001 27FEB1999:08:00:02 120
 5 0002 30JAN1999:08:01:02 10
 6 0002 06FEB1999:08:01:01 30
 7 0002 20FEB1999:08:01:00 90
 8 0002 06MAR1999:08:01:00 120
 9 0003 30JAN1999:08:02:01 10
10 0003 06FEB1999:08:02:01 30
11 0003 13FEB1999:08:02:01 60
12 0003 27FEB1999:08:02:08 90
13 0003 06MAR1999:08:02:01 120
...
RUN ;
```

In einer weiteren SAS-Tabelle `ad_event` sind alle Safety-Parameter wie z.B. unerwünschte Ereignisse (adverse events) oder weitere Variablen mit ihrem Auftreten dokumentiert:

```
DATA ad_event ;
INPUT Obs PatNo AE AEDtTM ;
DATALINES ;
 1 0001 NERVOUSNESS      30JAN1999:06:00:00
 2 0001 TACHYCARDIA      30JAN1999:12:15:00
```

```
 3 0001 NAUSEA            06FEB1999:16:20:00
 4 0001 DIZZINESS         20FEB1999:09:20:00
 5 0002 HEADACHE          06FEB1999:14:10:00
 6 0002 NAUSEA            06FEB1999:17:40:00
 7 0002 VASODILATATION    20FEB1999:14:30:00
 8 0002 ASTHENIA          20FEB1999:18:20:00
 9 0002 ANXIETY           27FEB1999:20:01:00
10 0002 CHILLS            28FEB1999:06:00:00
11 0002 CONSTIPATION      28FEB1999:22:00:00
12 0003 PALLOR            30JAN1999:10:17:00
13 0003 SWEATING          30JAN1999:12:17:00
14 0003 NAUSEA            06FEB1999:15:17:00
15 0003 DIZZINESS         06FEB1999:17:17:00
16 0003 HYPESTHESIA       13FEB1999:09:30:00
17 0003 PAIN              13FEB1999:12:30:00
18 0003 CHILLS            27FEB1999:16:45:00
19 0003 EUPHORIA          06MAR1999:10:15:00
20 0003 DIZZINESS         06MAR1999:18:46:00
...
RUN ;
```

Mit Hilfe des folgenden SAS-Makros können beide Data-Sets kombiniert werden:

```
%MACRO ae_match(dosedata, safedata, hrs, patno, dosetime,
                                          safetime, outdata) ;
PROC SORT DATA=&dosedata;
BY &patno &dosetime;
RUN ;
PROC FREQ DATA=&dosedata ORDER=FREQ NOPRUNT ;
TABLES &patno / OUT=temp1;
RUN ;
DATA _NULL_;
SET temp1;
IF _n_=1 THEN call symput('MaxDose',count);
RUN ;
PROC TRANSPOSE DATA=&dosedata OUT=temp2 (DROP=_name_)
PREFIX=dosen;
BY &patno;
VAR &dosetime;
RUN ;
%LOCAL num;
%DO num=1 %TO &maxdose;
PROC SORT DATA=temp2
OUT=dose&num (KEEP=&patno dosen&num);
```

```
BY &patno;
RUN ;
DATA ndata&num ;
MERGE &safedata(IN=a) dose&num(IN=b);
BY &patno;
IF a AND b;
dosen=&num;
hrs=(&safetime - dosen&num)/3600;
&dosetime=dosen&num;
%IF &hrs ne %THEN   %DO;
                    IF 0=<hrs<=&hrs THEN output ndata&num;
                    %END ;
%ELSE %IF &hrs eq %THEN %DO;
                        IF 0=<hrs THEN output ndata&num;
                        %END ;
DROP dosen&num;
RUN ;
%END ;
DATA &outdata;
SET ndata1;
RUN ;
%DO num=2 %to &maxdose;
DATA &outdata;
SET &outdata ndata&num;
RUN ;
%END ;
PROC SORT DATA=&outdata;
BY &patno dosen &safetime;
RUN ;
PROC PRINT ;
FORMAT &safetime &dosetime datetime18.;
TITLE3 "Output Data Set: &outdata";
RUN ;
%MEND ae_match;
```

Die wichtigsten Namen für die Verwendung des Makros sind: Dosis Data Set (`doseadmin`), Adverse Events Data Set (`ad_event`), Zeitfenster Dosis Variablen (z.B. 6, 12, etc. in Stunden), Patienten-ID (`PatNo`), Dosis Zeitvariable (`DoseDtTM`), Adverse Events Zeitvariable (`AeDtTM`) und der Name der Ausgabe Data Set. In vier Schritten werden im Makro die Daten aufbereitet:

Schritt 1:
Was ist die größte Zahl von Dosiswerten?
`PROC FREQ` → Makrovariable `maxdose`

Schritt 2:
Pro Patient alle Dosiswerte in einer Datenzeile
PROC TRANSPOSE

Schritt 3:
1:Many Merging
Zeitdifferenz: `Dosis-Zeit - Safety-Zeit`
Vergleich berechnete Zeitdifferenz mit definiertem Zeitfenster
Schleife erzeugt pro Dosiswert Untermenge

Schritt 4:
Zusammenführung aller Untermengen zur Ausgabe-Data-Set

Beispiel-Aufruf für ein Zeitfenster von sechs Stunden

```
%ae_match (doseadmin,ad_event,6,patno,dosedttm,aedttm,outdata)
```

Mit `PROC PRINT` angewandt auf die Ausgabedatei erhält man dann beispielsweise folgende Ausgabe (Ausschnitt):

Adverse Events
Zeitdifferenz: 6 Stunden

Pat No	Adverse Event	AeDtTM	DoseDtTM	Dose	Hours
0001	TACHYCARDIA	30JAN1999:12:15:00	30JAN1999:08:00:01	1	4.25
0001	DIZZINESS	20FEB1999:09:20:00	20FEB1999:08:00:00	3	1.33
0003	PALLOR	30JAN1999:10:17:00	30JAN1999:08:02:01	1	2.25
0003	SWEATING	30JAN1999:12:17:00	30JAN1999:08:02:01	1	4.25
0003	HYPESTHESIA	13FEB1999:09:30:00	13FEB1999:08:02:01	3	1.47
0003	PAIN	13FEB1999:12:30:00	13FEB1999:08:02:01	3	4.47
0003	EUPHORIA	06MAR1999:10:15:00	06MAR1999:08:02:01	5	2.22

5 Mehrere Matching-Variablen, m von n müssen erfüllt sein

Manchmal gibt es Situationen, dass es zwar viele Matching-Variablen gibt, aber nicht alle müssen für eine erfolgreiche Zuordnung erfüllt sein. Mit den folgenden SQL-Statements können im Prinzip solche Situationen gelöst werden. Die Namen der Variablen und Data Sets wurden selbsterklärend gewählt. Das Beispiel behandelt den Fall *n=3* und *m=2*:

```
PROC SQL;
CREATE TABLE PAIRS AS
SELECT  case.ADNUM    AS ADNUM1,
        control.ADNUM AS ADNUM2
FROM    RAWDATA case,
        RAWDATA control
WHERE ( (ABS(case.weight - control.weight)<=5 )
+       (case.sex   EQ control.sex)
+       (case.agegr EQ control.agegr) ) >= 2
AND     (case.ADNUM < control.ADNUM)
ORDER BY ADNUM1, ADNUM2 ;
QUIT;
```

Diese Vorgehensweise kann zu Problemen führen**:** SAS/SQL bildet ein kartesisches Produkt. Dies führt sehr schnell zur Ausschöpfung der Computer-Ressourcen. Es gibt schnellere und weniger Ressourcen verzehrende Alternativen auf die aber hier nicht weiter eingegangen werden soll.

6 Fall-Kontroll-Matching mit einem Propensity-Score

Bei Fall-Kontroll-Studien können durch unvollständiges oder nicht exaktes Matchen Verzerrungen auftreten. Unvollständiges Matching tritt auf, wenn Fälle ausgeschlossen werden müssen, weil keine entsprechenden Kontrollen in der Datenbasis vorhanden sind (Ziel: „Maximierung der Matches“). Nicht exaktes Matching tritt auf, wenn die Anzahl der Fälle maximiert werden soll. Die Verwendung eines Propensity Scores kann helfen Verzerrungen dieser Art zu verringern (vgl. [2], [4] und [5]). Im folgenden soll skizziert werden, wie mit Hilfe der Prozedur LOGISTIC ein Propensity Score erstellt und damit Fälle und Kontrollen gematcht werden können.
Ein Algorithmus für das Propensity Score-Matching könnte etwa folgendermassen aussehen:

- Bestimmung des Propensity Scores mit Hilfe der SAS Prozedur LOGISTIC,
- Greedy Matching mit Hilfe eines fünf-stelligen Scores,
- sucht die nächste passende verfügbare Kontrolle,falls mehrere Kontrollen passen, wird eine zufällig ausgewählt.Der Vorteil dieser Vorgehensweise ist,

dass bei mehreren Kovariablen die Abstände als Differenz der P-Scores zwischen Fällen und Kontrollen angesehen werden können. Das Verfahren liefert in der Regel gute „Matched-Pairs“. Beim sog. Greedy-Matching werden *n* Fälle auf eine Menge von *m* Kontrollen gematcht auf der Basis von *n* Entscheidungen. Nach der Zuordnung gibt es keine Änderungen mehr. Im Gegensatz dazu steht das optimale Matching: Man geht hier zunächst wie beim Greedy-Matching vor um anschließend die vorangegangenen Zuordnungen wieder (bzw. weiter) zu untersuchen, bevor eine Zuordnung stattfindet. Nach Bestimmung des Propensity Scores kann die Matching-Zuordnung analog wie in den vorangegangenen Abschnitten vorgenommen werden.
Das Programm zur Bestimmung des Propensity Scores könnte lauten:

```
PROC LOGISTIC DATA=dset DESCENDING ;
MODEL depvar = age sex
               ... weitere Variablen ...
/ SELECTION = STEPWISE RISKLIMITS LACKFIT RSQUARE PARMLABEL ;
OUTPUT OUT= ps_set PROB=prob ;
RUN;
```

Hierbei ist zu beachten, dass die abhängige Variable (`depvar`) eine 0-1-Variable sein muss (1=Fall, 0=Kontrolle). Der Propensity Score ist die Vorhersage-Wahrscheinlichkeit, dass ein Fall vorliegt. Die Greedy Zuordnung erfolgt über die Anzahl Stellen des Scores. Man beginnt mit 5 Stellen und geht erforderlichenfalls über zu 4 Stellen oder → 3 Stellen, ... usw.

7 Weitere alternative Vorgehensweisen für Matching

Abschließend seien noch ein paar weitere Methoden erwähnt, die dem Matching verwandt sind und verwendet werden könnten:

- die Benutzung von Distanzmaßen / Ähnlichkeitsmaßen für die Beschreibung der Ähnlichkeit einer Kontrolle zu einem Fall,
- Fuzzy Matching (ähnlich wie das Merging von Files mit nur ungefähr gleichen Schlüsselvariablen).
  Die Anwendung der SAS-Prozedur SURVEYSELECT.verlangt mehr als einfache Programmierkenntnisse und ist eher als eine Kunst anzusehen.

8 Zusammenfassung

Mit Hilfe des SAS-Systems können bei einiger Programmiererfahrung zahlreiche Varianten für Matching-Verfahren realisiert werden. Speziell die Verwendung von Propensity Scores können helfen, auftretende Verzerrungen zu vermindern sofern individuelle Charakteristiken dies erfordern.

Literatur

[1] Clayton, D. and Hills, M.: Statistical Models in Epidemiology. Oxford, Oxford University Press 1993

[2] D'Agostino, RB., Jr. (1998): Tutorial in Biostatistics: Propensity Score Methods for Bias Reduction in the Comparison of a Treatment to a Non-Randomized Control Group. Statistics in Medicine 17, 2265-2281

[3] Newman, S.C.: Biostatistical Methods in Epidemiology. New York, Wiley 2001

[4] Rosenbaum, P. and Ruben, D., "The Bias Due to Incomplete Matching", Biometrics, March 1985, 41, 103-116.

[5] Rubin, DB. (1997): Estimating Causal Effects from Large Data Sets Using Propensity Scores. Annals of Internal Medicine 127, 757-763