The tvtools package for R

David Shilane

August 23, 2012

1 Introduction

Modern digital recording systems have led to a vast increase in the types and amounts of data that can be collected. As these systems evolve, a variety of new data formats may arise. Electronic medical data are often organized in *long–form time–varying* records. These longitudinal records may also be called *panel data*. The hallmark of time–varying data is that a single subject may contribute many observations over time. These records are often compiled into data sets with multiple rows per subject in mutually exclusive time intervals. The traditional tools for descriptive statistics may not necessarily apply to time–varying data. For instance, the column average of a variable is not necessarily relevant when the data include multiple identical observations per subject. Furthermore, these descriptive statistics should take the temporal nature of panel data into account. Data visualizations also grow more complex in a longitudinal setting, and the sheer size of a time–varying data set may also pose challenges. In light of these concerns, I devised the tvtools software package for R [2] as a means to simplify the descriptive analysis of time–varying data. In this manuscript, I will discuss a variety of settings in which these data may be difficult to analyze and introduce the computational tools I implemented to address them.

2 Example of Time–Varying Data

	ID	time1	time2	age	drug	death
Row 1	1	0	5	65	1	0
Row 2	1	5	6	65	1	1
Row 3	2	0	3	60	1	0
Row 4	2	3	10	60	0	0
Row 5	2	10	12	60	1	0
Row 6	3	0	8	70	0	0
Row 7	3	8	9	70	1	1
Row 8	4	0	1	85	0	1
Row 9	5	0	6	55	1	0
Row 10	5	6	15	55	0	0

Table 1: Example of a long–form time–varying data set tracking 5 unique patients.

Table 1 provides an example of time–varying data. Its salient features include:

- **ID**: a unique subject identifier to link records in multiple rows.
- **time1, time 2**: the interval of time for the observation.
- **Constant variables**: values are assigned at baseline and do not change. For instance, the patient's age at index is a fixed quantity across rows in these data.
- **Time–varying variables**: values may change. So, for instance, the patient's drug treatment status and survival outcome may change over time.
- **Outcomes**: Binary variables that occur at the beginning of the interval.

The **tvtools** package was motivated by a variety of research projects arising from an analysis of time–varying data for patients with coronary heart disease treated by Kaiser Permanente of Northern California, a large, integrated health care delivery system. These data followed $n = 65,565$ patients in the years 2000 through 2008 from their initial diagnosis of heart disease until death, loss of follow–up through changing insurance providers, or administrative censoring on January 1, 2009. The data arised from Kaiser's electronic health records. The available information includes a rich profile of covariates, including demographics, comorbidities, inpatient and outpatient laboratory measures, prescription fills, and a variety of cardiovascular outcomes such as mortality, myocardial infarction, and procedures such as angioplasty and bypass surgery. With many opportunities to update the patients' profiles in follow–up, the overall data include approximately 1.7 million rows and roughly 100 columns. The average patient had 27 rows of data, while the maximum value was 497 updates for a single patient.

3 Visualizing Longitudinal Subject Histories

The sheer volume of information contained in time–varying data provides ample opportunity for coding errors and misinterpretation. Data visualization of subject histories can facilitate the investigation of quality assurance. Moreover, physicians can make use of a visual patient history to guide treatment decisions. The **timeplot** method within the **tvtools** package provides an automatic means of displaying an individual subject's data. Binary exposure variables may be tracked across time with broken lines indicating a gap in exposure or treatment. Patient outcomes such as procedures and medical events are indicated by vertical lines.

Figure 1 displays a visual record for one patient in the year between entry into the cohort for angioplasty and death. This graph tracks the prescription records for five medications, provide the time intervals in which the patient was hospitalized, and display the patient's complete record for procedures such as angioplasty and bypass surgery and outcomes such as unstable angina, myocardial infarction, and death. Using the **exposure** method (Section 5), we computed the patient's overall medical possession ratio for the drug treatments and hospitalization within the figure's legend.

Figure 1 identifies two major problems with the data we were analyzing. First, the patient was not hospitalized for the initial angioplasty at baseline. While still indicative of heart problems, the study seeked to follow patients after an initial hospitalized episode of coronary heart disease. As it turns out, roughly 5% of the overall cohort did not meet this

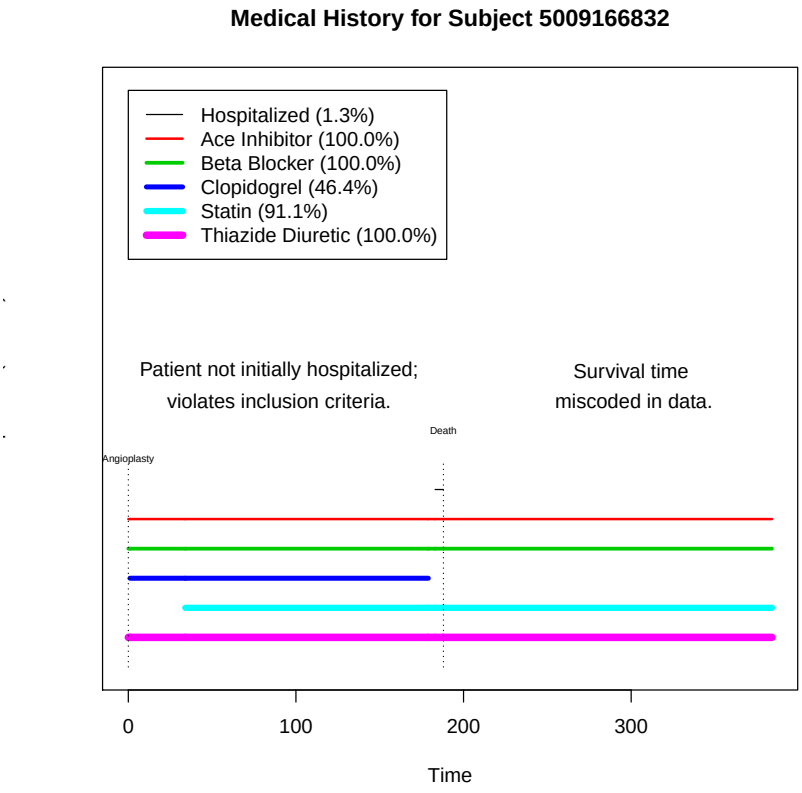

Figure 1: One patient's medical history illustrates multiple data issues.

inclusion criterion. Second, Figure 1 suggests that the patient died after approximately 6 months and then continued to receive 4 prescription drug treatments for the remainder of the year. This could either indicate a miscoding of the pharmacy records or the patient's time of death. After some investigation, we determined that all patient death times were systematically miscoded. Among those patients who died during follow–up, the time of death actually occurred at the end of the patient's final recorded interval. This is in contrast to the convention that all events and other updates occur at the beginning of the time interval; in survival studies, a Kaplan–Meier estimator or Cox proportional hazards model would systematically record deaths earlier than they occurred. In this particular case, the data's coding mistake would have resulted in a difference of 6 months in the patient's time to event. The mistake is easily corrected by adding an additional row for each patient and updating the time of death.

Fortunately, in drawing these patient histories, we discovered these data quality issues early on and corrected them. Our data set has served as the basis for a variety of manuscripts in progress. Without uncovering these issues, all of these studies would have utilized faulty survival data and included patients who should have been excluded from the cohort.

4 Descriptive Statistics with Time–Varying Data

	TVclopidogrel = 0	TVclopidogrel = 1	Total
< 0.25 years	666	89	755
0.25–1 years	737	108	845
>= 1 year	595	47	642
All follow-up	1989	244	2233

Table 2: **Death totals** by era.

	TVclopidogrel = 0	TVclopidogrel = 1	Total
< 0.25 years	2672.0	1450.5	4122.5
0.25–1 years	8317.2	2367.6	10684.9
>= 1 year	9712.3	1346.9	11059.1
All follow-up	20701.5	5165.0	25866.5

Table 3: **Person–years of follow–up** by era.

Because each subject includes repeated measures over multiple rows, traditional descriptive statistics such as the column average are not necessarily meaningful in the analysis of time–varying data. We can take the data's temporal component into account by computing crude event rates relative to the overall length of follow–up.

As an example, we can consider a study arising from a subset of Kaiser patients after their first episode of acute coronary syndrome. In particular, we sought to study the effect of Clopidogrel, an anti–platelet drug, on survival. A total of 17351 patients were considered,

	TVclopidogrel = 0	TVclopidogrel = 1	Total
$<$ 0.25 years	24.9	6.1	18.3
0.25–1 years	8.9	4.6	7.9
$>=$ 1 year	6.1	3.5	5.8
All follow-up	9.6	4.7	8.6

Table 4: **Crude death rates** per 100 Person-Years of follow–up by era.

including some who continuously used Clopidogrel for a year or longer, many who never filled a prescription, and numerous patients who were treated for shorter durations or repeatedly switched on and off of the medication. Clinically, we were interested in the survival benefits of Clopidogrel in several time periods: the first three months (0.25 years) after index, the remainder of the first year, and all times thereafter.

Table 2 provides death totals within each treatment category and time period, along with overall counts. The variable TVclopidogrel tracked time–varying medication possession based upon pharmacy records. Each death was classified based upon the current status of TVclopidogrel and the era in which it occurred. However, these death counts are not directly comparable because relatively few patients took Clopidogrel, and even those who did were not necessarily continuous users. Table 3 provides the overall **person–years of follow–up** for each era and treatment category. These quantities sum the overall time on and off of treatment within each era to get a sense of the overall exposure to the drug. Finally, the death counts of Table 2 can be weighed against the exposure time in Table 3 to compute the **crude rate of death per 100 person–years** in table 4. Because it incorporates the length of exposure, these crude rates are more directly comparable, and their ratio across treatment categories provides an estimate of the unadjusted hazard ratio of death on Clopidogrel versus no treatment. These three tables may be computed directly within the **cruderates** method of the **tvtools** package by specifying the outcome of interest, the (categorical) treatment variable, and the eras of interest.

Fundamentally, crude event rates are relatively easy to compute. However, time–varying data are not necessarily structured to properly estimate these rates within user–selected eras. Any observation that overlaps multiple eras may induce opportunities for miscalculation. To address this concern, any overlapping observations can be subdivided into multiple rows, each in time intervals contained in a single era. The **era.splits** method within the **tvtools** package performs this task automatically. In practice, it is a good first step to run prior to any descriptive analysis or fitting a model that relies upon era effects.

We subsequently modeled these data in a Cox proportional hazards regression model to estimate the era–specific effects of Clopidogrel while adjusting for potentially confounding factors. Table 5 shows the model estimates of the hazard ratio for Clopidogrel treatment within each era. Fortunately, Cox regression is already equipped to handle time–varying data [3, 1], so this model required no additional software development. However, the era–specific estimates required that we use **era.splits** to properly format the data. An accurate estimate of these era effects would otherwise be difficult to obtain.

	Hazard Ratio	Lower 0.95	Upper 0.95	p-value
Clopidogrel $<$ 90 Days (tv)	0.43	0.34	0.54	0.0000
Clopidogrel 90-365.25 Days (tv)	0.71	0.58	0.87	0.0011
Clopidogrel $>=$ 365.25 Days (tv)	0.93	0.69	1.25	0.6380
Age (Decade)	2.04	1.96	2.14	0.0000
Male	1.19	1.09	1.30	0.0001
African American	0.98	0.83	1.16	0.8373
Asian	0.81	0.69	0.95	0.0106
Other Race	0.69	0.57	0.85	0.0004
Index 2004	0.98	0.86	1.11	0.7274
Index 2005	1.02	0.90	1.16	0.7828
Index 2006	0.84	0.74	0.97	0.0174
Index 2007	0.90	0.78	1.04	0.1593
Index 2008	0.72	0.59	0.89	0.0020
Initial NSTEMI	2.74	2.41	3.10	0.0000
Heart Failure	1.95	1.76	2.15	0.0000
Bleed	1.27	1.07	1.51	0.0053
Hypertension	1.07	0.97	1.19	0.1662
Dyslipidemia	0.81	0.73	0.91	0.0002
Diabetes	1.58	1.44	1.74	0.0000
PAD	1.70	1.44	2.01	0.0000
Valvular Disease	1.16	1.02	1.32	0.0281
PPI Initiated	1.19	1.08	1.31	0.0003
ACE/ARB Initiated	0.84	0.77	0.92	0.0002
Beta Blocker Initiated	0.81	0.73	0.89	0.0000
Statin Initiated	0.71	0.65	0.79	0.0000

Table 5: Time-varying Cox regression model of mortality with 2-year follow-up with Clopidogrel era effects. Because many patients had follow–up in intervals that overlapped different eras, this model would not accurately fit without first splitting the data into mutually exclusive eras. The software package accomplishes this by adding additional rows for each patient record that crosses the 90 day or 1 year boundary.

Overall, the regression results suggest a strong association between Clopidogrel use and improved survival in the first 90 days after acute coronary syndrome. The effect is more modest but still significant in the remainder of the first year, and not significant thereafter. These results roughly correspond with medical guidelines and randomized trial results for Clopidogrel usage after acute coronary syndrome. The purpose of this study was to assess the impact of the drug in the broader population of patients who do not necessarily meet the selection criteria for a trial.

5 Calculating Exposure Rates

Collecting time–varying exposure to treatments and sources of disease can provide more specific information about a subject's likelihood for future events. In baseline medical studies, it is often considered sufficient to ascertain whether a patient was initially prescribed a

medication. A time–varying analysis can link the patient's medical outcomes to his or her record of prescription re–fills over time. The *medical possession ratio* is the percentage of a time interval in which the patient possessed a medical treatment. This ratio may be used as a proxy for the patient's overall drug adherence.

Drug possession and other exposures are typically encoded as time–varying binary variables. The medical possession ratio or exposure rate within a specific interval is simply the total possession divided by the length of time in question. The **exposure** method within the **tvtools** package provides an automatic means to calculate exposure rates for each subject and exposure.

6 Event Times

Many studies focus on the time to a patient's first event. Time–varying data may include multiple events for the same subject (e.g. all myocardial infarctions for a single patient). Iteratively searching for these first event times across outcomes and patients using **for** loops can be quite laborious in R. However, these searches can often be quite efficient using the appropriate calls to **sapply**. The **tvtools** package automates these calculations in the **firstevent** method.

In practice, the **firstevent** method can be quite flexible. In addition to computing the time to the first occurrence of a traditional outcome or procedure, it can also track the first time a patient goes on or off a new medication or when missing values first arise.

The distribution of follow–up times across subjects is also an important quantity. Within **tvtools**, the **followuptime** method is used to compute the maximum observed time for each subject. This in turn allows for the computation of traditional descriptions such as histograms or the median follow–up time in the sample.

7 Missing Data

Section 4 demonstrated how descriptive statistics and analyses could take the temporal component of time–varying data into account. Likewise, missing data may be a cause for concern, especially if the overall rate of missingness changes with time. The **missingness** method within the **tvtools** package calculates this rate relative to the overall number of observations available at the specified time. These rates may also be graphed using the **missingness.plot** method. By default, the plot will sample missing variable rates at regular time intervals and interpolate a piecewise linear trend. However, the times may also be specified. Missingness plots may be generated for a user–define subset of covariates or for all variables in the data set. Figure 2 provides an example tracking the missingness rate of LDL cholesterol in the Kaiser data.

8 Cross–Sectional Analysis

Not every study necessarily requires all of the information contained within a time–varying structure. In these settings, it may be sufficient to rely upon baseline information for the

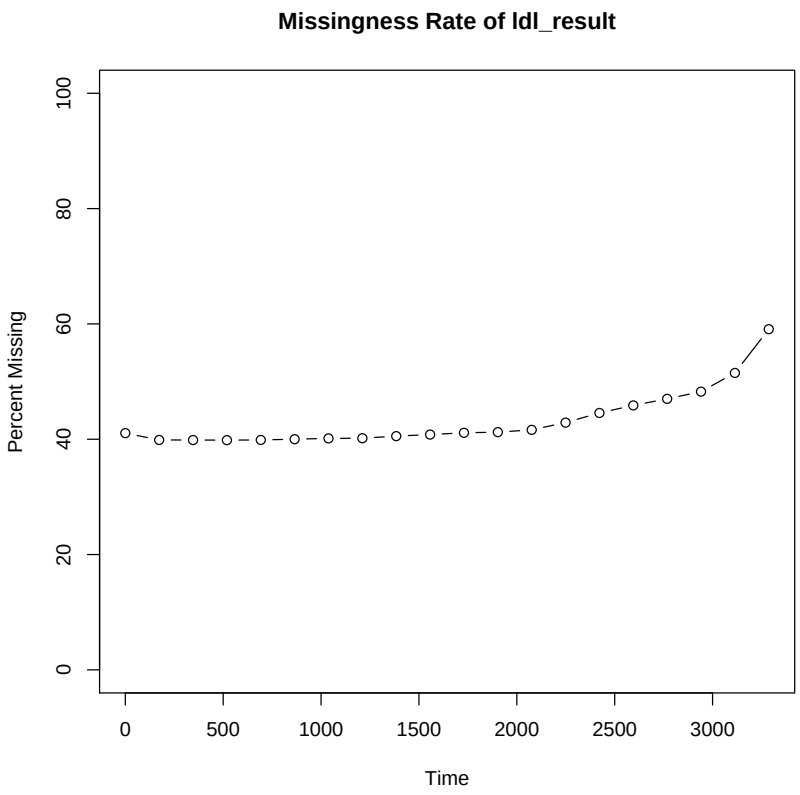

Figure 2: Tracking the rate of missing LDL cholesterol measurements over time. The trend was interpolated based upon interval sampling of missingness rates.

explanatory covariates and to compute the time to first events for outcomes, the censoring time, and quantities such as which medications were initiated within 30 days or the medical possession ratio in the first year. Other studies may focus on a cross–sectional analysis of the available data at a specific time point. In any of these cases, the user would prefer to work with a more tractable data file including only a single row per patient. The **tvtools** package automates the creation of these data subsets through the **create.baseline** and **cross.sectional.data** methods, which provide equivalent functionality. The user may specify the cross–sectional time (with baseline defaulting to time 0) and the outcomes on which to compute the time to first event. The end result is a baseline data set that includes one row per patient with explanatory covariates extracted at the selected time and the outcome and censoring times calculated from the time–varying stream. These methods can greatly facilitate the generation of specific data sets for individual research projects from a broad repository of time–varying data.

9 Conclusion

Time–varying data structures present unique challenges and add complexity to traditional statistical analysis. The **tvtools** package provides a variety of methods to facilitate data visualizations and the computation of descriptive statistics. Furthermore, it provides a computational infrastructure that aids in quality assurance and more detailed statistical models. Without the use of graphical patient timelines such as Figure 1, my research group may not have realized that our data set included patients who did not qualify for the study or that the survival times were systematically miscoded. Furthermore, models such as the Cox regression in Table 5 can not be accurately estimated without first subdividing the observations overlapping different eras into multiple units. The analytical methods presented here such as the calculation of crude event rates per person–year of follow–up or tracking the rate of missing variables over time have been utilized for quite some time. However, the computational tools for these calculations have not been developed for R in any systematic way. Time–varying data structures may increasingly become routine with the development of modern data collection mechanisms. The **tvtools** package provides a set of useful methods that can simplify the description, visualization, and analysis of time–varying data.

References

[1] F. E. Harrell. *Regression Modeling Strategies with Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer, 2010.

[2] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.

[3] T. M. Therneau and P. M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer, 2010.