**Tutorial on Design of a Matched Case Control Comparative Effectiveness Study**

Farrokh Alemi, Ph.D.,

Austin Brown

Friday, January 03, 2014

Corresponding author: Farrokh Alemi, Ph.D. 

This project was funded by appropriation #3620160 from VA Office of Geriatrics and Extended Care. The contents of this paper do not represent the views of the Department of Veterans Affairs or the United States Government.”

# Summary

This paper describes one method of conducting comparative effectiveness studies using matched case control. This method can be used to evaluate the impact of a program, such as Veteran Administration (VA) Medical Foster Home (MFH) program. Cases are selected from the program. Controls are selected from outside the program using the data available in the electronic health record. Controls are matched to cases in relevant characteristics. The impact of the program is examined by comparing cases to matched controls. This paperdescribes a nested, matched case control design using retrospective data. It defines enrollment, observation and follow-up time periods. It describes how cases and controls are matched. Finally it describes statistical procedures for verification of matching and evaluation of the statistical significance of the impact of the program.

# Background

In recent years there has been a growing interest in comparative effectiveness studies. This interest is partially due to the increased use of electronic health records which for the first time have made these techniques more accessible to a wider group of practitioners.

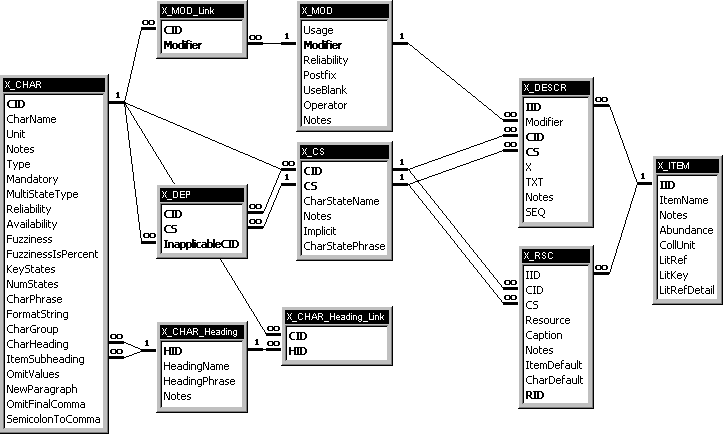
The gold standard of medical research is the randomized clinical trial – a rigorous approach that provides unbiased information about the impact of the intervention but (1) involves costly data collection, (2) restricts study to pre-defined eligible populations – typically those without comorbidities, and (3) denies access to some level of care for patients in the control group. By comparison, retrospective comparative effectiveness provides less clear conclusions but reflects the patient population being served with all of idiosyncratic comorbidities and characteristics. No data are collected for evaluation purposes and electronic medical records are used to assess the impact of the intervention. Although there are limitations, these techniques have yielded surprising and important insights in clinical care.

Many different techniques have been developed to conduct comparative effectiveness studies [[[1]](#endnote-1)]; none are without their critics [[[2]](#endnote-2)]. The chief complaint is often that different comparative effectiveness approaches can lead to contradictory findings [[[3]](#endnote-3)]. Contradictions can arise because findings are based on nonrandom data and observations drawn from a wide variety of disparate sources including databases for insurance claims, prescription histories, national registries, as well as patient treatment records. This illustrates both the problem and its solution - lacking true random sampling, studies must be carefully designed to ensure that data are representative of the larger population for the characteristics being assessed; moreover, it must be possible to measure outcomes with variables available in the database [[[4]](#endnote-4)]. In order to standardize the conduct of comparative effectiveness studies, this paper describes the procedures for a retrospective matched case control comparative effectiveness study.

# Source of data

Data for retrospective comparative effectiveness studies is usually obtained from electronic health records. These data may include prescriptions, diagnoses, records from hospitalizations and outpatient care, clinician’s notes and dates of encounters. Data are usually obtained for well-defined number of recent years that exceed both the planned observation prior to enrollment in the program and the follow-up years after enrollment in the program.

**Figure 1: Example of a Relational Database**[[5]](#endnote-5)



Statisticians are used to matrix data structures with cases in rows of a table and variables in columns. These types of data structures have sparse entries since many variables are not relevant to every case. In contrast, data in electronic health records are distributed in numerous smaller but dense tables. For example, all information about patient characteristics (e.g. date of birth, date of death) is available in one table (see left side of Table 1); information about encounters is available in other tables (see right side of Table 1), and still another table provides information about laboratory findings. In a modern electronic health record millions of data elements can be distributed in thousands of individual tables. The analysis of data starts with making yourself familiar with the structure of the data. The first challenge in performing a comparative effectiveness study is to aggregate the data in a format that can be used for statistical analysis [[[6]](#endnote-6)].

**Table 1: Patient Data & Visit Data Are in Two Different Tables**



Standard Query Language, SQL, is used to prepare the data for analysis. In an SQL, the investigator specifies the address of a table where the data can be found. If data are in multiple tables, as often they are, the investigator uses a Join command to include data from multiple tables. Tables are joined using the primary key of the table. A primary key is selected so that the table is a set of information about its primary key. So, a table on diagnoses codes has the code as primary key and has the description of the code as information about the primary key. A table on patient has the patients’ medical record number as the primary key and patients name and birthday as other variables (see left side of Table 1). A table on visits (see right side of Table 1) has encounter ID, and diagnoses ID but not the description of the diagnosis, a patient ID but not the patient characteristics, a provider ID but no other information about the provider. A “Join” command in SQL would allow the investigator to connect the visit table to the patient table and thus be able to read the date of birth of the patient. It would allow one to join the visit table to the diagnoses table and thus be able to read the description of the patient’s diagnoses. Knowledge of SQL is necessary for preparing data electronic health records for statistical analysis.

Besides join, SQL allows a handful of other commands including procedures to filter, count or average the data. SQL allows very few commands. One can learn these commands quickly. Repeated uses of these commands allow preparation of complex data in formats suitable for statistical analysis. Detailed instructions on use of SQL can be found at different locations on the web, including at <http://openonlinecourses.com/databases>. Perhaps more interesting, almost all common errors and methods of combining data can be found on Google and there are many sites where experienced SQL programmers will help novices solve data transformation problems.

# Study Design and Methods

In observational studies, such as studies of data in electronic health record, there is no random assignment of patients to groups. The observed outcomes may be due to patients’ conditions and not related to treatment. A matched case control study provides a comparison group for patients who have received the treatment and thus reduces the possibility of erroneous attributions.

The approach taken in case control studies has a long history. One of the earliest examples comes from the famous 1854 cholera epidemic in London in which it was demonstrated that most of those who died drew water from the same Broad Street pump [[[7]](#endnote-7)]. The approach was used in several studies in the 1920s but truly came to prominence in the 1950s with studies that demonstrated the unexpected relationship between smoking and cancer. [[[8]](#endnote-8)] These days, the use of matched case control studies in analysis of data from electronic health records is common.

## Definition of Cases and Controls

Patients who have received the intervention are referred to as “cases.” Patients who did not receive the intervention are referred to as “controls.” For example, patients who were admitted to the Medical Foster Home program (an alternative to nursing home care) may be considered cases and patients in the traditional nursing home program may be considered controls. The medical foster home allows patients to rent their own room in a community home while receiving medical and social services from the Veteran Administration in this community setting.

The identification of cases in a medical record is problematic as these databases report utilization of services and not necessarily participation in a program or need for care. There are at least two methods of identifying a case. First, a case could be identified by examining the medical record for a unique clinical event of interest. A clinical event could be a physician office visit, an inpatient admission, or an emergency room visit. For a study of heart failure, for example, a clinical event could be an initial congestive heart failure. Typically these events are defined using codified nomenclatures such as the International Classification of Diseases (ICD-9/10). The Healthcare Cost & Utilization Project of the Agency for Healthcare Research and Quality has defined how various diagnoses codes correspond to common disease categories [[[9]](#endnote-9)]. For example, heart failure can have one of the following ICD-9 codes: 402.01, 402.11, 402.91. 425.1, 425.4, 425.5, 425.7, 425.8, 425.9, 428.0, 428.1, 428.2, 428.21, 428.22, 428.23, 428.3, 428.31, 428.32, 428.33, 428.4, 428.41, 428.42, 428.43, or 428.9. Other examples include falls [[[10]](#endnote-10)], injuries [[[11]](#endnote-11)], medication errors [[[12]](#endnote-12)], mood and anxiety problems [[[13]](#endnote-13)], and hospitalization encounters.

Second, a case could be identified by examining admission to a program. For example, in the Medical Foster Home project, the providers gave the evaluators list of patients they had cared for. Patients’ social security numbers were used to identify them within the electronic health record. These patients were compared to patients in nursing homes, as medical foster home is an alternative to nursing home care. Nursing home patients were identified through admission and discharge dates for the nursing home, information available in the medical record of the patients.

## Measurement of Exposure to Treatment

In defining cases and controls, attention should be paid on the extent of exposure to the intervention. A sufficient exposure should be allowed so that the change in outcomes can be expected. For example, the day after enrollment in a Medical Foster Home (MFH) care one cannot expect any changes in patient outcomes. A person enrolled for one day is not considered to receive the full benefit of enrollment. Sometimes, patients enroll and dis-enroll shortly afterwards. We assume that 3 months of enrollment is necessary before the patient is considered to be a Medical Foster Home patient. A similar timeframe is used for controls in nursing homes. This excludes short stays – those that reside in nursing homes for less than three months.

Some patients receive both the intervention and the control programs. For example, a patient may enroll for MFH at first but after months of enrollment leave it for care in a nursing home. Patient’s enrollment in the case or control group is for a specific time period. Since the same patient has spent time in both groups they may appear to be ideal match for themselves. The case and control match many features, with one exception - the case and control are examined in different timeframes. Unfortunately, transition from one intervention to another is almost always accompanied with a major crisis that affects patient’s health. In these situations, the same patient before and after has a different health status. For example, in Figure 2, we see information on blood pressure of one patient. For 7 years this patient was in a nursing home. At end of the 7th year there was a hospitalization, shown as a circle. Following this hospitalization the patient was discharged to the Medical Foster Home. The blood pressure values during year 8 indicate the patient’s condition in the Medical Foster Home program. The values immediately prior year 8 indicate blood pressure when the patient was in the nursing home. The patient’s condition has worsened right before the transfer from nursing home to medical foster home program. If we have an accurate measure of how much the patient’s condition has worsened, we can use this information to compare blood pressure before and after transfer. Without it, we cannot compare the same patient at two different times.

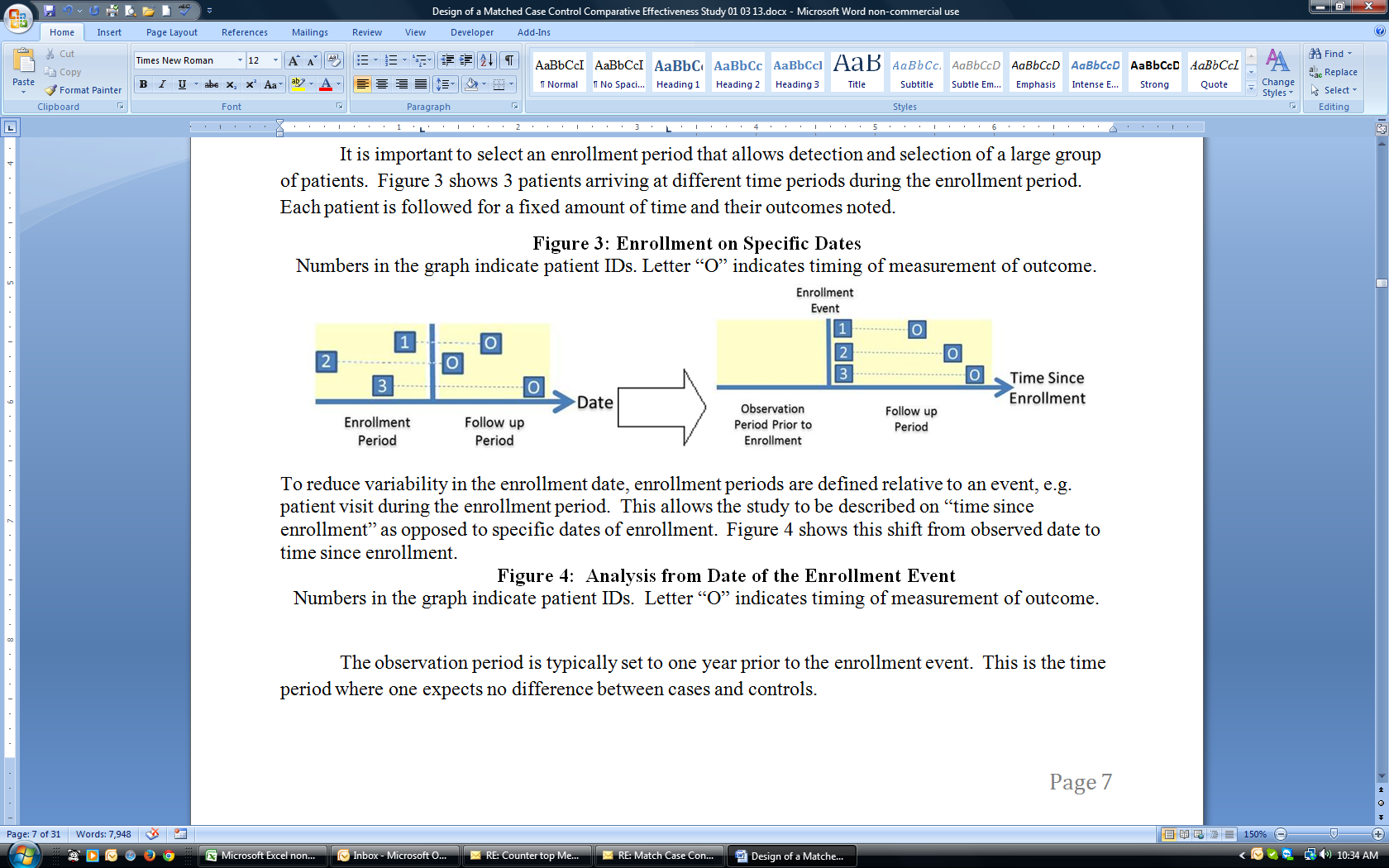
**Figure 2: Patient Transitions among Care Venues**



## Enrollment and Observation Period

It is important to select an enrollment period that allows detection and selection of a large group of patients. On the left side of Figure 3, patients arrive at different time periods during the enrollment period. Each patient is followed for an amount of time and their outcomes noted. To reduce variability in the enrollment date, enrollment periods are defined relative to an event, e.g. patient’s first visit during the enrollment period. This allows the study to be described on “time since enrollment” as opposed to specific dates of enrollment. The left side of Figure 3 shows a graph of data based on date of visits. The right side of Figure 3 shows the same data based on time since first visit. On the right side, we can see clearly that patients are followed for different intervals until the outcome of interest occurs.

**Figure 3: Enrollment on Specific Dates**Numbers in the graph indicate patient IDs. Letter “O” indicates timing of measurement of outcome.



The observation period is typically set to one year prior to the enrollment event. This is the time period where one expects no difference between cases and controls. In fact, by design controls are selected so that there would not be any major difference between cases and controls prior to enrollment. Then one can attribute the differences in outcomes to enrollment and not some pre-enrollment differences.

As an example, consider the data in Table 2. Two columns of data are given. Dates of various events are given in one column and the second column calculates time since enrollment in the program. Suppose enrollment is any diabetic visit after January 1; the follow up period is also scheduled for one year after enrollment. Table 2 lists time to ED visits after the patient was enrolled. The patient depicted in Table 2 visits the physician on January 9. This is the first diabetic visit during the enrollment period so this becomes the enrollment event. Over the next thirteen months the patient has several encounters with their physician as well as two ED visits. On both occasions the ED visits were the results of falls. The time since enrollment in the study is the number of days between these fall events and the enrollment date. As Table 2 shows, one of these fall events was within one year of enrollment and therefore in the follow up period. In analyzing data from electronic health records, it is important to clearly define the enrollment event and follow up and observation periods as these time intervals change the data.

**Table 2: A Sample Patient and His Encounters**



## Matching Criteria

Controls differ from cases in distinct ways. For example, controls may be older or have different comorbidities. These differences affect intervention outcomes. To make sure that these variations do not get attributed to the intervention, controls should be matched to cases. The variables used to make these matches are different from study to study. The following is some of the more typical variables used in matching cases to controls:

1. Same observation period. Controls are often selected so that they had the same time observation period prior to enrollment. When this is not done, the differences in case and controls could be due to changes that have occurred over time in the environment, or provider’s learning.
2. Same age on admission. Controls are often randomly selected so that their age matches the case.
3. Same gender on admission. The gender of the control and the case is often set to match.
4. Same comorbidities. Controls are randomly selected so that they match the comorbidities of the cases.

Another way of controlling for differences prior to enrollment is to select controls so that they match cases in outcomes observed in the period prior to enrollment. In this fashion, case and controls have the same history. The variables typically used in this fashion include history of hospitalization, medical errors, falls and use of mental health services.

For continuous variables, a control can be said to match a case, if controls are within one standard deviation of the case. For discrete variables, the two match if the control and the case have the same value. For each case, typically five matching controls are selected. If more than five controls are available, the selection among the matched controls is done randomly. Cases that fail to match to at least five controls are not included in the analysis.

To illustrate an example of how matching occurs, examine the data in Table 3. This table shows data for 2 cases (identified as cases where the variable treated is “Yes) and 7 controls (where the variable treated is “No”). These patients are to be matched based on age, where for simplicity is assumed to have two categories: “Young” and “Old”. The outcomes are shown as either “Positive” or “Negative.” Table 3 also shows a random number that can be used for selecting among several controls that match the case.

**Table 3: A Procedure for Matching Controls to Cases**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Patient ID** | **Received Intervention** | **Age** | **Months to Fall** | | **Random Number** | **Case or control?** |
| **Observation Period** | **Follow-up Period** |
| 1 | Yes | 65 |  | 3 | 0.24 | Case |
| 2 | No | 60 |  | 2 | 0.85 |  |
| 3 | Yes | 84 | 2 |  | 0.64 | Case |
| 4 | No | 82 | 4 |  | 0.7 | Control |
| 5 | No | 78 |  |  | 0.87 |  |
| 6 | No | 80 | 3 |  | 0.72 | Control |
| 7 | No | 79 |  |  | 0.86 |  |
| 8 | No | 64 |  |  | 0.16 | Control |
| 9 | No | 70 |  | 2 | 0.17 | Control |
|  | | | | | | |
| **Standard Deviation** | | 8.89 |  | | | |

Suppose we need to randomly select two controls for each case. For case 1, we select the two controls with the lowest random number, these are cases 8 and 9. For case 2, we also select the two controls with the lowest random number; these are cases 4 and 6. Let’s look at the calculations for this example step-by-step.

**Step 1**: Calculate the standard deviation for age across all nine patients.

The average age in Table 2 is 73.6 years. The standard deviation, S, can be calculated using the formula:

Where is the ith observation from n observations. is the average of the observations.

**Step 2**: For each patient that received treatment calculate a one standard deviation interval based on their age (see Table 4).

**Table 4: Sample Calculations of Matching Controls to Age of Cases**

|  |  |  |  |
| --- | --- | --- | --- |
| Case Patients | Control Case Age | Age – One Std. Deviation | Age + One Std. Deviation |
| Patient 1 | 65 | 65 – 8.89 = 56.1 | 65 + 8.89 = 73.9 |
| Patient 3 | 84 | 84 – 8.89 = 75.1 | 84 + 8.89 = 92.9 |

**Step 3**: Select as control, patients who fall within one standard deviation of the age of the case.

* Patient 1 is the first Case Patient with an age of 65. Patients 2, 8, and 9 all fall within one standard deviation of this age.
* Patient 3 is the second Case patient with an age of 84. Patients 4, 5, 6 and 7 fall within one standard deviation of this age.

These selections provide us with assignments of controls to cases. See column listed in the right hand side of Table 3. Note that in this example we have matched cases and controls on age. Other variables could also be used for matching and the decision on what variable cases and controls should be matched is a crucial decision for analysis. One common method to reduce uncertainty about the appropriate variable to match is to rely on the outcomes observed during the observation period, i.e. the period from enrollment till start of follow-up period. In this fashion, cases are matched to the control on history of the outcomes that is being monitored.

## Measurement of Outcomes

Outcomes of interest are typically defined over a range of values over time. Patients may be in and out of the therapeutic range on different days. It is important to calculate percent of time patients are within therapeutic range. This is done by a procedure first proposed by Rosendaal [[[14]](#endnote-14)]. Figure 5 shows an example of a patient going in and out of range for blood pressure. In Figure 5, days when systolic blood pressures are within the range of 120 mmHG to 140 mmHG has been indicated.

**Figure 5: Examples of Blood Pressures in and out of Therapeutic Range**

For two consecutive values that are in range, the calculation is straightforward. For two consecutive values, where one is in range and another is out of range a linear extrapolation is made to determine the day the patient moved to out of range. A preset maximum (e.g. 60 days) is used to reduce the influence of the linear extrapolation for two values that are very far apart. The percent of days the patient is in therapeutic range is calculated as the sum of all estimated days in range divided by the number of days from first to last measure. For two values at two consecutive measurements, the extrapolation is based on the following formula:

“Unknown days” refer to the difference of “days in range” and the maximum number of days that can be estimated from two measures. These are days that are neither in nor out of range. The percent of days in range is then calculated using the following approximation:

For example, in Table 5 we see six blood pressures, some in and some out of range. In the first two measurements, the patient goes from blood pressure of 125 to out of range pressure of 150. These two measures are 20 days apart; this patient is estimated to be in range for 12 out of these 20 days. Similar, estimates are made for every two consecutive blood pressure. For thesecond and third measurement the estimate of days in range is 20 days. If we assume that days more than 2 weeks away from a measurement are unknown days, then 6 out of these 20 days are unknown days. In the last two measurements, both measures are within range, there are 30 days in between these measures, 16 of which are unknown days. The total number of days the patient is in range is 77 days, the total number of unknown days is 22 and the percent of days in therapeutic range is 81%.

**Table 5: Calculation of Percent Days in Therapeutic Range**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Days Since Enrollment** | **BP** | **Therapeutic Range** | | **In range** | **Days in Range** | **Unknown Days > Max of 14 Days** |
|
| **Lower** | **Upper** |
| 0 | 125 | 120 | 140 | Yes | 12 | 0 |
| 20 | 150 | 120 | 140 | No | 20 | 6 |
| 60 | 130 | 120 | 140 | Yes | 5 | 0 |
| 70 | 110 | 120 | 140 | No | 10 | 0 |
| 90 | 130 | 120 | 140 | Yes | 30 | 16 |
| 120 | 135 | 120 | 140 | Yes |  |  |
|  |  | **Total Days in Range** | | | 77 | 22 |
|  | **Percent Days in Therapeutic Range** | | | | 81% |  |

At least two observations over time are needed before the “Percent Days in Therapeutic Range” can be calculated. A detailed example of these calculations is provided in Appendix A.

## Verification of Matching

To verify that the matching has been successful, the cases and controls are compared before the cases were enrolled in the intervention. The null hypothesis is that the cases and the matched controls have the same rate for outcomes during the pre-enrollment period. The alternative hypothesis is that there is a statistically significant difference between these two groups. First the difference between each case and the mean of the controls that matched to the case is calculated. Then the average and the standard deviation of the differences are calculated. The t-statistic is used to test if the mean of the differences is significantly different from zero. If so, the null hypothesis is rejected and the matching process has not led to equivalent groups. If equivalent groups are not available, the procedure is repeated until two statistically equivalent groups are identified.

In many studies, the calculation of the statistics is done in error [[[15]](#endnote-15)]. A paired t-tests is done when cases have 1:1 match to controls and have a symmetrical distribution, the Wilcoxon signed ranks test is used for measured non-parametric outcomes with 1:1 matching, linear or logistic regression is used to handle matched sets other than 1:1 matching [[[16]](#endnote-16)]. The section below shows how to test the statistical significant of outcomes when many controls have been matched to one case. Applying the same procedures to pre-intervention period is one way of testing that the procedure for matching controls to cases has been effective.

## Analysis of Outcomes

Outcomes of patients in case and controls can be analyzed using the odds ratio of observing these outcomes. First, the number of outcomes in cases and controls during the follow-up period are counted. Table 6 shows the presence of these outcomes among case and controls with positive or negative symbols.

**Table 6: Contingency Table for Adverse Outcomes**

|  |  |  |
| --- | --- | --- |
| **Outcome =Fall** | **Cases** | **Controls** |
| **Yes** | = 1 | = 1 |
| **No** | = 1 | = 3 |
| **Total** | = 2 | = 4 |

The probability of the outcomes in the cases is calculated as:

We calculate the relative risks associated with the cases using the odds ratio associated with the outcomes. In particular, given Table 5, the test statistic, L, is calculated as:

If there are a large number of cases in the study, the distribution of the log odds ratio is approximately normal. In a Normal distribution, 95% of the data would fall within 1.96 standard deviation of the mean. Therefore, the approximate 95% confidence interval for the population log odds ratio was estimated as [[[17]](#endnote-17)]:

If this confidence interval does not include 1, then the hypothesis that cases and controls have the same outcomes has less than 0.05 chance of occurrence. Since the probability is low, therefore one could reject the hypothesis with relative confidence.

The above procedure shows how the statistical significance of difference between cases and controls are examined. Table 7 applies these concepts to our data. Obviously, we have very few cases and controls and one would not expect the findings to be significant but we apply the proposed method to these data to illustrate the use of the formulas. We can examine the outcomes during the follow-up period among cases (i.e. 1 and 3) and controls (i.e. 2, 4, 8, and 9). From Table 5, the test statistic L is calculated as 3:

The 95% upper and lower control limits for these data are set at 6.58 and a negative number, which is displayed as 0 as no negative numbers are possible.

The confidence interval for the ratio includes 1. Therefore, the hypothesis of significant difference between cases and controls cannot be rejected.

**Table 7: Calculation of Odds Ratio for Example Cases and Controls**

|  |  |  |
| --- | --- | --- |
| **Outcome** | **Cases** | **Controls** |
| **Yes** | 1 | 1 |
| **No** | 1 | 3 |
| **Total** | 2 | 4 |
|  |  |  |
| **L=** | 3 |  |
| **Variance=** | 3.33 |  |
| **Upper Limit=** | 6.58 |  |
| **Lower Limit=** | 0.00 |  |

Many policymakers also prefer to see a visual display. Typically one plots the data using the probability of the outcome before and after enrollment. Before enrollment in the program, for cases, the rate is 50% positive and for controls the rate is also 50% positive. This is the overlapping point in Figure 6 and shows that before enrollment in the program cases and controls had the same rate of falls. In the follow-up period, the percent of patients with fall changes. Both groups have a decline in rate of falls but the decline is more in the cases than controls.

**Figure 6: Impact of Enrollment in Program on Patient Falls**

The problem with Figure 6 is that it does not show the period of time that lapses before a fall occurs. Since adverse outcomes eventually happen (we all die), it is important to trace the time to the event.

## Analysis of Time to Event

Sometime the outcome of interest is not a count of events (e.g. falls) but days or in our case months to the event. The analysis of outcomes over time is done using Kaplan-Meier estimator [[[18]](#endnote-18)]. It is used to report the time until adverse outcomes occur. An advantage of using this estimator is that it can accommodate changes in patients available to the study. For example, if patients die before the outcome of interest is measured, they can be left out of the calculation for periods that they were not available. For another example, if patients change their place of care, then they can be excluded from the analysis from the time when they changed provider. At each time period one can recalculate the number of patients at risk and therefore patients who drop out of the study do not affect the calculations.

At each time period, the number of patients at risk of falling (number of patients in study minus the patients that are censored) and the number of patients that do not fall (number at risk minus number that have fallen) is calculated. This is the conditional probability of falling given that the patient has not fallen in previous time periods. The probability of falling in a time period is calculated as the probability of falling in the previous time period times the conditional probability of not falling. Table 8 shows the calculation of these probabilities for our data. Among the 2 cases, there was one fall in the 3rd month. Among the 4 controls there was 1 fall in the 2nd month. There was one patient in the control group that died in the 1st month and therefore was not available to the study for the following months.

**Table 8: Probability of Falling at Different Time Periods**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Period** | **Number at risk** | **Number censored** | **Number Fell** | **Number No Fall** | **No Fall Probability** |
| **Cases** | 1 | 2 | 0 | 0 | 2 | 1 |
| 2 | 2 | 0 | 0 | 2 | 1 |
| 3 | 2 | 0 | 1 | 1 | 0.5 |
| **Controls** | 1 | 4 | 1 | 0 | 4 | 1 |
| 2 | 3 | 0 | 1 | 2 | 0.67 |
| 3 | 2 | 0 | 0 | 2 | 0.67 |

Figure 7 shows the resulting percent of patients having no falls at different time intervals. Cases had a fall sooner than the controls. The statistical significance of differences in rates of falls over time can be established by using the variance for Kaplan Meier statistic [[[19]](#endnote-19)].

**Figure 7: Impact of Enrollment on Months to Falls**

## Measurement of Severity

In comparative effectiveness analysis it is important to statistically control for differences among severity of illness of cases and controls. Without adequate control for severity of illness, one may mistakenly attribute differences in cases and controls during the follow up period to the intervention as opposed to initial differences in severity of illness. The matched case control tries to control for possible differences in the two groups prior to the intervention. It is reasonable to include severity as one of the variables in which cases and controls are matched on. Using the history of patient outcomes is another way of also measuring severity as these two variables are often related. No matter what is used for matching, it is important to verify that the cases and controls did not differ at start in their severity of illness.

There are many approaches to measuring severity of illness from patient information available in medical records. One approach uses patients laboratory findings and another approach uses patients diagnoses as coded into International Classification of Diseases [[[20]](#endnote-20),[[21]](#endnote-21)]. Since data warehouses have a large number of cases, it is often possible to statistically derive the relationship between the patient features and outcomes from the data. In these circumstances, the features and the statistical model can be used as a measure of severity of illness.

## Propensity Scoring

We have shown how matched case control studies are organized. An alternative to matching is to use all data but weight the control patients to match the proportion of the cases on specific variables. Propensity scoring allows one to use all of the data while matching discards patients that do not match to cases.

# Discussion

This paper has described how matched case controls can be organized and applied to data within electronic health records. These types of studies control for a number of possible ways in which observational studies may lead to erroneous conclusions. Like a quasi-experimental study, the matched case control protects against variables that affect both cases and controls, and the effect of time and enrollment period. There are a number of ways in which these approaches could lead to erroneous conclusions. These include:

1. Poor selection of variables that cases and controls are matched on
2. Inadequate measurement of severity of illness of patients
3. Incomplete medical records, when patients visits outside the healthcare system is not reflected in the record
4. Inadequate measurement of outcomes over time.

# Problem Sets

1. Matching Controls to Cases

Hypertension is a pervasive medical condition in the U.S. that can often be managed through diet. A study was performed to examine the impact of diet in controlling hypertension by reducing sodium intake. A control group was given a diet rich in fruits and vegetables, while a randomly selected study group was given a diet that also included low-fat dairy products. Both groups initially participated in a 3-week period in which they consumed a ‘typical’ American diet. During this period the baseline data for sodium intake were obtained (see Table 9).

Table 9: Hypertension and Salt Intake



* 1. Participants 7, 8 and 12 were randomly selected to participate in the study group. Create a data table from which to perform the matching analysis. The table should include the Participant Number, Study or Control Group Flag, Sodium Level, and Random Number Assignment.

* 1. Matching on Daily Sodium Intake, which three participants would be the best controls for Participant 7?
  2. Which is the single best control for Participant 12?

1. Verification of Matching

For the cases and controls selected in Problem 1.1, verify that the matching was successful.

* 1. State the hypothesis to be tested.
  2. Calculate the mean and standard deviation for the differences of sodium intake between the case and control participants.
  3. Calculate and interpret the t-statistic. Is there a statistical difference in sodium intake between participants in the case and control groups?

1. Days in Range

Venous thromboembolism (VTE) is a medical condition in which blood clots form within a vein that can block the vein or break off and become a life-threatening pulmonary embolism. One treatment for VTE is the use of anticoagulants such as warfarin. Anticoagulants are effective, but their level in the patient’s blood must be actively monitored. If it is too high, the patient is at risk for internal bleeding; too low and anticoagulant won’t have a therapeutic effect. Dosing for anticoagulants is monitored using INR, a measure that represents the ease with which blood clots. The target INR for effective treatment of VTE is between 2.0 and 3.0. Clinical protocols suggest that patients with a transient risk for VTE be put on warfarin for a three-month period. During this time the INR is monitored weekly. Table 10 represents the experience of a typical patient.

**Table 10: INR Values over a 90 Day Treatment Regimen**



* 1. Plot the observation data. Calculate the days on which the patient crossed control limits.
  2. Calculate the number of days the patient was in therapeutic range.
  3. Calculate the unknown days, and the percentage of days in range assuming a 3-day therapeutic horizon.

1. Odds Ratio for Cases and Controls

Prostate cancer develops in the prostate gland in males and is common among the elderly. Most often prostate cancer causes moderate discomfort; however, aggressive prostate cancer can be life-threatening.

The presence of prostate cancer is indicated by high levels of Prostate-Specific Antigen (PSA). Men at low risk for prostate cancer have a PSA level of less than 10 ng/ml (nanograms per milliliter); those at intermediate risk have PSA levels between 10 and 20 ng/ml.

Radiation therapy is one of the most successful treatments to combat localized prostate cancer. In these treatments, tumors are irradiated, destroying or weakening cancer cells. Dosing for radiation is measured in Grays (Gy). Typical radiation dosing for prostate patients is 70.2 Gy.

There is evidence that higher doses of radiation might be more effective in managing prostate cancer. Table 11 shows the experience of fifteen participants in a study in which a control group was given the usual dosage of 70.2 Gy of radiation, while an experimental group was given 79.2 Gy. Participants were enrolled over one year in an oncology clinic, and then monitored for six months thereafter. A relapse of cancer was indicated if the PSA level rose to 11 or higher.

**Table 11: Response to Treatment for Prostate Cancer**



* 1. Complete the following Odds Ratio Table.

Relapse of Prostate Cancer Among Study Participants

|  |  |  |
| --- | --- | --- |
| Outcome | Cases | Controls |
| Yes |  |  |
| No |  |  |
| Total |  |  |

* 1. What are the odds of having a relapse among Case patients? Among Control Patients?
  2. Calculate a) the L Statistic, b) the Standard Deviation for L.
  3. Calculate a 95% confidence interface for L.

1. Kaplan Meier Survival Graph
   1. For the data presented in Table 11, complete the table below for Control patients.



* 1. Draw the Kaplan Meier Survival Graph

# Problem Solutions

1. Matching Controls to Cases
   1. Participants 7, 8 and 12 were randomly selected to participate in the study group. Create a data table from which to perform the matching analysis. The table should include the Participant Number, Study or Control Group Flag, Sodium Level, and Random Number Assignment.



* 1. Matching on Daily Sodium Intake, which three participants would be the best controls for Participant 7?



* 1. Which is the single best control for Participant 12?

1. Verification of Matching
   1. State the hypothesis to be tested.



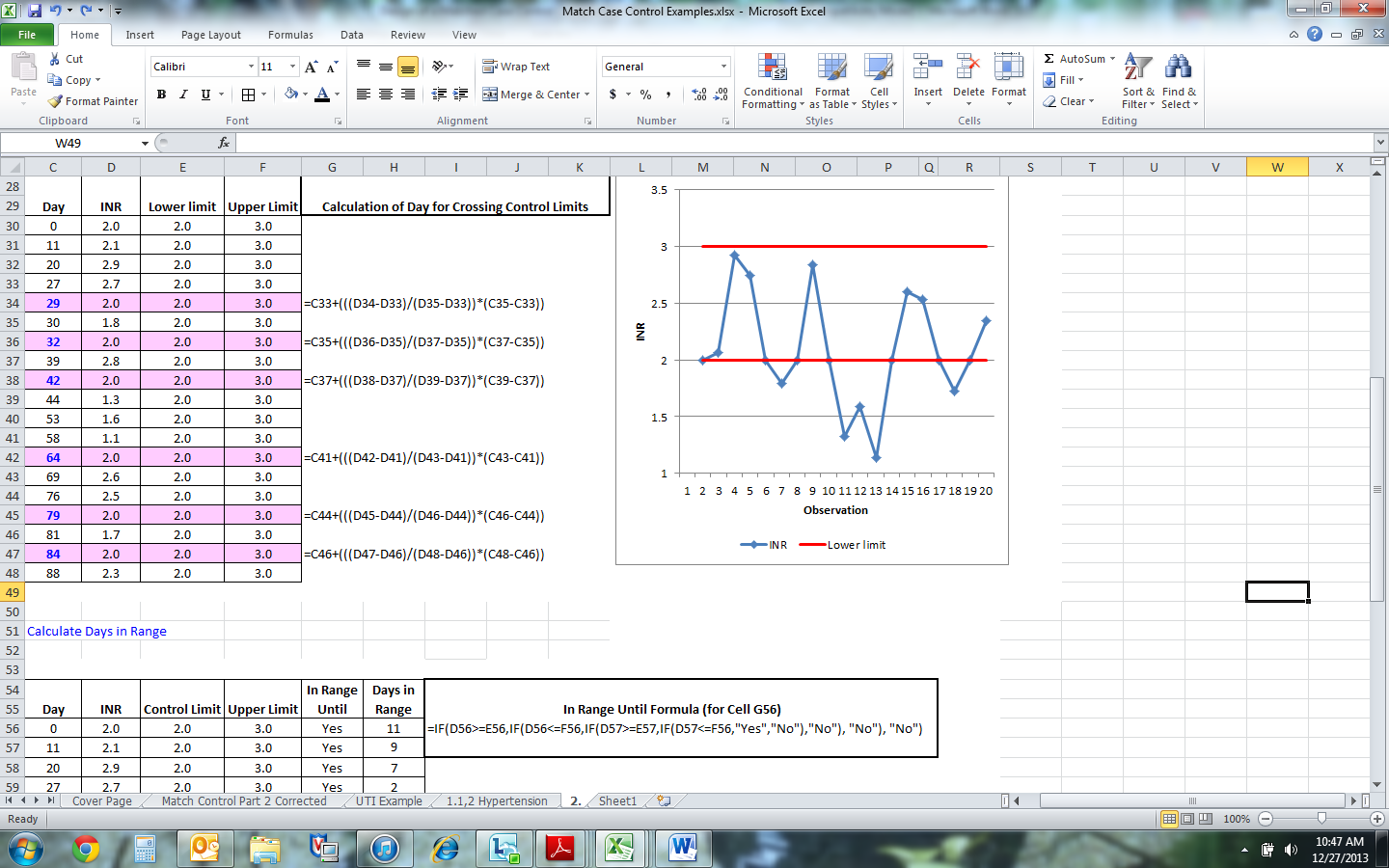
* 1. Calculate the mean and standard deviation for the differences of sodium intake between the case and control participants.



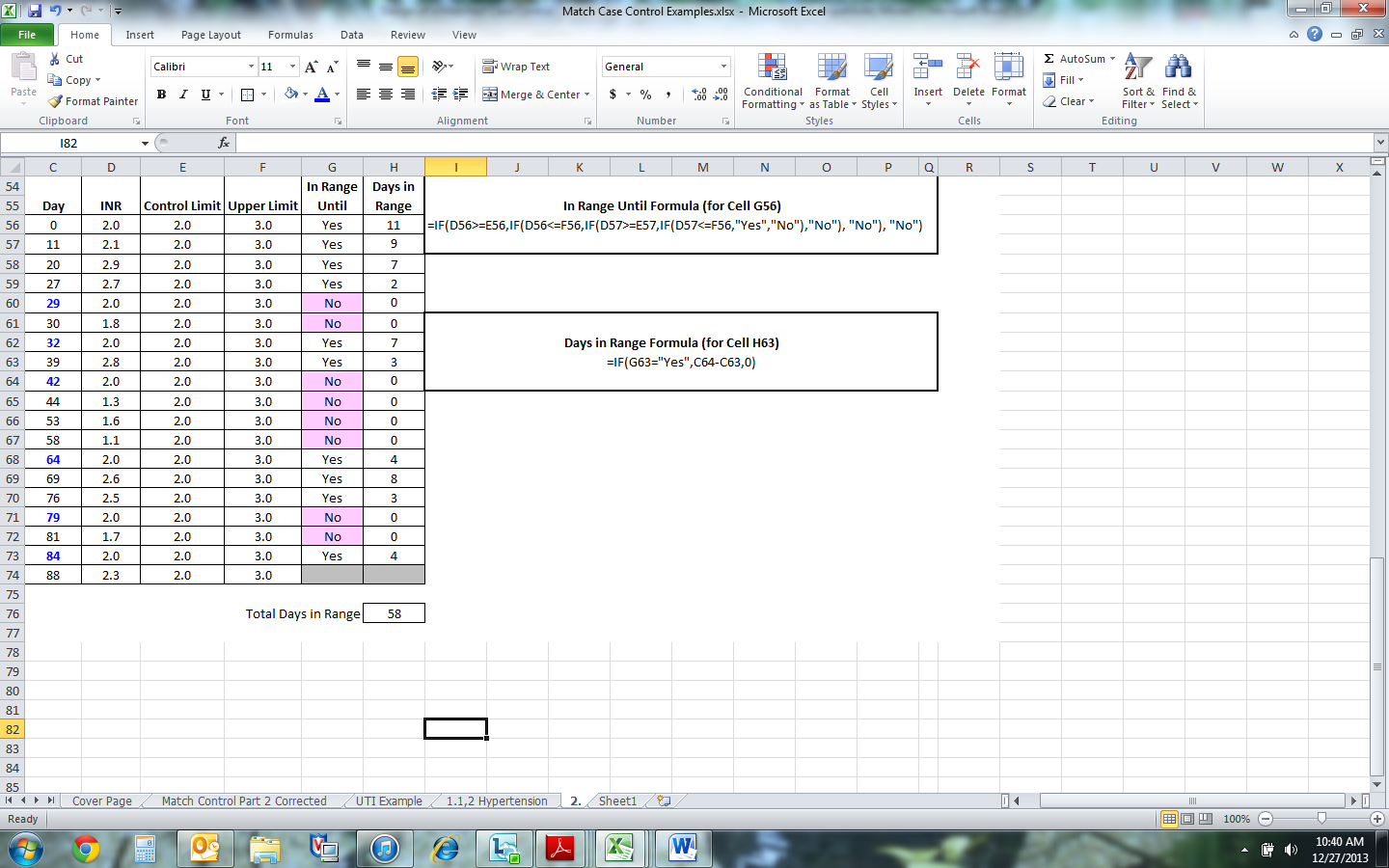
* 1. Calculate and interpret the t-statistic. Is there a statistical difference in sodium intake between participants in the case and control groups?



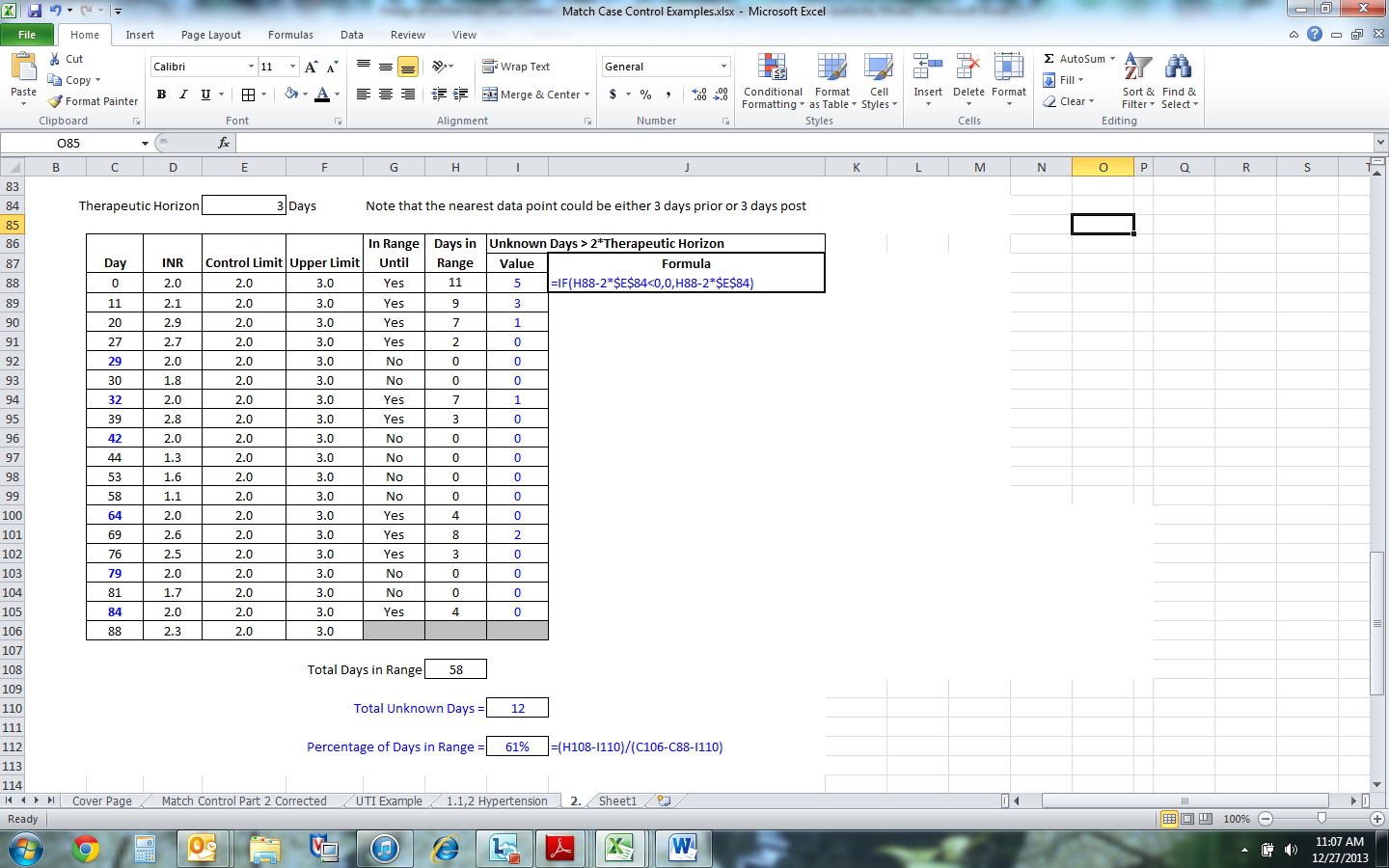
1. Days in Range
   1. Plot the observation data. Calculate the days on which the patient crossed control limits.



* 1. Calculate the number of days the patient was in therapeutic range.



* 1. Calculate the unknown days, and the percentage of days in range.



1. Odds Ratio for Cases and Controls
   1. Complete the following Odds Ratio Table.

|  |  |  |
| --- | --- | --- |
| Outcome | Cases | Controls |
| PSA>10 | 2 | 6 |
| PSA<10 | 3 | 4 |
| Total | 5 | 10 |

4.2 What are the odds of having a relapse among Case patients? Among Control Patients?

* 1. Calculate a) the L Statistic, b) the Standard Deviation for L.

1. = (2\*4)/(6\*3) = 0.44
2. = = 1.1

4.3 Calculate a 95% confidence interface for L..

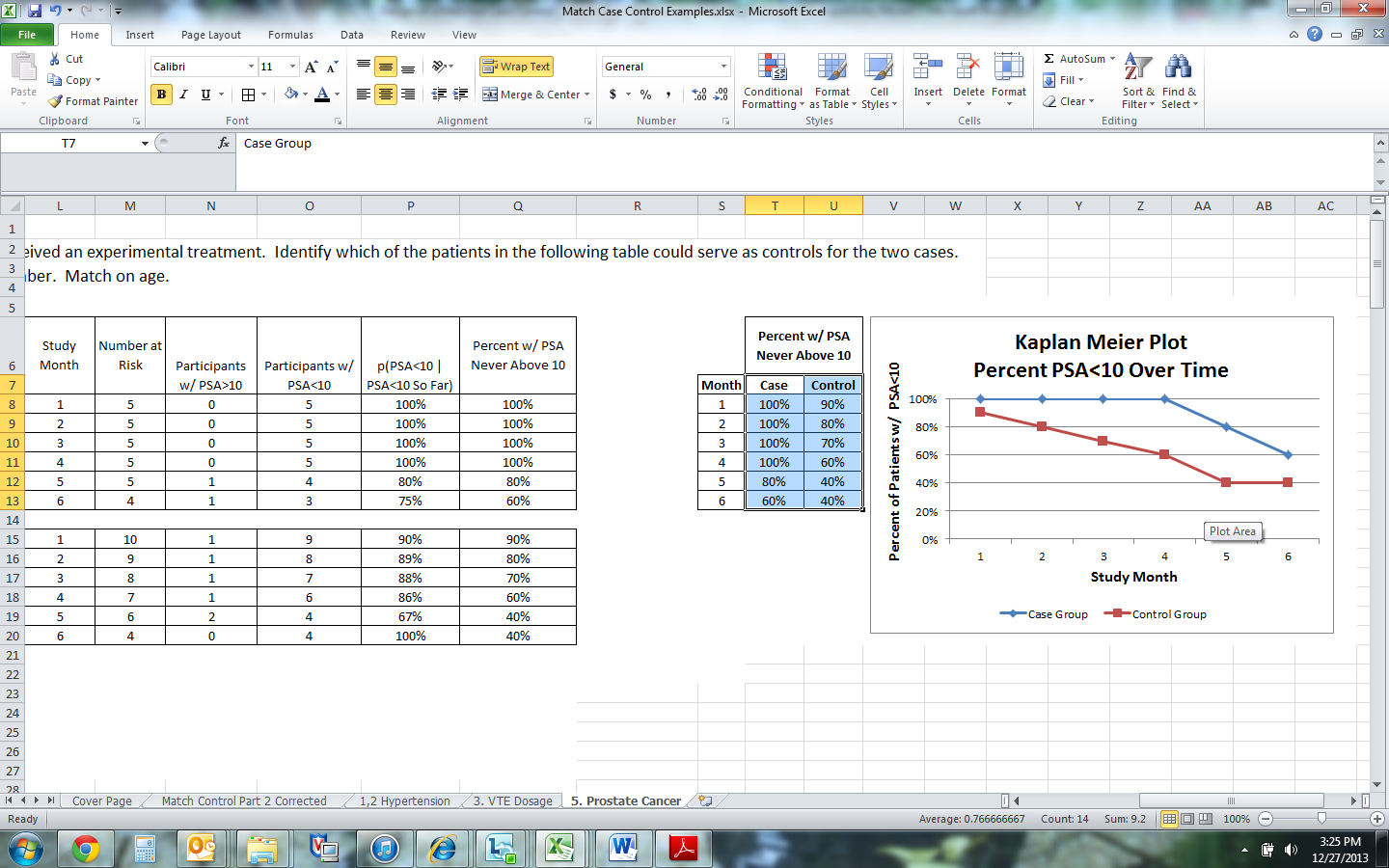
= (1.7, 2.6)

Kaplan Meier Survival Graph

* 1. Complete the table below for Control patients.



* 1. Draw the Kaplan Meier Survival Graph



Step 1: Create a table listing the Percent of Patients per month who have never had a PSA>10 ng/ml.

Step 2: Highlight the cells for case and control values, as illustrated at right.

Step 3: In Excel, Choose Insert 🡪 Chart 🡪 Line. Use normal Excel functions to add titles.

**Appendix A: Details of Therapeutic Range Calculations**

Consider the case of a patient whose blood pressure is taken as outlined in Figure 5. Establish the percent of days the patient is in therapeutic range. On days when blood pressure was not measured, assume that the blood pressure is the same as the nearest data point less than 7 days away. For all other days assume that the blood pressure is unknown.

Step 1: Plot the Observation Data

1. There are four points at which the plot of observed data crossed the control limits.
2. The first two points crossed the upper control limit; the second crossed the lower.

Step 2: Calculate Where the Lines Cross the Control Limits

1. Insert rows for each point at which the plotted data crossed the control limit.
2. Estimate the day on which the blood pressure moved out of range as a linear interpolation.

The first out of range blood pressure occurred on Day 20, registering 150 mmhg. What day was it likely that the patient’s blood pressure first moved out of range? This is calculated as a linear interpolation.



For this first point,

A similar approach is used to calculate the other three dates on which the patient’s blood pressure crossed control thresholds.

Step 3: Calculate Days in Range

1. Calculation of Days in Range is done by inspection for each observation. If both the starting and ending observations are in range.

*Days in Range = Ending Observation Date – Starting Observation Date.*

For the first point,

*Days in Range = 12 – 0 = 12.*



Step 4: Calculate Unknown Days

1. On days when blood pressure was not measured, assume that blood pressure is the same as the nearest data point less than 7 days away. For all other days assume that the blood pressure is unknown.

This means that for any given day we look 7 days forward and 7 days back for the nearest point. If we don't find one, the value is null.



Step 5: Calculate Percentage of Unknown Days

1. The Percentage of Days in Range is given by the following:

*Percentage Days in Range = (Total Days in Range – Total Unknown Days) / (Ending Observation Date – Starting Observation Date – Total Unknown Days)*

For this example,

*Percentage Days in Range = (77 – 22) / (120 – 0 – 22)= 56%*



**Appendix C: Age Comparison Verification**

To verify that the matching has been successful, the ages of cases and controls are compared. The null hypothesis is that the cases and the matched controls are the same age. The alternative hypothesis is that there is a statistically significant difference between ages of patients in the case and control groups.

Step 1: State the Hypothesis

H0: Average AgeCase - Average AgeControl = 0

Ha: Average AgeCase - Average AgeControl <> 0

Step 2: Calculate the difference between each case and the mean of the controls that match to that case.

Case 1 Controls Age Case 3 Controls Age

Patient 8 64 Patient 4 82

Patient 9 70 Patient 6 80

Average 67 Average 81

Age of Case 1 - μ(Case 1 Controls) = (65-67) = -2

Age of Case 2 - μ(Case 2 Controls) = (84-81) = 3

Step 3: Calculate the average and standard deviation of the differences.

μ(Difference) =(-2+3)/2 = 0.5

Standard Deviation (Difference) = 3.5

Step 4: The t-statistic is used to test if the mean of the differences is significantly different from zero.

t-statistic = [ μ(Difference) - 0 ] / STD(difference) = (0.5-0) / 3.5 = 0.14

t(0.14, 4 *df*) = 0.64

t(Two-Tail Test, 0.05 Significance, 4 Degrees of Freedom) = 2.78

The t-statistic of 0.64 is less than the t-value of 2.78 so we fail to reject the null hypothesis that there is no statistical difference between the average ages of patients in the case and control groups.

# References

1. IOM. Digital data improvement priorities for continuous learning in health and health care. The National Academic Press, Washington DC, 2013. [↑](#endnote-ref-1)
2. Cuttino LW, Khan A, Wazer DE, Arthur DW, Vicini FA. When retrospective comparative effectiveness research hinders science and patient-centered care. J Clin Oncol. 2013 Jun 10;31(17):2226-7. [↑](#endnote-ref-2)
3. Lohr KN. Comparative effectiveness research methods: symposium overview and summary. Med Care. 2010 Jun; 48(6 Suppl): S3-6. [↑](#endnote-ref-3)
4. Berger M., Mamdani M, Atkins D., Johnson M. 2009. “Good Research Practices for Comparative Effectiveness Research: Defining, Reporting and Interpreting Nonrandomized Studies of Treatment Effects Using Secondary Data Sources: The ISPOR Good Research Practices for Retrospective Database Analysis Task Force Report—Part I,” Value in Health, Volume 12, Number 8. [↑](#endnote-ref-4)
5. Hagedorn, G. (2002). *Preview of deltaaccess* . Retrieved from http://diversityworkbench.net/OldModels/Descriptions/Docu160/PREVIEW\_DELTAACCESS.HTML [↑](#endnote-ref-5)
6. Schwarz , J. (2010, October 12). [Web log message]. Retrieved from http://healthcaresecurity.wordpress.com/2010/10/12/why-are-hierarchical-databases-like-mumps-still-popular-in-healthcare/ [↑](#endnote-ref-6)
7. Newsom, S. W. B. (2006). Pioneers in infection control: John snow, henry. *Journal of Hospital Infection*, *64*, 210-216. doi: 10.1016/j.jhin.2006.05.020 [↑](#endnote-ref-7)
8. Paneth, N., Susser, E., & Susser, M. (2002). Origins and early development of the case-control study:. *Journal of Epidemiology and Community Health*, *47*, 359–365. Retrieved from http://www.epidemiology.ch/history/papers/SPM 47(6) 359-65 Paneth et al. \_ Part 2.pdf [↑](#endnote-ref-8)
9. Elixhauser A., Steiner C., Harris R., Coffey R.M. 1998. "Comorbidity Measures for Use with Administrative Data," Medical Care 36:8-27. [↑](#endnote-ref-9)
10. Patient falls were classified using the following E codes in the International Classification of Disease: E8800 E8801 E8809 E8810 E8811 E882 E8830 E8831 E8832 E8839 E8840 E8841 E8842 E8843 E8844 E8845 E8846 E8849 E885 E8850 E8851 E8852 E8853 E8854 E8859 E8860 E8869 E888 E8880 E8881 E8888 E8889 E9681 E9870 E9871 E9872 E9879. [↑](#endnote-ref-10)
11. Patient injuries were measured using the following codes: 7960 7990 79901 79902 9071 9073 9074 9075 9079 9085 9086 9089 9092 9094 9099 9300 9301 9302 9308 9309 931 932 9330 9331 9340 9341 9348 9349 9350 9351 9352 936 937 938 9390 9391 9392 9393 9399 9500 9501 9502 9503 9509 9510 9511 9512 9513 9514 9515 9516 9517 9518 9519 9530 9531 9532 9533 9534 9535 9538 9539 9540 9541 9548 9549 9550 9551 9552 9553 9554 9555 9556 9557 9558 9559 9560 9561 9562 9563 9564 9565 9568 9569 9570 9571 9578 9579 9580 9581 9582 9583 9584 9585 9586 9587 9588 95890 95891 95892 95893 95899 9590 95901 95909 9591 95911 95912 95913 95914 95919 9592 9593 9594 9595 9596 9597 9598 9599 990 9910 9911 9912 9913 9914 9915 9916 9918 9919 9920 9921 9922 9923 9924 9925 9926 9927 9928 9929 9930 9931 9932 9933 9934 9938 9939 9940 9941 9942 9943 9944 9945 9946 9947 9948 9949 9951 9955 99550 99551 99552 99553 99554 99555 99559 99580 99581 99582 99583 99584 99585 99589 99590 99591 99592 99593 99594 V155 V1551 V1559 V156 V1588 V713 V714 V715 V716 V9010 V9011 V9012 V902 V9031 V9032 V9033 V9039 V9081 V9083 V9089 V909. [↑](#endnote-ref-11)
12. The following codes were used to determine if the patient has had an adverse medication effect: E9300 E9301 E9302 E9303 E9304 E9305 E9306 E9307 E9308 E9309 E9310 E9311 E9312 E9313 E9314 E9315 E9316 E9317 E9318 E9319 E9320 E9321 E9322 E9323 E9324 E9325 E9326 E9327 E9328 E9329 E9330 E9331 E9332 E9333 E9334 E9335 E9336 E9337 E9338 E9339 E9340 E9341 E9342 E9343 E9344 E9345 E9346 E9347 E9348 E9349 E9350 E9351 E9352 E9353 E9354 E9355 E9356 E9357 E9358 E9359 E9360 E9361 E9362 E9363 E9364 E9370 E9371 E9372 E9373 E9374 E9375 E9376 E9378 E9379 E9380 E9381 E9382 E9383 E9384 E9385 E9386 E9387 E9389 E9390 E9391 E9392 E9393 E9394 E9395 E9396 E9397 E9398 E9399 E9400 E9401 E9408 E9409 E9410 E9411 E9412 E9413 E9419 E9420 E9421 E9422 E9423 E9424 E9425 E9426 E9427 E9428 E9429 E9430 E9431 E9432 E9433 E9434 E9435 E9436 E9438 E9439 E9440 E9441 E9442 E9443 E9444 E9445 E9446 E9447 E9450 E9451 E9452 E9453 E9454 E9455 E9456 E9457 E9458 E9460 E9461 E9462 E9463 E9464 E9465 E9466 E9467 E9468 E9469 E9470 E9471 E9472 E9473 E9474 E9478 E9479 E9480 E9481 E9482 E9483 E9484 E9485 E9486 E9488 E9489 E9490 E9491 E9492 E9493 E9494 E9495 E9496 E9497 E9499. [↑](#endnote-ref-12)
13. The following codes were used to determine mood disorders: 29383 29600 29601 29602 29603 29604 29605 29606 29610 29611 29612 29613 29614 29615 29616 29620 29621 29622 29623 29624 29625 29626 29630 29631 29632 29633 29634 29635 29636 29640 29641 29642 29643 29644 29645 29646 29650 29651 29652 29653 29654 29655 29656 29660 29661 29662 29663 29664 29665 29666 2967 29680 29681 29682 29689 29690 29699 3004 311. The following codes were used to identify anxiety disorders: 29384 30000 30001 30002 30009 30010 30020 30021 30022 30023 30029 3003 3005 30089 3009 3080 3081 3082 3083 3084 3089 30981 3130 3131 31321 31322 3133 31382 31383. [↑](#endnote-ref-13)
14. Rosendaal FR, Cannegieter SC, van der Meer FJ, Briet E. A method to determine the optimal intensity of oral anticoagulant therapy. Thromb Haemost. 1993 Mar 1;69(3):236-9. [↑](#endnote-ref-14)
15. Niven DJ, Berthiaume LR, Fick GH, Laupland KB. Matched case-control studies: a review of reported statistical methodology. Clin Epidemiol. 2012;4:99-110. [↑](#endnote-ref-15)
16. Breslow NE, Day NE. Statistical Methods in Cancer Research; Volume 1 – The Analysis of Case-Control Studies. Lyon, France: International Agency for Research on Cancer; 1980. [↑](#endnote-ref-16)
17. Morris and Gardner; Gardner, MJ (1988). "Calculating confidence intervals for relative risks (odds ratios) and standardized ratios and rates". British Medical Journal 296 (6632): 1313–1316. [↑](#endnote-ref-17)
18. Kaplan, E. L.; Meier, P.: Nonparametric estimation from incomplete observations. J. Amer. Statist. Assn. 53:457–481, 1958. [↑](#endnote-ref-18)
19. Greenwood M. The natural duration of cancer. Reports on Public Health and Medical Subjects. London: Her Majesty's Stationery Office 1926;33:1–26. [↑](#endnote-ref-19)
20. Alemi F, Walters SR. A mathematical theory for identifying and measuring severity of episodes of care. Qual Manag Health Care. 2006 Apr-Jun;15(2):72-82. [↑](#endnote-ref-20)
21. Alemi F, Uriyo M. Accuracy of Claims-Based Measures of Severity of Childhood Illnesses. Health Outcomes Research in Medicine 2012, 2 (2): e71-e78. [↑](#endnote-ref-21)