Paper GH-12

Development of a SAS® Macro for Automated Data Cleaning of Major Outcomes of Interest in Hematopoietic Cell Transplantation

Peigang Li, Min Chen and Zhiwei Wang, CIBMTR, Milwaukee, WI

ABSTRACT

Previously we have developed a set of SAS macros to run against clean outcomes data so that univariate summary statistics can be automatically generated for major outcomes of interest, such as relapse, treatment related mortality, progression/disease free survival and overall survival (Li, Zhu and Chen, MWSUG Paper 177-2010). Among the outcomes, the data cleaning of relapse is the most time-consuming due to (a) evolving definitions of relapse; (b) applicable monitoring methodologies for acute or chronic hematological diseases (National Cancer Institute Relapse Workshop November 2009); (c) multiple sources of relapse information from comprehensive report forms (CRFs), transplant essential data (TED) forms, and legacy forms; (d) insufficient reporting by the transplant centers.

We have designed and developed a SAS macro to automate and standardize the process of relapse cleaning for Acute Lymphoblastic Leukemia (ALL), Acute Myelogenous Leukemia (AML), Chronic Myelogenous Leukemia (CML), and Myelodysplastic Syndrome (MDS). We validated relapse status against clean outcomes data from early studies. Both sensitivity and specificity have achieved 99% from the most recent test, and a few misclassified non-relapse cases are likely due to hardcoding in early studies. The initial design required input data to include relapse-related key variables in addition to the patient unique identifiers. The final version only requires patient unique identifiers. The macro will greatly speed up CIBMTR studies from protocol development to creation of statistical analysis datasets.

KEYWORDS

Hematopoietic cell transplantation (HCT), outcomes research, observational database, randomized controlled trial (RCT), the Blood and Marrow Transplant Clinical Trials Network (BMT CTN), clinical research center (CRC), fluorescence in situ hybridization (FISH), donor leukocyte infusion (DLI), multiple myeloma (MM)

INTRODUCTION

THE CENTER FOR INTERNATIONAL BLOOD AND MARROW TRANSPLANT RESEARCH (CIBMTR)

The objectives of CIBMTR are to develop and maintain a comprehensive database of clinical information on recipients of hematopoietic cell and bone marrow transplants, to conduct clinical studies addressing important issues in transplantation and cancer treatment, and to serve as an information resource for issues related to hematopoietic cell and bone marrow transplantation. Hematopoietic cell transplantation has been used to treat patients diagnosed with a variety of diseases including leukemia, severe aplastic anemia, Hodgkin's disease, non-Hodgkin's lymphoma, multiple myeloma, a number of blood disorders, and some solid tumor cancers. HCT studies are usually proposed by physicians with well defined objectives, scientific justification, patient eligibility criteria, study design and outcomes. Typical outcomes include treatment-related mortality (TRM), relapse (REL), disease free survival (DFS), and overall survival (OS).

Relapse has become a major cause of treatment failures in HCT over the years. Long-term outcomes of interest can be closely followed up in a observational database like the CIBMTR registry, which has designed a set of forms including comprehensive report forms (CRFs), transplant essential data (TED) forms, disease-specific forms, and follow-up forms of day 100, 6 months, and yearly, in both paper format and the newly implemented electronic FormsNet. A lot of efforts have been made to ensure clinical data were properly reported and collected at the data entry stage and during clinical outcomes studies at the CIBMTR. Since 2004, CIBMTR has affiliated with the National Marrow Donor Program (NMDP) dedicated to creating an opportunity for all patients to receive the bone marrow or umbilical cord blood transplant. NMDP has rich information of unrelated donors such as the human leukocyte antigen (HLA) typing that is used to match patients and donors for bone marrow or cord blood transplants and is a good prognostic factor in predicting survivals of transplantation patients.

Major efforts have been made to reconcile/harmonize report forms between CIBMTR and NDMP. At the same time, the expanded outcomes research via such a partnership aims at improving clinical services and helping provide better service to the transplantation community. CIBMTR shares its database and research results with transplant physicians and the public.

CIBMTR data facilitates retrospective studies or even prospective phase 4 clinical trials with cohort design while linking to national registry to conduct cost utilization analysis (e.g. Nationwide Inpatient Sample - NIS). Long-term outcome and late effects in transplantation patients have been initiated in CIBMTR studies, RCTs or ancillary clinical trials under parent trials (BMTCTN; Friedrichs, Tichelli, Bacigalupo et al. 2010).

RESPONSE AND RELAPSE DEFINITIONS

Relapse event is defined as a clinical relapse/disease recurrence or persistent disease post-transplant for patients not in complete remission at transplant. In November 2009 National Cancer Institute hosted a Relapse Workshop on The Biology, Prevention, and Treatment of Relapse. Applications of monitoring technologies are well summarized for both acute (ALL/AML/MDS) and chronic diseases (CML/MM) (Kroger, Bacher, Bader et al. 2010; Kroger, Bacher, Bader et al. 2010; Pavletic, Kumar, Mohty et al. 2010). Sensitive technologies have been used to detect relapse, including molecular, cytogenetic/FISH, and hematological methods. Molecular method has the highest diagnosis sensitivity, but not all diseases have molecular target for disease monitoring. Relapse of CML is easily detectable only by using molecular method by indication of minimum residual disease, and other diseases could use the molecular diagnosis once appropriate biomarkers deem clinical useful. In the new CIBMTR comprehensive report forms, molecular methods are included for most hematological diseases.

The outcomes of transplant patients who relapsed remain poor. Analysis of large cohorts via observational databases or registries at CIBMTR is essential to advancing the outcomes research (Pavletic, Kumar, Mohty et al. 2010). Common factors affecting risk of relapse include disease status, patient and donor characteristics (especially HLA matching), conditioning regimen, transplant characteristics (e.g., chimerism), and comorbidities. Clinical therapies for relapse include second transplant and DLI.

STANDARD ANALYSIS PROCESS

The initial analysis consists of a description of the patient population. A series of summary tables for patient characteristics and outcomes of both univariate and multivariate analyses were then generated as well as related figures for graphical visualization and data interpretation(Gratwohl, Baldomero, Aljurf et al. 2010). In contrast to prospective studies where numbers of patients to be studied is determined by the power desired, CIBMTR studies generally focus on a defined number of patients available in the database and the questions that can be answered, therefore, are determined by the numbers of patients available. The BMT CTN is actively conducting clinical trials with detailed study design and sample size calculation in the protocols.

COMPETING RISKS OF RELAPSE

TRM event is defined as death in continuous remission. TRM is considered as a competing risk for relapse, i.e., one event prevents the occurrence of the other one. The proper summary curve is the cumulative incidence function (CIF) which is a standard method of univariate analysis in most CIBMTR studies (Klein, Rizzo, Zhang et al. 2001; Li, Zhu and Chen 2010; Kumar, Zhang, Li et al. 2011). In addition to TRM, relapse could be a competing risk for incidence of acute or chronic graft-vs-host-disease (aGVHD & cGVHD) (Gooley, Leisenring, Crowley et al. 1999; Klein, Rizzo, Zhang et al. 2001). It's been demonstrated that cGVHD is associated with lower risks of relapse. Depending on the aims of a study, specific analysis methods may be applied in multivariate analysis. (a) Fine and Gray model or directly adjusted CIF in competing risks settings (Varadhan, Weiss, Segal et al. 2010; Zhang and Zhang 2010; Zhang, Zhang and Fine 2011); (b) pseudo-value approach to assess the effect of covariates on the risk of cGVHD (Klein, Gerster, Andersen et al. 2008; Alsultan, Giller, Gao et al. 2011).

The data cleaning macro is aimed to: (a) automate and standardize the process of relapse cleaning; (b) accurately identify patients who relapsed after transplantation; (c) minimize the amount of missing outcomes data to be checked by the CRC.

METHODS

RULES OF RELAPSE CLASSIFICATION - CLINICAL STANDARDIZATION

Relapse is a frequent outcome in CIBMTR analyses. In clinical trials (e.g., BMT CTN), primary endpoints often include relapse, TRM, DFS, and OS. While analysis datasets in RCTs were frozen after certain point, outcomes in CIBMTR studies constantly get updated. Sometimes it is necessary to apply the most recent outcomes information.

Most relapses are known at the time of therapies such as second HCT and DLI. The relapse event can be determined by looking at the following fields in the database:

- Disease-specific forms: did the patient relapse after transplantation? If yes, provide date of relapse.

- Acute disease (ALL, AML, MDS)
    - Disease diagnosis status: hematological and cytogenetic methods
    - Extramedullary relapse counts as relapse
    - Reason for re-transplant = persistent malignancy or recurrent malignancy
- Chronic disease (CML only)
    - Disease diagnosis status: hematological, cytogenetic, and *molecular* methods
    - Treatment for relapse was given
    - Reason for re-transplant = persistent malignancy (does not apply for multiple myeloma)
    - Reason for re-transplant = recurrent malignancy
- Cause of death = primary disease (e.g., leukemia), rel=1
- Relapse information in the TED database

For relapse date post-HCT, clinical knowledge plays a major role in addition to the information reported on the forms.

- if a patient died within 28 days post-HCT, then it was regarded as “no relapse”, i.e., rel=0; trm=1;
- if a patient died of primary disease (rel=1) and relapse date was missing, then relapse date was 1 day before the last contact date
- for persistent disease, relapse date was 1 day after transplantation
- if there was a second transplantation and relapse date was missing, then and relapse date was 1 day before the second transplantation
- If the event indicator is missing but the date of relapse is valid, then the patient has relapsed.

If either event indicator or date of relapse is still missing, then those cases are sent to the clinical research center (CRC) for review and the results are merged back to form the final analysis dataset. The case could get very complex when second transplantation or DLI was involved. The current implementation of the SAS macro focuses on first transplantation only.

MACRO DESIGN

The macro design involves two parts. In the protocol development stage of a CIMBTR study, a set of patient-, disease-, transplantation-, and outcomes-related variables have been retrieved. The macro could run against such a dataset to generate the expected variables of outcomes of interest, fast and simple. However, key outcomes-related variables were sometimes not retrieved or have been renamed, and the macro would stop due to missing variables. The macro was designed to retrieve those key variables regardless of whether the input data include them or not. The only required variables in the input data are unique identifies: CIBMTR recipient ID (crid) and transplantation number (txnum). Specifically, the macro first retrieves key outcomes-related variables from the latest research database (CRF) and replaces (overwrites) whatever included in the input data. This is preferred because new CRF data is supposed to be more accurate and updated. Then, the macro gets information from the registration database (TED) to provide extra relapse information or cover the missing values in the newly retrieved key variables. Note that research database is a subset of the registration database but with much more information for outcomes research.

Ideally, the input dataset should not include variables irrelevant to a specific study. Often times, all variables in the database are retrieved to avoid missing values being generated in computing new variables from existing ones. This would become an undue overhead. For this reason, the macro was designed to generate/reuse temporary datasets (both research and registration) when the macro is run the first time and subsequent runs will reuse them. New temporary datasets will be generated when they were physically deleted or new filenames were given. The outcomes variables (rel, trm, and dfs) are appended to input dataset where lower case differentiates those in the test datasets.

```
%macro outcomes_by_diseases(
    libnamedata,  /* Path of input SAS dataset */
    indata,       /* Input SAS dataset */
    txnum,        /* disease type */
    distype,      /* disease type */
    yeartxstart,  /* start of year tx */
    yeartxend,    /* end of year tx */
    crfdata,      /* Input CRF dataset */
    crfdatatmp,   /* Existing CRF subset, without accessing the retrieval */
    teddata,      /* Input PRE-TED dataset */
    teddatatmp,   /* Input PRE-TED subset, without accessing the retrieval */
    outdata,      /* Output outcomes dataset */ );
```

Besides the input data and path, the macro takes the general selection criteria as inputs to reduce the overhead.

Code samples: ALL/AML/MDS – with legacy forms

```
if disease=10 and amstathi in (2:4 14 15 33) then rel=1;
else if disease=20 and alstathi in (2:4 14 15 33) then rel=1;
else if disease=50 and mdstathi in (2:4 14 15 43) then rel=1;
else if persishi=1 then rel=1;
else if disease=10 and amstathi in (1) then rel=0;
else if disease=20 and alstathi in (1) then rel=0;
else if disease=50 and mdstathi in (1) then rel=0;
```

Code samples: ALL/AML/MDS – with registration (TED) and new CRF forms

```
if amstathi in (2:4 14 15 33)    /* old research form */
or disrelhi=1 or cldzashi=1 or fsdzashi=1 or cydzashi=1 /* new CRF form */
or (relprghi=1 and hemrelhi=1)   /* registration clinical */
or (relprghi=1 and cytrelhi=1)   /* registration cytogenetic */
or (dthprim=70 or death2_crf=70 or death3_crf=70 or death4_crf=70
or death5_crf =70 or death6_crf=70)         /* cause of death */
```

VALIDATION

We have tested the macro with a number of clean datasets from prior studies where relapse information of a number of cases was hard coded. We used sensitivity and specificity criteria to assess the goodness of the outcomes cleaning macro. The sample sizes vary from 2,000 to 16,000. The rationale is that once a patient has relapsed, the relapse status stays unchanged (rel=1). We are not overly concerned with a non-relapsed patient being classified as relapsed (false positive = *new relapse*) because new follow-up information has changed the patient disease status. We are more concerned with a relapsed patient being classified as non-relapsed (*false negatives*). In other words, we are looking to minimize *false negatives* and maximize *sensitivity*. The test data will include crid and txnum, as well as clean outcomes (REL and TRM).

RESULTS

HOW TO RUN THE MACRO

```
%include 'outcomes_by_diseases_mcr.sas';   *replace key vars;

%let indata=%str(inputdata);
%let crfdatatmp=%str(crf_tmp);
%let teddatatmp=%str(ted_tmp);
%let crfdata=%str(crf_Jan2011);
%let teddata=%str(ted_Jan2011);
%let txnum=1;
%let distype=%str(10, 20, 40, 50);
%let yeartxstart=1995;
%let yeartxend=2007;
%let outcomesdata=%str(outcomesdata);

%outcomes_by_diseases(%str(sasdata), &indata, &txnum, &distype, &yeartxstart,
&yeartxend, &crfdata, &crfdatatmp, &teddata, &teddatatmp, &outcomesdata);
```

Note: Jan 2011 was the most recent data release date at the time of macro testing. Generalized selection criteria can be used here, such as transplantation year from 1985 to 2010.

RUNTIME

We compare the runtime with and without the temporary datasets, and the runtime was 15 seconds vs. 3 minutes for a large dataset with 700 variables.

OUTCOMES TO BE CHECKED BY THE CRC

After running the macro, there are still about 3% cases with missing relapse event or date. The CRC can request additional information from the transplant centers and provide updated outcomes information to be incorporated into the final analysis dataset.

SENSITIVITY AND SPECIFICITY

1. Run the macro against new datasets and compare how many overlap with the gold standard that is a clean outcomes dataset (n=5343).

(a) input data #1: n=2891; overlap=633

	Test/Validation	REL=1	REL=0	Subtotal	
TP	rel=1	113	81	194	FP
FN	rel=0	0	439	439	TN
	Subtotal	113	520	633	
		REL=NA		**0**	Standard
	GVHD STUDY	2891		**633**	5343
			Total	**Overlap%**	21.90

Specificity	84.42 %	TN/(TN+FP)
Sensivity	100 %	TP/(TP+FN)

(b) input data #2: n=16037; overlap=2870

	Test/Validation	REL=1	REL=0	Subtotal	
TP	rel=1	457	122	579	FP
FN	rel=0	2	2289	2291	TN
	Subtotal	459	2411	2870	
		REL=NA		**0**	

Specificity	94.94 %	TN/(TN+FP)
Sensivity	99.56 %	TP/(TP+FN)

2. Run the macro against clean datasets and compare classification accuracy.

(a) Clean data #1 (n=2582):

	Test/Validation	REL=1	REL=0	Subtotal	
TP	rel=1	663	25	688	FP
FN	rel=0	1	1893	1894	TN
	Subtotal	664	1918	2582	

Specificity	98.70 %	TN/(TN+FP)
Sensivity	99.85 %	TP/(TP+FN)

(b) Clean data #2 (n=5020):

	Test/Validation	REL=1	REL=0	Subtotal	
TP	rel=1	1637	152	1789	FP
FN	rel=0	9	3176	3185	TN
	Subtotal	1646	3328	4974	
	missing	n=2	dup key	44	
	Specificity	95.43 %		TN/(TN+FP)	
	Sensivity	99.45 %		TP/(TP+FN)	

(c) Clean data #3 (n=1521):

	Test/Validation	REL=1	REL=0	Subtotal	
TP	rel=1	381	102	483	FP
FN	rel=0	11	1021	1032	TN
	Subtotal	392	1123	1515	
	missing	n=1	dup key	5	
	Specificity	90.92 %		TN/(TN+FP)	
	Sensivity	97.19 %		TP/(TP+FN)	

CONCLUSION

We have achieved the aim of automating the data cleaning process while preserving high accuracy of relapse classification. The final analysis dataset must account for the outcomes information returned by the CRC which we hope to standardize the coding and merging process. Future improvement may deal with complex issues such as relapse after second transplantation or DLI, or relapse prior to transplantation.

REFERENCES

Alsultan, A., R. H. Giller, D. Gao, et al. (2011). "GVHD after unrelated cord blood transplant in children: characteristics, severity, risk factors and influence on outcome." Bone Marrow Transplant **46**(5): 668-75.

BMTCTN "https://web.emmes.com/study/bmt2/."

Friedrichs, B., A. Tichelli, A. Bacigalupo, et al. (2010). "Long-term outcome and late effects in patients transplanted with mobilised blood or bone marrow: a randomised trial." Lancet Oncol **11**(4): 331-8.

Gooley, T. A., W. Leisenring, J. Crowley, et al. (1999). "Estimation of failure probabilities in the presence of competing risks: new representations of old estimators." Stat Med **18**(6): 695-706.

Gratwohl, A., H. Baldomero, M. Aljurf, et al. (2010). "Hematopoietic stem cell transplantation: a global perspective." JAMA **303**(16): 1617-24.

Klein, J. P., M. Gerster, P. K. Andersen, et al. (2008). "SAS and R functions to compute pseudo-values for censored data regression." Comput Methods Programs Biomed **89**(3): 289-300.

Klein, J. P., J. D. Rizzo, M. J. Zhang, et al. (2001). "Statistical methods for the analysis and presentation of the results of bone marrow transplants. Part I: unadjusted analysis." Bone Marrow Transplant **28**(10): 909-15.

Kroger, N., U. Bacher, P. Bader, et al. (2010). "NCI First International Workshop on the Biology, Prevention, and Treatment of Relapse after Allogeneic Hematopoietic Stem Cell Transplantation: report from the Committee on Disease-Specific Methods and Strategies for Monitoring Relapse following Allogeneic Stem Cell Transplantation. Part I: Methods, acute leukemias, and myelodysplastic syndromes." Biol Blood Marrow Transplant **16**(9): 1187-211.

Kroger, N., U. Bacher, P. Bader, et al. (2010). "NCI first international workshop on the biology, prevention, and treatment of relapse after allogeneic hematopoietic stem cell transplantation: report from the committee on disease-specific methods and strategies for monitoring relapse following allogeneic stem cell transplantation. part II: chronic leukemias, myeloproliferative neoplasms, and lymphoid malignancies." Biol Blood Marrow Transplant **16**(10): 1325-46.

Kumar, S., M. J. Zhang, P. Li, et al. (2011). "Trends in allogeneic stem cell transplantation for multiple myeloma: a Center for International Blood and Marrow Transplant Research (CIBMTR) analysis." Blood 2011 Jun 20. [Epub ahead of print].

Li, P., X. Zhu and M. Chen (2010). "Tips for Automating Univariate Outcomes Analysis in Hematopoietic Stem Cell Transplantation." MWSUG Paper 177-2010.

Pavletic, S. Z., S. Kumar, M. Mohty, et al. (2010). "NCI First International Workshop on the Biology, Prevention, and Treatment of Relapse after Allogeneic Hematopoietic Stem Cell Transplantation: report from the Committee on the Epidemiology and Natural History of Relapse following Allogeneic Cell Transplantation." Biol Blood Marrow Transplant **16**(7): 871-90.

Varadhan, R., C. O. Weiss, J. B. Segal, et al. (2010). "Evaluating health outcomes in the presence of competing risks: a review of statistical methods and clinical applications." Med Care **48**(6 Suppl): S96-105.

Zhang, X. and M. J. Zhang (2010). "SAS macros for estimation of direct adjusted cumulative incidence curves under proportional subdistribution hazards models." Comput Methods Programs Biomed **101**(1): 87-93.

Zhang, X., M. J. Zhang and J. Fine (2011). "A proportional hazards regression model for the subdistribution with right-censored and left-truncated competing risks data." Stat Med **30**(16): 1933-51.

ACKNOWLEDGMENTS

We thank the Center for International Blood and Marrow Transplant Research (CIBMTR) for providing us with the data sets and the Scientific Directors of the CIBMTR Working Committees for clinical advice. Thanks to our terrific Biostatistician colleagues for the discussions and support.

RECOMMENDED READING

1. Munker R, Lazarus HM and Atkinson K (eds). The BMT Data Book. Second Edition, 2009. Cambridge University Press, Cambridge, UK and New York, USA, ISBN: 978-0-521-71100-5.
2. Pasquini MC, Wang Z. Current use and outcome of hematopoietic stem cell transplantation: CIBMTR Summary Slides, 2010. Available at: **http://www.cibmtr.org/**

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Peigang Li, Biostatistician
Center for International Blood and Marrow Transplant Research (CIBMTR)
Froedtert and the Medical College of Wisconsin Clinical Cancer Center
9200 W. Wisconsin Avenue, Suite C5500
Milwaukee, WI 53226 USA
Telephone: 414-805-0700
Fax: 414-805-0714
E-mail: **peigang@mcw.edu**
Web: **http://www.cibmtr.org**