Paper SD-05

Matched Case-Control Techniques: A Data Manager's Case Study

Matthew P. Shevrin, MM,
Maria J. Silveira, MD, MA, MPH, Anamaria Segnini Kazanis MA, MA

Abstract

To accurately test outcome differences between retrospectively defined case and control samples, it is advisable to match the samples on characteristics that could otherwise bias the results. Successful application of this method, known as matched case-control, involves the following steps: Collection of descriptive information by treatment groups, means testing and analysis of key attributes, identification of outlying characteristics, and the random exclusion of controls using data-specific criteria. This paper will describe base SAS programming techniques to effectively and impartially normalize a control group to a case group for an epidemiological study.

Introduction

Using medical record data, we sought to retrospectively test the hypothesis that patients dispensed a particular class of drugs who died of certain diseases were more likely to have their drug treatment halted sooner than patients who died without having these diseases. A relatively small number of potential candidates met the study criteria. Therefore, data were first divided into the two distinct groups (case and control) based on the presence or absence of a defined group of diagnoses, and then the indirect attributes were tested for comparability. The data manager's task was to render the two groups as similar as possible in terms of the indirect attributes while maintaining the integrity of the common attribute, which was the particular drug therapy. It was also expected that the size of the case group be reduced as little as possible, thereby restraining the elimination process to the control group.

The indirect attributes were age, categorized in deciles, and co-morbidities, counted as ?, and benefits status served as a proxy for socio-economic status. Frequency distributions and means testing were run to exclude occurrences where outlying values of indirect attributes were identified. Then, frequency distributions describing the proportion of categories within each attribute were generated and measured. Repeated application of randomized exclusions of the control group using PROC SURVEYSELECT normalized the proportions of the indirect attributes with the case group.

The CONSORT (Consolidated Standards of Reporting Trials) diagram in Appendix A depicts the number of patients identified and filtered as a result of the entire process.

Once the cases and controls were defined, it was possible to run a survival analysis with the ability to assess the results of the discontinuation rates for drug therapy controlled for other differences between the groups.

Data Setup (Case-finding and Categorization)

Pharmacy, inpatient, outpatient, enrollment, and death records were extracted from the Veterans Affairs Information Resource Center (VIReC), which includes the National Patient Care Database (NPCD), the VA Decision Support System (DSS), and the Beneficiary Identification & Records

Locator System, or Death File (BIRLS). First, all patients who died between July 1, 2004 and June 30, 2005 were identified. Then, any of these patients with a supply of statin drugs within nine months of their date of death dispensed from a VA facility within one of the Veterans Integrated Service Networks were retained. Next, inpatient and outpatient records for this group were extracted and assessed for admissions and/or encounters with the diagnostic are of interest within six months of their date of death. The patient age at death and a tally of co-morbidities were categorized and used as indirect attributes. Age was grouped by decade and the co-morbidities were categorized using the ICD-9CM diagnosis codes from the Charlson Co-morbidity Index[1]. Enrollment priority status was used as a proxy for socio-economic status and grouped by whether the patient had no drug co-payments, had service connected drug co-payments only, or always had drug co-payments. Veterans' health coverage eligibility is based on a combination of their financial standing and whether they are being treated for injuries incurred while on active duty or were a prisoner of war. Among other things, their enrollment status defines the co-payment obligation for prescription drugs, if any.

The working data set was built at the patient level, using scrambled identifiers, with indicators established for the cohort grouping, and demographic and clinical attributes used for the case-control process.

Table 1. Sample of cohort grouping and demographic and clinical indicators

ID	Cohort	Age	Age Category	Co-morbidity Count	Enrollment Category
1	0	62	2	1	1
2	0	73	3	3	2
3	0	86	4	4	2
4	1	83	4	5	1
5	0	83	4	0	3

Methodology (Matched Case-Control)

Category matching of cases to controls was done stratifying independent variables that could confound the discontinuation rate[2]. To improve the precision with which the relative risk is estimated, the distributions were continuously updated to create an approximately balanced case-control ratio across the strata.

PROC MEANS, PROC UNIVARIATE and PROC FREQ were used to report the statistical ranges and value distributions of age and co-morbidity count, grouped by the case and control cohorts. See Appendix B for the output of statistical results.

```
/* Means testing of patient age at death and co-morbidity count */
/* By cohort      */

   PROC MEANS Data = dataset1 n min max range mean;
      Var AGE CHRLCT;
      Class STUDY_GROUP; /* Cohort */
      Output out = work.output1;
   RUN;
```

```
PROC UNIVARIATE Data = dataset1;
   Var AGE CHRLCT;
   By STUDY_GROUP;
RUN;
```

Patients who were younger than 51 years or older than 90 years at death were considered to be outside of the statistical norm, resulting in the exclusion of three case patients and 34 control patients.

```
PROC FREQ Data = dataset2;
   Tables AGE_CAT*STUDY_GROUP CHLRCT*STUDY_GROUP / chisq;
RUN;
```

Three additional cases and 180 additional controls were excluded because they had no co-morbidities or more than eight distinct conditions.

```
PROC FREQ Data = dataset1;
   Var CHRLCT;
   By STUDY_GROUP; /*Cohort*/
RUN;
```

Seven control group patients were excluded because they did not have a valid enrollment priority value.

```
PROC FREQ Data = dataset2;
   Tables PRIORITY_CAT*STUDY_GROUP / chisq;
RUN;
```

Once outlying values were excluded from both cohorts, PROC SURVEYSELECT was applied solely to sub-sections of the control cohort to randomly identify patients with indirect attributes to match the distribution proportions of the case cohort.

```
PROC SURVEYSELECT Data = work.dataset3
   Method = SRS
   Out = work.random3
   N = 145;
   Where STUDY_GROUP=0 and CHRLCT in (1,2) and AGE_CAT=4;
RUN;
```

A SQL query with a nested sub-query was run to exclude the patients identified above:

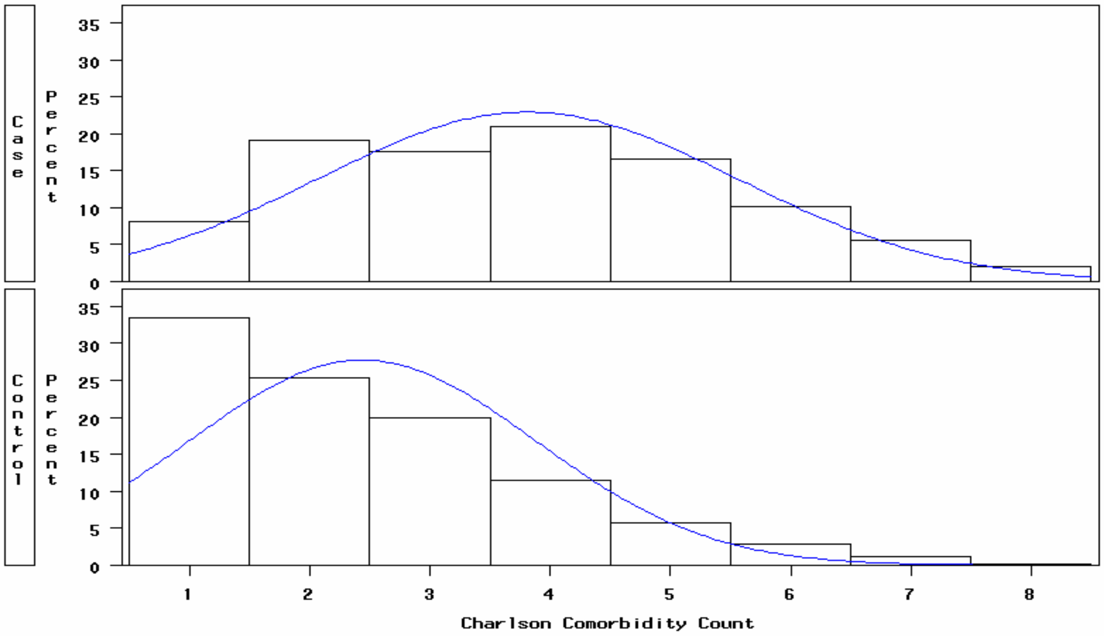
```
PROC SQL;
   Create Table work.dataset4 As
    Select * From work.dataset3
     Where ID Not In (Select ID From work.random3);
QUIT;
```

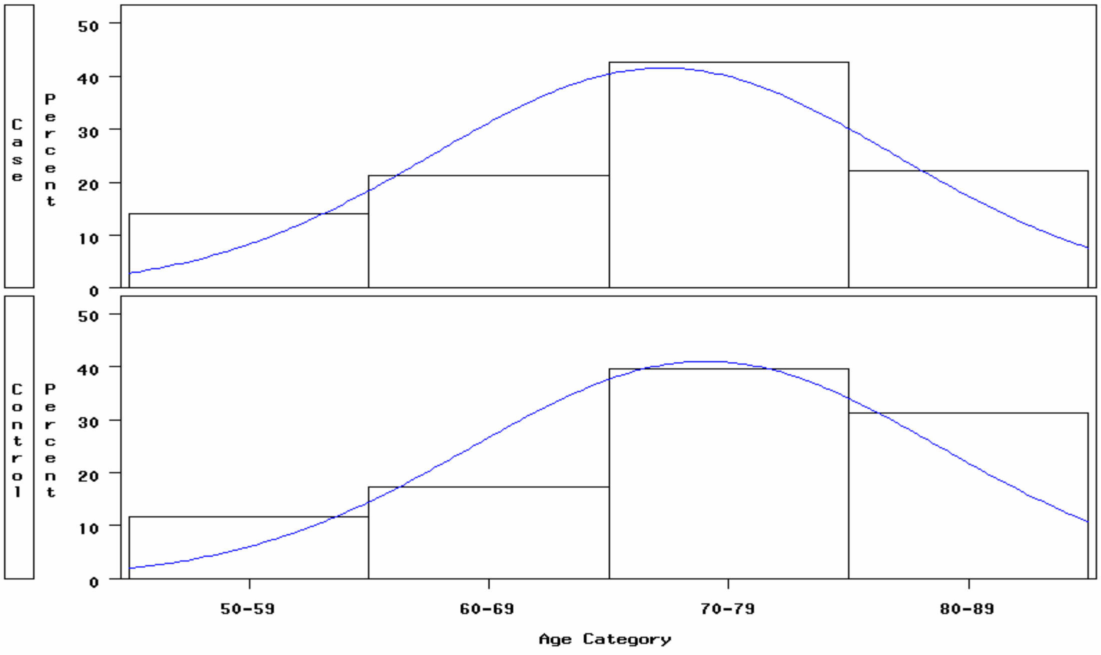
These two steps were repeated with combinations of co-morbidity count and/or age category until the two groups were similarly balanced across the strata.

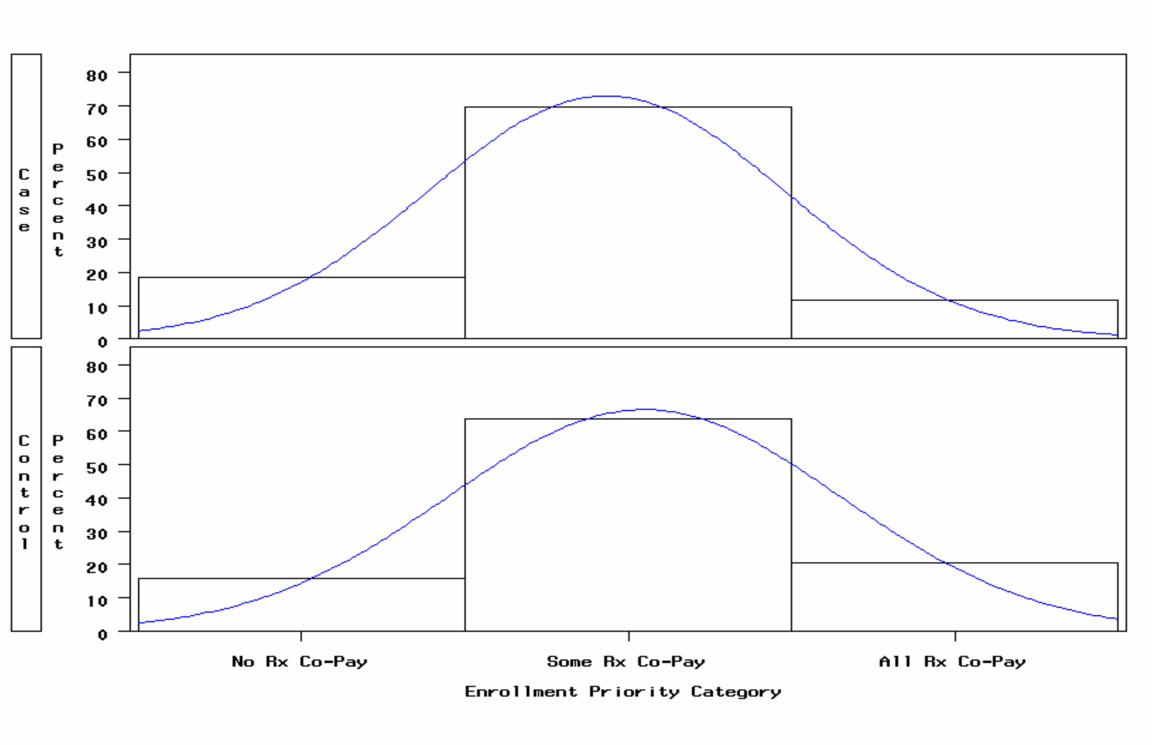
Results

Prior to the randomized exclusions, the frequency distributions of Charlson counts, age groupings and enrollment priority categories were graphically described using a histogram option within PROC UNIVARIATE.

Figures 1-3. Percent distributions of co-morbidity counts, age group and enrollment priority category by cohort, prior to randomized exclusions:



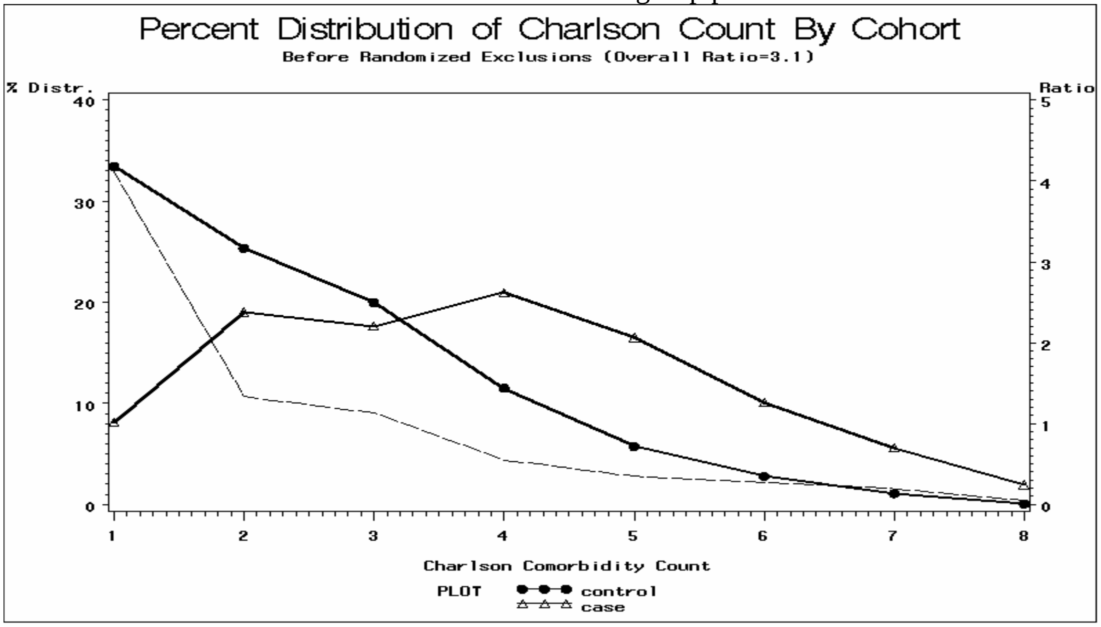


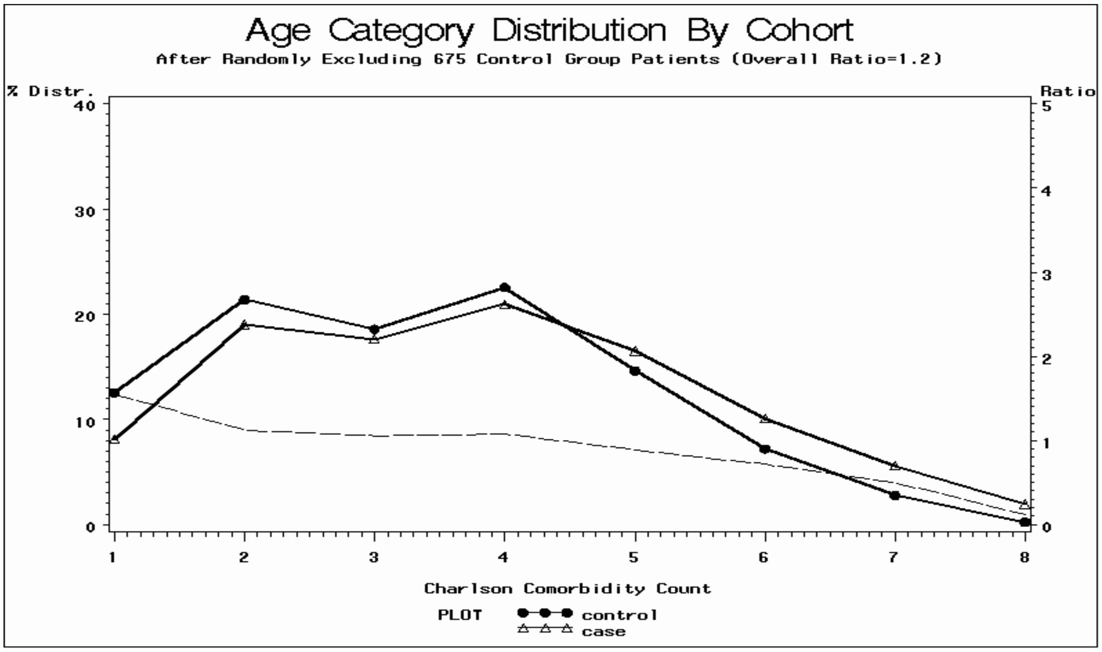


Based on the Charlson count and age group distributions, opportunities to exclude control group patients were identified among patients with low Charlson counts and those who were in the oldest age group. Since the distributions for all other values in addition to the targeted ones could change once exclusions were initiated, repeated random exclusions along with assessments of the full distributions and the ratios of proportion were assessed until the two groups were as similar as possible in regards to these two attributes. Enrollment priority category, the socio-economic status proxy, was observed throughout the process and was not affected by the randomized exclusions.

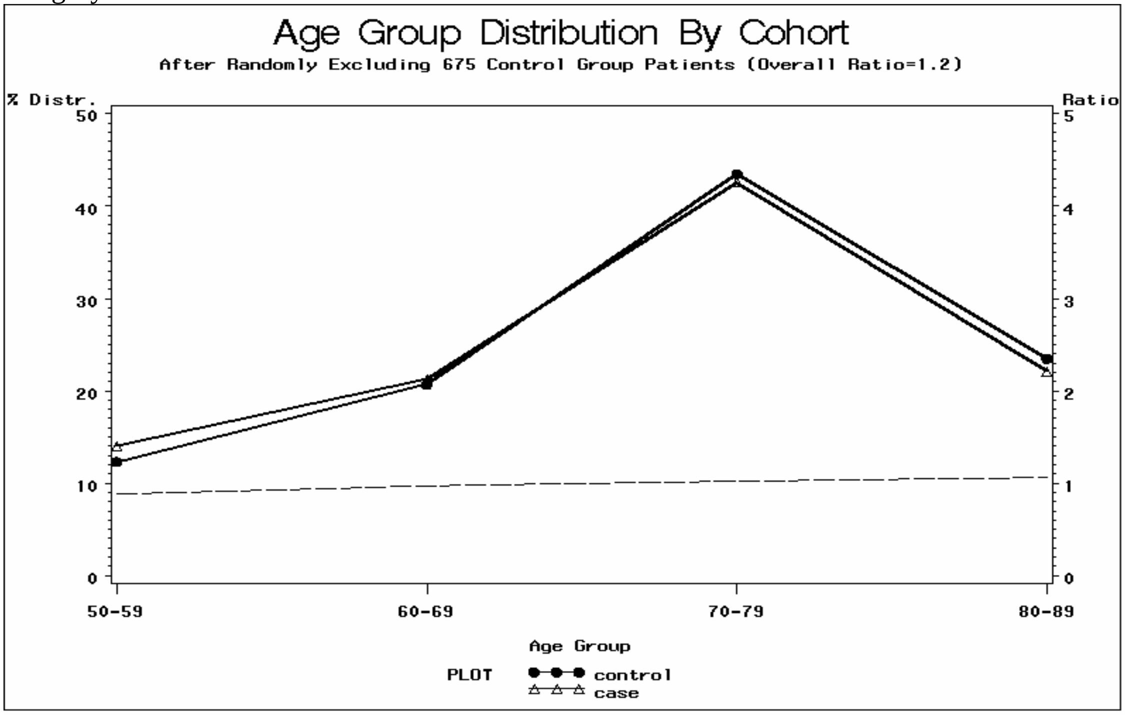
To enhance the analysis beyond observing the numeric and graphic contrast between the percentages of the two groups, the data manager included the calculation of the ratio of the proportion of the control to the case groups (ratio=control percent/case percent). The resulting ratios for each category were assessed for their proximity to 1 before and after the randomized exclusions were done, in which 1 represents a perfect match. PROC GPLOT was used to allow for the overlaying of the ratios onto the distributions of the two groups.

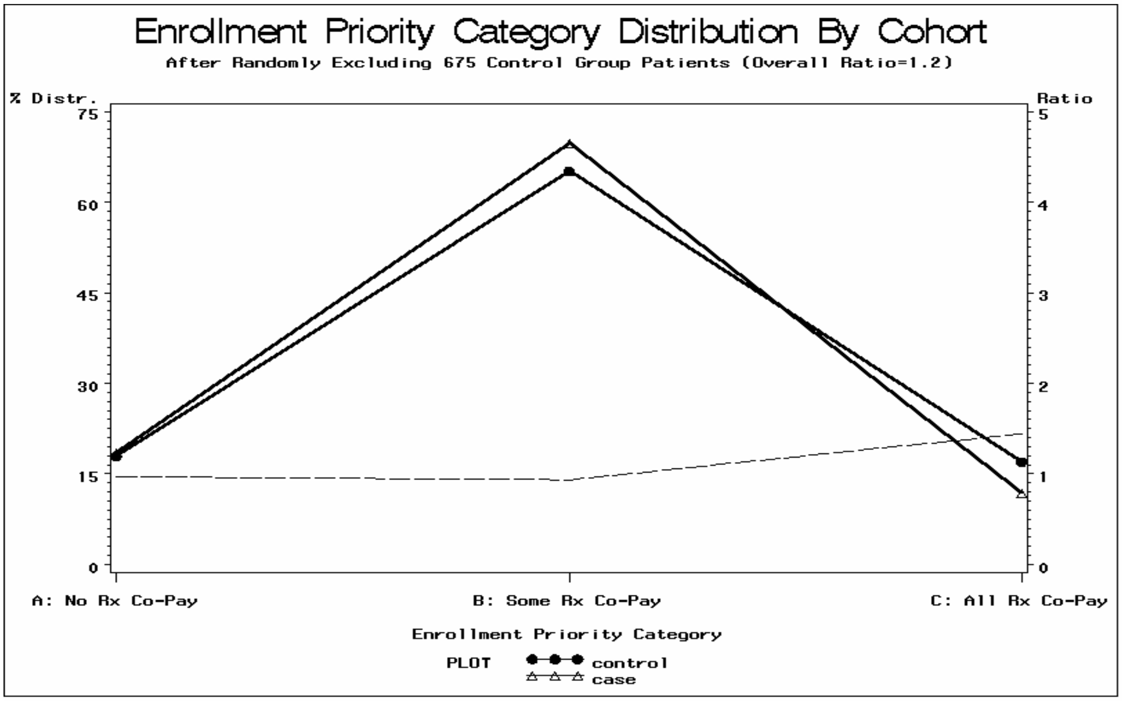
Figures 4-5. Percent distributions (solid lines) and ratios of proportion (dashed line) by cohort before and after randomized exclusion of 675 control group patients: Charlson count





Figures 6-7. Percent distributions (left legend) and ratios of proportion (right legend) by cohort after randomized exclusion of 675 control group patients: Age group and enrollment priority category





Discussion

The role of the data manager in support of exploratory data analysis for this project presented several challenges. Since the direction and progression of data extraction and translation was not fully prescribed from the start, programming adjustments and revisions occurred regularly. Descriptive reports and means testing were performed at various stages and a variety of scenarios in which data was categorized and summarized differently were pursued. Therefore, accuracy, promptness, organization, patience and clear communication were needed in addition to good technical skills.

The data manager was involved early in the discussion phase and throughout the analysis phase to point out the strength, and weaknesses, of the data itself and to recommend potential analytic solutions. For example, drug dispensed dates were not linked to clinical encounters nor were the supplies prescribed to patients always the same. The experience the data manager had regarding these variables alone had a critical impact on the validity of the overall results.

With respect to the case-matching process, the professional collaborative approach with the statistician facilitated a productive iterative relationship. One example was the data manager's suggestion to use enrollment priority as a socio-economic covariate. Lastly, the data manager created graphic descriptives in the form of bar graphs and plot charts to display the frequency distribution of key attributes in the data. In addition, he included a comparative measure, the ratio of proportions, between the two cohorts. As a result, the research investigator was assured that the ultimate goal of testing the hypothesis was not biased by differences across the study populations.

Conclusion

Using base SAS procedures, PROC MEANS, PROC FREQ and PROC SURVEYSELECT, we were able to accurately and impartially apply matched case-control methods in a retrospective study. This involved the identification and exclusion of outlying values, including the application of randomized exclusions.

Authors

Matthew Shevrin is a data manager with the Center for Practice Management of Outcomes Research, Veterans Affairs Ann Arbor Healthcare System.

Maria Silveira is a research investigator affiliated with the Center for Practice Management of Outcomes Research, Veterans Affairs Ann Arbor Healthcare System and the Division of General Medicine, University of Michigan.

Anamaria Kazanis is a data analyst with the Center for Practice Management of Outcomes Research, Veterans Affairs Ann Arbor Healthcare System.

References

[1]Quan H, Sundararajan V, Halfon P, et al. (2005), "Coding Algorithms for Defining Co-morbidities in ICD-9-CM and ICD-10 Administrative Data", Medical Care 43(11), pp. 1130-1139

[2]Schlesselman, James J., "Case-Control Studies. Design, Conduct, Analysis", Oxford University Press, New York, 1982, pp. 105-123

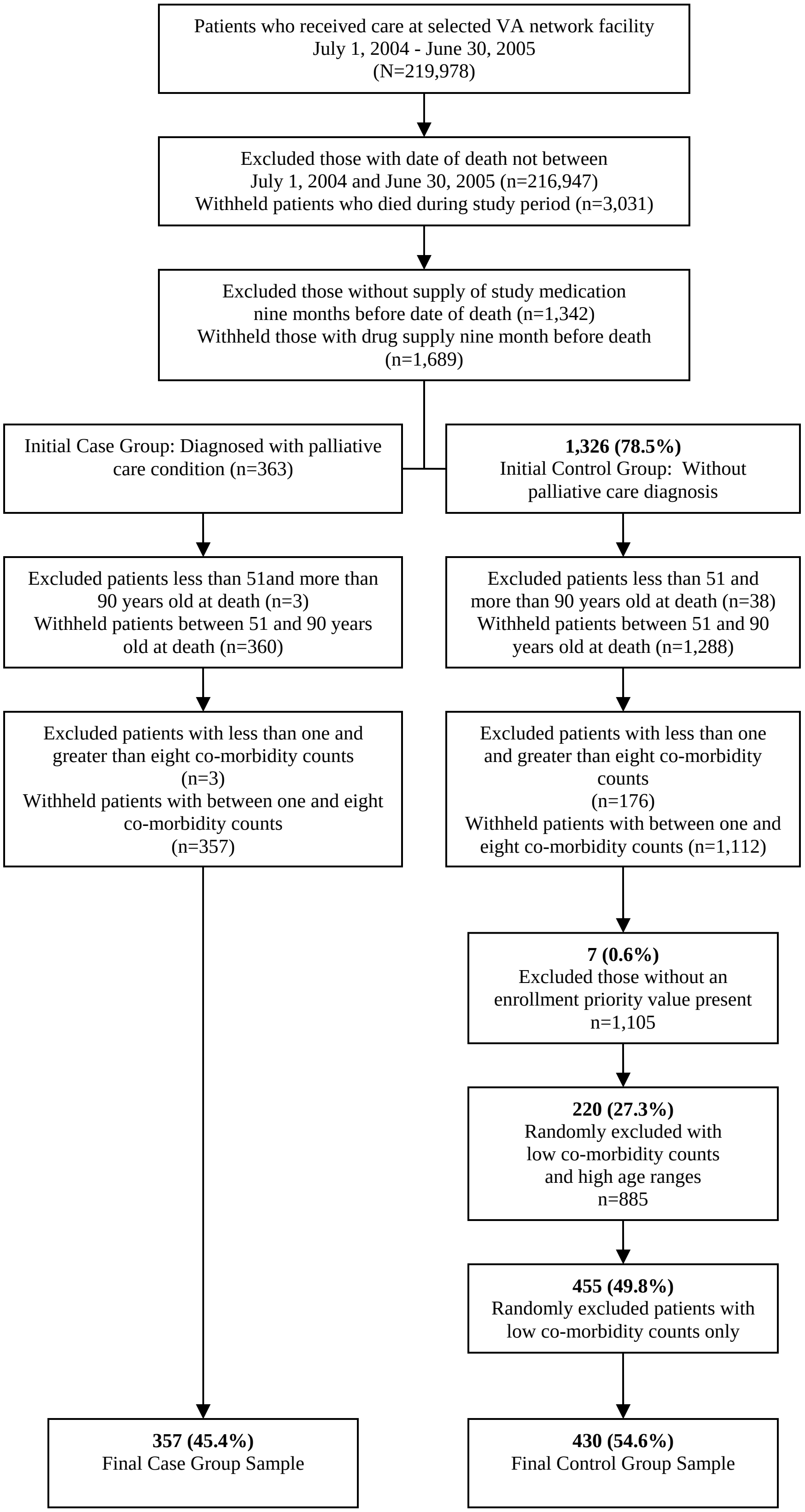
Trademarks

SAS ® version 9.1 (TS1M3) for Windows
SAS Institute Inc., Cary, NC, USA

Contact Information

Matthew Shevrin
HSR&D Center of Excellence
HSR&D/SMITREC
Department of Veterans Affairs
P.O. Box 130170
Ann Arbor, MI 48113-0170
(Email: matt.shevrin@med.va.gov)

Appendix A: CONSORT diagram



Patients who received care at selected VA network facility
July 1, 2004 - June 30, 2005
(N=219,978)

Excluded those with date of death not between
July 1, 2004 and June 30, 2005 (n=216,947)
Withheld patients who died during study period (n=3,031)

Excluded those without supply of study medication
nine months before date of death (n=1,342)
Withheld those with drug supply nine month before death
(n=1,689)

Initial Case Group: Diagnosed with palliative care condition (n=363)

1,326 (78.5%)
Initial Control Group: Without palliative care diagnosis

Excluded patients less than 51and more than 90 years old at death (n=3)
Withheld patients between 51 and 90 years old at death (n=360)

Excluded patients less than 51 and more than 90 years old at death (n=38)
Withheld patients between 51 and 90 years old at death (n=1,288)

Excluded patients with less than one and greater than eight co-morbidity counts
(n=3)
Withheld patients with between one and eight co-morbidity counts
(n=357)

Excluded patients with less than one and greater than eight co-morbidity counts
(n=176)
Withheld patients with between one and eight co-morbidity counts (n=1,112)

7 (0.6%)
Excluded those without an enrollment priority value present
n=1,105

220 (27.3%)
Randomly excluded with
low co-morbidity counts
and high age ranges
n=885

455 (49.8%)
Randomly excluded patients with low co-morbidity counts only

357 (45.4%)
Final Case Group Sample

430 (54.6%)
Final Control Group Sample

Appendix B: Statistical results

Before case-control matching

STUDY_GROUP	N Obs	Variable	N	Minimum	Maximum	Range	Mean
Control	**1326**	**age**	1326	42.01	96.41	54.40	74.57
		chrlct	1326	0	8	8	2.08
Case	**363**	**age**	363	41.61	93.07	51.46	72.30
		chrlct	363	1	10	9	3.85

After case-control matching

STUDY_GROUP	N Obs	Variable	N	Minimum	Maximum	Range	Mean
Control	**430**	**age**	430	51.16	89.16	38.00	73.67
		chrlct	430	1	8	7	3.39
Case	**357**	**age**	357	51.43	89.79	38.36	72.38
		chrlct	357	1	8	7	3.81

Appendix C: Programming Examples

Histograms

```
   ODS rtf file="W:\shevrin\MWSUG\hist.rtf";

   PROC UNIVARIATE Data=histogram_chrlct;
      Class study_desc;
      Histogram chrlct       / normal midpoints=1 to 8 by 1;
      Histogram age_cat      / normal midpoints=2 to 5 by 1;
      Histogram priority_cat / normal midpoints=1 to 3 by 1;
   RUN;

   ODS rtf close;
```

Graph Plot

```
/* Save percent output from PROC FREQ */

   PROC FREQ Data=gplot_chrlct;
      Tables chrlct / out=chrlct_freq1;
      By study_group;
   RUN;

   PROC SORT Data=chrlct_freq1; By chrlct; RUN;

/* TRANSPOSE to columns by Charlson count category */

   PROC TRANSPOSE Data=chrlct_freq1
      Out=chrlct_freq1a (rename=(col1=control col2=case));
      Var percent;
      By chrlct;
   RUN;
```

```
/* Create ratio of proportion measure */

   DATA Chrlct_freq1b;
      FORMAT c_ratio 4.2;
   Set Chrlct_freq1a;
      c_ratio=control/case;
      LABEL c_ratio='Ratio';
   RUN;

/* Set graphics environment */

   GOPTIONS reset=all border;

/* Create symbol definitions */

   Symbol1 i=j v=dot      l=1 c=black w=3;
   Symbol2 i=j v=triangle l=1 c=black w=3;
   Symbol3 i=j            l=4 c=black w=1;
   Symbol4 i=j v=dot          c=black w=3;

/* Create axis definitions */

   Axis1 order=(1 to 8 by 1);
   Axis2 label=('% Distr.') order=(0 to 40 by 10);
   Axis3 label=('Ratio') order=(0 to 5 by 1);

/* Create legend definition */

   Legend1 value=(tick=3 'Ratio') across=1;

/* Create title */

   Title1 'Percent Distribution of Charlson Count By Cohort';
   Title2 'Before Randomized Exclusions (Overall Ratio=3.1)';

/* Produce plot */

   ODS rtf file="W:\shevrin\MWSUG\plot.rtf";

   PROC GPLOT Data=chrlct_freq1b;
      Plot control*chrlct=1 case*chrlct=2
            / overlay haxis=axis1 vaxis=axis2 legend=legend1;
      Plot2 c_ratio*chrlct=3
            / vaxis=axis3 legend=legend1;
   RUN;
   QUIT;

   ODS rtf close;
```