Resilience Needed, or Just Good Old Testing?
By Woody Epstein

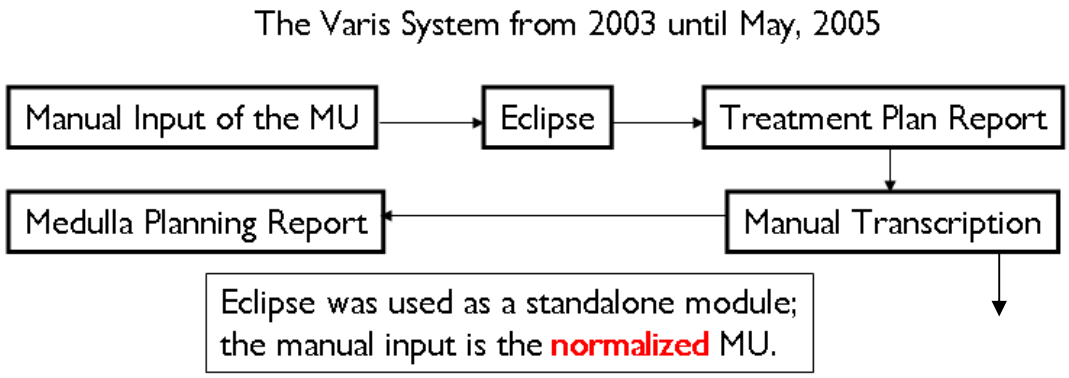
1. Introduction

The overexposure of Lisa Norris to radiation did not come about because the Beatson staff was un-resilient to a new threat to its standard, vigilant safe operations. The overexposure happened because the most elementary precautions of user testing of a computer system upgrade, before system acceptance and actual use, were never, I believe, carried out. In fact, it is difficult to find in the investigation report any reference to software testing, acceptance criteria, test cases, or user training on a live system.

The actual data on failure of the *Varis 7* computer system for the type of treatment under question shows that of the five times the system was used, there were four failures of planning personnel to use the systems correctly, with the first use of the system being contrary to the procedures then in place, and thus successful. It is my belief, that if "dry testing" had been performed, these failure events would not have happened during operations; certainly 100 "dry tests" would have shown at least one failure of the type committed, to wit, the forgettery of normalization, and the accident would most probably never have taken place.

2. What Went Wrong?

Prior to May, 2005, the *Varis* computer system used only the *Eclipse* software module for planning, schedule, and delivery of radiotherapy treatment.



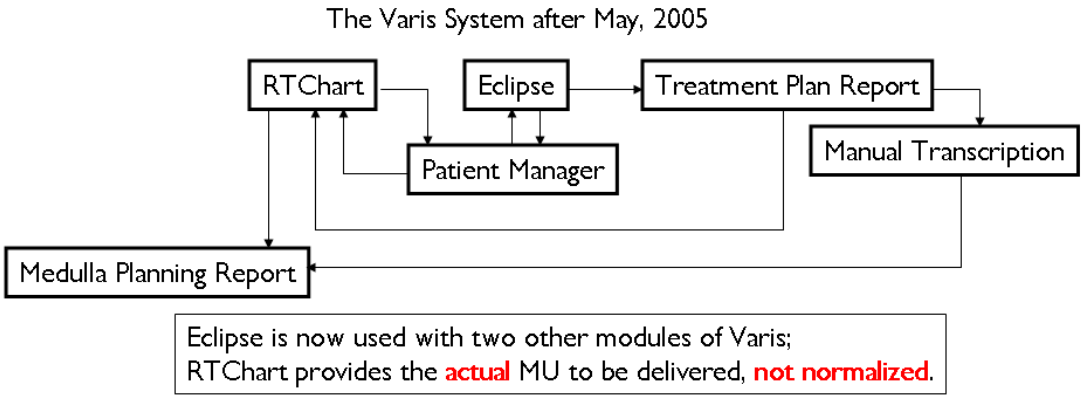
There is not much information in the incident report about how the *Varis* system was used prior to May, 2003. It is not clear if, or how, either the *RTChart* or the *Patient Manager* modules were used. It is clear that the values for treatment dose and number of fractions were input manually to *Eclipse*. But prior to May, 2005, the treatment dose was always entered in units of MU per 100 CentiGrays, and therefore the Treatment Plan Report printed by *Eclipse* was always in units of MU per 100 CentiGrays.

Were there ever any incidents of miscalculation of MUs prior to May, 2005? I would hope that if there had been any noticed occurrences, they would have been mentioned in the incident report; so let us assume that operations had been perfect with respect to calculation and checking of MU values. It is important to note that the best information would differentiate and include calculation errors which were corrected by planning checks before actual treatment. In these cases, there was a system error (incorrect calculation) which was caught by a backup system (plan verification). This is not a success of the system as a whole, but a failure of the frontline system and rectified by a backup system. In counting system failures, we must count this is a failure of the frontline.

Why was not the *Varis* system used as an integrated whole, why was only the *Eclipse* module used? The only information we have is related in section 4.8 of the report, "… a decision had been taken at the BOC to use the *Eclipse* planning system as a standalone module … for a number of operational and technical reasons" [page 6]. There is no further elucidation.

In May, 2005 things changed. A decision was made by the BOC to use other *Varis 7* modules:

> After the upgrade in May 2005 to *Varis 7*, a decision was taken to integrate the *Eclipse* module more fully with other *Varis* software modules. [section 4.8, page 6].



Was this decision really made, as stated, after the upgrade? If so, does that mean that any testing that was done to verify that *Varis 7* worked correctly was done before the decision to more fully integrate the modules? And after the decision, was a new round of testing done to make sure that the man-machine interactions would still produce a safe and reliable system? Did the BOC realize that a computer system incorporates both human and machine elements, and must be tested as such?

With the upgrade, it was possible to transfer information electronically between modules, including the treatment dose in terms of MUs. In this case, the *Patient Manager* module could import the treatment dose from *RTChart* directly to *Eclipse*, and then to the Treatment Plan Report for review, then to the Medulla Planning Form for treatment delivery.

And this is what happened. The MU from *RTChart* was transferred electronically to *Eclipse* and to the Treatment Planning Report.

However, it was the actual MU which was transferred, not the normalized MU, because now *RTChart* was transferring actual units, while before the upgrade, the manual transfer to *Eclipse* was always normalized units, causing Miss Norris to receive almost twice the intended dose of radiation. No one, in this case, noticed the error.

The frontline system failed (incorrect calculation). The backup systems failed (treatment verification). The *Varis* 7 system calculated correctly, given the inputs, and printed information out nicely. No one noticed the error.

Why was the decision made to integrate the modules? Again, from the report:

> … '*Manual transfer of data either from planning to treatment units or between treatment units is associated with a high risk of transcription error'* and recommends, therefore that '*The transfer of treatment data sets should be by local area IT network as far as is possible*'.[section 8.3, page 33]

Changing from manual transcription to electronic transcription will lower the risk of transcription error, but will this lower the risk to the patient? I do not believe we can make this inference without some type of evidence and theory to stand behind the claim; it is not axiomatic. In an electronic system, an error in input, which is propagated electronically to other "treatment units", will absolutely proliferate through all data; with manual systems, a human has many chances to look at the data, and unlike a machine, may even think about the data, or notice other information, such as instructions to normalize data. We tend to believe numbers on beautiful printouts or screen displays without questioning; this is not so when we work with our hands and minds.

Why was the procedure of entering normalized dose changed to entering actual dose? We only know that it was done to " … optimize the benefit of the change to electronic data transfer …"[section 8.2, page 33]. Certainly it shows a lack of communication between the users of the system and the software developers. Certainly the software could have been customized so that the users could continue entering normalized doses.

Off-the-shelf software rarely works in the same way as an individual organization. For software developers, if, on the average, the software works in the same way as the organization in 90% of the functions, it is considered very successful. But "means" are fictions; variation is the rule.

The *Varis 7* system was used as an integrated whole, even though some of its features were not used, some features were used in ways not intended, and instead of complimenting the way people worked, it caused a major change in the way they worked, using normalized doses, with disastrous results.

Probably the software developers never imagined that the system would not be used as a completely integrated whole. They probably did not imagine, for example, that information generated by *Eclipse* would <u>not</u> be transferred electronically to *RTChart*. In this case, because

the plan was a whole CNS plan, the BOC had made a decision to transfer data manually. When the system was used in this way, it was necessary to mark the Treat Plan Report status as "*Rejected*", so, I speculate, the data would not transferred electronically. I imagine that the easiest way to stop data transfer from *Eclipse* to *RTChart* was to mark the Treatment Plan as "Rejected".
Ironically, in this case, it was precisely because the plan was marked as "*Rejected*" that a radiologist and a senior planner discovered the errors being made [section 5.42, page 18].

Using a system in this way is sometimes called "walking hobbled": we use a feature of a system (being able to reject a treatment plan) in a way never intended (to stop the plan from automatically going to *RTChart*). In its worst incarnation, "walking hobbled" becomes "using a 'bug' as a 'feature'", when we "trick" a computer system to get the output we need. When these "bugs" are subsequently fixed, there is no telling how the new system, old data, and users' habits will interact to cause a system failure.

Computer systems are brittle, not resilient. They do not respond well in situations for which they were not programmed; they respond unpredictably when used in ways unforeseen (at least by the developers). And they cannot improvise when data is incorrect. A simple example here: if you have been typing "Y", for yes, or "N", for no, hundreds of times on an entry form, and suddenly you type "U", instead of "Y" (remember that on anglo keyboards the "u" key is adjacent to the "y" key), will the software recognize this as a simple typographical error? I think not.

I imagine a surly, middle-aged East European bureaucrat (and I have much experience in this arena) in place of *Eclipse*, and the joy he would feel by pointing out to me that the new rules specify normalization, and with a pen indicating where the changes must be made.

3. How Likely Was It?

In the absence of test data, we are left only with operational experience with which to estimate the likelihood of failure from non-normalization of units for MU.

During the period from May 2005 and February 2006 there were five whole CNS procedures planned.

For the first of these, in August 2005, the prescribed radiation dose was not included in the data input to *Eclipse*. This could mean two things: either the normalized radiation dose was input in place of the prescribed (actual) dose, or no radiation dose at all was entered. In either case, the procedures in effect for the upgraded system were not followed. Ironically, a failure of the system in place (not following procedures) contributed to a safe delivery of radiation treatment. There is no comment in the report as to why the prescribed dose was not entered. [section 5.22, page 13]

For the second plan, in November 2005, the prescribed (actual) dose was entered into *Eclipse*. But because the prescribed dose per fraction was 100 CentiGrays, the normalized values and the actual values are the same. So again, even though there was a failure of the system, accidentally, no adverse consequences resulted. [section 5.22, page 13]

The third plan, December 2005, was the treatment plan for Miss Norris. As we know, not only did the frontline system (data entry according to procedures) fail, but the backup systems (input verification) did also.

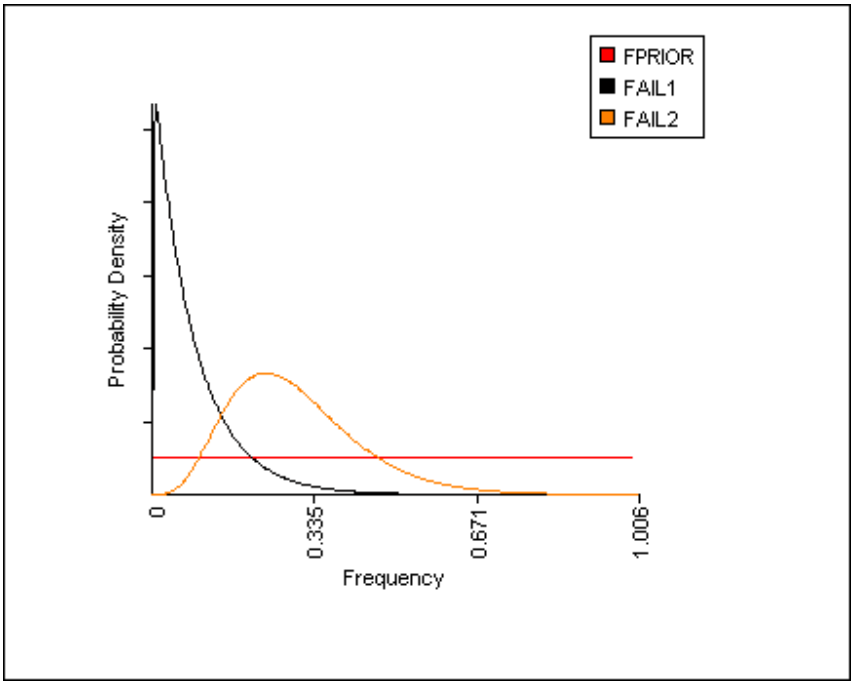
The fourth plan, a medulla plan, was done in January 2006. The normalization procedure was necessary in this case. A senior planner noticed that the unit of measurement on the Treatment Plan (175 CentiGrays) was different from the unit of measurement on the Medulla Planning Form, and made the appropriate re-calculation. It should be noted here that the senior planner in this case had never done a whole CNS before, and was unaware of the changes to procedures. His previous experience therefore did not blind him. He noticed things during transcription instead of proceeding in a familiar, but non-questioning, way. [section 5.4, page 19]

The last plan, February 2006, was fortuitously the plan which brought to light the previous errors. In this case, a question was raised by a radiographer as to why the Treatment Plan was marked as "Rejected". A senior planner looked at the Treatment Plan to remind himself as to why the status "Rejected" was used; and in this second, more focused, examination discovered the original errors. [section 5.42, page 18]
We can summarize the data in this table:

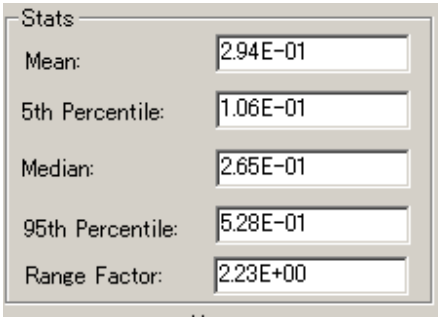
Event	Failure to Normalize	Failure of Verification
August 2005	1	?
November 2005	1	?
December 2005	1	1
January 2005	0	0
February 2006	1	0

As a Bayesian, I did a "back of the envelope" calculation. As my prior distribution, I used a flat prior, called FPRIOR, which indicates that I have no knowledge before operations as to the failure rate of the first *Varis* system used by the BOC. A flat prior indicates that every failure rate is equally possible.

I first updated FPRIOR with the data from the period 2003 to May 2005. The BOC estimates that 4-6 whole CNS treatments were performed each year. Since no incidents of failure to normalize were mentioned, we can conservatively say that there were 0 failures in 12 treatments. This distribution is named FAIL1. Then we must update FAIL1 with the data from after May 2005. Using the above table, I updated the distribution with 4 failures in 5 treatments, which results in the following posterior distribution, FAIL2:



and with distribution statistics for FAIL2 as follows:



Given the resultant posterior, we have better than a 25% chance that an incident like this will occur per whole CNS procedure, or about once a year. Quite unacceptable.

I find it difficult to believe that acceptance testing or a "dry test" was performed by the BOC. If they had, then their performance would have shown a similar type of failure rate, before the possibility of an accident could occur.

It can be said that during testing, many subsequent errors do not surface, because the testers take care while using a system and concentrate on doing things correctly. My response is that one must insure, as closely as possible, that the test conditions are similar to operational conditions, using a broad spectrum of testers, test situations, and with various time constraints.

Moreover, it has been said that with mature software systems, and seven versions of the *Varis* software indicates that it has been around, the Pareto Principal applies: 80% of the

new errors will actually be caused by updates to the system. Computer software is brittle, and slight changes in one part of the software can cause errors which were never imagined. Part of the Therac-25 tragedy, where 6 known accidents involved massive radiation overdoses, was caused by a simple update to the user interface allowing users to edit individual data fields, where before all data fields had to be re-entered if a data entry error occurred [Leveson, Turner 1993].

4. What Were the Consequences?

A person died.