Ethical Issues in Getting it Right: Using Big Data to Determine Medical Treatment

John Crowley[1], Pingping Qu[1], Bart Barlogie[2]

[1]Cancer Research And Biostatistics, 1730 Minor Avenue, Suite 1900, Seattle, WA 98101
[2]Myeloma Institute for Research and Therapy, University of Arkansas for Medical Sciences, 4301 West Markham, Little Rock, AR 72205

Abstract

Medical tests have long been used to determine treatment, either through eligibility criteria for a clinical trial, or to decide on treatment once a patient is on a trial. What is new with Big Data is that there are many more opportunities for getting it wrong. I will use examples to demonstrate approaches to minimize such risks.

Important steps in getting it right include careful data management, adherence to principles of reproducible research, and validation, validation, validation. Data management methods and reproducible research will be discussed in the context of SWOG, one of the national cancer cooperative groups. Our work with gene expression profiles (gep) of the tumors of patients with multiple myeloma will be used to illustrate the principle of validation. Myeloma patients treated at the Myeloma Institute for Research and Therapy of the University of Arkansas for Medical Sciences are now treated with different protocols depending on a gep risk score developed on one set of patients at Arkansas, validated in another, and further validated using data from several institutions around the world.

Key Words: Big Data, medical decisions, gene expression profiling, data management, reproducible research, validation

1. Introduction

The increasing use of Big Data in medical decision making raises several important ethical issues. Data security and patient confidentiality are certainly among those issues, but of at least equal importance is the ethical imperative to make the right decision based on the data. There is nothing new in Big Data with this imperative, except that the number of ways to get it wrong is multiplied by several orders of magnitude. In order to maximize the chances of getting it right, several areas need careful attention; among these are data management, reproducible research, and validation. We illustrate these concepts through examples of our work with SWOG, one of 5 national cancer cooperative groups, and with the Myeloma Institute for Research and Therapy, a premier referral center for patients with multiple myeloma.

2. The importance of data management

2.1 SWOG

SWOG, formerly known as the Southwest Oncology Group, was organized in 1956 as a national multi-institution consortium to study new cancer therapies through controlled clinical trials. SWOG now has hundreds of participating treatment centers around the world, and thousands of associated physicians and clinical research associates (data managers). Approximately 5,000 patients enter SWOG clinical trials annually. The Statistics and Data Management Center for SWOG has been in Seattle Washington since 1984, and is co-located at Cancer Research And Biostatistics (CRAB) and the Fred Hutchinson Cancer Research Center. The data center and the data management functions for SWOG are at CRAB and constitute a major focus of activity. The computer systems in the data center are qualified according to guidelines set forth in the Code of Federal Regulations in 21 CFR Part 11 and certified according to standards in the Federal Information Security Management Act of 2002 (FISMA). Software is developed to be compliant with 21 CFR 11. Data have been transmitted via secure and encrypted web-based data collection systems since the early 2000s; the systems feature an extensive system of automatic edit checks along with manual review for consistency. Data are stored in a commercial relational data base management system (Oracle), which has in recent years included pointers to medical images (CT scans, MRIs). The data flow process is depicted in Figure 1. The SWOG data base is described in (1).

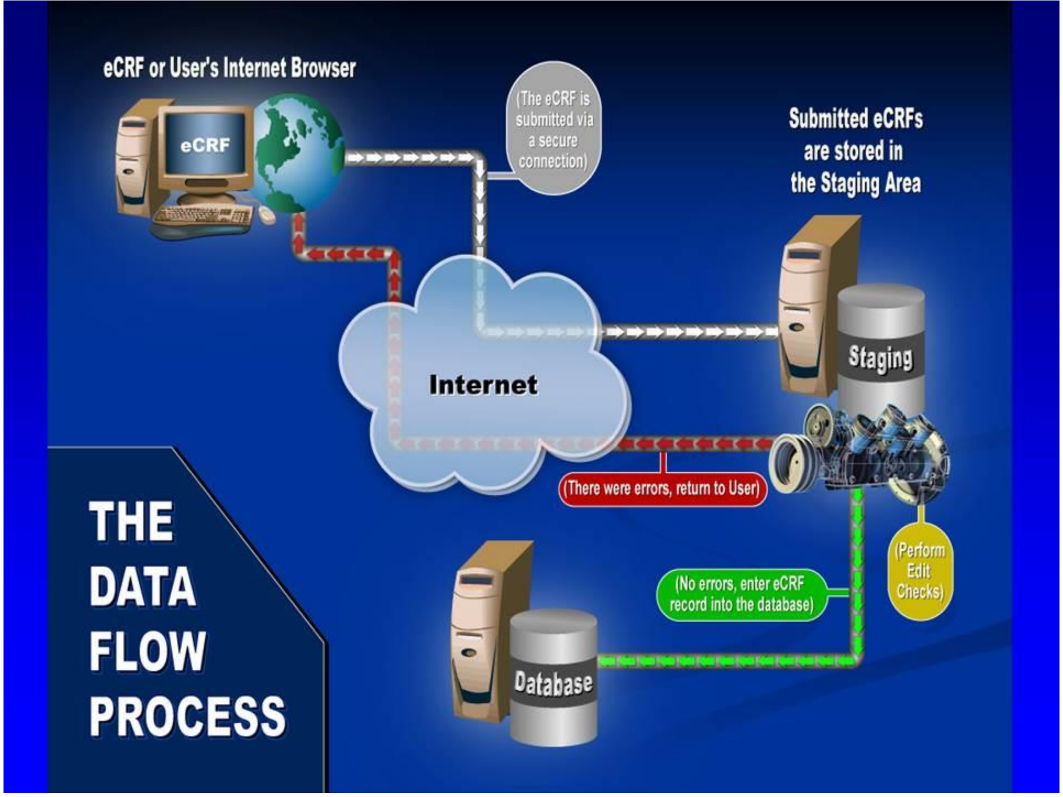

eCRF or User's Internet Browser
eCRF
(The eCRF is submitted via a secure connection)
Submitted eCRFs are stored in the Staging Area
Internet
Staging
(There were errors, return to User)
(Perform Edit Checks)
(No errors, enter eCRF record into the database)
Database
THE DATA FLOW PROCESS

Figure 1: Data flow process in SWOG. At the treating center data are entered through browser based electronic case report forms (eCRFs) via the internet to the data center in Seattle. At the data center a series of automated edit checks are applied and errors fixed before data are committed to the data base, where further automated and manual checks are performed.

2.2 MIRT

The Myeloma Institute for Research and Therapy at the University of Arkansas for Medical Sciences in Little Rock is one of the premier center for the treatment of myeloma in the world. Over 500 newly diagnosed myeloma patients are treated yearly, and extensive clinical and translational data are collected, including Big Data in the form of gene expression profiling, most recently using the Affymetrix U133 array with 54,000 probes (parts of genes) interrogated. Protocol required clinical data are entered into a research data based called the Multiple Myeloma Data Base (MMDB), which pulls research data from various parts of the medical record (Figure 2).

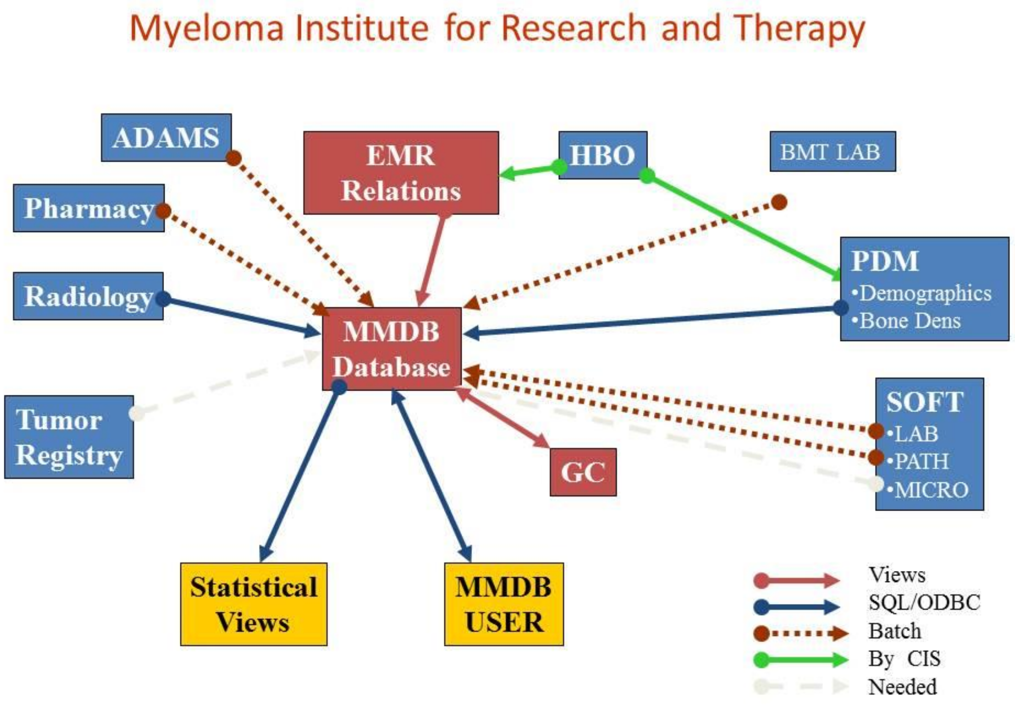


Figure 2: The Multiple Myeloma Data Base (MMDB) pulls research data from various medical record sources at the University of Arkansas for Medical Sciences, and also contains protocol driven data entered by data managers. Users are provided controlled access through views.

A limited subset of data is transmitted nightly using secured and encrypted technology to CRAB for statistical analysis. This data set does not contain what is referred to as Protected Health Information (PHI) and conforms to requirements for patient confidentiality contained in the Health Insurance Portability and Accountability Act (HIPAA). The data flow from MIRT to CRAB is outlined in Figure 3.

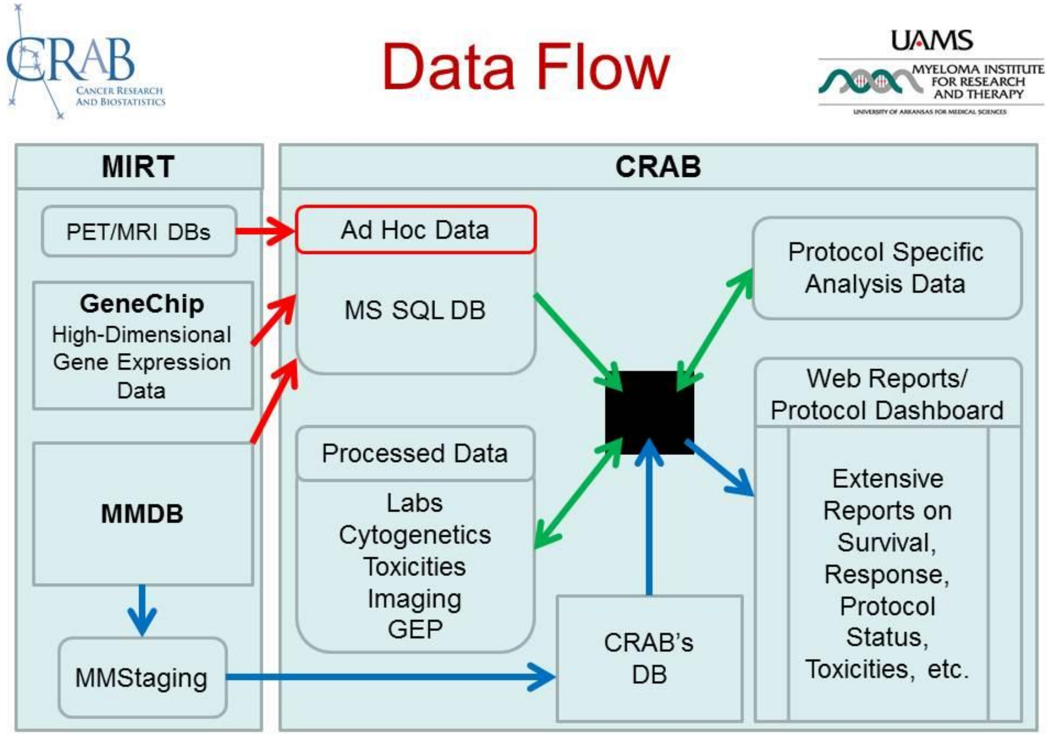
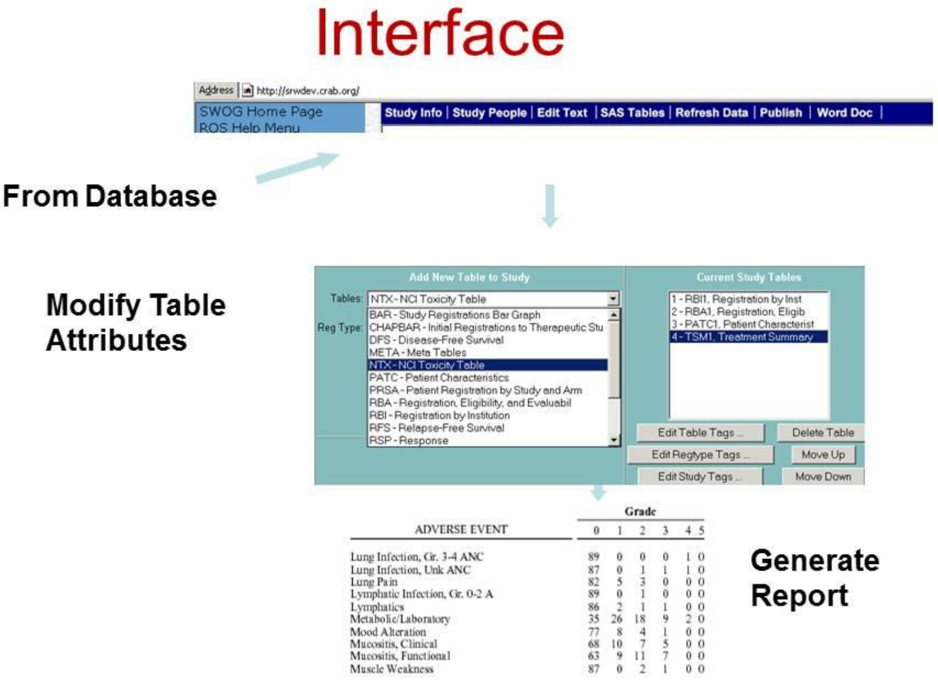

CRAB
Cancer Research and Biostatistics
Data Flow
UAMS
Myeloma Institute for Research and Therapy
University of Arkansas for Medical Sciences
MIRT
CRAB
PET/MRI DBs
GeneChip
High-Dimensional Gene Expression Data
MMDB
MMStaging
Ad Hoc Data
MS SQL DB
Processed Data
Labs
Cytogenetics
Toxicities
Imaging
GEP
CRAB's DB
Protocol Specific Analysis Data
Web Reports/ Protocol Dashboard
Extensive Reports on Survival, Response, Protocol Status, Toxicities, etc.

Figure 3: A subset of the data from the Multiple Myeloma Data Base (MMDB) is transferred nightly to CRAB, including gene expression profiling data on each patient.

3. An example of reproducible research

An important research principle is that the essential statistical analyses resulting from the work are able to be reproduced easily by others. Most discussion of reproducible research centers on the use of open source software, though this is certainly not a requirement for research to be reproducible. The SWOG Statistics and Data Management Center developed a platform for reproducible research in the 1980s, based on extracting data from the Oracle database into SAS files, and converting output into both web pages and a Word document. We called the platform the Statistician’s Report Worksheet or SRW (2).

At any one time SWOG has 75 or more clinical trials open for accrual, and dozens more in active follow-up prior to publication. We provide a standardized summary (excluding key outcome information) of each trial every 6 months in a volume called the Report of Studies, which forms the content for the twice-yearly meeting of the investigators of the Group. A key to SRW is that there is a consistency of data definitions where possible across all the trials (so that for example sex is coded and stored the same way in the database for all trials). This makes it possible to have a common interface to pick from a number of possible Tables and Figures for all trials, with the ability to modify each in a defined way (by choosing all randomized patients to include, or all eligible patients). The interface is illustrated in Figure 4, and output from a given set of choices is shown in Figure 5.


Address
http://srwdev.crab.org/
SWOG Home Page
ROS Help Menu
Study Info | Study People | Edit Text | SAS Tables | Refresh Data | Publish | Word Doc |
Add New Table to Study
Tables: NTX - NCI Toxicity Table
Reg Type:
BAR - Study Registrations Bar Graph
CHAPBAR - Initial Registrations to Therapeutic Stu
DFS - Disease-Free Survival
META - Meta Tables
NTX - NCI Toxicity Table
PATC - Patient Characteristics
PRSA - Patient Registration by Study and Arm
RBA - Registration, Eligibility, and Evaluabil
RBI - Registration by Institution
RFS - Relapse-Free Survival
RSP - Response
Current Study Tables
1 - RBI1, Registration by Inst
2 - RBA1, Registration, Eligib
3 - PATC1, Patient Characterist
4 - TSM1, Treatment Summary
Edit Table Tags ...
Delete Table
Edit Regtype Tags ...
Move Up
Edit Study Tags ...
Move Down

Interface

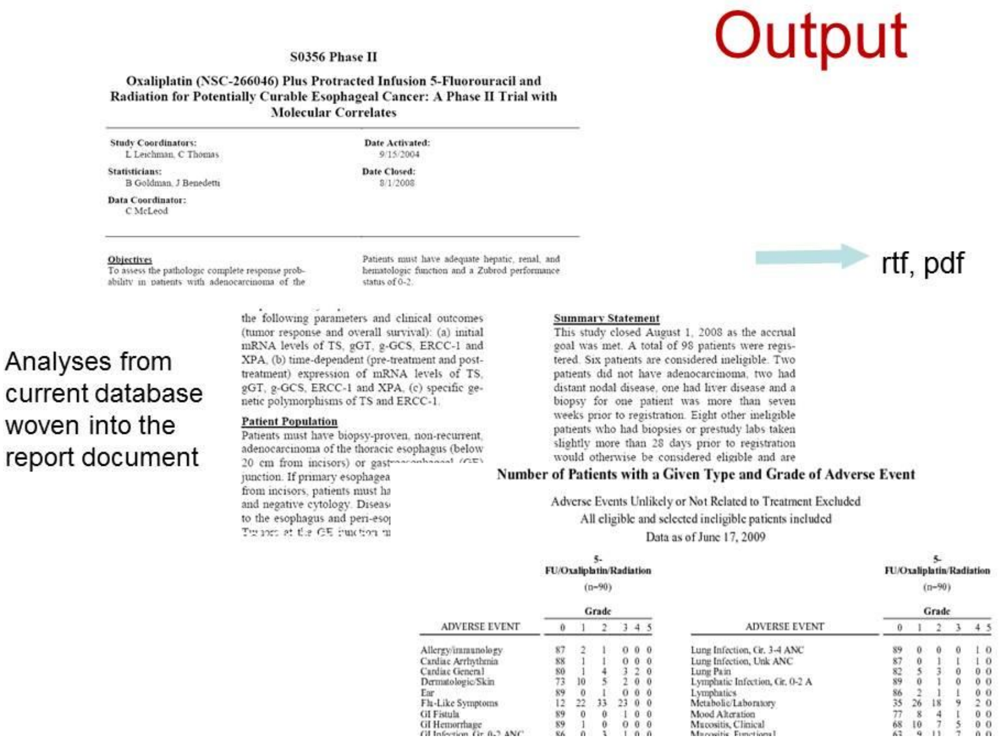
From Database

Modify Table Attributes

Generate Report

	Grade					
ADVERSE EVENT	0	1	2	3	4	5
Lung Infection, Gr. 3-4 ANC	89	0	0	0	1	0
Lung Infection, Unk ANC	87	0	1	1	1	0
Lung Pain	82	5	3	0	0	0
Lymphatic Infection, Gr. 0-2 A	89	0	1	0	0	0
Lymphatics	86	2	1	1	0	0
Metabolic/Laboratory	35	26	18	9	2	0
Mood Alteration	77	8	4	1	0	0
Mucositis, Clinical	68	10	7	5	0	0
Mucositis, Functional	63	9	11	7	0	0
Muscle Weakness	87	0	2	1	0	0

Figure 4: Interface for Statistician’s Report Worksheet, or SRW. Choices are made from among a standard set of Tables and Figures.



Output

rtf, pdf

S0356 Phase II

Oxaliplatin (NSC-266046) Plus Protracted Infusion 5-Fluorouracil and Radiation for Potentially Curable Esophageal Cancer: A Phase II Trial with Molecular Correlates

Study Coordinators:
L Leichman, C Thomas

Statisticians:
B Goldman, J Benedetti

Data Coordinator:
C McLeod

Date Activated:
9/15/2004

Date Closed:
8/1/2008

Objectives
To assess the pathologic complete response probability in patients with adenocarcinoma of the

Patients must have adequate hepatic, renal, and hematologic function and a Zubrod performance status of 0-2.

the following parameters and clinical outcomes (tumor response and overall survival): (a) initial mRNA levels of TS, gGT, g-GCS, ERCC-1 and XPA, (b) time-dependent (pre-treatment and post-treatment) expression of mRNA levels of TS, gGT, g-GCS, ERCC-1 and XPA, (c) specific genetic polymorphisms of TS and ERCC-1.

Patient Population
Patients must have biopsy-proven, non-recurrent, adenocarcinoma of the thoracic esophagus (below 20 cm from incisors) or gastroesophageal (GE) junction. If primary esophagea from incisors, patients must ha and negative cytology. Diseas to the esophagus and peri-eso Tumors at the GE junction m

Summary Statement
This study closed August 1, 2008 as the accrual goal was met. A total of 98 patients were registered. Six patients are considered ineligible. Two patients did not have adenocarcinoma, two had distant nodal disease, one had liver disease and a biopsy for one patient was more than seven weeks prior to registration. Eight other ineligible patients who had biopsies or prestudy labs taken slightly more than 28 days prior to registration would otherwise be considered eligible and are

Number of Patients with a Given Type and Grade of Adverse Event

Adverse Events Unlikely or Not Related to Treatment Excluded
All eligible and selected ineligible patients included
Data as of June 17, 2009

	5-FU/Oxaliplatin/Radiation (n=90)							5-FU/Oxaliplatin/Radiation (n=90)					
	Grade							Grade					
ADVERSE EVENT	0	1	2	3	4	5	ADVERSE EVENT	0	1	2	3	4	5
Allergy/immunology	87	2	1	0	0	0	Lung Infection, Gr. 3-4 ANC	89	0	0	0	1	0
Cardiac Arrhythmia	88	1	1	0	0	0	Lung Infection, Unk ANC	87	0	1	1	1	0
Cardiac General	80	1	4	3	2	0	Lung Pain	82	5	3	0	0	0
Dermatologic/Skin	73	10	5	2	0	0	Lymphatic Infection, Gr. 0-2 A	89	0	1	0	0	0
Ear	89	0	1	0	0	0	Lymphatics	86	2	1	1	0	0
Flu-Like Symptoms	12	22	33	23	0	0	Metabolic/Laboratory	35	26	18	9	2	0
GI Fistula	89	0	0	1	0	0	Mood Alteration	77	8	4	1	0	0
GI Hemorrhage	89	1	0	0	0	0	Mucositis, Clinical	68	10	7	5	0	0
GI Infection, Gr. 0-2 ANC	86	0	3	1	0	0	Mucositis, Functional	63	9	11	7	0	0

Analyses from current database woven into the report document

Figure 5: An example of output generated from SRW. The program creates a SAS data set and runs procedures to create Tables and Figures, which are then formatted into a document in Word, along with study specific descriptions created by the statistician.

The SAS data sets are archived along with the output, so that the results can be easily reproduced by others. The standardized reports can also be supplemented with more sophisticated analyses based on the same data set.

4. Validation, validation, validation

Newly diagnosed myeloma patient at MIRT are treated according to a philosophy called Total Therapy, aimed at cure. Many drugs are involved, but the backbone consists of tandem transplants, two cycles of high dose melphalan followed by rescue of the patient’s immune system by infusion of the patient’s own peripheral blood stem cells (autologous transplant). Al illustrative treatment schema is shown in Figure 6.

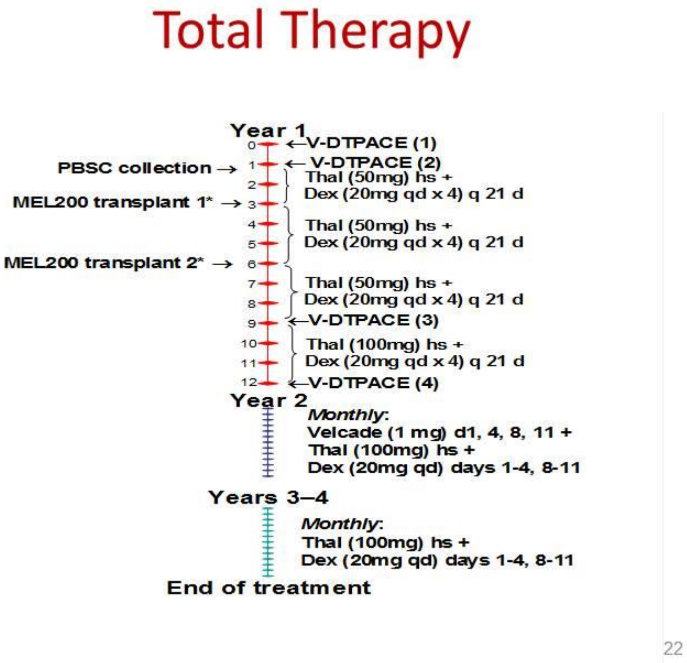


Figure 6: Treatment schema for Total Therapy 3 at MIRT. Induction chemotherapy is followed by two cycles of high dose melphalan and autologous stem cell rescue and then maintenance chemotherapy.

Total Therapy 1 (3) demonstrated the benefits of this approach to treatment. In Total Therapy 2 a randomization was done to assess whether the addition of thalidomide improved long term outcomes, with positive results (4). In Total Therapy, the new agent bortezomib (Velcade) was added to the thalidomide arm of Total Therapy 2, with promising results as reported in (5). Progression-free survival for these 3 trials, including a separation by arm for Total Therapy 2, is given in Figure 7, showing dramatic improvements in outcome with each successive addition to the backbone.

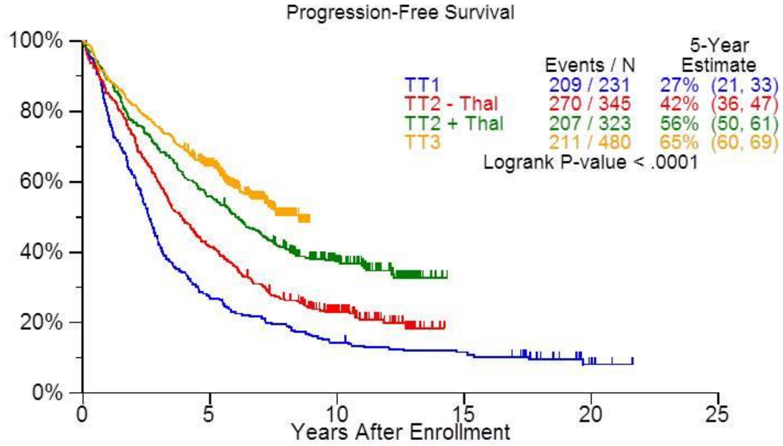
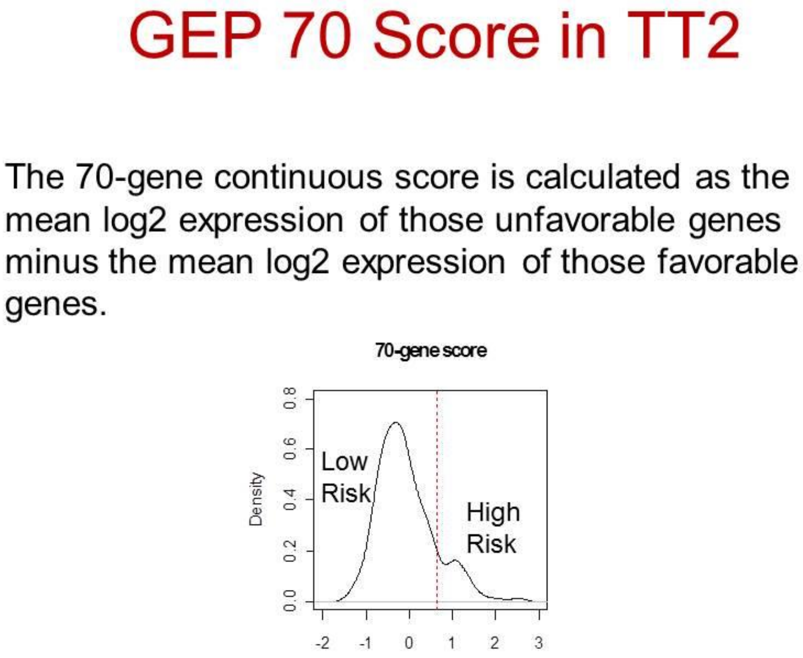


Figure 7: Progression-free survival for Total Therapy by trial and trial arm, demonstrating improved outcomes by adding new agents.

Beginning partway through Total Therapy 2, each patient has had a bone marrow biopsy so that the genetics of their tumor can be assessed, using technology developed by Affymetrix. The most recent version of the chip in use at MIRT is the Affymetric U133, which measures the abundance of 54,000 probes (gene portions) in the sample. These expression levels can then be studied in various ways to deepen biologic understanding of the disease. At CRAB we have developed a method for generating prognostic risk scores (predictive of outcome) based on these gene expression profiles. The philosophy features principles of simplicity and parsimony, to arrive at interpretable and generalizable models. There are exploratory steps and data reduction methods, then model building with attention to issues of multiple comparisons. We use cross-validation for choosing model size, but ultimately rely on validation, using other data sets from MIRT (other than that used for model building) and truly external data sets if at all possible. Algorithmically, we

- Rank probes by univariate associate with outcome, such as survival or progression-free survival
- Calculate a score as the mean log2 expression of positively correlated probes minus the mean log2 expression of negatively correlated probes
- Dichotomize the score by finding an optimal cutpoint based on for example the log rank test
- Use cross-validation to determine the number of probes to use in the score
- Develop the final score and cutpoint based on all the data
- Test the final model on a validation set.

Using this basic methodology, we developed a 70 gene score (gep 70) and optimal cutpoint for predicting survival using patients treated on Total Therapy 2, and validated the score and cutpoint on Total Therapy 3 (6). The distribution of the score in the training set is given in Figure 8, which also shows a hint of truly separate populations defined by the score.



GEP 70 Score in TT2

The 70-gene continuous score is calculated as the mean log2 expression of those unfavorable genes minus the mean log2 expression of those favorable genes.

Figure 8: Distribution of the 70 gene score for predicting survival, gep 70, in the training set, Total Therapy 2.

Survival results for the risk groups defined by the gep 70 cutpoint are given in Figure 9, for the training set Total Therapy 2 and the validation set Total Therapy 3. Further truly external validation has been carried out by others (7,8,9). As a result of this work patients at MIRT are now stratified into different treatment approaches (though still based on tandem transplantation), with gep 70 low risk patients being treated on Total Therapy 4, and high risk patients with a more aggressive strategy on Total Therapy 5 (Figure 10).

Training and Validation Set for GEP 70 Risk Score

OS by GEP 70 in TT2
Training Set

OS by GEP 70 in TT3
Validation Set

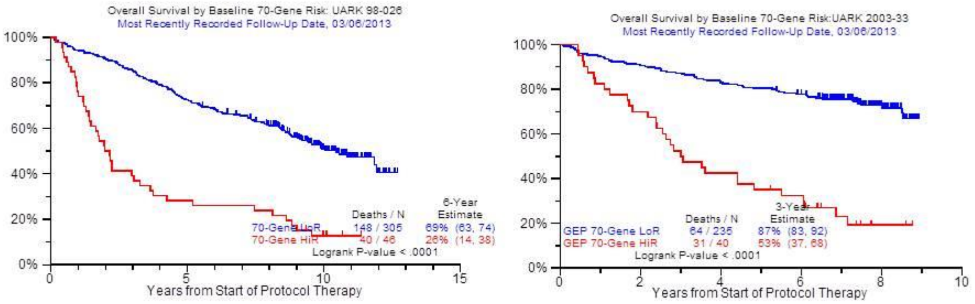
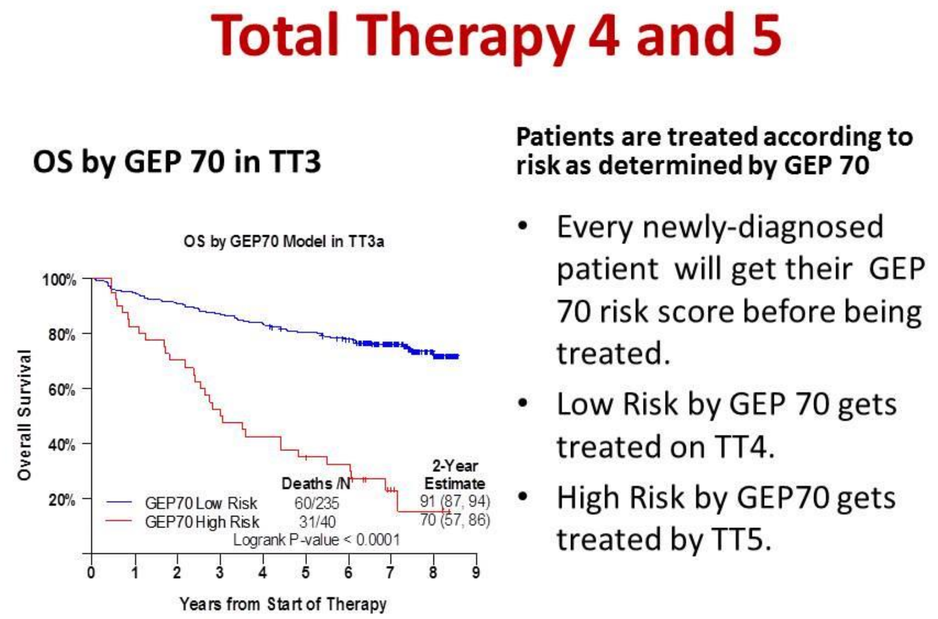


Figure 9: Overall survival by gep 70 risk score, for the training set Total Therapy 2 (left panel) and the validation set Total Therapy 3 (right panel).



Total Therapy 4 and 5

OS by GEP 70 in TT3

Patients are treated according to risk as determined by GEP 70

- Every newly-diagnosed patient will get their GEP 70 risk score before being treated.
- Low Risk by GEP 70 gets treated on TT4.
- High Risk by GEP70 gets treated by TT5.

Figure 10: The gep 70 risk stratification validated in Total Therapy 3 is now used to define different treatment strategies in Total Therapy 4, for gep 70 low risk patients, and Total Therapy 5, for high risk patients.

Summary

The use of Big Data has made medical decision making orders of magnitude more prone to error. With the ethical imperative to maximize the chances of making good medical decisions comes the statistical imperatives to use careful data and database management methods (an Excel spreadsheet is not a data management tool), to emphasize reproducible research, and to insist on validation of results.

Acknowledgements

Data managers and database managers in SWOG and at MIRT are responsible for the high quality processes summarized in this report.

References

1. Blumenstein B. A. The relational database model and multiple multicenter clinical trials. Controlled Clinical Trials 10:386-406, 1989.
2. Goodman P. J., Crowley J. and Benson C. Creation of a semiannual report for a multicenter co-operative clinical trials group. Statistics in Medicine 11:1367-1376, 1992.
3. Barlogie B., Jagannath S., Vesole D. *et al*. Superiority of tandem autologous transplant over standard therapy for previously untreated multiple myeloma. Blood 89:789-793, 1997.
4. Barlogie B., Tricot G., Anaissie E., *et al*. Thalidomide and hematopoietic-cell transplantation for multiple myeloma. New England Journal of Medicine 348:2609-2617, 2006.
5. Barlogie B., Anaissie E., van Rhee F., *et al*. Incorporating bortezomib into upfront treatment for multiple myeloma: early results of total therapy 3. British Journal of Haemotology 138:176-185, 2007.
6. Zhan F, Huang Y, Colla S, *et al*. The molecular classification of multiple myeloma. Blood 108:2020-2028, 2006.
7. Chng W. J., Kuehl W. M., Bergsagel P. L. *et al*. Translocation t(4;14) retains prognostic significance even in the setting of high-risk molecular signatures. Leukemia 22:459-462, 2008.
8. Mulligan g>, Mitsiades C., Bryant B. *et al*. Gene expression profiling and correlations with outcome in clinical trials of the proteasome inhibitor bortezomib. Blood 109:3177-3188, 2007.
9. Zhan F., Barlogie B., Mulligan G. *et al*. Highirisk myeloma: a gene expression based risk-stratification model for newly diagnosed multiple myeloma patients treated with hogh-dose therapy is predictive of outcome in relapsed disease treated with single-agent bortezomib or high-dose dexamethasone. Blood 111:968-969, 2008.