

PRIMA: A Case Study of Using Information Visualization Techniques for Patient Record Analysis

Donna L. Gresh and David A. Rabenhorst*
IBM T.J. Watson Research Center

Amnon Shabo†
IBM Haifa Research Laboratory

Shimon Slavin‡
Hadassah University Hospital

Abstract

We have created an application, called PRIMA (Patient Record Intelligent Monitoring and Analysis), which can be used to visualize and understand patient record data. It was developed to better understand a large collection of patient records of bone marrow transplants at Hadassah Hospital in Jerusalem, Israel. It is based on an information visualization toolkit, Opal, which has been developed at the IBM T.J. Watson Research Center. Opal allows intelligent, interactive visualization of a wide variety of different types of data. The PRIMA application is generally applicable to a wide range of patient record data, as the underlying toolkit is flexible with regard to the form of the input data. This application is a good example of the usefulness of information visualization techniques in the bioinformatics domain, as these techniques have been developed specifically to deal with diverse sets of often unfamiliar data. We illustrate several unanticipated findings which resulted from the use of a flexible and interactive information visualization environment.

CR Categories: I.3.6 [Computer Graphics]: Methodology and Techniques—Interaction techniques

Keywords: visualization, information visualization, bioinformatics, medical records

1 Introduction

Mining the complex data collections in the fields of biology and bioinformatics is an increasingly important task. The data may be collections of genomic data, along with results of gene expression experiments, or large sets of patient records which are to be mined for relationships between patient care and outcome. The collections are often large, with dozens if not hundreds of variables. In addition, the data is often of varied type. It may include information with a spatial context (*e.g.* radiology images, gene location, molecular structures) and information with no spatial context at all (*e.g.* white blood cell count, drug family, experimental protocol). It is common for the data to include both categorical (*e.g.* medical diagnoses) and numerical (*e.g.* patient age, gene expression level) values.

Two of the important tasks in making these data useful and accessible are to, first, organize the data in a meaningful way so that queries can be launched against it and information returned, and second, to present the results of the queries to the user in a way

which enables him or her to understand the data intuitively, and allows intelligent and interactive exploration of the data.

In the Visual Analysis Group at IBM T.J. Watson Research Center, we have been developing tools to enable interactive exploration of data of varied type. We have developed a Java™ toolkit, Opal, which can be used to build custom applications, of which PRIMA is an example. The toolkit includes a wide variety of visual presentations of data, some of which are particularly suited to categorical data, and some of which are more appropriate to numerical data. However all views are linked so that exploration can occur regardless of the underlying form of the data. Meanwhile, in the Integrated Medical Records (IMR) group at IBM Haifa Research Laboratory, we have been creating a framework for organizing, storing, and retrieving medical records of varied type. Thus a natural collaboration was born to combine these efforts to enhance understanding and analysis of medical record data.

We note that the LifeLines work [2] also addressed the issue of visualizing medical records data. It was based on an interactive timeline for understanding the course of patient disease and treatment. It is most akin to a structured patient chart and is thus distinct from the problem we are addressing, which is the correlation and analysis of patterns in a large number of different patients. It would seem a natural extension to PRIMA to allow a lifelines type presentation for a given selected patient.

2 Underlying Visualization Technology

The visualization technology behind PRIMA is the Opal Java™ Toolkit. This toolkit is a successor to the Diamond [3] libraries which were created in our group at the IBM T.J. Watson Research Center in the 1990s and written in C. As with Diamond, the library provides a set of data representations which include brushing and linking to allow a user to browse, explore, and interact with a set of tabular data. Opal also includes novel data compression techniques so that it is practical to use in a client-server configuration on large data sets over a network.

Opal also includes numerous mechanisms for data “cleansing,” which we found to be particularly useful, and in fact, required for the bone marrow transplant data we were considering. Given the real world conditions in which the data was collected and entered into the database, and then converted to a form analyzable by PRIMA, there are situations where some adjustment of the data is necessary. For example, in some cases the death date was not known precisely, and was thus entered as simply a month and year, or only a year. Sometimes the date even had a clearly erroneous entry such as “18-JUN-200.” As another example, patient age was typically an integer, but included a few fields such as “6 MO.,” and “29-AP.” While of course such information can be hand-corrected (or purged if necessary) by an intelligent human, we wanted to provide a system where the input data set could be changed at any time, and intelligent modifications to the data could be automatically made. Opal provided us with this capability, for instance by having built-in date conversion capabilities, which can handle a variety of input

*gresh,drab@us.ibm.com, IBM T.J. Watson Research Center, P.O. Box 704 Yorktown Heights, NY 10598

†shabo@il.ibm.com, IBM Research and Development Labs, Haifa University, Mount Carmel, Haifa, Israel

‡Head, Bone Marrow Transplantation and Immunobiology Research Center, Hadassah University Hospital, Jerusalem, Israel

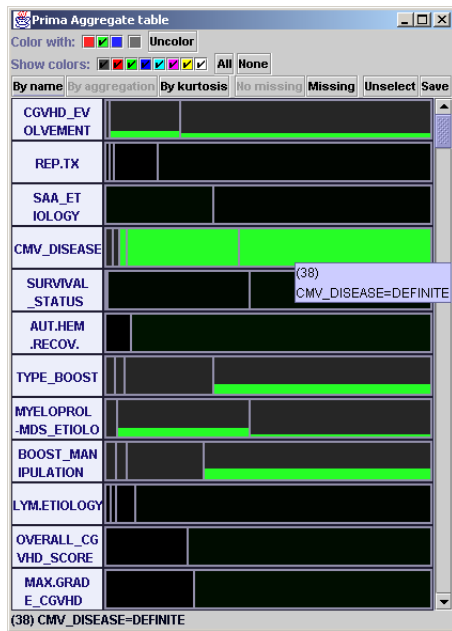


Figure 1: The PRIMA Aggregate window. All patients for whom the record indicated CMV_DISEASE of “YES”, “SUSPECTED,” or “DEFINITE” have been colored green. The user has let the mouse linger in the right-most green block of the “CMV_DISEASE” variable, resulting in the pop-up window indicating the contents of that block.

date formats, and the ability to filter fields which are *nearly* all numeric to remove non-numeric data (the age example above). While in this case a small amount of data is lost, it is preferable to the common alternative of treating the entire age field as a string variable, resulting in age 5 and age 50 being considered “close” in any resulting picture representation.

It is also the case that this data set contains many missing values, which must be handled by the application. (We note that the data set under consideration is, in addition, only a small subset of the entire data set in the hospital database.) In some cases, this is because the information is not known; for example, for a few of the patients, the survival status is not known. In other cases, missing data is due to the field being inapplicable to that particular patient; for example CGVHD evolution (chronic graft vs. host disease history) is only relevant to a patient who actually developed this syndrome, and relapse date is only relevant for patients who had a relapse.

The underlying toolkit allows the PRIMA application to be delivered either as a stand-alone application or as an applet. The PRIMA application is currently delivered as a stand-alone application, which can directly access the IMR servlet to collect the most recent data set.

3 Views

PRIMA is based on the fundamental premise of the value of multiple, linked views of the data under consideration. The initial window allows the user to choose which of several views of the data to present. Some of these choices are available at all times, while others require a variable to have been previously chosen, and thus are inactive at the initial startup of the application.

The Aggregate table is the first presentation we will discuss. The Aggregate view for the bone marrow transplant data is shown in

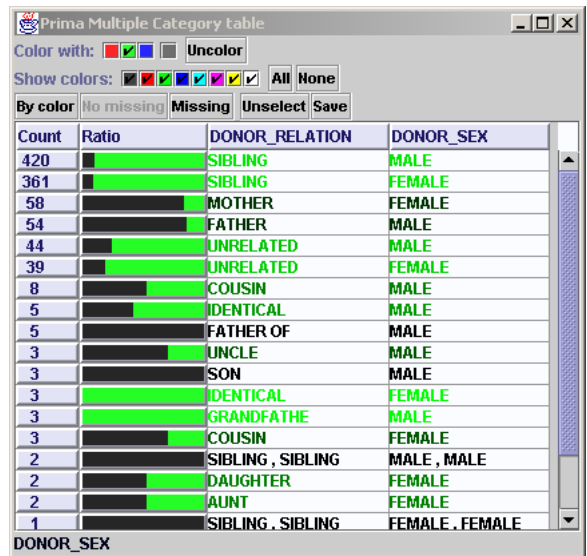


Figure 2: The PRIMA multiple category window, for the two variables DONOR_RELATION and DONOR_SEX. Patients for which the variable DONOR_MATCH was FULL_MATCH are colored green. In addition, the text indicates by color the proportion of green cases to be found in that combination.

Figure 1. This view shows the proportion of cases (patients) in each category for that variable. This presentation is particularly suited for “categorical” (rather than numerical) data, where the data fall into one of several groups. By default, the variables are ordered by a statistic we call “aggregation,” which is computed automatically by Opal. This places the most “categorical” variables at the top of the scrolling window. It is also possible to order the variables by other measures. Note that this window (as well as other windows as appropriate) also allows the user to choose a color “brush,” whether or not to show missing values, and to save the current picture as an image. In Figure 1, the user has already colored in green all patients whose record indicated that CMV (Cytomegalovirus) disease was “YES,” “SUSPECTED,” or “DEFINITE.” (Note the typical dispersion of human-entered data, with “YES” and “DEFINITE” presumably meaning the same thing.) Figure 1 shows the aggregate table when the user has let the mouse linger in the rightmost of the colored blocks; the result is the temporary pop-up window which indicates the contents of the block. The status line at the bottom of the window also indicates this information immediately as the mouse is moved around the window. The proportion of patients with CMV disease YES, SUSPECTED or DEFINITE is also shown by the colored blocks in all of the other variables. In this view, variables with missing values are not shown, explaining why some variables show no colored cases. Because this data set contains numerous missing values, our users requested the ability to optionally not show those cases. The Aggregate window is an excellent way to get an overview of the categorical variables in a data set, and to see correlations between variables through interactive coloring.

Another view of the data is called a “Histogram stack,” which shows a histogram of each variable. Again it has a default ordering, appropriate for numerical (as opposed to categorical) data, which can be changed by the user. In addition to the Histogram stack, the user can also ask for a histogram on a single variable, simply by selecting that variable (by clicking on its name in any picture) and then requesting a “Single Histogram” from the initial window.

The Category tables allow a user to see the values which occur

in a data set for either a single variable, or for multiple categories. For the multiple category case, one can see all of the unique combinations of values. For example, the category table for the variables “DONOR_RELATION” and “DONOR_SEX” is shown in Figure 2 (with all cases of “FULL MATCH” currently colored green). We see that the data are consistent, in that donors who should be female (e.g. mothers) are female, with 58 cases of this combination present in the data set. (The presence of both “FATHER” and “FATHER OF” is due to truncation of the value “FATHER OF FETUS” when the data was transcribed into the subset of the data used here.) We also see the (expected) relationship of a high proportion of sibling donors being a full match compared with parental donors.

A final, and important view of the data, requested by our users, is the Kaplan-Meier survivability curve [1]. These curves express the probability of survival given potentially incomplete data (such as patients for whom the current survival status is unknown). Kaplan-Meier handles the problem of incomplete data through a procedure called “censoring,” whereby patients who are still surviving (whose eventual survival length is not known), and patients who are lost to study (again, whose true survival time is not known) are *censored* at the time of their “disappearance” from the study, and survival is based on the the number of deaths divided by the number of remaining cases at that time span. It was important for our users to be able to compare K-M curves for different classes of patients in an easy way. We will discuss the use of the K-M curves further in the next section.

We note that the category table is also useful for quickly separating classes of patients for further analysis. For example, one can create a category table on the DIAGNOSIS variable, and then color two diagnoses differently and compare the K-M survivor curves. Further refinements are possible using *colorblending*, which allows the superposition of colors to see which cases exhibited two or more characteristics. For example, one could color all DIAGNOSIS = BREAST CANCER patients red, and then color all DONOR_MATCH = MIS-MATCH cases green. Resulting red cases would represent non-mis-matched breast cancer cases, green cases would represent mis-matched, non-breast cancer cases, and resulting yellow cases (since yellow is green plus red) would represent cases which were *both* breast cancer *and* mis-matched. These colors would appear as separate curves in the Kaplan-Meier presentation as well.

We also provide the facility for a two-dimensional scatter plot of any two variables. We will show the use of the scatter plot in an example in the following section.

4 Results

An example scenario will show how the PRIMA application might be used in practice, and how the natural and interactive nature of the application can quickly yield answers to questions about the clinical records. We also found that the interactive visualization made possible by PRIMA led to important findings about some of the limitations of the collected data and caveats for its interpretation.

Suppose one is interested in how the survival statistics vary for patients who received transplants recently compared with those who received transplants a longer time ago. A user could bring up a single histogram on the variable “TransplantTime” which is a numerical measure of years since January 1, 2000 (both positive and negative), and is created automatically on startup from a standard date such as “20-APR-1986.” One could then color recent transplants (say in the last 7 years) green, and less recent transplants red, simply by dragging the mouse in the histogram window. One could then choose to bring up a K-M survival curve, with the result shown in Figure 3 (left side). Perhaps somewhat surprisingly, we see markedly lower survival for the recent (green) transplant patients. Further investigation shows that in the Aggre-

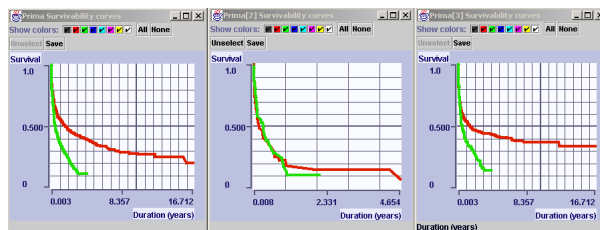


Figure 3: (left) Kaplan-Meier survivability for patients receiving transplants in the last five years (green), and prior to the last five years (red). (center) The same curves for the subset of patients who received a mis-matched transplant. (right) The same curves for the subset of patients who received a fully matched transplant.

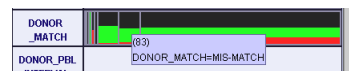


Figure 4: (A portion of the) PRIMA Aggregate window, showing the differences in proportions for early (red) and later (green) transplant patients, in particular for the DONOR_MATCH variable. The user has let the mouse dwell in the second block from the right, resulting in the pop-up window describing the contents of the block. We note that there are significantly higher percentage of recent transplants in the MIS-MATCH category than in the FULL-MATCH category to its right.

gate plot (Figure 4), a larger proportion of mismatched donors were recent, leading to a hypothesis that perhaps recently, more difficult (mis-matched) transplants are being undertaken. We investigate this further by coloring *only* mis-matched transplants, and *reinvoking* on only the colored cases (an option available from the initial window). The result is a new version of PRIMA considering only the mismatched transplants. The experiment with coloring early and late transplants differently is repeated on this subset of the data, with the result shown in Figure 3 (center). Here we see that the difference between survival probability has decreased. However, if we repeat the experiment, looking at only the subset of patients with a fully-matched transplant, we still see a significant difference in survival between the recent and early transplants.

Further investigation can easily be performed, looking at various subsets of patients; for example, those with only a certain disease, and the result is consistently found that recent transplants have lower survival probability than early transplants. This was a surprising result, both to ourselves and our users. However the availability of a variety of presentations through which to analyze the data meant that the reasons behind this result could be probed. For example we created a two-dimensional scatter plot of a variable we call “duration” vs. the time of transplant. The variable “duration” is simply the length of time between either transplant and death, or transplant and last-followup-visit, as appropriate. This scatterplot is shown in Figure 5. Here, surviving patients have been colored green and deceased patients have been colored red. We notice two linear features in the green patients parallel to the diagonal. These can only be due to large numbers of patients who had their last followup visit *around the same date*. A histogram on the follow-up visit date immediately shows this, as shown in Figure 6. We see two prominent peaks. Further investigation finds that these dates correspond to the fall of 1990 and late 1996, when much larger numbers of patients than typical had their *last* follow-up visit.

When using Kaplan-Meier curves to infer survivability, it is very important to ensure that the data do not violate the Kaplan-Meier

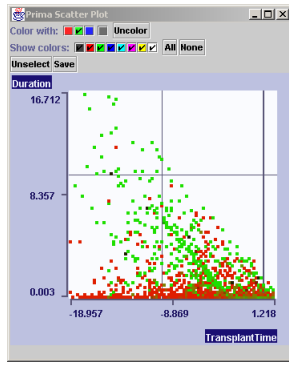


Figure 5: Two-dimensional scatterplot of duration vs. transplant date, where duration is time from transplant to death, or transplant to last followup visit, as appropriate. Living patients are colored green, and deceased patients are colored red. Note the two linear features in green parallel to the diagonal. These correspond to large numbers of patients with approximately the *same* last followup visit date.

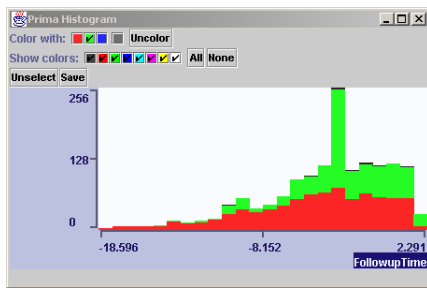


Figure 6: A histogram of followup visit dates, showing the two peaks corresponding to the two linear features in Figure 5. Further, simply accomplished investigation using coloring and the category table show that these times correspond to the fall of 1990 and late 1996. Green represents surviving patients, red deceased patients.

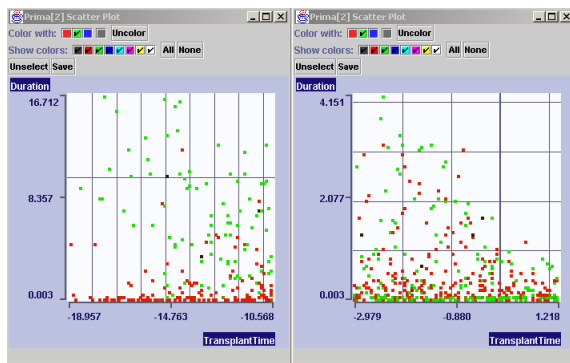


Figure 7: Two-dimensional scatterplots of duration vs. transplant date for *early transplants* (left), and *recent transplants* (right). Living patients are colored green, and deceased patients are colored red. Note for recent transplants the large number of surviving patients who are lost to followup essentially immediately after the transplant (the preponderance of green points near duration=0). This strongly implies that Kaplan-Meier curves are suspect due to the heavy censoring of patients, and lack of knowledge of their true survival time.

assumptions. For example, it is required that *implicit factors* do not confound the results. For example, if other factors, such as diet, which are not measured, affect survivability, then the subgroups we consider are not from a single population. Another important assumption is that the pattern of censoring is independent of survival time. For example, if patients who are more ill tend to withdraw from the study, survival estimates will be too high. If patients who survive longer tend to drop out of the study, survival estimates will be too low. The scatterplot shown in Figure 5 clearly shows a pattern which violates the assumption that censoring is independent of survival time, since a large number of patients were censored on one particular date, and thus lost to further followup. Even more striking is the plot shown in Figure 7, which compares the duration vs. transplant time scatterplot for early transplants vs. recent transplants, again colored green for surviving patients and red for deceased. Note the large number of surviving patients (at least as far as is known) in the recent transplant group who are lost to followup essentially immediately following the transplant (the large number of green points near the bottom of the plot). This non-uniform censoring of the “successful” transplants will significantly reduce apparent survival in the Kaplan-Meier curves.

Following this investigation, we discussed the findings with the physicians, and learned of two reasons for these results. First, as we had surmised from the visual investigation, the center tends to be better at documenting death dates than followup visits of surviving patients, leading to a bias in Kaplan-Meier curves. Their belief is that the patients listed as surviving can be assumed to be surviving at the current date (and not just until the last followup visit date), so we plan to make this choice an option in PRIMA for K-M analysis. Thus the interactive visual analysis capabilities of PRIMA allowed the deficiencies of the available data to be recognized and corrected for more accurate analysis. A second, medically based, reason for the differences suggested by the physicians is that the Hadassah Hospital bone marrow transplant center was the first in Israel some 20 years ago, but now several other centers exist, and Hadassah now often receives the more serious and complex cases, and is the hospital which tends to do more experimentation with new treatment protocols.

5 Conclusions

PRIMA provides a fast, interactive interface to diverse medical records, and incorporates an important view of the data, Kaplan-Meier survivability curves. The ability to quickly construct multiple views of the data has also led to a better understanding of the limitations of this particular data set as far as interpreting survival information, thus pointing to the importance of high quality data collection, and the value of interactive visualization in probing trends and patterns in a complex data set.

References

- [1] E. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457–481, 1958.
- [2] C. Plaisant, R. Mushlin, A. Snyder, J. Li, D. Heller, and B. Shneiderman. Lifelines: Using visualization to enhance navigation and analysis of patient records. In *American Medical Informatic Association Annual Fall Symposium*, pages 76–80, 1998.
- [3] D. A. Rabenhorst. Interactive exploration of multidimensional data. *Proceedings of the SPIE Symposium on Electronic Imaging*, February 1994.