

Evaluation of proposed countermeasures for long-term missions

Prepared by Alan H. Feiveson, Ph.D.

1.0 Introduction

A comprehensive program for evaluating, certifying or rejecting proposed countermeasures (CMs) for long-term space missions should include the following basic elements: 1) a consistent approach to defining standards of CM effectiveness, 2) implementation of the best possible experimental test program given logistic and cost constraints, 3) implementation of appropriate statistical analyses of test data with the objective of deciding whether or not the standards have been met, taking uncertainty into account

Because of the limited numbers of astronauts expected to participate in long-term missions in the near future, proposed CMs can be tested only as part of packages. We will not have the luxury of being able to separate out the effects of each CM via traditional control vs. treatment experimental designs. This presents a challenging problem of how to decide which packages to test; however a more realistic scenario is that nothing less than the current best guess at the most effective package will ever be used for any given long-term mission. There is also a possibility that CMs may interact with each other in the sense that one may be effective only in the presence of others; or conversely, that one may be ineffective because of the presence of some others. If interactions are substantial, it may be that particular CMs could not be appropriately certified without repeated trials using the same CM package.

In general, there are three main levels of data potentially available for evaluation of CMs. At the highest level is data collected from long-term missions. This is the only data, being actual field results, for which we can confidently assume there is no bias. At the next level is short-term mission results. Depending on the physiological system in question, these data may have a utility ranging from virtually worthless to essentially as informative as actual long-term results. Finally, there may be data available from studies which do not directly involve, but are assumed to be related to, the performance of astronauts in long-term space flight. In particular, this category includes ground studies with human subjects under presumed “analog” conditions (e.g. bedrest) and animal studies (space and ground). In order to achieve the most efficient decision process for certifying or rejecting CMs, we will have to develop procedures for effectively integrating information from all of these disparate sources.

2.0 Standards of effectiveness

The cornerstone of an objective CM certification process is to establish standards of effectiveness *before* any test data is gathered. Because it is unlikely that any package of CMs will protect everyone on all missions, the evaluation and certification process should be viewed in the light of risk-assessment; i.e. we attempt to identify a CM package which will minimize the risk of serious debilitation. Standards of effectiveness must therefore be probabilistic in nature, reflecting variability between missions and especially potential astronauts with regard to a) their ability to withstand the debilitating conditions of long-term space flight, b) their propensity for being helped by a CM and c) their consistency in applying a CM over long periods of time.

2.1 Astronaut population

Assume a loosely-defined population of current or potential astronauts, any of whom could be selected for participation in a long-term space mission. Taking (a) - (c) above into account, the effectiveness of a CM package is best evaluated in terms of the probability of success or failure of the package with respect to this population. Possible mission effects could be accounted for as concomitant information or could be regarded as uncontrollable random variation. In particular, we would like to know how likely it is that a given package would protect an unspecified astronaut for an upcoming mission, or conversely, how likely is it that the package would fail in certain areas. To do this, it is first necessary to define success/failure criteria for a particular astronaut on a given mission.

2.2 Test Measurements

For a CM package to be effective it must consistently protect all body systems thought to be affected by long-term space flight. For convenience, we shall refer to this collection of body systems, as the “protected system set” (PSS). The raw information used to evaluate the performance of PSS members, (and by inference, the CMs) would be obtained from preflight and postflight batteries of tests. These data would be processed and converted to a collection of diagnostic measurements $\mathbf{M} = \{M_1, M_2, \dots, M_N\}$. In an idealized setting, each CM would be designed to protect a particular body system, whose performance would be reflected by an associated group of one or more of the M_i . The M_i would thus provide information as to the extent of which individual CMs are working, how to modify them, or decide whether to replace them with an entirely different procedure. All body systems in the PSS should be “covered” in the sense that if each of the M_i fell within acceptable limits, we could assume the subject did not experience a medically significant debilitation in any relevant system.

In practice, at each postflight observation session for a subject, only a limited amount of testing is feasible, hence the number and types of raw measurements would be also limited. The test battery would have to be carefully designed so that “coverage” of the PSS (as defined above) is still maintained, but in so doing, information about any particular body system might be sacrificed. This creates increased uncertainty as to which CM to adjust or replace if results are not satisfactory. Despite this uncertainty however, a set of rules based on the values of \mathbf{M} , would have to be devised for deciding whether CMs performed successfully for that subject-mission.

2.3 Success criteria

An objective evaluation of a CM or CM package must rely on pre-set rules for determining a success rating from the values of \mathbf{M} for a given astronaut-mission. To facilitate comparison across CMs and aggregation over astronaut-missions, the rating should be a relatively simple scale. In its simplest form, the rating for a single CM would be binary: 1 = “success”; 0 = “failure”. Another possibility is a 3-level scale: 0 = “failure”, 1 = “inconclusive”, 2 = “success”. In the rest of this paper, it will be assumed that a 3-level scale is to be used; however, modifications for a more detailed scale are certainly possible.

Example 2.3: An exercise CM against bone loss in the femoral neck is to be rated as follows for a given astronaut-mission:

- “S” if there is a gain or if the relative loss of bone mineral density (BMD) in the femoral neck does not exceed 1%.
- “F” if at least one of the following occur:
 - a) postflight BMD is less than 0.86 g/cm^2 in the femoral neck.
 - b) relative BMD loss in the femoral neck exceeds 3%.
- “I” if neither “S” nor “F” above apply.

(Note: It is assumed that the potential astronaut population does not include anyone that has an initial BMD of less than 0.87 g/cm^2 . Thus, even with a 1% loss, the postflight BMD would still be above the threshold of 0.86 g/cm^2 .) In this example, the raw measurements are pre- and postflight BMD, and the diagnostic measurements are $M_1 = \text{postflight BMD}$ and $M_2 = \text{percent change in BMD}$. Stated more mathematically, if Y_0 and Y_1 are the respective pre- and postflight BMD measurements, then the above is equivalent to:

- $M_1 = Y_1$; $M_2 = 100(Y_1 - Y_0)/Y_0$
- “S” if $M_2 \geq -1.0$ and $M_1 \geq 0.86$
- “F” if $M_1 < 0.86$ or $M_2 < -3.0$

2.4 Overall effectiveness

Once rules for determining whether a CM is “a success” (S), “a failure” (F) or “inconclusive” (I) for a single astronaut-mission, simple criteria for overall effectiveness can be established in terms of the probabilities $P(S)$, $P(F)$ of S and F respectively, under a scenario in which astronauts are selected at random from a large pool, so that the probability of the same astronaut flying more than one mission is negligible. In general, a CM would be considered “effective” if $P(S) > p_S$ and $P(F) < p_F$, where p_S and p_F are fixed thresholds. Recognizing that CMs are unlikely to “succeed” for everyone, the lower limit for $P(S)$ should be somewhat liberal, perhaps allowing p_S to be as low as 0.80 or so. Conversely, we would not want to experience the utter failure of a CM for anyone, but out of practical necessity, we might settle for something like $p_F = 0.025$. It should be emphasized that these standards apply only for long-term missions, even though data for estimating $P(S)$ and $P(F)$ may be gathered in part from short-term missions or analog ground studies. The problem of how to adjust for non-field data is substantial (see Section 3).

3.0 Statistical aspects of CM assessment

3.1 Acceptance/rejection strategy

Since a CM or CM package can only be tested a finite number of times, there can never be 100% assurance that it satisfies the standard of effectiveness of Section 2.4. In general, the less information there is available about a CM’s performance, the more uncertain we are about whether it meets (or fails to meet) specifications. We therefore apply statistical criteria for accepting a CM, taking this uncertainty into account. To do so, we propose to use the method of “repeated confidence intervals” which is used in some applications of quality control or sequential clinical trials (for example, see Whitehead, (1992)). In this method, acceptance/rejection criteria are based on confidence limits for performance parameters, which in our application, are $P(S)$ and $P(F)$.

After each mission, confidence limits (PS1, PS2) for P(S) and (PF1, PF2) for P(F) would be calculated using the current and relevant historical data. A CM would be rejected or changed if PS2, the upper confidence limit for P(S) is less than p_S , *or* if PF1, the lower confidence limit for P(F) exceeds p_F . Conversely, the CM would be “accepted” (i.e. certified) if PS1, the lower confidence limit for P(S) exceeds p_S , *and* if PF2, the upper confidence limit for P(F) is less than p_F . If neither of the above occurs, no certification decision is made and testing continues.

The values of p_S , and p_F as well as the levels used to construct the upper and lower confidence bounds directly affect the probability of reaching a decision (either positive or negative) in a given number of “trials” (i.e. astronaut-missions) and of three types of errors that we may make:

Type 1: Rejecting a CM when in fact $P(S) > p_S$ and $P(F) < p_F$.

Type 2: Certifying a CM when in fact $P(S) < p_S$.

Type 3: Certifying a CM when in fact $P(F) > p_F$.

In general, setting a confidence level high reduces the probability of the associated type of error, but makes it less likely that a decision will be reached by a given number of trials. In deciding which confidence levels to use, care must be taken to balance the risk of the three error types, taking the specified values of p_S and p_F into account and also how many trials we are likely to be able to support. There is no requirement of symmetry however. Different confidence levels may be used for each of the separate bounds PS1, PS2, PF1 and PF2. As a general rule, confidence levels for rejection criteria should be lower than those for certification. This prevents an unnecessary large number of trials with ineffective CMs. Conversely, certification confidence levels should be high. We do not wish to claim a CM is effective without a strong assurance that such is the case. In practice, some CMs may never attain a “certification” level of performance within the time scope of this project; however they can continue to be used (and tested) as long as they do not qualify for rejection.

3.2 Parametric vs. non-parametric estimation

In order to have a reasonable chance at arriving at a decision (rejection or certification) after a moderate number of long-term missions, it is imperative that the most efficient estimates possible be made of P(S) and P(F). In particular, parametric models for the probability distributions of the diagnostic measurements (M_i) will have to be used to narrow the confidence limits to workable lengths. These models can be developed and verified through ground studies and/or short-term flight testing, but would then have to be assumed to be valid for representing long-term data. The latter would be used to adjust the values of a limited number of parameters in the model to account for differences between long-term and short-term or ground analog missions. Without any model assumptions about the distribution of the M_i , the only way to estimate P(S) and P(F) would be by the direct counting of “successes” and “failures” using only the long-term mission results. In quality control applications where many trials are possible, this non-parametric approach is often preferred because it avoids making assumptions which may or may not be true. However we do not anticipate enough long-term astronaut-missions in the near future to make this methodology feasible for CM certification.

Example 3.2:

To illustrate the decision process, suppose we require the following conditions for rejecting or certifying the bone loss CM example of Section 2.4:

Reject if the 75% upper one-sided confidence bound for $P(S)$ is less than 0.80 ($p_S = 0.80$)

OR if the 75% lower one-sided confidence bound for $P(F)$ is greater than 0.025 ($p_F = 0.025$)

Accept if the 90% lower one-sided confidence bound for $P(S) > p_S (= 0.80)$

AND if the 95% upper one-sided confidence bound for $P(F) < p_F (= 0.025)$

Recall that the diagnostic measurements M_1 and M_2 are respectively postflight BMD measured in g/cm^2 and percent change in BMD (postflight relative to preflight). Suppose it has been verified with bedrest studies that M_1 and M_2 have approximately a bivariate normal distribution over subjects (the parametric model). Using the long-term mission data, we obtain estimates of the means of M_1 and M_2 and their covariance matrix. Then using the properties of the bivariate normal distribution, we can estimate $P(F)$ and $P(S)$ as well as obtain the 1-sided confidence bounds PF_1 , PF_2 , PS_1 and PS_2 .

In this example, ground data has been used to establish the distribution of the M_i (in this case, bivariate normal; but it could just as well have been some other distributional family). The long-term flight data was then used to obtain the parameters (means and covariance matrix) specifying *which* normal distribution was to be used for calculating $P(S)$, $P(F)$ and confidence limits. A possible variation of this scenario would be that the parametric model restricts the covariance matrix of the M_i to be the same for both bedrest and flight data, but that there is a shift in the mean for flight. Then all the data (bedrest and flight) could be used to obtain a pooled estimate of the covariance matrix, but only the flight data would be used to estimate the mean. Knowing that the M_i have a particular distributional form or family is what enables the calculation of point estimates and confidence limits for $P(S)$ and $P(F)$. Without the larger samples sizes provided by ground studies, it would not have been credible to assume a distributional family using flight data alone. This is the essence of the “parametric” approach.

Suppose now that a sequence of missions have been completed. The hypothetical data shown in the following two examples illustrates two possible outcome scenarios. For each scenario, the values of M_1 and M_2 are shown for each astronaut-mission under the heading “Results for each trial”. Entries in the “S”-column indicate whether the CM was “successful” (1) or otherwise (0), according to the criteria of Example 2.3. Similarly, “failures” are indicated in the “F”-column. For example, in Scenario 1 (a “good” CM) the CM was successful for all trials except for Subject 3 in Mission 1, because postflight BMD (M_1) exceeded 0.86 gm/cm^2 and the percent change (M_2) exceeded -1.0% . For Subject 3 in Mission 1, there was no “success”, ($M_2 = -1.08$) but also no “failure” was indicated because M_2 was not worse than -3.0% . On the other hand for Scenario 2 (a not-so-good CM), there was only one “success” because losses usually exceeded 1% . However the only “failure” occurred on Subject 8, Mission 3, because the postflight BMD was less than 0.86 gm/cm^2 .

The second table for each scenario shows some cumulative statistics after each mission. In this table “N” is the cumulative number of subjects tested, “ave M_1 ” (“ave M_2 ”) is the average value of M_1 (M_2) to date, and “sd M_1 ” (“sd M_2 ”) is the sample standard deviation of M_1 (M_2) to date. The headings “est $P(F)$ ” and “est $P(S)$ ” indicate best estimates of $P(F)$ and

P(S) respectively to date, and PF1, PF2, PS1 and PS2 are the confidence limits as defined previously, calculated using all the data to date.

In Scenario 1 we can certify after the 7-th mission because PS1 exceeded $p_S = 0.80$, for the first time. Also PF2 was less than $p_F = 0.025$, for the first time after the 7-th mission. In Scenario 2 we would have rejected the CM after the third mission because PF1 exceeded .025. There was no need to continue testing after this.

3.3 Calculation of confidence limits

Exact methods for obtaining confidence limits for P(S) and P(F) with anticipated complex multivariate criteria for “success” and/or “failure” and/or non-normal measurements do not exist in general. However the parametric bootstrap as discussed in Efron (1980), provides a simple method for obtaining approximate confidence limits which should be more than adequate for the purpose of this application.

4.0 Experimental design issues

Numbers of subjects needed

Numbers and/or timing of sessions per subject (pre, post, in)

Numbers of replications (trials) per session per subject

The role of control data

References

Efron B (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. SIAM Regional Conference Series in Applied Mathematics 38; Bristol, England: J. W. Arrowsmith Ltd.

Whitehead, J. (1992). *The Design and Analysis of Sequential Clinical Trials*. Chichester, England: Ellis Horwood.

Scenario 1 - A good CM**Results for each trial**

| Mission | Subject | M ₁ | M ₂ | S | F |
|---------|---------|----------------|----------------|---|---|
| 1 | 1 | 0.9368 | 0.32 | 1 | 0 |
| 1 | 2 | 0.9286 | -0.46 | 1 | 0 |
| 1 | 3 | 0.9098 | -1.08 | 0 | 0 |
| 2 | 4 | 0.9210 | 0.75 | 1 | 0 |
| 2 | 5 | 0.9394 | 0.86 | 1 | 0 |
| 2 | 6 | 0.8846 | 1.18 | 1 | 0 |
| 3 | 7 | 0.9007 | 0.06 | 1 | 0 |
| 3 | 8 | 0.9233 | -0.25 | 1 | 0 |
| 3 | 9 | 0.9554 | 0.74 | 1 | 0 |
| 4 | 10 | 0.9292 | 0.03 | 1 | 0 |
| 4 | 11 | 0.9210 | -0.62 | 1 | 0 |
| 4 | 12 | 0.8864 | -0.81 | 1 | 0 |
| 5 | 13 | 0.8681 | 0.74 | 1 | 0 |
| 5 | 14 | 0.9027 | -0.57 | 1 | 0 |
| 5 | 15 | 0.9448 | 0.42 | 1 | 0 |
| 6 | 16 | 0.8874 | -0.49 | 1 | 0 |
| 6 | 17 | 0.9024 | 0.94 | 1 | 0 |
| 6 | 18 | 0.9278 | 1.04 | 1 | 0 |
| 7 | 19 | 0.9092 | -0.09 | 1 | 0 |
| 7 | 20 | 0.9074 | 0.58 | 1 | 0 |

Cumulative Results

| Mission | N | ave. M ₁ | sd M ₁ | ave M ₂ | sd M ₂ | est P(F) | PF1 | PF2 | est P(S) | PS1 | PS2 |
|---------|----|---------------------|-------------------|--------------------|-------------------|----------|--------|--------|----------|--------|--------|
| 1 | 3 | 0.9251 | 0.0138 | -0.4067 | 0.7015 | 0.0000 | 0.0000 | 0.4780 | 0.7680 | 0.1200 | 0.9912 |
| 2 | 6 | 0.9200 | 0.0204 | 0.3281 | 0.7499 | 0.0017 | 0.0001 | 0.3109 | 0.8617 | 0.2766 | 0.9877 |
| 3 | 9 | 0.9222 | 0.0215 | 0.2813 | 0.6485 | 0.0019 | 0.0005 | 0.1872 | 0.9359 | 0.4212 | 0.9842 |
| 4 | 12 | 0.9197 | 0.0212 | 0.0943 | 0.6747 | 0.0024 | 0.0011 | 0.0927 | 0.9476 | 0.5677 | 0.9750 |
| 5 | 15 | 0.9168 | 0.0245 | 0.1151 | 0.6526 | 0.0102 | 0.0066 | 0.0512 | 0.9563 | 0.6793 | 0.9708 |
| 6 | 18 | 0.9150 | 0.0237 | 0.1789 | 0.6767 | 0.0101 | 0.0072 | 0.0349 | 0.9593 | 0.7636 | 0.9678 |
| 7 | 20 | 0.9143 | 0.0225 | 0.1856 | 0.6496 | 0.0079 | 0.0065 | 0.0178 | 0.9660 | 0.8307 | 0.9700 |

Scenario 2 - A not-so-good CM**Results for each trial**

| Mission | Subject | M ₁ | M ₂ | S | F |
|---------|---------|----------------|----------------|---|---|
| 1 | 1 | 0.9447 | -1.48 | 0 | 0 |
| 1 | 2 | 0.9453 | -1.40 | 0 | 0 |
| 1 | 3 | 0.9279 | -1.45 | 0 | 0 |
| 2 | 4 | 0.8824 | -1.31 | 0 | 0 |
| 2 | 5 | 0.8860 | -1.25 | 0 | 0 |
| 2 | 6 | 0.9682 | -2.18 | 0 | 0 |
| 3 | 7 | 0.9036 | -0.90 | 1 | 0 |
| 3 | 8 | 0.8566 | -1.62 | 0 | 1 |
| 3 | 9 | 0.9261 | -1.27 | 0 | 0 |

Cumulative Results

| Mission | N | ave. M ₁ | sd M ₁ | ave M ₂ | sd M ₂ | est P(F) | PF1 | PF2 | est P(S) | PS1 | PS2 |
|---------|---|---------------------|-------------------|--------------------|-------------------|----------|--------|--------|----------|--------|--------|
| 1 | 3 | 0.9393 | 0.0099 | -1.4446 | 0.0388 | 0.0000 | 0.0000 | 0.4310 | 0.0000 | 0.0000 | 0.9762 |
| 2 | 6 | 0.9257 | 0.0347 | -1.5130 | 0.3373 | 0.0290 | 0.0012 | 0.2787 | 0.0641 | 0.0087 | 0.8210 |
| 3 | 9 | 0.9156 | 0.0360 | -1.4299 | 0.3453 | 0.1113 | 0.0257 | 0.1922 | 0.1066 | 0.0245 | 0.6727 |

Countermeasure Evaluation (overview)

- Establish rules for deciding whether CM succeeds (S), fails (F) for a particular astronaut on a particular mission.

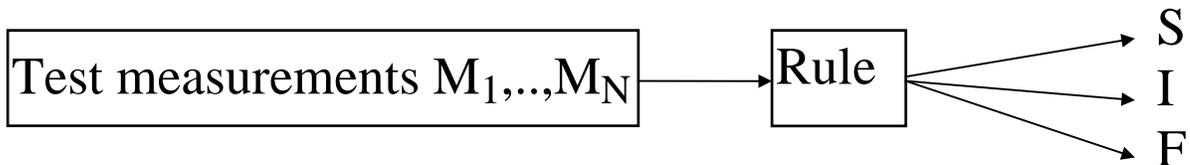


← bone density change →

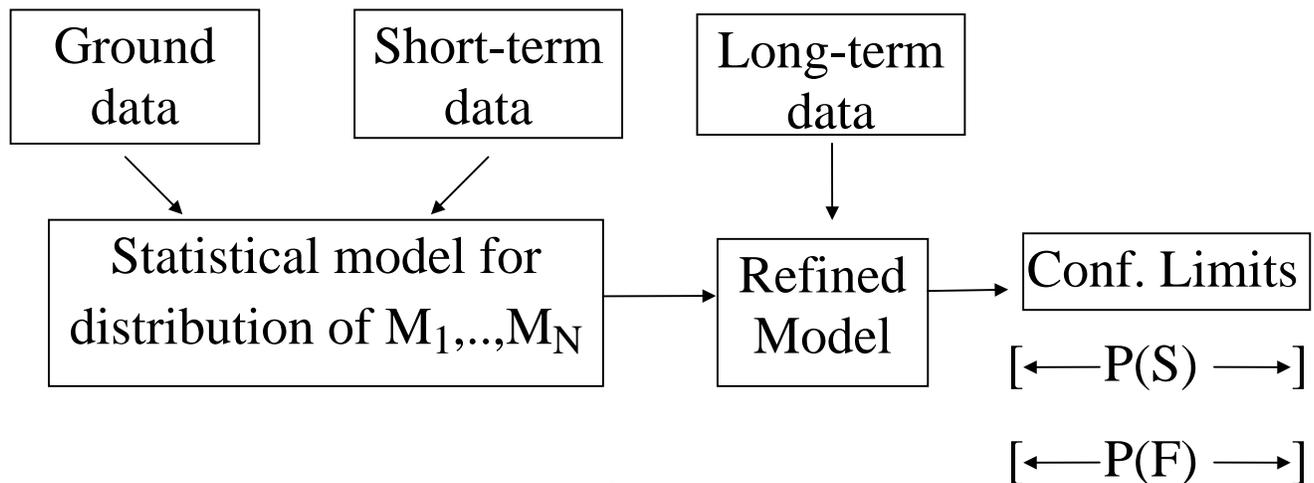
- Establish criteria for success and failure rate percentages that a certifiable CM must satisfy. For example:

CM must succeed on at least 80% of astronaut missions
 CM must not fail on more than 5% of astronaut missions

- After each mission, decide whether CM satisfies or does not satisfy criteria, taking uncertainty into account.
- Establish rules for declaring CM a “success” (S), a “failure” (F) or “indeterminate” (I) for each “trial” (astronaut-mission).



- Develop procedure for obtaining confidence limits for
 - proportion of astronauts for which CM is successful $P(S)$
 - proportion of astronauts for which CM fails $P(F)$



Countermeasure Evaluation (overview) (cont'd.)

- Sequential decision process: Method of Repeated Confidence Intervals
 - Based on methodology used in some applications of quality control and sequential clinical trials.
 - Allows for uncertainty about P(S) and P(F).
 - Uncertainty reduces with each new mission.
 - Ref. Whitehead, J. (1992). *The Design and Analysis of Sequential Clinical Trials*. Chichester, England: Ellis Horwood.

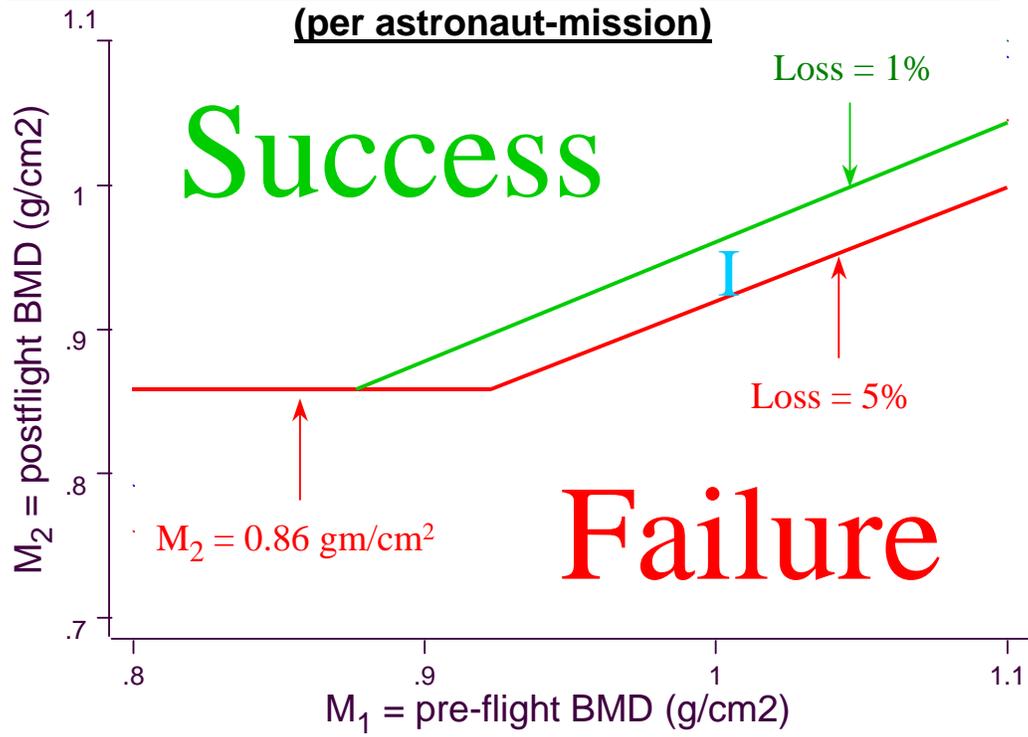
- After each mission, re-calculate confidence limits for P(S), P(F).

- Reject CM if lower limit for P(F) too high *or* upper limit for P(S) too low.

- Accept CM if upper limit for P(F) below threshold *and* lower limit for P(S) above threshold.

Bone Loss CM: Success/Failure Criteria

(per astronaut-mission)



Sequential Decision Process

