SURVIVAL ANALYSIS AND MAXIMUM LIKELIHOOD TECHNIQUES AS APPLIED TO PHYSIOLOGICAL MODELING

Seattle, Washington
May 19, 1998

Chaired and Edited by

Paul K. Weathersby
Gales Ferry, CT

and

Wayne A. Gerth*
Duke University Medical Center
Durham, NC

Sponsored by

Divers Alert Network
National Aeronautics and Space Administration
United States Air Force
United States Navy
Undersea and Hyperbaric Medical Society

* Present address: Navy Experimental Diving Unit, Panama City, FL.

TABLE OF CONTENTS

Preface

Wayne A. Gerth

In 1984, Dr. Paul Weathersby, Dr. Lou Homer and Dr. Edward Flynn published a seminal paper in which they introduced survival analysis into the study of decompression sickness (DCS). The approach they outlined, and continued to develop with colleagues Shalini Survanshi, Erich Parker and others at the Naval Medical Research Institute (NMRI) in a subsequent series of published papers and reports, gave new direction to the way we reconcile theory with experience in this field. First, it explicitly recognizes that a given physiological outcome is not an inevitable result of a particular environmental history, but instead is only a probabilistic function of that history. Second, the approach includes rigorous means to make one or more candidate expressions of that probabilistic function each provide its best possible, or optimum, correlation of observed outcomes in actual experience. The optimized models that emerge from such work are consequently quantitative generalizations of that experience, which renew the analyst's focus on the data he or she has in hand. Finally, the approach allows quantitative assessments to be made of how well a given model accounts for observed behavior in specific sets of data, so that the best of a collection of candidate models can be selected. This selection process allows models that are more complex than the data warrants to be identified and deselected, helping to separate necessary theoretical complexity from speculation.

Workers interested in other undersea and aerospace physiological problems soon recognized the analytic advantages of survival modeling. Adoption of the techniques in these areas has lead to development of application-specific functions describing responses to ever more complex patterns in the independent variables, and to use of meta-analytic approaches to build data sets with analytically tractable numbers of occurrences of the adverse events of interest. As these applications have ventured farther from those described in standard statistical tests, there has been a growing need to pause and distill their underlying principles, critically evaluate their merits, and outline directions for further development and application. The present Proceedings of the Workshop on Survival Analysis and Maximum Likelihood Techniques as Applied to Physiological Modeling is both an attempt to meet that need, and a salute to the NMRI workers who originally introduced us to this promising line of inquiry.

Welcoming Remarks

Paul K. Weathersby

Welcome.

Within the professional interests of this Society, there are a number of environmental stressors that carry risks to humans. The study of those stressors, and how people survive them, has been subject to an increasing amount of analysis.

The conduct of these analyses is not something that any of us had the benefit of reading a good textbook about. That book probably is yet to be written. But the scope of endeavors in this direction have become widespread enough, that the Society thought that taking a day for some degree of review of the subject would be valuable.

The program is quite full. You see that it is structured to start with an original presentation by Dr. Gerth, who will try to cram all of what we did not learn in school into a mere 30 minutes. There will then be a number of presentations up through mid-afternoon by people that I would refer to as "practitioners" in this area. We practitioners have been usually driven by an application that we needed to fix. When you are committed to taking care of an applied problem, sometimes you lose sight of the rigor in the techniques that are available to you. We have asked the practitioners to follow an outline of: what is your data?, how do you fit your data?, and how do you assure success?

We are hoping that during the prolonged critique session this afternoon, Dr. Harrell and Dr. Homer will be able to help us get re-oriented if we have lost our way somewhat. Following that, they have each graciously agreed to make a short presentation on something that may be useful to us in the future.

I would note that because of the structure of the Workshop, we will not have time for any extensive question session along with each paper. There should be a little time for discussion following all the presentations.

Workshop Origin

Edward D. Thalmann, M.D.
Captain, Medical Corps, U.S. Navy (Retired)

This workshop started out as an idea that I had after coming to NMRI (Naval Medical Research Institute) from NEDU (Navy Experimental Diving Unit) and learning about something they were doing called probabilistic modeling. I spent the next seven years immersed in the technique to model decompression sickness occurrence, and discovered that I didn't understand it as well as I would have liked. There were some standard texts to which one could refer (many very heavy going), but they really did not cover the methods that we were using -- especially the design of risk functions.

I also began to notice that there were a lot of papers appearing in which the technique of maximum likelihood was used. In reading some of them, it began to occur to me that there wasn't any good way to tell if the technique was being applied correctly or not, mainly because the published work appeared to be beyond the scope of the standard texts.

Talking with Wayne Gerth and Paul Weathersby, it seemed like the idea of a workshop was the way to go. So, I took it upon myself to make some phone calls and see if anybody was interested in actually funding this thing. Sure enough, we did get some funding, and I'd like to mention those organizations and their points of contact, without whose support this workshop would not have taken place.

Captain Marie Knafelc, the Senior Medical Officer at the Navy Experimental Diving Unit in Panama City, FL, convinced her Commanding Officer to provide some funding. Dr. Andy Pilmanis convinced his boss at Brooks Air Force Base in San Antonio, TX, to participate. Dr. Mike Powell managed to get the folks at NASA-JSC (Johnson Space Center, Houston, TX) to contribute some funding, and then Dr. Peter Bennett at the Divers Alert Network (DAN, Durham, NC) also graciously decided to provide funding. These four organizations provided enough funding to have what I expect will be a first class Workshop.

It is notable that although this workshop is occurring in conjunction with an annual scientific meeting of the UHMS, two of our sponsors are in the aerospace community and two are in the diving community. This reflects the balance that we sought. If you look at the program, while it's heavily weighted towards diving, you'll also notice we're going to be talking about applying this technique to altitude exposure, hypothermia, to some respiratory problems, and oxygen toxicity.

One of the first things I did after getting this funding was to fall back on my training in the Navy and completely delegate all responsibility to Paul and Wayne. They have really been working very hard to put this together, and I think that any kudos about this workshop should go to them.

One note. It was the intent of the workshop to focus on the methodology of maximum likelihood and not focus on the actual physiological models themselves. What is of interest is how the technique can be used to take a model that somebody has conceived and "fit it to data". I hope everybody will stick with the spirit of that. There are plenty of other sessions in the upcoming meeting where we can discuss the actual models themselves, but here we're concerned with application of this technique to whatever model you have come up with.

Overview of Survival Functions and Methodology

Wayne A. Gerth
F.G. Hall Laboratory
Center for Hyperbaric Medicine and Environmental Physiology
Duke University Medical Center
Durham, NC 27710

Table of Contents

1. Introduction

I want to join Dr. Weathersby and Dr. Thalmann in thanking our sponsors and all of you for making possible this workshop on "Survival Analysis and Maximum Likelihood Techniques as Applied to Physiological Modeling." My overall objective here is to review the basic principles of survival analysis and how maximum likelihood techniques are used in such analyses. If successful, this overview will provide a background for the presentations that follow, and give the environmental physiologist with a strong background in mathematics and a limited background in statistics information sufficient to develop an intuitive and quantitative understanding of how survival analysis might be applied in his or her own work.

What is survival analysis? Survival analysis is the study of the time courses of responses to a provocation in a population of individuals. The outcome variable is the time until occurrence of a particular event of interest, or the time until observation of an individual is terminated without occurrence of the event. This variable is defined as the survival time, T. The term "failure time" is used synonymously with "survival time." Survival time is a continuous, positive-valued random variable, any particular value of which is denoted by t ($t>0$).

The response in survival analysis is categorical; occurrence or non-occurrence of an event of interest; and can be one of two types. The first type is the univariate response (also called a dichotomous or binary response), which consists of a single, nonrepetitive, all-or-nothing event that occurs or does not occur in any one individual. An individual is typically considered to either fail or survive, but the event marking failure can be as innocuous as a change in position or as final as death. Competing risk problems involve a subset of this class of responses in which an individual may experience one of two or more events. In these problems, an individual may still fail only once, but in more than one way or by more than one mechanism. In problems involving univariate responses, the probability of an observation at any time when an individual is under study is unity, consisting of the sum of the probabilities of the two mutually exclusive outcomes; that an event has occurred (E) or not occurred (0):

$$P(0) + P(E) = 1 \qquad (1)$$

The other type of categorical response in survival analysis is the multivariate response (also called a polychotomous or multinomial response) in which each of two or more events can occur or not occur in any one individual. With this type of response, it is possible to observe more than one failure time on an individual, so that *Eq.* (1) does not apply. Problems involving multivariate responses are beyond the scope of this Workshop and will not be considered further here.

The goals of survival analysis are to develop and evaluate quantitative descriptions of survival experience in a population and identify the important risk factors. In the work we will review today we also wish to emerge with a generalization of that experience that can be used to manage individual risk in future exposures. We consequently seek to transcend simple

description of the data to develop tools by which future behavior can be prescribed, recognizing that such applications are controversial and require great care [21].

The overall approach is schematized in Figure 1. At the outset of the process, we have on one hand a collection of experience or data that consists of actual observations of survival experience in an exposed population, including all information required to describe each exposure and its corresponding outcome. On the other hand, we have a collection of theories or models that we think explain in some sort of abstract fashion the relationships between independent variables[1] that describe the exposures and occurrences of the response of interest. The problem is that we seek a theory that provides the best-possible representation of that experience. Likelihood maximization is a tool that helps this to be achieved. Using this tool, each model or theory among the collection of candidates is optimized about the available data, meaning that it is made to provide its own best representation of that data. Optimized models are then evaluated so that the best among them can be selected. This evaluation is undertaken using byproducts of the optimization processes that allow the correlation provided by each model to be directly and indirectly compared to the correlations provided by other models. If none of the optimized models is found to provide a satisfactory representation of the available experience, the models are refined by correcting identified theoretical weaknesses, reoptimized about the data and re-evaluated until a satisfactory model is obtained.

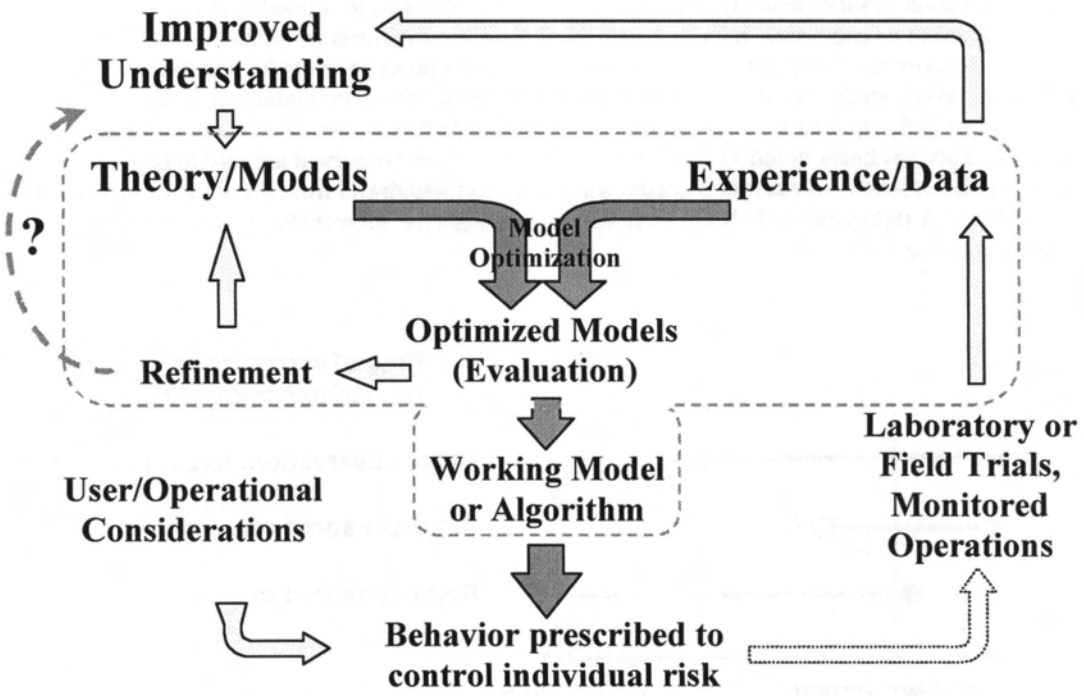


Figure 1. Schematic of the probabilistic modeling enterprise. Elements within the dashed line are the focus of this Workshop. Experience/Data includes all independent variables (covariates) required to describe each exposure and its corresponding outcome.

The model that finally emerges from this loop becomes a working model that can be used to prescribe future exposures in which individual risk is allowed to reach but not exceed specific target or acceptable risk*s*. These prescriptions can then be taken into the laboratory and tested in controlled trials or forwarded into the field for operational use under carefully

[1] Also called covariates

monitored conditions. Such testing or usage yields more experience that feeds back into the whole process, so that the models can learn from experience as it continues to accumulate.

The aspects of this process that are the focus of this Workshop are indicated within the dashed line in Figure 1. My purpose in what follows is to review these particular aspects with rigor sufficient to illuminate the quantitative relationships that are most widely used in this work. At risk of being overly pedantic for some, I have specifically sought to avoid planting important equations at the end of abbreviated derivations that can be difficult if not impossible to follow.

2. Properties of Survival Data

Modeling survival experience requires careful consideration of the properties of survival data that distinguish it from data obtained in other types of study.

2.1. Censoring

A data point or observation in survival analysis is the survival time from a well-defined time of origin until the occurrence of a particular event or end-point. Survival times recorded for individuals in which the event of interest occurs are called *uncensored* observations. These are *exact* if the observed failure times are singular and distinct. If the event is observed in all individuals under study, the set of observed survival times for those individuals is said to be *complete*. Most often, however, particularly in the applications that are the focus of this Workshop, the event is not observed in all individuals under study. Survival times of individuals in which the event is not observed are said to be *right censored*. As illustrated in Figure 2, a right-censored observation occurs when a subject withdraws from the study or is lost to follow-up during the study period (L). A right-censored observation also occurs when the subject simply does not experience the event before the study ends (C).

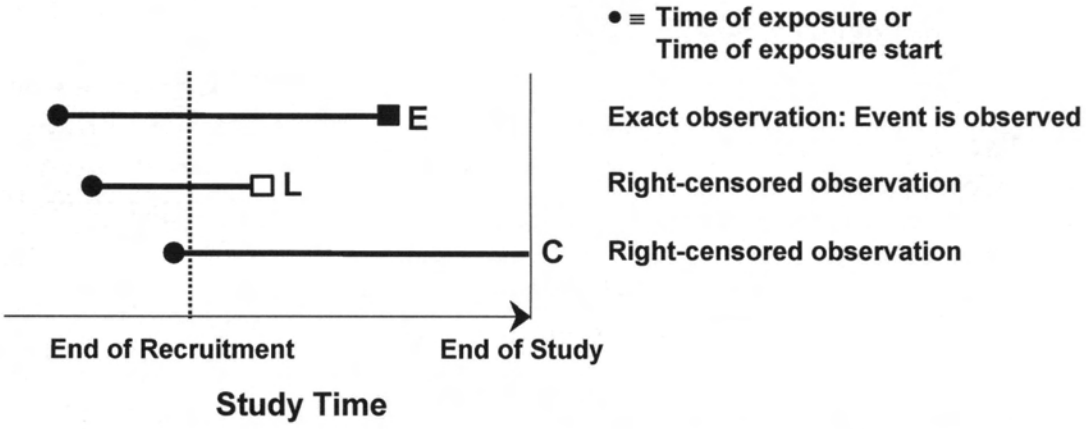


Figure 2. Illustration of an uncensored, exact observation (E), a right-censored observation in which the subject withdraws from study or is lost to follow-up during the study period (L), and a right-censored observation in which the subject does not experience the event before the study ends (C). An exposure may be an event that is complete at essentially one point in time or the start of a process that continues for some period.

Subjects in Figure 2 are shown to be recruited into the study at different times, and right-censored at the same time if they are not accidentally lost and do not suffer occurrence of the event. This is Type III or *progressive* censoring. The relevant survival time for each subject is then the time-on-study, or "subject time," obtained by arithmetically adjusting the study time for the individual to begin at t=0. Subject times for the examples in Figure 2 are shown in Figure 3.

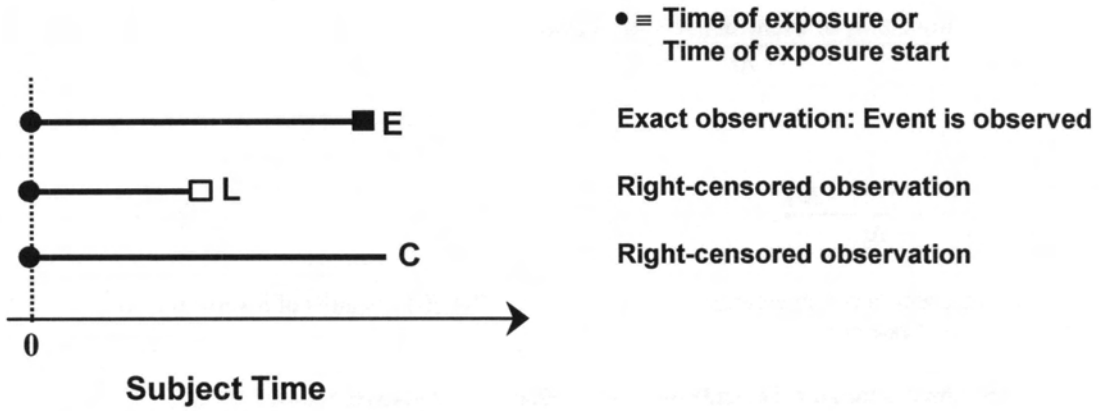


Figure 3. Observations in Figure 2 shifted to give survival times with reference to a common t=0 start time.

In contrast to the case illustrated in Figure 2, many studies are conducted by placing all subjects under study at the same time, t=0. In such studies, Type I censored observations are obtained by right-censoring all surviving individuals at a pre-specified time, yielding equal survival times for those individuals. Alternatively, the study may continue until a pre-specified portion of the individuals have failed, at which time all surviving individuals are censored (Type II censoring), or pre-specified fractions of surviving individuals may be right-censored at various ordered failure times as the study progresses (progressive Type II censoring). More complex censoring schemes may also be used that depend arbitrarily on various aspects of the study as it unfolds. Regardless of which censoring mechanism is used, however, care must be taken to ensure that it remains independent of the failure mechanism that governs the survival time of any individual. In other words, an individual in a group of individuals who have the same values of all relevant independent variables, but who is censored at time t, must remain representative of all other individuals that survive to that time.

2.2. Failure Time Distributions

Failure time is a random variable, the statistical properties of which become clear as its value is observed on an increasing number of individuals drawn at random from a population. These properties are established by arranging the observations into, or by assuming that the observations follow one of four different but interrelated distributions.

2.2.1. Probability Density Distribution

If N individuals from a homogeneous population are exposed to a given provocation beginning at time t=0, the number of individuals that experience the event of interest in any time interval t to $t+\Delta t$, $t>0$, can be determined. The *probability density distribution* of failure times, $f(t)$, is then given by:

$$f(t) = \lim_{\Delta t \to 0} \frac{\left(\dfrac{\#\text{events in } (t, t+\Delta t) \text{ interval}}{\Delta t} \right)}{N}; \qquad 0 \le t < \infty. \tag{2}$$

Note that the domain of the distribution includes all positive values of t. Because the incidence of an event divided by the sample size is the probability of the event, *Eq.* (2) becomes

$$f(t) = \lim_{\Delta t \to 0} \frac{\text{probability of event in } (t,\ t + \Delta t) \text{ interval}}{\Delta t},$$

which is expressed as:

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t)}{\Delta t}. \qquad (3.a)$$

The probability of the event in the numerator is dimensionless, so that $f(t)$ has units of inverse time (t^{-1}) and can be considered as the instantaneous event rate at t.

In some studies, observations can be made on an individual only at discrete times, t_j, j=1,… ; $t_1 < t_2 <$, …, $< \infty$. Conversely, any finite set of survival data, and therefore any set of survival data in actual practice, can be considered to be a sample of discrete observations from a continuous distribution. The associated discrete probability density distribution is then discontinuous, given by the probability of the event at each t_j:

$$f(t_j) = P(T = t_j). \qquad (3.b)$$

Under these conditions, $f(t_j)$ is a dimensionless quantity. The means by which a continuous distribution defined by *Eq.* (3.a) can give rise to a discrete distribution defined by *Eq.* (3.b) is considered in Appendix B. Unless otherwise noted, all distributions considered in the remainder of this overview are continuous on t.

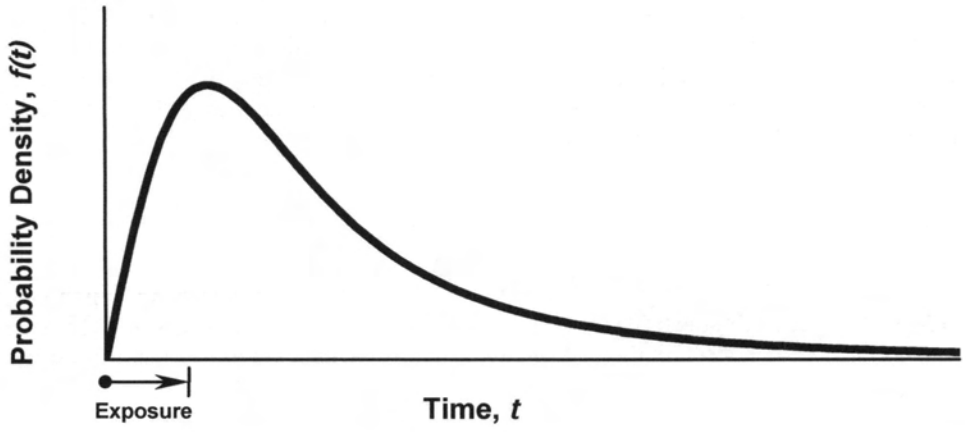


Figure 4. Hypothetical probability density function, $f(t)$, of failure (or event) times in a large number of individuals exposed to a given set of failure-provoking conditions during the period indicated.

If the sample size is large, a plot of $f(t)$ *vs.* time might appear as illustrated in Figure 4. As illustrated in this example, $f(t)$ is typically right-skewed. This is in contrast to the distributions typical of other types of data, which tend to be symmetric about their mean values.

2.2.2. Cumulative Distribution and Survivor Functions

The probability density function provides the basis for the definition of two other important functions in survival analysis, the *cumulative distribution function* and the *survivor function.* The value of the cumulative distribution function at

each t, $F(t)$, is the probability that the survival time T is less than t; i.e., that the event will be observed in the interval between 0 and t ($0 \le T < t$):

$$F(t) = P(T < t), \tag{4}$$

or in terms of the probability density function[2],

$$F(t) = \int_0^t f(u)du. \tag{5}$$

The value of the survivor function at each t, $S(t)$, is the probability that the survival time is greater than or equal to t:

$$S(t) = P(T \ge t), \tag{6}$$

which in terms of the probability density function is:

$$S(t) = \int_t^\infty f(u)du. \tag{7}$$

Through application of *Eqs.* (5) and (7), the cumulative distribution and survivor functions are readily visualized in terms of areas under the probability density function, $f(t)$, as shown in Figure 5.

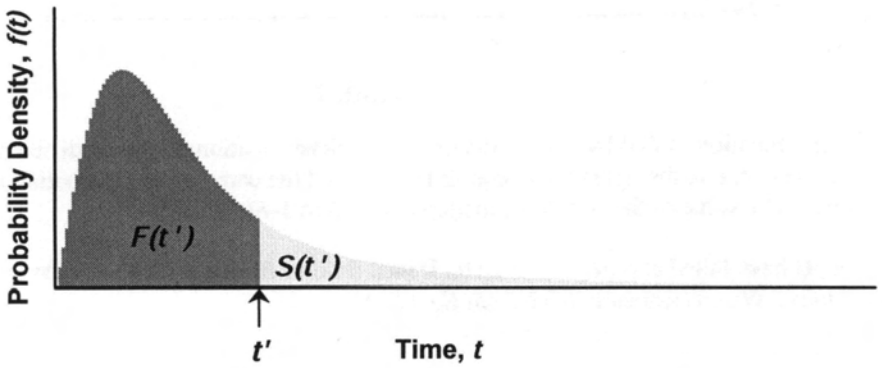


Figure 5. The cumulative distribution function, $F(t)$, and survivor function, $S(t)$, as areas under the probability density function of survival times, $f(t)$. The value of $F(t)$ at time $t = t'$ is the darkly shaded area under the probability density function to the left of t'. The value of $S(t)$ at time $t = t'$ is the lightly shaded area under the probability density function to the right of t'.

The value of the cumulative probability distribution function at time t, $F(t)$, is the area under the probability density function to the left of t. Similarly, the value of the survivor function at time t, $S(t)$, is the area under the probability density function to the right of t. Note that $F(t)$ and $S(t)$ are dimensionless because they are both probabilities.

Because $P(0)$ and $P(E)$ in *Eq.* (1) are simply $S(t)$ and $F(t)$, respectively, *Eq.* (1) can be rewritten and rearranged to obtain:

[2] Throughout this overview, u is used as a dummy variable of integration.

$$S(t) = 1 - F(t)\,. \tag{8}$$

Figure 6 illustrates how *Eq.* (8) and the relationships illustrated in Figure 5 are used to construct $F(t)$ and $S(t)$:

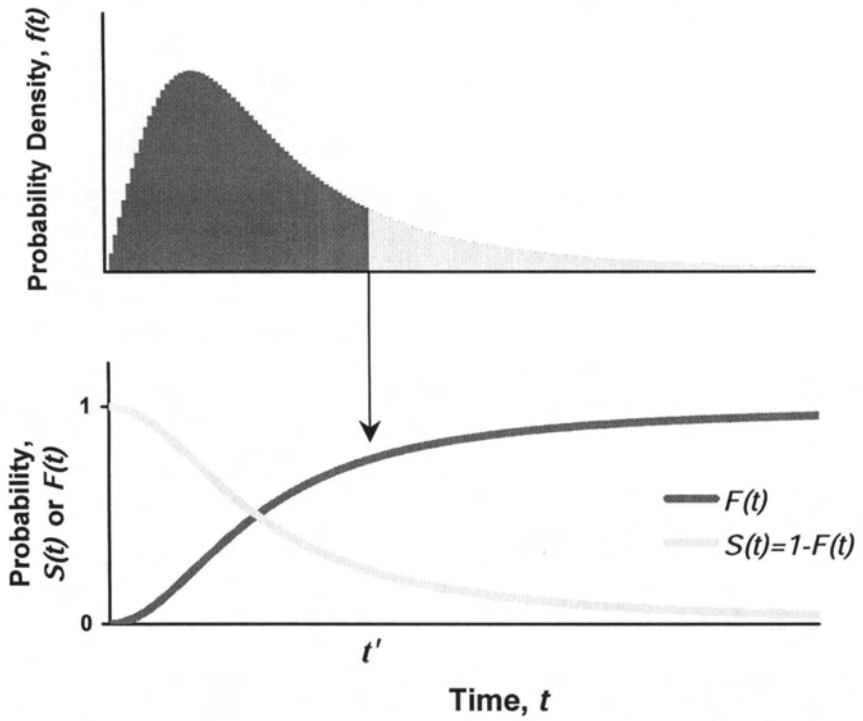


Figure 6. Determination of $F(t)$ by integration of $f(t)$ and determination of $S(t)$ by difference. Area under $f(t)$ to the left of the arrow at t' in the upper panel equals the value of the cumulative distribution function at t', $F(t')$, in the lower panel. The value of the survivor function at t' is then 1-$F(t')$.

Because no one can yet have failed at time t=0, $F(0)$=0. It then follows from *Eq.* (8) that $S(0)$=1; i.e., the probability of survival at t=0 must be unity. We consequently have from *Eq.* (7) that

$$\int_{t=0}^{\infty} f(u)du = 1\,. \tag{9}$$

Referring to *Eq.* (5), *Eq.* (9) implies that the probability of failure at infinite time, $F(\infty)$, must be unity. In other words, the event must be assumed to *eventually occur* in *all* individuals under study. This is not a troublesome requirement if the event is death. However, other events that we wish to model do *not* inevitably occur in all individuals under study. An example of such an event is decompression sickness, which usually does not occur in many individuals regardless of how long they are observed after decompression. Fortunately, the analytical requirement for eventual occurrence of the event of interest is no problem because of the vitally important role played by right-censoring. We will see that the form of the distribution function after the highest right-censored time in any data set does not affect the results and is hence arbitrary. We are only interested in the form of the distribution function *up to* the highest right-censored time, and can allow the function thereafter to assume whatever form might be necessary to satisfy *Eq.* (9). In effect, then, the density distribution is considered to consist of two parts, $f_a(t)$ and $f_b(t)$, defined with respect to an arbitrarily high time, T_r, at which all possible events have occurred:

$$f(t) = f_a(t) + f_b(t)\,, \tag{10}$$

where

$$f_a(t) = 0\,;\, t > T_r$$
$$f_b(t) = 0\,;\, t < T_r\,.$$

We are thus ordinarily concerned only with determination of $f_a(t)$, for which the following holds:

$$0 < \left\{ \int_0^{T_r} f_a(u)du \right\} \leq 1\,. \tag{11}$$

Because $f_a(t)$ does not conform to *Eq.* (9), it is called an *improper* density distribution. As shown in Appendix A, this problem can also be addressed using proper distribution functions that conform to *Eq.* (9).

2.2.3. Hazard Function

The above density and distribution functions are defined with respect to the total number of individuals at experiment start. However, the number of individuals at risk in a given experiment decreases over time through action of the failure mechanism. A fourth function, the *hazard* (or *risk*) function, incorporates this information to make it particularly useful in the analysis of survival data. The hazard function $h(t)$ is defined as the instantaneous event rate at time t, *given that the individual has survived up to that time*:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}\,. \tag{12}$$

The hazard is thus a *conditional failure rate.*

The hazard function is intimately related to the survivor function, $S(t)$, based on the definition of conditional probability [2]. If we have two events, denoted A and B, with respective probabilities $P(A)$ and $P(B)$, the conditional probability of event B given occurrence of event A, $P(B \mid A)$, is:

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}, \tag{13}$$

where $P(A \cap B)$ is the probability of joint occurrence of events A and B. Thus, if we let $P(A) = P(T \geq t)$ and $P(B) = P(t \leq T < t + \Delta t)$, the numerator in *Eq.* (12) becomes:

$$P(t \leq T < t + \Delta t | T \geq t) = \frac{P(t \leq T < t + \Delta t)}{P(T \geq t)}, \tag{14}$$

where we have made use of fact that the sample space for event B consists wholly of individuals in which event A has occurred, so that $P(A \cap B) = P(B)$. We also have from *Eq.* (1) that the total probability of any given observation is unity. We therefore have

$$P(t \leq T < t + \Delta t) = 1 - P(t + \Delta t \leq T) - P(T < t)\,, \tag{15}$$

which, because $1 - P(t + \Delta t \le T) = P(T < t + \Delta t)$, can be re-written

$$P(t \le T < t + \Delta t) = P(T < t + \Delta t) - P(T < t). \tag{16}$$

Using the definition of the cumulative distribution function, $F(t)$, given in *Eq.* (4), *Eq.* (16) becomes:

$$P(t \le T < t + \Delta t) = F(t + \Delta t) - F(t). \tag{17}$$

Substituting this result into *Eq.* (14) and using *Eq.* (6) yields

$$P(t \le T < t + \Delta t | T \ge t) = \frac{F(t + \Delta t) - F(t)}{S(t)}, \tag{18}$$

so that *Eq.* (12) becomes

$$h(t) = \lim_{\Delta t \to 0} \left\{ \frac{F(t + \Delta t) - F(t)}{\Delta t} \right\} \frac{1}{S(t)}. \tag{19}$$

The first factor on the right of this equation is the definition of the derivative of $F(t)$ with respect to t. When *Eq.* (17) is substituted into *Eq.* (3.a), this factor is also seen to equal $f(t)$. Substitution of this latter equality yields an important intermediate result:

$$h(t) = \frac{f(t)}{S(t)} = \frac{\text{instantaneous event rate at } t}{\text{probability of surviving to } t \text{ or longer}}. \tag{20}$$

The hazard has units t^{-1}, unless $f(t)$ is discrete as defined by Eq. (3.b). In the latter case, Eq. (20) still applies, but $h(t)$ is then also discrete and dimensionless (Appendix B). In either case, it should be evident that the hazard is *not* a probability, but ranges over all positive real numbers between 0 and infinity ($0 \le h(t) < \infty$).

If we multiply both the numerator and denominator of *Eq.* (20) by the sample size; i.e., by the number of individuals that entered the experiment at t=0; we get

$$h(t) = \frac{\#\text{ individuals that fail in } (t, t + \Delta t) \text{ interval}}{(\#\text{ individuals that survive to } t) \cdot \Delta t}. \tag{21}$$

This definition also follows directly from the statement preceding *Eq.* (12), where the numerator is the number of individuals that fail in the infinitesimally small time interval between t and $t+\Delta t$ and the denominator is the number of individuals that remain of those that entered the experiment to be at risk of failing in that interval.

We proceed by re-writing the numerator in *Eq.* (20) using the derivative of $F(t)$ from *Eq.* (5):

$$h(t) = \frac{f(t)}{S(t)} = \frac{dF(t)/dt}{S(t)},$$

and then use *Eq.* (8) to obtain:

$$h(t) = \frac{d(1 - S(t))/dt}{S(t)}.$$

Simplification yields:

$$h(t) = -\left\{\frac{dS(t)/dt}{S(t)}\right\} = -\frac{d}{dt}\{\ln S(t)\},$$

from which our final, desired result follows:

$$S(t) = \exp\left[-\int_0^t h(u)du\right] = \exp[-H(t)], \tag{22}$$

where the *cumulative hazard*, $H(t)$, is defined:

$$H(t) = \int_0^t h(u)du = -\ln S(t). \tag{23}$$

Note that *Eq.* (22) provides an expression for the survivor function in terms of the hazard function in the $[0;t]$[3] interval, the only interval in survival analysis where observations can be made.

Proper and improper hazard functions

It follows from *Eq.* (22) that the hazard function must diverge for the survivor function to equal 0 at $t=\infty$ [$S(\infty)=0$] in accord with *Eq.* (9). This requirement is quantitatively expressed as follows:

$$\lim_{t\to\infty} \int_0^t h(u)\,du = \infty. \tag{24}$$

Hazard functions that meet this requirement are *proper* hazard functions that correspond to *proper* survival functions. As discussed following *Eq.* (9), however, survival analyses in the context of the present Workshop often involve events that do not occur in all individuals under study. Such analyses entail specification of a hazard function only for the density distribution $f_a(t)$ in *Eq.* (10). Such hazard functions do not comply with *Eq.* (24) and are called *improper*, corresponding to *improper* survival functions.

2.2.4. Inter-Relationships: Example

It should now be clear that the functional form of any one of the above distribution functions; $f(t)$, $F(t)$, $S(t)$ or $h(t)$; completely determines the forms of all others. This is readily illustrated for the simplest of cases in which the hazard is constant and equal to λ:

$$h(t) = \lambda. \tag{25}$$

The survivor function is then readily obtained from *Eq.* (22):

$$S(t) = \exp(-\lambda t). \tag{26}$$

---

[3] In interval notation, $[a;b]$ denotes a closed interval consisting of all x such that $a \le x \le b$.

The form of this function suggests its name, the *exponential distribution.* Because of its simplicity, this distribution is often used as a null model against which more complex distributions are compared. The corresponding probability density and cumulative distribution functions are, from *Eq.* (20):

$$f(t) = \lambda \exp(-\lambda t), \tag{27}$$

and from *Eq.* (8):

$$F(t) = 1\text{-}S(t) = 1\text{-}\exp(-\lambda t). \tag{28}$$

These functions for the exponential distribution are illustrated in Figure 7 for an example case in which $h(t)=\lambda=5$.

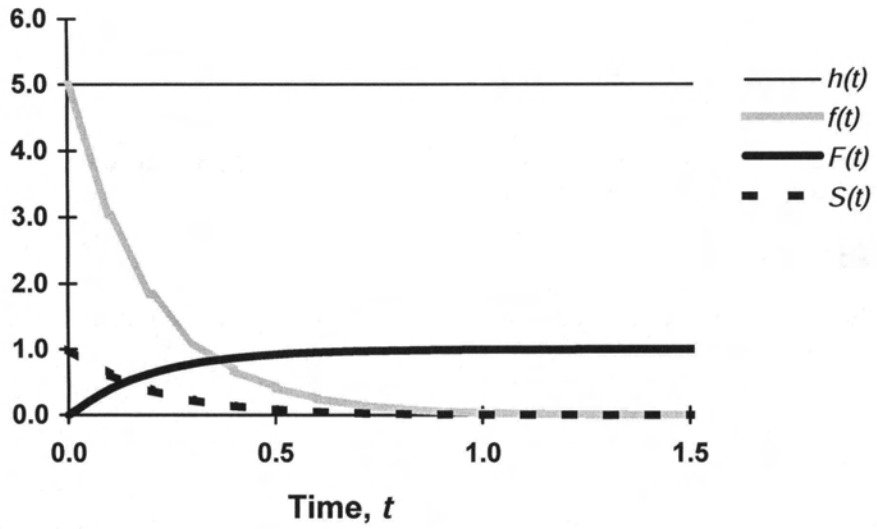


Figure 7. The hazard $h(t)$, probability density $f(t)$, cumulative probability $F(t)$ and survivor $S(t)$ functions for the exponential distribution.

3. Strategies for Model Development: Specification of the Hazard Function

Except for our illustration using the exponential distribution in Section 2.2.4, we have so far described the relationships between the various distribution functions in only general terms. Explicit specification of one of the distribution functions is a required step in the development of any model. This can be undertaken in one of two fundamentally different ways: by a non-parametric or by a parametric approach.

3.1. Non-Parametric Procedures

In non-parametric approaches, the form of the survivor function is estimated directly from the data, with no specific assumptions about the distribution of survival times. Life table and Kaplan Meier estimates of the survivor function are examples of this approach covered in standard texts. This type of approach is taken primarily for data description and factor identification, and can thus help choose an appropriate parametric model for subsequent parametric analysis.

3.2. Parametric Procedures

In parametric approaches, information is available to motivate specification of the mathematical form of the hazard function. The survivor, failure and probability density functions are then obtained from this hazard function. Parametric approaches are usually taken in modeling adverse responses to environmental stress. Candidates for the hazard function can be considered to fall into one of two categories:

1) functions that are based on well-characterized statistical distributions such as the exponential, Weibull, log-logistic and gamma distributions (These are all special cases of the generalized F-distribution), or;

2) functions with shape presumed to be representative of, or explicitly defined as, the output of modeled physiological/etiological processes. Such "mechanistic" (or "scientific" [14]) functions are particularly useful in modeling responses governed by independent variables that vary over time in complex patterns. The motivation for use of such functions in modeling the incidence and time of occurrence of decompression sickness (DCS) in man is the focus of another presentation in this Workshop [10].

3.3. Parameters in the Hazard Function

In general, the shape of the hazard function is governed by the values of a collection of p parameters, β_k (k=1, 2, …, p) that may serve a variety of purposes:

- Set location and shape properties of the function for a reference population characterized by zero values for all explanatory variables. The hazard, $h(t)$, in our illustration of the exponential distribution in *Eqs.* (25) – (28) and Figure 7 serves this purpose in describing the distribution of survival times for a homogeneous population.

- Scale the influences of explanatory variables that accommodate heterogeneity in the population of interest. Accommodation of heterogeneity in the population requires generalization of the hazard to be a function of a vector of independent explanatory variables, or covariates, **z**. For example, for our hypothetical example in Figure 4, it was stipulated that all individuals in the sampled population were exposed to the same provocation. However, models are usually constructed to examine how the distribution of survival times varies as particular aspects of the provocation, or individual involved, are varied. Samples in such cases are consequently drawn from populations that include groups of individuals distinguished from others by having characteristic values of particular independent variables; e.g., dive depth, dive bottom time, gender, etc. The hazard for a given group in the population is then a function of the values of the independent variables, or covariates, for that group. In order to generalize the exponential distribution, for example, $h(t)$ is replaced by $h(t;\mathbf{z})$, which remains constant for any given **z** but depends on **z**. The $h(t;\mathbf{z})$ function can be formulated in any fashion, but additional parameters are almost inevitably required to scale the influences of added elements in **z**.

- Serve as required constants in a "mechanistic" hazard function. Parameters of this type perform the same functions as the two types of parameters described above. However, such parameters are associated with specific biophysical properties, such as gas solubilities and diffusivities, which govern how the hazard varies with changes in the independent variables.

4. Calibrating the Hazard Function

The hazard function is fit to a data set of observed survival times by adjusting, or calibrating, the parameter values, β_k, (k=1, 2, …, p). The model's best fit to the data, and the optimum values of the β_k, are obtained by maximizing $L(\boldsymbol{\beta})$, a likelihood function of the parameter vector, $\boldsymbol{\beta}$. This function is defined as the joint probability of the observed data, given the specified hazard model, specific values for the parameters of the hazard model, and action of the censoring mechanism. An understanding of how the likelihood function is defined clarifies how its maximization optimizes model fit to the data.

4.1. Likelihood Definition

With dichotomous responses, one can make only one of two possible observations on any one exposure, *i*:

a) the event (E_i) is observed by or at failure time t_i, with a corresponding probability $P(E_i)$, or;
b) the event is not observed (0_i) by right-censored survival time t_i with probability $P(0_i)$.

The likelihood, $l_i(\beta)$, or probability of the outcome *actually observed* on the i^{th} exposure, is thus *either* $P(E_i)$ or $P(0_i)$:

$$l_i(\beta) = P(0_i)^{1-\delta_i} P(E_i)^{\delta_i}, \tag{29}$$

where

$\delta_i = 1$ if t_i is the time at which an event occurred (t_i is a failure time),[4] and;
$\delta_i = 0$ if t_i is a right-censored time (the event was not observed).

$l_i(\beta) = P(E_i)$ if t_i is a failure time, or $l_i(\beta) = P(0_i)$ if t_i is a right-censored time.

The outcome of each exposure is assumed to be independent of the outcomes of other exposures. The joint probability of the observations in a data set of *N* exposures is thus the product of the *N* individual likelihoods:

$$L(\beta) = \prod_{i=1}^{N} l_i(\beta) = \prod_{i=1}^{N} P(0_i)^{1-\delta_i} P(E_i)^{\delta_i}. \tag{30}$$

We can now consider how each of the factors on the right of *Eq.* (29) contributes to the overall likelihood, $L(\beta)$, in *Eq.* (30).

4.1.1. Definition of $P(0_i)$ in the Likelihood

If the event was not observed, the recorded survival time t_i for the exposure is a right-censored time. The likelihood for the exposure $l_i = P(0_i)$ is the probability of surviving to t_i. This probability is given in terms of the hazard function by *Eq.* (22):

$$P(0_i) = S(t_i) = \exp\left[-\int_0^{t_i} h(u)du\right]. \tag{31}$$

The importance of the $P(0_i)$ value for an observed right-censored time t_i is readily envisioned in terms of areas under candidate probability density functions, using the definition of $S(t_i)$ given by *Eq.* (7). Two candidate density functions are illustrated in Figure 8 for an observed t_i, differing only in the values of their respective parameters. The function with the higher likelihood, l_i, in Case 1 is in better accord with the observed t_i.

$P(0_i)$, and the contribution of a given right-censored observation at t_i to the overall likelihood L, increase as the area under the probability density distribution to the right of t_i increases. Thus, when maximizing the likelihood of right-censored observations, parameter values in the hazard function are favored that cause risk to manifest to the right of the censoring times. Note, however, that the distribution of that risk beyond the censoring time is unimportant and, in fact, can be of any arbitrary form. Only the *area* under the density function to the right of t_i is important, not the *shape* of the function. As a

[4] In some applications, the contributions of "marginal" outcomes to the likelihood are weighted by assigning such outcomes a fractional δ_i value, such as 0.1. A marginal event is the observation of an intrinsically graded response that is just sufficiently severe to be diagnosed as an "event" versus a "no-event".

result, a right-censored observation tells us nothing about the shape of the distribution of survival times. It only provides *location* information, indicating that the risk is beyond the right-censored time where observations were not made.

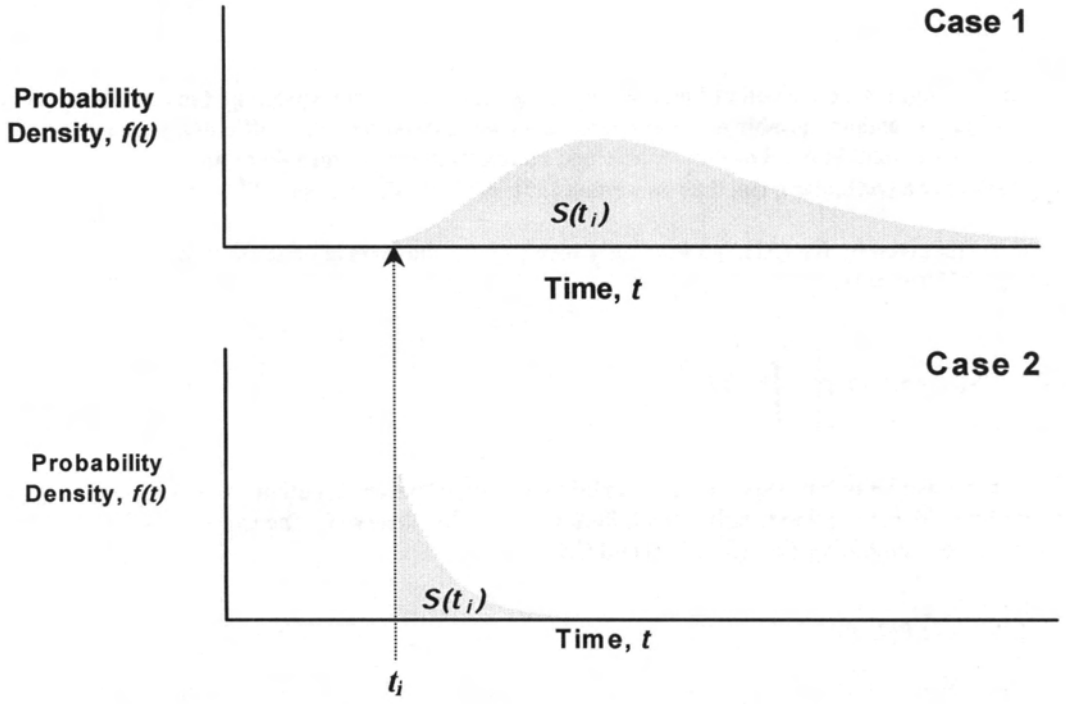


Figure 8. Probabilities of a right-censored observation at t_i as areas under different probability density distributions. Case 1: The area under the probability density function to the right of the right-censored time, t_i, is high, leading to a high $S(t_i)=P(0_i)$ and high l_i. Hazard function parameterizations that lead to such cases are favored in likelihood maximization over parameterizations such as those in Case 2, below. In Case 2, the area under the probability density function to the right of t_i is low, leading to a low $P(0_i)$ and low l_i for the same right-censored t_i. Note that only the area under $f(t)$ to the right of t_i is important, not the shape of $f(t)$ in this region.

4.1.2. Definition of $P(E_i)$ in the Likelihood

Three definitions of $P(E_i)$ have been used for dichotomous outcomes depending on the nature of the observed failure time:

1) Failure time is exact;
2) Failure time is interval-censored; i.e., only known to have occurred within some time interval between t_1 and t_2; $0 < t_1 \leq T < t_2$, or;
3) Failure time is unknown or not used, leading to "incidence-only" or binary quantal response assays.

In the following, we consider each of these definitions in some detail.

4.1.2.1. P(E_i) for an Exact Failure Time

If we assume the probability density function to be discrete, the probability of failure at a particular time t_i is given by *Eq.* (3.b):

$$P(E_i) = f(t_i)\,. \tag{32}$$

However, we generally assume the distribution functions to be continuous. Strictly speaking, the probability of any exact observation on a continuous random variable is always zero. In order to overcome this difficulty, such a probability is evaluated over an arbitrarily small interval of the variable and understood to have meaning only with respect to that interval. The probability of failure at a particular t_i can thus be written [17]: $F(t_i + 0)$-$F(t_i)$; where $F(t_i + 0) \equiv \lim_{\Delta t \to 0^+} F(t_i + \Delta t)$. This probability is equal to that given by *Eq.* (32), but implicitly incorporates the inverse of the limiting Δt.[5] *Eqs.* (20) and (22) then hold, so that *Eq.* (32) becomes:

$$P(E_i) = h(t_i)S(t_i) = h(t_i)\exp\left[-\int_0^{t_i} h(u)du\right] \tag{33}$$

This probability is very sensitive to the shape of the probability density or hazard function. When maximizing the likelihood, hazard functions are favored that maximize risk *when* failure was actually observed. The expression for the likelihood of N observations is obtained by combining *Eqs.* (30), (31) and (33):

$$\begin{aligned} \mathrm{L}(\boldsymbol{\beta}) &= \prod_{i=1}^{N} P(0_i)^{1-\delta_i}\, P(E_i)^{\delta_i} \\ &= \prod_{i=1}^{N}\left\{\exp\left[-\int_0^{t_i} h(t)dt\right]\right\}^{1-\delta_i} \cdot \left\{h(t_i)\cdot\exp\left[-\int_0^{t_i} h(t)dt\right]\right\}^{\delta_i} \end{aligned} \tag{34}$$

where δ_i is as defined in *Eq.* (29). Note that *Eq.* (34) simplifies to:

$$\mathrm{L}(\boldsymbol{\beta}) = \prod_{i=1}^{N}\left\{h(t_i)^{\delta_i}\cdot\exp\left[-\int_0^{t_i} h(t)dt\right]\right\}. \tag{35}$$

4.1.2.2. $P(E_i)$ for Interval-Censored Failure Time

An event known to occur only between two times, $0 < t_{1i} \le T < t_{2i}$, is actually the composite of two sub-events:

- Sub-event A_i: individual i remains event-free to t_{1i}, with probability $P(A_i)=P(T \ge t_{1i})$, and;
- Sub-event B_i: individual i experiences the event in the ensuing $[t_{1i};t_{2i})$ interval[6], with probability $P(B_i)=P(t_{1i} \le T < t_{2i})$.

$P(E_i)$ is thus the probability of joint occurrence of these two sub-events:

$$P(E_i) = P(A_i \cap B_i). \tag{36}$$

---

[5] This is shown in Appendix B using a discrete distribution with interval probabilities at t_i given by the continuous distribution.

[6] In interval notation, $[a;b)$ denotes a half-open interval consisting of all x such that $a \le x < b$.

However, because sub-event B_i can only occur with sub-event A_i, we have, as before, that

$$P(A_i \cap B_i) = P(B_i). \tag{37}$$

Therefore, $P(E_i)$ is the unconditional probability of sub-event B_i:

$$P(E_i) = P(B_i). \tag{38}$$

By reasoning analogous to that used to obtain *Eq.* (17), we obtain:

$$P(E_i) = P(t_{1i} \leq T < t_{2i}) = F(t_{2i}) - F(t_{1i}). \tag{39}$$

Applying the definition of $F(t)$ given in *Eq.* (5):

$$P(E_i) = \int_0^{t_{2i}} f(t)dt - \int_0^{t_{1i}} f(t)dt, \tag{40}$$

which simplifies to:

$$P(E_i) = \int_{t_{1i}}^{t_{2i}} f(t)dt. \tag{41}$$

The probability $P(E_i)$ can thus be considered directly in terms of the area under the probability density function, $f(t)$, between the t_{1i} and t_{2i} times, as illustrated in Figure 9.

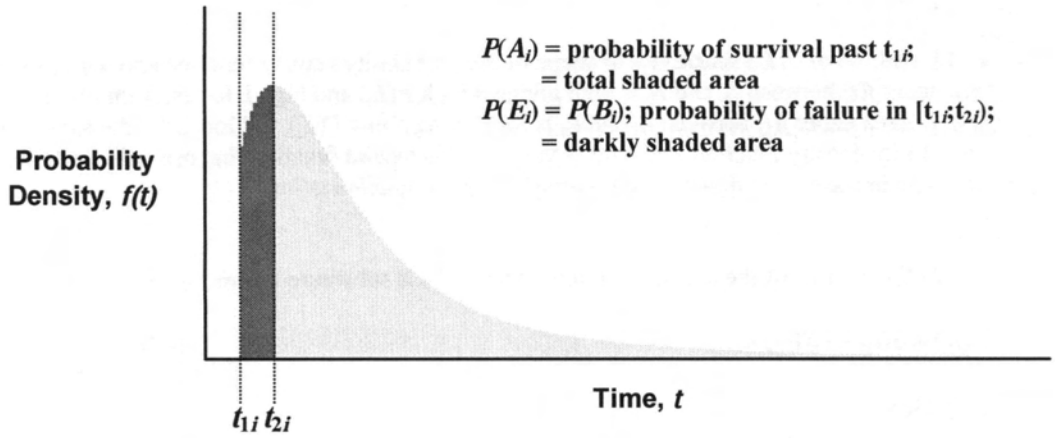


Figure 9. $P(E_i)$ for an event in the $[t_{1i};t_{2i})$ interval as area under the probability density function between t_{1i} and t_{2i}.

$P(E_i)$ is high only when the darkly shaded area is high. Figure 10 illustrates the sensitivity of $P(E_i)$ to the shape of the hazard function when the $[t_{1i};t_{2i})$ interval[7] is not too long. As a result, functions that maximize risk in the $[t_{1i};t_{2i})$ interval are favored in likelihood maximization.

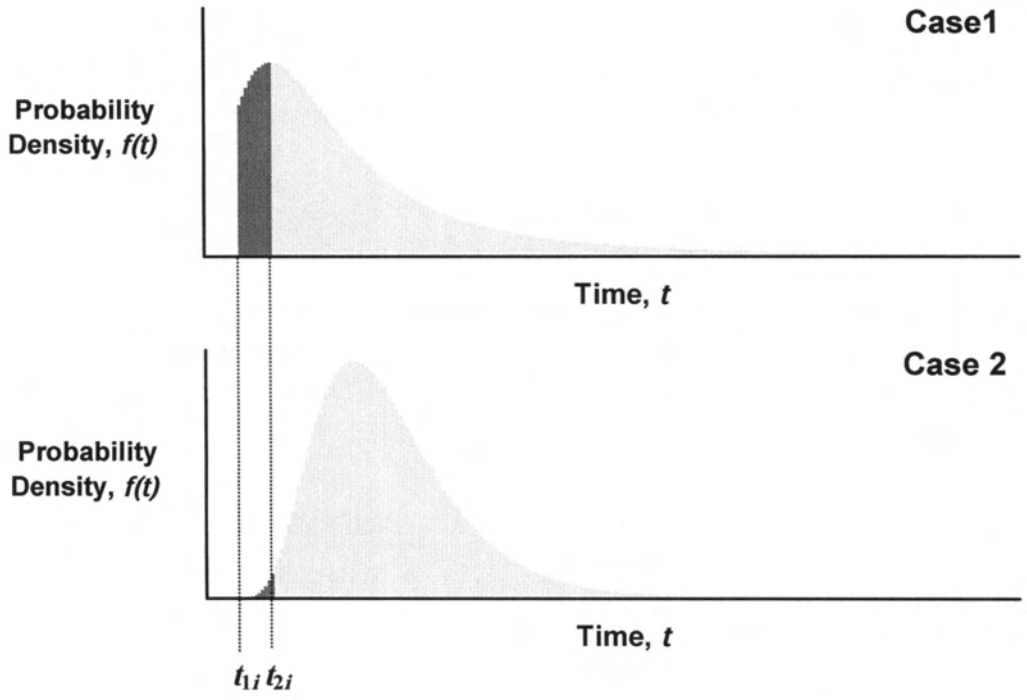


Figure 10. Illustration of $P(E_i)$ sensitivity to shape of the probability density function and hence to shape of $h(t)$. Case 1: Area under $f(t)$ between t_{1i} and t_{2i} is high giving a high $P(E_i)$ and high l_i for an event observed between t_{1i} and t_{2i}. Case 2: Area under $f(t)$ between t_{1i} and t_{2i} is low giving a low $P(E_i)$ and low l_i for the same event but different probability density function. Parameter values in the hazard function that drive the underlying $f(t)$ towards the shape in Case 1 are thus favored during likelihood maximization.

In order to express $P(E_i)$ in terms of the hazard function, $F(t)$=1-$S(t)$ is substituted from *Eq.* (8) into *Eq.* (39), yielding:

$$P(E_i) = S(t_{1i}) - S(t_{2i}), \tag{42}$$

which, using *Eq.* (22), yields:

$$P(E_i) = \exp\left[-\int_0^{t_{1i}} h(u)du\right] - \exp\left[-\int_0^{t_{2i}} h(u)du\right]. \tag{43}$$

The right side of *Eq.* (43) is factored to obtain our final result:

[7] The interval is written here as a half-open interval to remain consistent with the definition of the probability in *Eq.* (39). The distinction between half-open and closed has no effect on evaluation of the integral in *Eq.* (41).

$$P(E_i) = \left\{ \exp\left[-\int_0^{t_{1i}} h(u)du \right] \right\} \cdot \left\{ 1 - \exp\left[-\int_{t_{1i}}^{t_{2i}} h(u)du \right] \right\}. \tag{44}$$

Eq. (44) can also be obtained using the definition of conditional probability, which allows these analyses to be generalized to accommodate an arbitrarily large number of time intervals (See Appendix B).

The expression for the likelihood of N independent observations is obtained by combining *Eqs.* (30), (31) and (44):

$$\begin{aligned} L(\boldsymbol{\beta}) &= \prod_{i=1}^{N} P(0_i)^{1-\delta_i} P(E_i)^{\delta_i} \\ &= \prod_{i=1}^{N} \left\{ \exp\left[-\int_0^{t_i} h(u)du \right] \right\}^{1-\delta_i} \cdot \left\{ \left\{ \exp\left[-\int_0^{t_{1i}} h(u)du \right] \right\} \cdot \left\{ 1 - \exp\left[-\int_{t_{1i}}^{t_{2i}} h(u)du \right] \right\} \right\}^{\delta_i} \end{aligned} \tag{45}$$

where δ_i is as defined in *Eq.* (29). Note that t_i is the right-censored time $\neq t_{1i}$.

The mechanism by which t_{1i} and t_{2i} times are determined in any given study must be undertaken mindful of additional terms that are induced in the likelihood to account for dependence of the observations on the censoring mechanism. These terms cancel out of likelihood comparisons between models only if they are equal in the different likelihoods; i.e., only if the censoring remains noninformative (*c.f.*, Section 4.2).

4.1.2.3. $P(E_i)$ in Absence or Neglect of Failure Time Information: "Incidence-Only" Assay

Incidence-only assay can be viewed as a specialization of the interval-censored form of analysis in which there is only a single interval. In these analyses, the event is only known to have occurred or not between t=0 and an arbitrary observation time τ_i. Analyses based on incidence-only probabilities can be important in the evaluation of a fitted model, even if the model was fit using survival time information.

The probability of failure between t=0 and an arbitrary time τ_i is given by *Eq.* (5) as the integral of the probability density function over this period:

$$P(E_i) = \int_0^{\tau_i} f(u)du\,, \tag{46}$$

or, from *Eqs.* (8) and (22):

$$\begin{aligned} P(E_i) &= 1 - S(\tau_i) \\ &= 1 - \exp\left[-\int_0^{\tau_i} h(u)du \right]. \end{aligned} \tag{47}$$

The expression for the likelihood of N observations is obtained by combining Equations (30), (31) and (47):

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{N} P(0_i)^{1-\delta_i} P(E_i)^{\delta_i}$$

$$= \prod_{i=1}^{N} \left\{ \exp\left[-\int_0^{\tau_i} h(u)du \right] \right\}^{1-\delta_i} \cdot \left\{ 1 - \exp\left[-\int_0^{\tau_i} h(u)du \right] \right\}^{\delta_i} \tag{48}$$

where δ_i is as defined in *Eq.* (29). Note that this expression is identical to *Eq.* (45) with $t_i = \tau_i$ for right-censored observations, and $t_{1i} = 0$ and $t_{2i} = \tau_i$ for exact observations. Thus, *Eq.* (48) is a special case of *Eq.* (45).

As illustrated in Figure 11, incidence-only analyses are relatively *in*sensitive to the shape of the probability density function (or hazard function) when τ_i is so high that the area under any candidate function to the right of τ_i is low. Areas under the probability density functions to the left of the indicated τ_i in the illustration are very nearly equal, despite different function shapes. Thus, both probability density functions give a similar $P(E_i)$ value so that neither is favored in likelihood maximization.

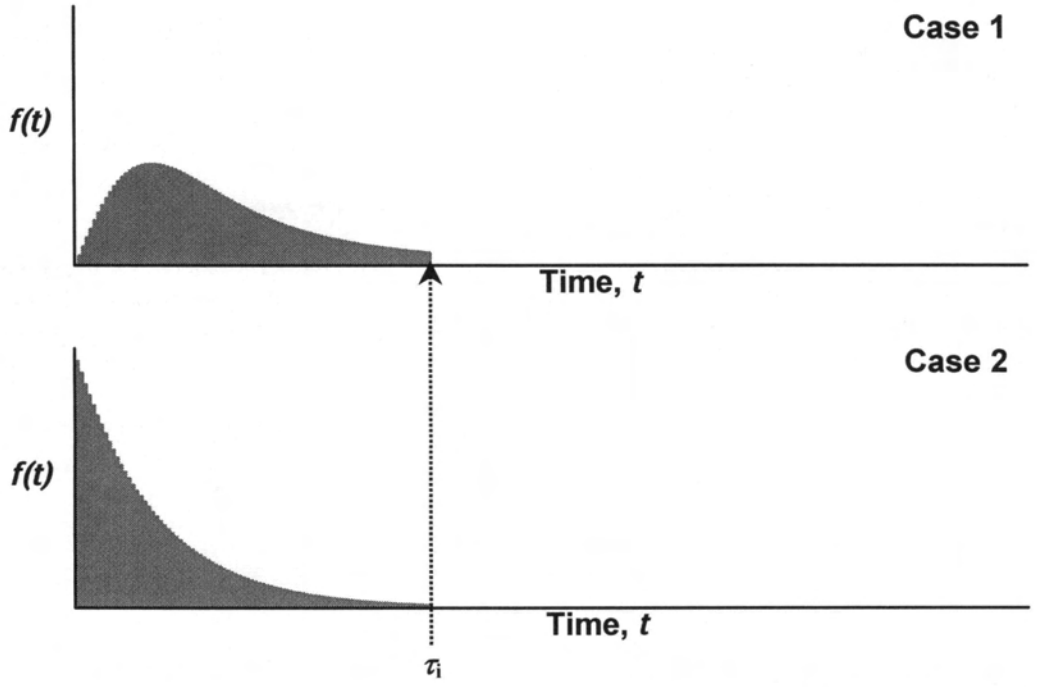


Figure 11. Illustration of $P(E_i)$ insensitivity to shape of the probability density function in "incidence-only" assays when the observation time, τ_i, is high.

However, the situation is much different if τ_i is low. Areas under candidate probability density functions to the left of τ_i, and hence results from optimization, can become dependent on the value of τ_i. The extreme example in the Figure 12 shows areas under the same probability density functions in the previous figure, but to the left of a much lower τ_i. The areas in the two illustrated cases, which were nearly the same at higher τ_i, are now very different.

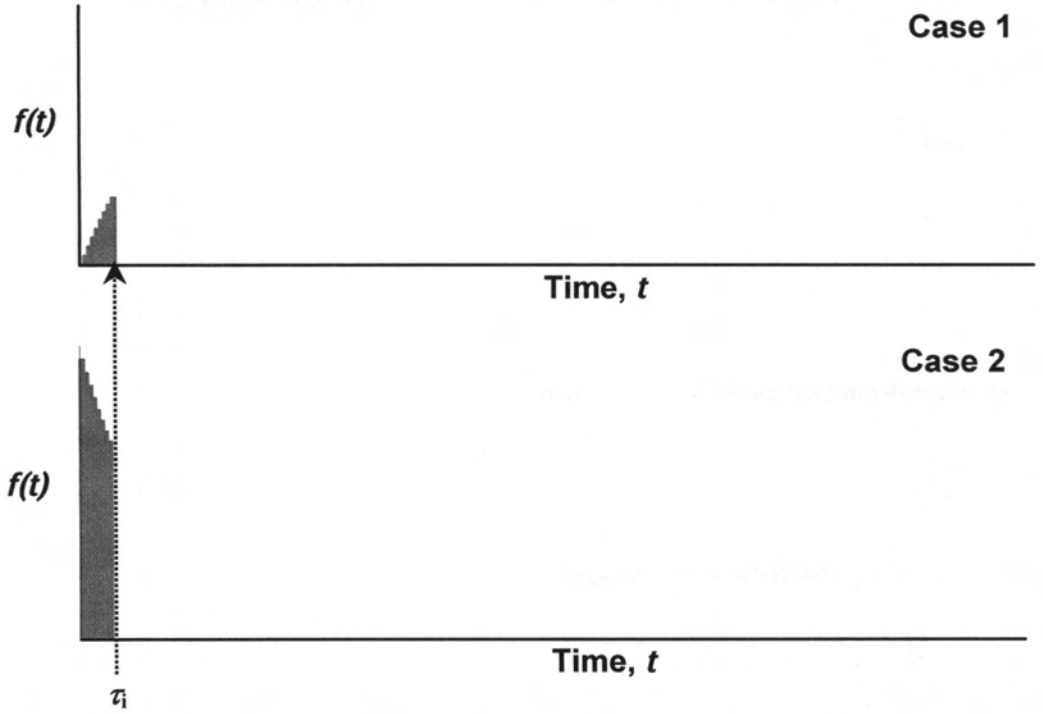


Figure 12. Illustration of $P(E_i)$ sensitivity to shape of the probability density function in "incidence-only" assays when the observation time, τ_i, is low.

Kalbfleisch and Prentice [17] show that this type of analysis is equivalent to classical quantal response assay under certain conditions. Note that in *Eq.* (46), the distribution of survival times, $f(t)$, with domain $0 \le t < \infty$, is integrated from 0 to τ_i in order to obtain the probability of failure between 0 and τ_i. More generally, the distribution of a response is given by a function $f(w)$ over the domain $-\infty \le w \le \infty$. Integration of this function over the range $-\infty$ to w then yields the probability of failure at w [7]:[8]

$$P(E_i) = \int_{-\infty}^{w} f(u)du\,. \tag{49}$$

Several of the classical distributions in survival analysis are based on a log-linear model of the survival time in which it is presumed that the response, y, is the logarithm of the survival time given by:

$$y = \ln t = \mu + \mathbf{z}\gamma + \sigma w\,, \tag{50}$$

where μ and σ are linear parameters, $\mathbf{z}$ is a vector of time-invariant independent variables (covariates), γ is a vector of parameters associated with $\mathbf{z}$, and as before w is a variable associated with the error distribution of y. Note that $\mu + \mathbf{z}\gamma$ gives

---

[8] The probability density function of survival times, $f(t)$, is readily obtained from $f(w)$: $f(t) = f(w)\left(\frac{dw}{dt}\right)$.

the deterministic component of y, while the σw term gives the random component of y. If a logistic distribution is assumed for w;

$$f(w) = \frac{e^w}{\left(1+e^w\right)^2}; \tag{51}$$

Eq. (49) is solved to yield:

$$P(E_i) = \frac{e^w}{1+e^w}. \tag{52}$$

This expression is the familiar logistic model for $P(E_i)$, where

$$w = \ln\left[\frac{P(E_i)}{1-P(E_i)}\right] \tag{53}$$

is the *logit* of $P(E_i)$. Solving *Eq.* (50) for w at $t = \tau_i$ yields

$$w = \frac{\ln \tau_i - \mu}{\sigma} - \frac{\mathbf{z}\gamma}{\sigma}. \tag{54}$$

If we define $\alpha' = \frac{\ln \tau_i - \mu}{\sigma}$ and $\boldsymbol{\beta} = -\frac{\gamma}{\sigma}$, then $w = \left(\alpha' + \mathbf{z}\boldsymbol{\beta}\right)$, which when substituted into *Eqs.* (49) and (52) yields:

$$P(E_i) = \int_{-\infty}^{\alpha'+\mathbf{z}\boldsymbol{\beta}} f(w)dw = \frac{\exp\left(\alpha' + \mathbf{z}\boldsymbol{\beta}\right)}{1+\exp\left(\alpha' + \mathbf{z}\boldsymbol{\beta}\right)}. \tag{55}$$

When τ_i values for all observations are the same or are so high that occurrence of the event beyond even the lowest τ_i has only negligible probability, $P(E_i)$ is independent of τ_i. Either time is omitted from the analysis altogether, or α' is a constant that folds into a constant term in $\mathbf{z}\boldsymbol{\beta}$. The outcome of interest is then simply whether the event occurs or does not occur in any given individual. This outcome y for a given $\mathbf{z}$ is either 0 (no event) or 1 (event), so that the expression for y in *Eq.* (50) is replaced by $y=\pi(\mathbf{z})+w$, where $\pi(\mathbf{z})$ is the probability of $y=1$ at $\mathbf{z}$. Because w can then assume only one of two possible values; $-\pi(\mathbf{z})$ if $y=0$ or $1-\pi(\mathbf{z})$ if $y=1$; w follows a binomial distribution with mean zero and variance equal to $\pi(\mathbf{z})[1-\pi(\mathbf{z})]$, and the analysis is simply a binary quantal response assay [16].

However, when the above conditions on the τ_i do not hold, results are dependent on the distribution of τ_i values in the data. If the analysis is pursued as a quantal response assay, τ_i emerges as an independent variable or factor in $w = \left(\alpha' + \mathbf{z}\boldsymbol{\beta}\right)$. At the same time, if we let $-\mu = \ln\lambda$ and $p = \sigma^{-1}$, *Eq.* (55) becomes:

$$P(E_i) = \frac{(\lambda\tau_i)^p \exp(\mathbf{z}\beta)}{1+(\lambda\tau_i)^p \exp(\mathbf{z}\beta)}. \tag{56}$$

This is the expression for the cumulative log-logistic distribution of survival time at a particular time τ_i. We therefore see that the analysis under these conditions is equivalent to a single-interval censored survival analysis ($t_{1i}=0$, $t_{2i}=\tau_i$) in which each τ_i serves as an effective survival time.

This type of analysis has been used to model occurrence of DCS during altitude exposure, with the factor τ_i taken as the planned time at altitude [3,24]. In these applications, τ_i is not a fixed observation time for all subjects under study and is

usually low enough for many observations that substantial DCS risk remains at $t > \tau_i$. Such analysis is survival analysis with a likelihood biased by the distribution of τ_i in the data.

It should be clear that, although illustrated using a log-linear model for T with a particular error distribution, the above conclusions are general. No assumptions were made that violate the fundamental presumption that event occurrence is arbitrarily distributed in time.

4.1.2.4. $P(E_i)$ with Competing Risks

In some survival problems, any one individual is able to fail only once, but in more than one way or by more than one mechanism. These problems are special cases of more general "competing risk" problems, and are analyzed using generalizations of the expressions presented above for problems in which only a single failure type is considered. Following a theoretical development presented by Kalbfleisch and Prentice [17], the overall hazard function, $h(t)$, or instantaneous conditional rate of failure by any type at time t is given as before by *Eq.* (20), which we reproduce following:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t | T \ge t)}{\Delta t}.$$

The instantaneous conditional rate of failure by a specific type j at time t in the presence of the other failure types is similarly defined:

$$h_j(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t, J = j | T \ge t)}{\Delta t}, \tag{57}$$

where J is a random variable for failure type and $P(t \le T < t + \Delta t, J = j | T \ge t)$ is the probability of failure by type j in the $[t;t+\Delta t)$ interval, given survival to time t while at risk for all types of possible failures. Invoking the definition of conditional probability in *Eq.* (13), with $P(A) = P(T \ge t)$ and $P(B)$ equal to the *un*conditional probability of failure by type j in the $[t;t+\Delta t)$ interval, *Eq.* (57) becomes

$$h_j(t) = \frac{\lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t, J = j)}{\Delta t}}{P(T \ge t)}. \tag{58}$$

The limit in the numerator of this expression is seen by reference to *Eq.* (3.a) to be the definition of $f_j(t)$, the partial probability function for failure type j. *Eq.* (58) can therefore be rewritten as:

$$h_j(t) = \frac{f_j(t)}{S(t)}. \tag{59}$$

Because any individual can fail only once and by only one of the possible failure types, the failure types are mutually exclusive. The overall (unconditional) instantaneous failure rate is therefore the sum of the type-specific failure rates:

$$f(t) = \sum_{j=1}^{m} f_j(t), \tag{60}$$

where m is the number of different possible failure types. It then follows with substitution of *Eq.* (59) that

$$f(t) = S(t) \cdot \sum_{j=1}^{m} h_j(t), \tag{61}$$

and from *Eq.* (20) that the overall hazard is the sum of the type-specific hazards:

$$h(t) = \sum_{j=1}^{m} h_j(t). \tag{62}$$

We then have from *Eq.* (22) that the overall survivor function at time t is

$$\begin{aligned} S(t) &= \exp\left[-\int_0^t \sum_{j=1}^{m} h_j(u)du \right] \\ &= \prod_{j=1}^{m} \exp\left[-\int_0^t h_j(u)du \right] \end{aligned} \tag{63}$$

It also follows from *Eq.* (60) that the overall cumulative probability of failure by all failure types at time t is the sum of the type-specific cumulative incidences, $I_j(t)$, at time t:

$$F(t) = \sum_{j=1}^{m} \left[\int_0^t f_j(u)du \right] = \sum_{j=1}^{m} I_j(t). \tag{64}$$

The individual $f_j(t)$ are thus seen to be improper distributions because the overall probability of failure cannot exceed unity.

If the i^{th} individual in a study fails by failure type J=j_i at time t_i, $P(E_i)$ for the individual is given by an analogue of *Eq.* (33):

$$P(E_i) = f_{j_i}(t_i) = S(t_i) \cdot h_{j_i}(t_i), \tag{65}$$

which, after substitution of *Eq.* (63), becomes:

$$P(E_i) = \left\{ \prod_{j=1}^{m} \exp\left[-\int_0^{t_i} h_j(u)du \right] \right\} \cdot h_{j_i}(t_i). \tag{66}$$

Using Eqs. (30), (31), (63) and (66), the likelihood for the i^{th} individual is seen to be given by an analog of the likelihood for the i^{th} individual in *Eq.* (35):

$$l_i(\boldsymbol{\beta}) = S(t_i) \cdot \left[h_{j_i}(t_i) \right]^{\delta_i}, \tag{67}$$

where the censoring indicator, δ_i, is as defined in *Eq.* (29); 1 for failure at t_i by any type and 0 for no failure through t_i. Thus, when failure occurs, $\delta_i = 1$ is singularly associated with the type of failure that occurs, j_i, while observations on the other failure types are effectively right-censored at t_i. When failure does not occur, observations on all failure types are right-censored and j_i does not enter the likelihood.

This type of analysis in its incidence-only form has been used in decompression sickness studies to examine the relationship between simulated gas bubble size and the observed profusion of ultrasonically-detectable central venous gas

bubbles [9]. As above, $P(E_i)$ for the i^{th} individual in these cases is governed by the partial density function for the type of failure that occurs, and is given by an analog of *Eq.* (46):

$$P(E_i) = \int_0^{\tau_i} f_{j_i}(u)du = I_{j_i}(\tau_i) . \tag{68}$$

The corresponding likelihood is

$$l_i(\boldsymbol{\beta}) = S(\tau_i)^{1-\delta_i} I_{j_i}(\tau_i)^{\delta_i} . \tag{69}$$

where, again, the censoring indicator, δ_i, is as defined in *Eq.* (29).

4.1.3. Combination of data with different definitions of $P(E_i)$ under a single model

Data may consist of individual exposures with different types of failure times. It is not uncommon, for example, to lack information about the times of occurrence of the event of interest for some exposures in which the event is known to have occurred, while having time of occurrence information for other such exposures. Provided that the same event constitutes failure in all of the exposures, and that appropriate information is available to define the covariates, these different data can be combined and analyzed under a single model. The overall likelihood for the "pooled" data, $L(\boldsymbol{\beta})$, is calculated in such cases by using *Eq.* (30) with the appropriate definition of the likelihood, l_i, for each of the individual exposures. Note that competing risk data generally cannot be combined with simpler binary response data because different events may constitute failure in competing risk data.

4.2. Fitting the Hazard Function: Likelihood Maximization

Likelihood maximization yields the highest probability of the observed data according to the specified hazard function and its parameters, β_k; k=1,2,..., p. Parameter values $\hat{\beta}_k$ that maximize the likelihood are the so-called *maximum likelihood estimates* of the true parameter values, β_k. The vector, $\hat{\boldsymbol{\beta}}$, of these $\hat{\beta}_k$ is called the *maximum likelihood estimator* (m.l.e.). In order to avoid numerical difficulties associated with extremely small numbers, it is convenient to work with the logarithm of L, or the log-likelihood (LL), rather than L. Because LL is a monotonically increasing function of L, maximizing LL maximizes L and yields the same values for the $\hat{\beta}_k$. Parameter values are systematically adjusted to maximize the log-likelihood of a model (hazard function) about a data set of exposures and observed survival times by solving the following p equations simultaneously:

$$U(\beta_k) = \frac{d \ln L(\boldsymbol{\beta})}{d\beta_k} = 0 , \qquad \text{for } k=1, 2, \ldots, p. \tag{70}$$

The solutions are usually obtained using well-described numerical techniques, such as the Newton-Raphson procedure [19] or a modified Marquardt algorithm [15,20].

The sum of the $U(\beta_k)$, $\sum_{k=1}^{p} U(\beta_k)$, is called the *total efficient score* statistic. The p-by-p matrix of second partial derivatives of the log-likelihood function is the *Hessian matrix*, $\mathbf{H}(\boldsymbol{\beta})$. When evaluated at the maximum likelihood estimator, the $(j,k)^{th}$ element of the Hessian matrix is given by:

$$\mathbf{H}(\hat{\beta}_j, \hat{\beta}_k) = \frac{\partial^2 \ln L(\hat{\boldsymbol{\beta}})}{\partial \beta_j \partial \beta_k}, \tag{71}$$

for j=1, 2, …, p, and k=1, 2, …, p. The *observed information matrix*, $\mathbf{I}(\hat{\boldsymbol{\beta}})$, is

$$\mathbf{I}(\hat{\boldsymbol{\beta}}) = -\mathbf{H}(\hat{\boldsymbol{\beta}}). \tag{72}$$

The variance-covariance matrix of the maximum likelihood estimator, written $\mathbf{cov}(\hat{\boldsymbol{\beta}})$, is then approximated by the inverse of $\mathbf{I}(\hat{\boldsymbol{\beta}})$;

$$\mathbf{cov}(\hat{\boldsymbol{\beta}}) = \mathbf{I}^{-1}(\hat{\boldsymbol{\beta}}). \tag{73}$$

It was noted earlier that the complete likelihood of a model on a data set containing right-censored observations includes factors that account for action of the censoring mechanism. These factors are omitted from the likelihood expressions in this overview. If the censoring scheme is deterministic, given the complete history of the study up to each censoring time, these factors equal unity and the total likelihoods equal those expressed here. The Type I and Type II censoring schemes outlined in Section 2.1 are in this category. However, if the censoring scheme is random, these factors contribute to the total likelihood. The likelihoods in the present expressions are then partial likelihoods that are proportional, not equal, to the total likelihoods. In such cases, the total likelihood is still maximized by solution of *Eq.* (70), and ensuing results still apply, as long as the censoring scheme, and hence its contributions to the total likelihood, does not depend on $\boldsymbol{\beta}$. Censoring schemes that meet this condition are independent of the failure mechanism, and like deterministic schemes, are said to be *noninformative.* Contributions of censoring mechanism to the complete likelihood are discussed in detail by Kalbfleisch and Prentice [17].

Under certain relatively mild conditions, the maximum likelihood estimates approach the true values of the parameters as the number of observations in the data set become infinitely large. $\hat{\boldsymbol{\beta}}$ is then the asymptotically unique solution to *Eq.* (70), and is multivariate normal with mean $\hat{\boldsymbol{\beta}}$ and variance-covariance matrix $\mathbf{I}^{-1}(\hat{\boldsymbol{\beta}})$. These asymptotic results form the basis for statistical inference about $\hat{\boldsymbol{\beta}}$. Conditions for their applicability include absence of isolated and extreme values of the model covariates in the data; i.e., as $N \to \infty$, the influence of any observation i on $\hat{\boldsymbol{\beta}}$ should vanish. Also, the components of $\boldsymbol{\beta}$ should not have unnecessary range restrictions that cause the likelihood to be maximized by parameter value(s) that are at the boundaries of their allowed ranges.

The likelihood surface is usually not concave over the ranges of the parameters, so that care must be taken to ensure that a likelihood maximum found by any given iterative procedure is in fact the global maximum, not a local maximum. This care is usually exercised by running the parameter optimization algorithm to completion from as many different combinations of initial parameter values selected from within the domains of the parameters as practicable. The time and tedium required by this process ultimately limit both the data set size and the numerical complexity of hazard functions that can be considered. In the end, one can only declare, not prove, that the global maximum has been achieved.

Finally, it should be noted from *Eq.* (30) that a model fails with a given $\boldsymbol{\beta}$ when, for any observation i in the calibration data, either $P(E_i)$=0 and δ_i=1, or $P(0_i)$=0 and δ_i =0. In such an instance, the likelihood of the observation, l_i, is zero, which propagates through the product of individual likelihoods to yield an overall likelihood, $L(\boldsymbol{\beta})$, of zero. (When working with the log likelihood, $\ln l_i$ is undefined in such a case.) Special provisions must be taken in the parameter estimation algorithm to trap and handle such instances to avoid frustrating and time-consuming run-time crashes. In the Marquardt or Newton-Raphson methods, such provisions include assigning them arbitrarily low non-zero likelihoods, which allows the algorithm to proceed to test an improved set of parameter values that may resolve the problem, or restarting the parameter estimation process from a new starting set of parameter values. Unresolved instances must be flagged for later assessment.

In summary, likelihood maximization yields the following analytic products:

- Maximum log likelihood achieved by the model on the data, LL_{max}.
- Vector of maximum likelihood estimates of the parameters, $\hat{\boldsymbol{\beta}}$.
- Variance-covariance matrix of the parameters, $\mathbf{cov}(\hat{\boldsymbol{\beta}})$.

4.3. Required Data Set Size: Meta-Analysis

Recall that right-censored observations provide no information about the shape of the distribution of survival times. It follows that determination of any one of the survival distribution functions depends only on the number of observed failures in the data. Such determination requires more failures as the complexity of the distribution increases. On the other hand, available data usually have low incidences of failure or relatively few events because hazardous exposures that would produce higher failure rates cannot usually be tested. All of these factors motivate a meta-analytic approach, which entails combination of data from many sources under a given model in order to accumulate a workable number of exposures in which the event of interest occurred.

Such pooling of available data requires careful consideration of biases that can arise due to sampling from different study populations and use of potentially different procedures, event/no-event criteria, censoring mechanisms, etc., in the different studies. These issues are covered in more detail by Dr. Paul Weathersby in a separate presentation in this Workshop [25].

5. Statistical Inference

All of these analyses are ultimately undertaken to reach conclusions about the population from which the model calibration data were drawn. Under the (usually assumed) asymptotic properties of $\hat{\beta}_k$, a given value of $\hat{\beta}_k$ determined from a sample of N observations is an estimate of the mean for the population. The precision of this determination would then be given by the standard deviation of means estimated from a large number of independent random samples of N individuals from the population. This standard deviation of means for a given parameter is the standard error of the parameter. However, we usually have only one sample to work with. The essence of statistical inference is to estimate standard errors and confidence intervals on the estimated parameters and model-estimated probabilities from properties of this single sample. Such inferences are made using the variance-covariance matrix of the maximum likelihood estimators.

5.1. Parameter Standard Errors

The square root of the $(k,k)^{th}$ diagonal element of $\mathbf{cov}(\hat{\boldsymbol{\beta}})$ approximates the standard error of the maximum likelihood estimate, $s.e.(\hat{\beta}_k)$, of the k^{th} adjustable parameter, $\hat{\beta}_k$:

$$s.e.(\hat{\beta}_k) = \mathbf{cov}(\hat{\beta}_k, \hat{\beta}_k)^{1/2}; \qquad k=1, 2, \ldots, p. \tag{74}$$

5.2. Confidence Intervals on the Parameters

A confidence interval on an estimated parameter is the interval about the estimate in which there is a prescribed probability $(1-\alpha)$ of including the true value of the parameter. If it is assumed that the estimate of interest is normally distributed with a mean value equal to the estimate, confidence intervals can be computed using the estimated standard error

of the estimate and percentage points of the standard normal distribution. The ±100(1-α)% confidence limits on $\hat{\beta}_k$ are then given by:

$$\hat{\beta}_k \pm \left[z_{\alpha/2} \cdot \text{s.e.}(\hat{\beta}_k) \right], \tag{75}$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ point of the standard normal distribution. At $z_{\alpha/2}$, the probability is $\alpha/2$ that the standard normal variable[9], Z, will have a value greater than $z_{\alpha/2}$; i.e., $P(Z > z_{\alpha/2}) = \alpha/2$. Similarly, the probability is $\alpha/2$ that Z will have a value less than $-z_{\alpha/2}$; i.e., $P(Z < -z_{\alpha/2}) = \alpha/2$. As shown in Figure 13, these probabilities are the areas under the standard normal distribution to the right of $z_{\alpha/2}$ and to the left of $-z_{\alpha/2}$, respectively. For α=0.05, for example, $z_{\alpha/2} = 1.96$, so the ±95% confidence interval on $\hat{\beta}_k$ is $\hat{\beta}_k \pm 1.96 \cdot \text{s.e.}(\hat{\beta}_k)$.

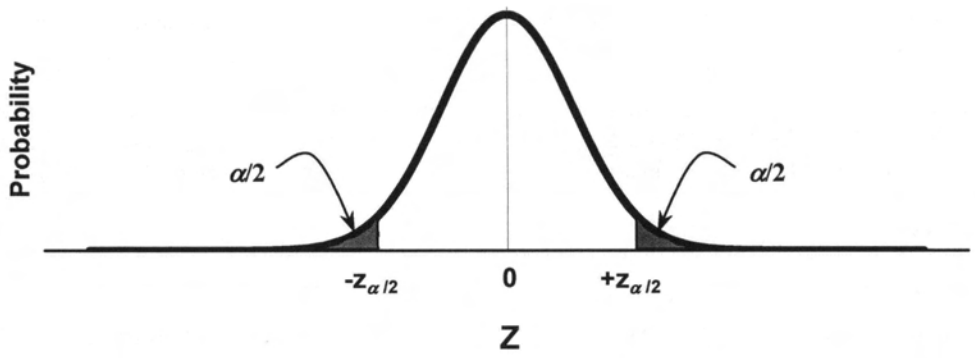


Figure 13. Position of the $z_{\alpha/2}$ percentage points in the distribution of the standard normal variate Z. Hatched area under the distribution to the right of $+z_{\alpha/2}$ is the probability $\alpha/2$ that Z will have a value greater than $z_{\alpha/2}$. Hatched area under the distribution to the left of $-z_{\alpha/2}$ is the probability $\alpha/2$ that Z will have a value less than $-z_{\alpha/2}$. Because the total area under the distribution is unity, the probability that Z will have a value between $-z_{\alpha/2}$ and $+z_{\alpha/2}$ is the remaining area under the distribution, 1-α.

5.3. Standard errors and confidence intervals on estimated probabilities, $\hat{F}(t)$ and $\hat{S}(t)$ [10]

Standard errors and confidence intervals on model-estimated probabilities are obtained through consideration of how errors in the parameters propagate through the model to influence model-estimated quantities. The propagation of errors formula is obtained from the multivariate Taylor series approximation to the variance of a function of p random variables. Elandt-Johnson and Johnson [4] show that if $g(\mathbf{X})$ is a twice-differentiable function of p continuous random variables, $\mathbf{X} = x_1, x_2, \ldots, x_p$, the variance of $g(\mathbf{X})$ is approximately

$$\text{var}[g(\mathbf{X})] = \sum_i^p \sum_j^p \left[\frac{\partial g(\mathbf{X})}{\partial x_i} \cdot \frac{\partial g(\mathbf{X})}{\partial x_j} \cdot \text{cov}(x_i, x_j) \right]. \tag{76}$$

[9] The standard normal variable follows a normal distribution with mean=0 and standard deviation=1.

[10] Because $\hat{F}(t)$ and $\hat{S}(t)$ only differ by a constant, their variances and standard errors are equal.

This expression was also presented by Ku [18]. After substitution of $\hat{F}(t)$ for $g(\mathbf{X})$, $\mathbf{X}=\hat{\boldsymbol{\beta}}$, and the standard error of $\hat{F}(t)$ is

$$s.e.[\hat{F}(t)] = \sqrt{\text{var}[\hat{F}(t)]}\,. \tag{77}$$

If it is assumed that $\hat{F}(t)$ is normally distributed about a mean equal to $\hat{F}(t)$, the $\pm 100(1-\alpha)\%$ confidence limits for $\hat{F}(t)$ are given by:

$$\hat{F}(t) \pm \left[z_{\alpha/2} \cdot s.e.[\hat{F}(t)]\right], \tag{78}$$

where, as above, $z_{\alpha/2}$ is the upper $\alpha/2$ point of the standard normal distribution.

Values of $\hat{F}(t)$ near 0 or 1 may not be normally distributed, so that incorrect assumption of normality in such cases can yield impossible values outside the permissible range [0;1]. Such values are avoided by transforming $\hat{F}(t)$ to a value in the unrestricted range $(-\infty, \infty)$, obtaining a confidence interval on the transform under the assumption that the transform is normally distributed over that range, and back-transforming the result to obtain the confidence interval on $\hat{F}(t)$. A convenient transform for these purposes is obtained from *Eq.* (76), which simplifies to the following if p=1:

$$\text{var}\left[g(\text{x})\right] = \left\{\frac{dg(\text{x})}{d\text{x}}\right\}^2 \cdot \text{var}\left[\text{x}\right]. \tag{79}$$

Substituting $g(\text{X})=\ln \text{x}$ and $\text{x} = \hat{S}(t)$, we then obtain

$$\text{var}\left[\ln \hat{S}(t)\right] = \frac{1}{\hat{S}^2(t)} \cdot \text{var}\left[\hat{S}(t)\right], \tag{80}$$

where $\text{var}\left[\hat{S}(t)\right]$ is given by *Eq.* (76). *Eq.* (80) is rearranged to obtain:

$$\text{var}\left[\hat{S}(t)\right] = \hat{S}^2(t) \cdot \text{var}\left[\ln \hat{S}(t)\right]. \tag{81}$$

Another substitution of $-\ln \hat{S}(t)$ for $\hat{S}(t)$ in *Eq.* (81) yields

$$\text{var}\left[-\ln \hat{S}(t)\right] = \left[\ln \hat{S}(t)\right]^2 \cdot \text{var}\left[\ln\left\{-\ln \hat{S}(t)\right\}\right]. \tag{82}$$

Noting that $\text{var}\left[-\ln \hat{S}(t)\right] = \text{var}\left[\ln \hat{S}(t)\right]$, *Eq.* (82) rearranges to:

$$\text{var}\left[\ln\left\{-\ln \hat{S}(t)\right\}\right] = \frac{\text{var}\left[\ln \hat{S}(t)\right]}{\left[\ln \hat{S}(t)\right]^2}. \tag{83}$$

Substituting the expression for $\text{var}\left[\ln \hat{S}(t)\right]$ from *Eq.* (80) then yields:

$$\mathrm{var}\left[\ln\left\{-\ln\hat{S}(t)\right\}\right]=\frac{\mathrm{var}\left[\hat{S}(t)\right]}{\hat{S}^2(t)\cdot\left[\ln\hat{S}(t)\right]^2}. \tag{84}$$

This final result gives the estimated variance of $\hat{v}(t)$, the log-log transform of $\hat{S}(t)$:

$$\hat{v}(t)=\left[\ln\left\{-\ln\hat{S}(t)\right\}\right]. \tag{85}$$

Note from *Eq.* (23) that $\hat{v}(t)$ is the logarithm of the maximum likelihood estimate of the cumulative hazard function, $H(t)$. Assuming that the distribution of $\hat{v}(t)$ is normal with mean equal to $\hat{v}(t)$, the $\pm 100(1-\alpha)\%$ confidence limits on $\hat{v}(t)$ are obtained in the usual fashion from the upper $\alpha/2$ point of the standard normal distribution. Thus, *Eq.* (85) is rewritten as:

$$\hat{v}(t)\pm z_{\alpha/2}\cdot\hat{\sigma}(t)=\left[\ln\left\{-\ln\hat{S}(t)\right\}\right]\pm z_{\alpha/2}\cdot\hat{\sigma}(t), \tag{86}$$

where the standard error of $\hat{v}(t)$, $\hat{\sigma}(t)$, is the square root of the variance of $\hat{v}(t)$ given by *Eq.* (84):

$$\hat{\sigma}(\mathrm{t})=\left\{\mathrm{var}\left[\hat{v}(t)\right]\right\}^{\frac{1}{2}}=\left\{\frac{\mathrm{var}\left[\hat{S}(t)\right]}{\hat{S}^2(t)\cdot\left[\ln\hat{S}(t)\right]^2}\right\}^{\frac{1}{2}}. \tag{87}$$

The confidence limits on $\hat{S}(t)$ are then obtained by back-transforming *Eq.* (86). Exponentiating, we obtain

$$\exp\left[\hat{v}(t)\pm z_{\alpha/2}\cdot\hat{\sigma}(t)\right]=-\ln\hat{S}(t)\cdot\exp\left(\pm z_{\alpha/2}\cdot\hat{\sigma}(t)\right),$$

the right side of which is rearranged using the identity $a\cdot\ln b=\ln b^a$ to obtain:

$$-\exp\left[\hat{v}(t)\pm z_{\alpha/2}\cdot\hat{\sigma}(t)\right]=\ln\left(\hat{S}(t)^{\exp\left(\pm z_{\alpha/2}\cdot\hat{\sigma}(t)\right)}\right)$$

and

$$\exp\left[-\exp^{\hat{v}(t)\pm z_{\alpha/2}\cdot\hat{\sigma}(t)}\right]=\hat{S}(t)^{\exp\left(\pm z_{\alpha/2}\cdot\hat{\sigma}(t)\right)}. \tag{88}$$

The ± 95% confidence limits on $\hat{S}(t)$ are thus given by:

$$\hat{S}(t)^{\exp[\pm 1.96\cdot\hat{\sigma}(\mathrm{t})]}. \tag{89}$$

Distributions of the log-log transform of selected values of $\hat{S}(t)$ are illustrated in Figure 14 for two different values of the standard error of the transform, $\hat{\sigma}(t)$.

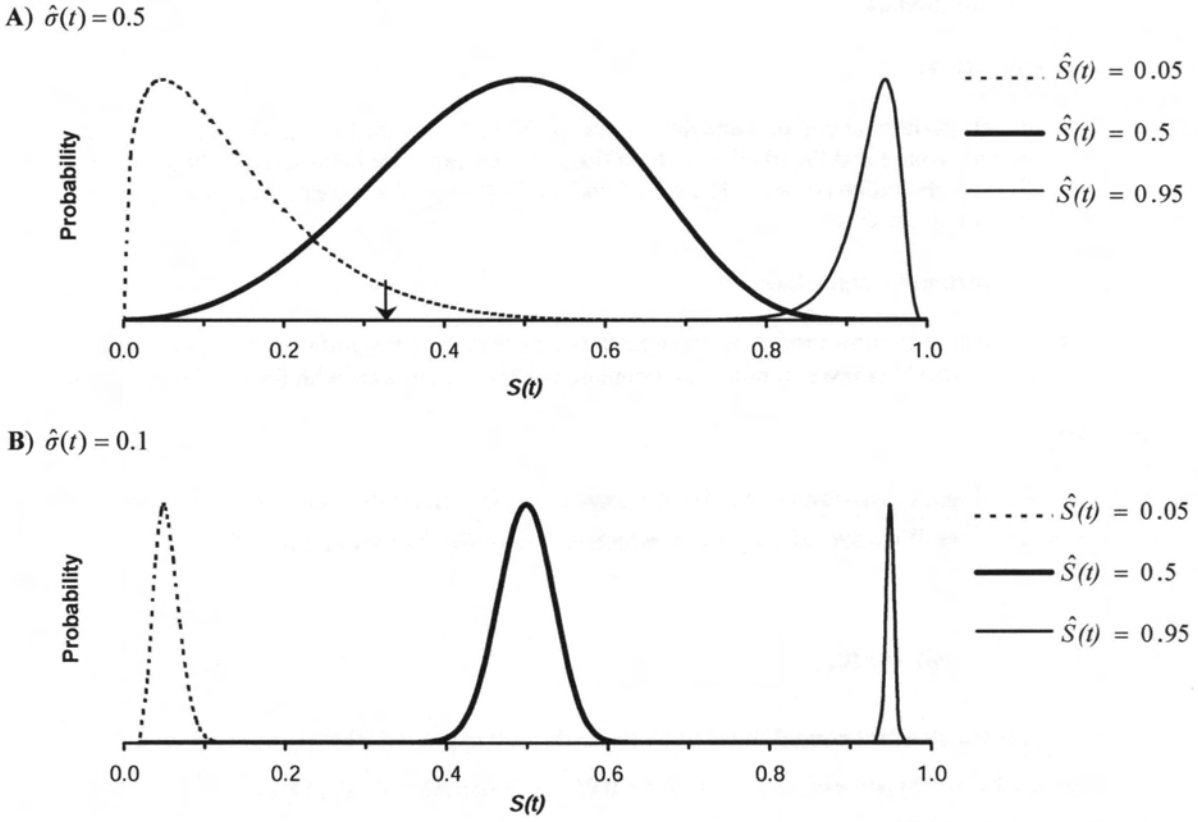


Figure 14. Distributions of the log-log transform of various values of $\hat{S}(t)$ for $\hat{\sigma}(t) = 0.5$ (top panel) and $\hat{\sigma}(t) = 0.1$ (lower panel). The upper 95% confidence limit on $\hat{S}(t) = 0.05$ is indicated by the arrow in the top panel. The corresponding lower 95% confidence limit, 0.0003, is graphically indistinguishable from 0. Note that the distribution for a value of $\hat{S}(t) < 0.5$ is not the mirror image of the distribution for $1 - \hat{S}(t)$. All distributions approach the normal distribution as $\hat{\sigma}(t)$ decreases, except at the limiting probabilities, $\hat{S}(t) = 0$ and $\hat{S}(t) = 1$, where the log-log transform of $\hat{S}(t)$ is undefined.

6. Goodness-of-Fit Assessment and Model Selection

The optimized parameter values, taken with the form of the hazard function, constitute the most important product of likelihood maximization. This product must be evaluated for its ability to actually reproduce the data to which it was fit. Such evaluation is undertaken by comparing different models, both informally and formally through tests of parameter significance, by comparing estimated and observed probability density functions, and by comparing incidence-only model predictions to observed incidences. Results of this evaluation are then used to select the "best-fitting" model from a collection of competing models. Although the selected model is considered to provide the best correlation of the calibration data, or to be most consistent with that data, its superiority in this regard cannot be construed as indication that it is the most *correct* of the models compared. For thorough discussions of this issue, see Oreskes, *et al.* [21] and Hilborn and Mangel [14].

6.1. Comparing different models.

6.1.1. Informal comparisons

How different models perform about the same data is often informally assessed by direct comparisons of model LL_{max} values. Such comparisons require that the likelihood definition for each exposure in the data be the same under the different models, except for different definitions of $h(t)$. The model with highest LL_{max} (or lowest $-LL_{max}$) is then concluded to provide the best correlation of the data.

6.1.2. Formal tests of parameter significance

Other formal statistical tests allow models with parameters that contribute insignificantly to goodness-of-fit to be rejected as "over-parameterized" in favor of more parsimonious models; i.e. models with fewer adjustable parameters, β_k.

6.1.2.1. Wald test

The Wald test is analogous to a t-test in analysis of variance, and is perhaps the simplest test of parameter significance. The Wald test is based on the Wald statistic, $Z_W(\hat{\beta}_k)$, which is defined for the maximium likelihood estimate of parameter β_k as:

$$Z_W(\hat{\beta}_k) = \hat{\beta}_k / s.e.(\hat{\beta}_k). \tag{90}$$

$Z_W(\hat{\beta}_k)$ has an asymptotic standard normal distribution under the null hypothesis that $\hat{\beta}_k = 0$. Thus, for example, $Z_W(\hat{\beta}_k) > 1.96$ rejects the null hypothesis of $\hat{\beta}_k = 0$ at $p < 0.05$. Alternatively, $Z_W^2(\hat{\beta}_k)$ is chi-square with one degree of freedom under the null hypothesis.

6.1.2.2. Likelihood Ratio Tests

The ratios of the likelihoods of model pairs are used to formally test significance of parameters added to one model to obtain another; e.g., to test significance of parameters added to a Null model.

6.1.2.2.1. Nested models, Likelihood ratio test

If Model (A) can be expressed as a reduced form of Model (B) by assigning zero values to $r \geq 1$ parameters in Model (B), Model (A) is said to be "nested" in Model (B). The significance of the r added parameters in Model (B); i.e. whether the r parameters in Model (B) afford significant improvement over Model (A), can be formally tested using a likelihood ratio test:

$$\chi^2 = -2 \ln \frac{L_{max}(\text{Model (A)})}{L_{max}(\text{Model (B)})} = -2 \ln \frac{L_{max}(\boldsymbol{\beta})}{L_{max}(\boldsymbol{\beta}; \boldsymbol{\alpha})} \; ; \; \text{d.f.} = r, \tag{91}$$

where $\boldsymbol{\alpha}$ is the vector of r parameters in Model (B) assigned zero values to obtain Model (A). The null hypothesis for this test, H_0, is that the r added parameters in Model (B) afford no significant improvement over Model (A), and that the elaboration of Model (A) obtained by addition of the r tested parameters is not statistically warranted. A high χ^2 motivates rejection of H_0; i.e., indicates that at least one nonzero component in $\boldsymbol{\alpha}$ provides a significantly improved model fit.

6.1.2.2.2. Nearly nested models, Approximate likelihood ratio test

If Model (A) is a form of Model (B) obtained by fixing $r \geq 1$ parameters in Model (B) at particular values, Model (A) is said to be "nearly nested" in Model (B). The null hypothesis, H_0, that the estimated values of the r parameters in Model B are not significantly different from their fixed H_0 values in Model A is tested using an approximate likelihood ratio test [4]:

$$\chi^2 \cong -2 \ln \frac{L_{max}(\text{Model (A)})}{L_{max}(\text{Model (B)})} = -2 \ln \frac{L_{max}(\boldsymbol{\beta}; \tilde{\boldsymbol{\alpha}})}{L_{max}(\boldsymbol{\beta}; \boldsymbol{\alpha})} \; ; \; \text{d.f.} = r \, , \tag{92}$$

where $\tilde{\boldsymbol{\alpha}}$ is the vector of r parameters in Model (B) assigned fixed values to obtain Model (A). A high χ^2 motivates rejection of H_0; i.e., indicates that at least one component of $\boldsymbol{\alpha}$ is significantly different from its fixed value in $\tilde{\boldsymbol{\alpha}}$.

6.1.3. Akaike Information Criterion (AIC)

Likelihood ratio tests allow discrimination between only two candidate models at a time, and require that one of the models in each pair be nested or nearly nested in the other. Selection among multiple models requires successive evaluations of the likelihood ratio in multiple hypothesis tests. Use of the Akaike information criterion, or AIC, reduces this process to a single statistical decision, and does not require that the candidate models be nested or nearly nested.[1]

The AIC for a given model is a measure of the discrepancy between the probability distribution estimated by that model and the true distribution, given by:

$$AIC = -2LL_{max} + 2p, \tag{93}$$

where LL_{max} is the log-likelihood of the model with its p optimized parameters. Note that by inclusion of the 2p term, the AIC contains an intrinsic penalization for increasing model complexity. The model with minimum AIC among a collection of candidate models is the model of choice.

6.2. Comparing estimated and observed probability density distributions

Construction of observed and model-estimated occurrence density distributions for relatively complex models of large and heterogeneous data has been well-described [12,23]. These distributions give the total observed and estimated numbers of failures per interval of time, allowing graphical assessment of a model's ability to correlate observed failure times as well as overall observed incidences of failure.

It is first ensured that subject times for all individual exposures in the data are defined with respect to a common reference zero time. In analyses of decompression sickness incidence and time of occurrence, for example, this reference time is conveniently chosen as the time at which the last decompression in an exposure is completed. The subject time is then arbitrarily divided into discrete time intervals on either side of this reference time, and the interval-censored form of $P(E)$ [*Eq.* (44)] is used to obtain an estimated probability of failure in each interval for each individual exposure in the data. The overall estimated number of failures in each interval is then the sum of the individual probabilities of failure in the interval. Comparison with the observed distribution provides a visual indication of model performance that illuminates the temporal properties of model predictions. Such comparisons can be made on the training data as a whole, on subsets of the training data to illuminate areas where model performance is weak, and on various validation data sets that were not included in the training data.

An example assessment of this type is given in Figure 15. While only failures appear in this example, division of each ordinate value by the total number of subjects in the data transforms each value into a probability, allowing examination of the temporal distributions of observed and estimated incidences of failure.

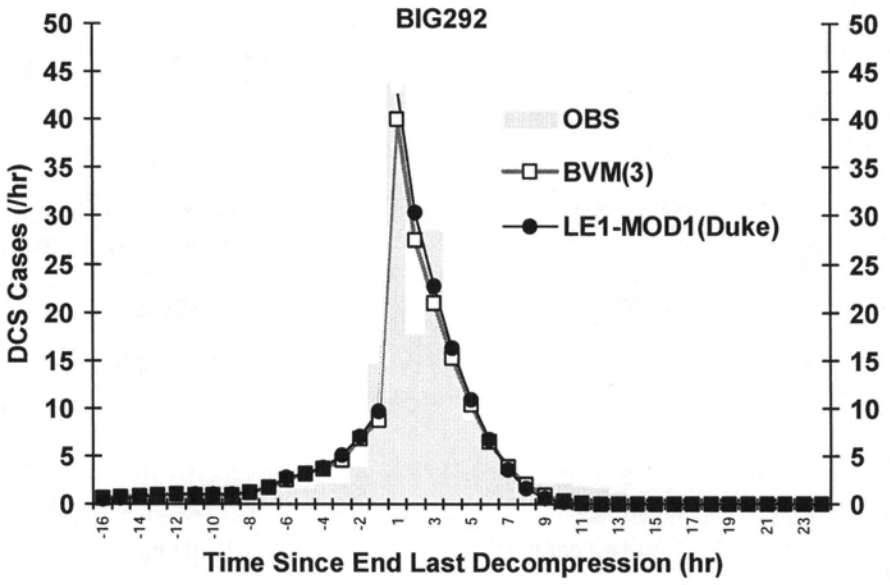


Figure 15. DCS occurrence density distributions estimated by two models of DCS occurrence compared to the observed distribution for the NMRI BIG292 data set of 3322 air and nitrox man-dives [11]. Lines between model-estimated points are drawn for clarity only.

6.3. Comparing incidence-only model predictions to observed incidences

Several forms of incidence-only analysis involve comparisons of observed incidences in various data sets to model-estimated incidences for the same data.

6.3.1. Quantitative: Chi-Square Tests

Chi-square tests can be used to provide both group-specific and global measures of fit.

6.3.1.1. Group-specific Chi-Square Tests

The observed number of events in a binomial experiment of N trials is $N\pi$, where π is the proportion of events observed in the N trials $\left(\pi = \frac{\#\,\text{events}}{N}\right)$. If π_e is the model-estimated or expected proportion of events in the experiment, the expected number of events is $N\pi_e$. Under the null hypothesis that $\pi_e = \pi$, the random variable Z, given by

$$Z = \frac{\text{observed mean} - \text{expected mean}}{\text{standard error of the mean}} = \frac{(N\pi - N\pi_e)}{\sqrt{N\pi_e(1-\pi_e)}}, \tag{94}$$

is distributed according to the standard normal distribution.

For the j[th] group of n_j individuals in a data set, the observed proportion of events is $\pi_j = \frac{\#\text{ events in group j}}{n_j}$, and the expected proportion of events is π_{ej}. The Pearson residual (PR_j) for the group is then the square of the Z statistic for the group, which follows a chi-square distribution with one degree of freedom under the null hypothesis that $\pi_{ej} = \pi_j$ [16]:

$$PR_j = Z_j^2 = \frac{(n_j\pi_j - n_j\pi_{ej})^2}{n_j\pi_{ej}(1-\pi_{ej})} \cong \chi_j^2 . \tag{95}$$

In terms of the censoring indices, δ_{i_j}, for the individuals in group j,

$$n_j = \sum_{i_j=1}^{n_j}(1-\delta_{i_j}) + \sum_{i_j=1}^{n_j}\delta_{i_j} = \eta_{0j} + \eta_{1j} = \sum_{\Delta=0}^{1}\eta_{\Delta j} ,$$

where i_j is the index for the i^{th} individual in the group, η_{0j} is the observed number of right-censored observations in the group, and η_{1j} is the observed number of event occurrences in the group. The index Δ is used here to represent binary outcome class 0 or 1. Omitting the group designation subscript for clarity, it follows from *Eq.* (95) that:

$$PR = \sum_{\Delta=0}^{1}\frac{(\eta_\Delta - \eta_{e\Delta})^2}{\eta_{e\Delta}}, \tag{96}$$

where $\eta_{e1} = n_j\pi_{ej}$ is the expected number of event occurrences in the group, and $\eta_{e0} = n_j(1-\pi_{ej})$ is the expected number of right-censored observations in the group. PR_j thus includes consideration of observed and expected numbers of both events and right-censored observations in the group.

The null hypothesis, H_0, that the estimated group incidence equals the observed group incidence can be tested using the Pearson residual as a chi-square (χ^2) statistic:

High $\chi^2 \Rightarrow$ rejection of H_0; estimate <u>not</u> consistent with observation;
model estimate for the group is unsatisfactory.
Low $\chi^2 \Rightarrow$ cannot reject H_0; estimate consistent with observation;
model estimate for the group is acceptable.

Examination of group-specific residuals is useful to identify areas in a data set over which a model performs well from those over which it performs poorly. Table 1, for example, shows the behavior of a model of altitude DCS incidence about its training data set of 1514 individual altitude exposures grouped according to the maximum altitude attained in each exposure. Note that the data are pooled from two different laboratory sources, and that some groups are defined with respect to single discrete maximum altitudes, while others are defined with respect to various ranges of maximum altitude. Pearson residuals for three groups in this data (enclosed in dotted lines in Table 1) stand out with particularly high values, identifying these groups as being only poorly handled by the model. Comparison of observed and estimated numbers of DCS cases indicates that the model underestimates the number of DCS cases in each of these groups.

Table 1. Residuals of a model of hypobaric DCS about its calibration data grouped according to maximum exposure altitude.

ALT_{max} (thousand ft)	# Exposures (n_j)	# DCS Cases OBS $n_j\pi_j$	EST $n_j\,\pi_{ej}$	95% C.I.	Pearson Residual $(n_j\,\pi_j - n_j\,\pi_{ej})^2/n_j\,\pi_{ej}(1-\pi_{ej})$
30.3*	291	52	55.440	(51.358 - 59.655)	0.264
30.0	111	74	46.794	(43.839 - 49.662)	*27.346*
29.5	322	149	143.261	(134.806 - 151.499)	0.414
27.5-27.6	98	72	39.049	(36.663 - 41.379)	*46.223*
25.5-25.0	176	74	52.029	(48.746 - 55.325)	*13.172*
22.8*	29	1	0.304	(0.277 - 0.334)	1.610
22.5	70	25	16.092	(15.012 - 17.192)	6.403
19.8-20.6	12	0	2.932	(2.709 - 3.162)	3.880
18.0-18.1	19	1	2.371	(2.124 - 2.621)	0.906
16.5	196	4	5.024	(4.243 - 5.959)	0.214
16.0	25	0	0.373	(0.329 - 0.404)	0.379
15.0	41	1	0.473	(0.437 - 0.529)	0.594
14.4	10	0	0.110	(0.101 - 0.122)	0.111
13.0	23	0	0.160	(0.147 - 0.178)	0.161
11.6	38	0	0.222	(0.197 - 0.240)	0.223
11.5	42	0	0.132	(0.123 - 0.154)	0.132
9.0-10.3	11	0	0.027	(0.022 - 0.027)	0.027
	1514	453	364.793	χ^2= (df=15)	102.060 p<<0.0001
With highlighted groups omitted:					
	1129	233	226.921	χ^2= (df=12)	15.319 p=0.2244

* Profiles from Laboratory A. (All others from Laboratory B.)

6.3.1.2. Global Chi-Square Tests

The sum of the Pearson Residuals over J groups in a data set is the summary chi-square statistic for the data set, with (J-2) degrees of freedom:

$$\chi^2 = \sum_{j=1}^{J} PR_j \,, \qquad (97)$$

where the PR_j are given by *Eq.* (95) or (96). In this case, the null hypothesis, H_0, is that the overall estimated incidence equals the overall observed incidence:

High $\chi^2 \Rightarrow$ rejection of H_0; estimate not consistent with observation;
model correlation of the entire data set is unsatisfactory.

Low $\chi^2 \Rightarrow$ cannot reject H_0; estimate consistent with observation;
model correlation of the entire data set is acceptable.

The global chi-square for our example correlation of hypobaric DCS data is shown in Table 1 for all of the groups. A global chi-square is also shown as obtained by omission of the three groups that are poorly fit by this model. The high global chi-square for all the groups corresponds to a p-value much less than 0.0001, motivating rejection of the null hypothesis that estimated and observed group incidences are the same. However, when the three poorly fit groups are omitted, the chi-square decreases to a corresponding p-value of 0.22. Model correlation of the restricted data would thus appear to be satisfactory.

The operation of grouping the data can significantly affect results of χ^2 tests, whether group-specific or global in type. Rather than attempt a quantitative illustration, a grouping effect for the above altitude DCS data is readily illustrated graphically. The illustration also provides an example of the last means of model evaluation that we will review here, namely graphical comparison of observed and estimated group incidences.

6.3.2. Qualitative: Graphical Comparisons

Observed *vs.* model-estimated group DCS incidences obtained from data in Table 1 are illustrated graphically in Figure 16. With the exception of the groups identified by Pearson Residual to be poorly fit by the model (indicated by arrows), the data visually conform well to the dashed identity line.

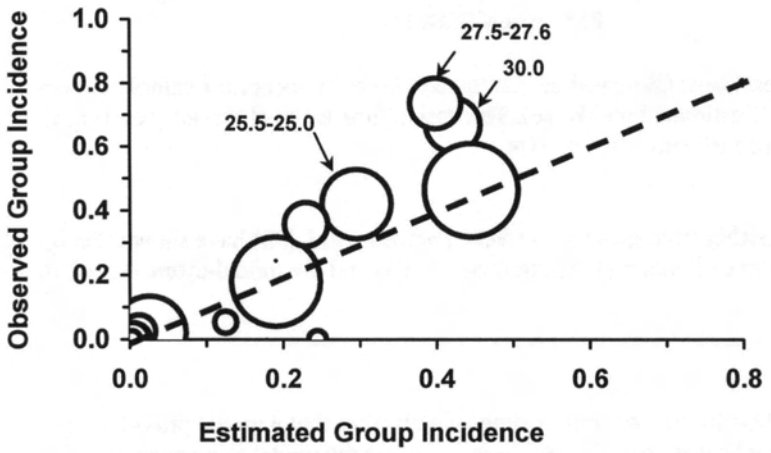


Figure 16. Graphical comparison of observed and estimated grouped DCS incidences, grouped by altitude. The dashed line is the observed = estimated identity line. Bubble area is proportional to group size. Altitude groups identified by Pearson Residual to be poorly fit by the model are indicated by arrows.

This satisfying situation deteriorates if the data are regrouped according to different criteria and redrawn, as shown in Figure 17. Here, the data are grouped by quintiles of estimated incidence. With estimated incidences in the data ranging from 0 to 70%, 14 groups emerge from the data under this grouping scheme.

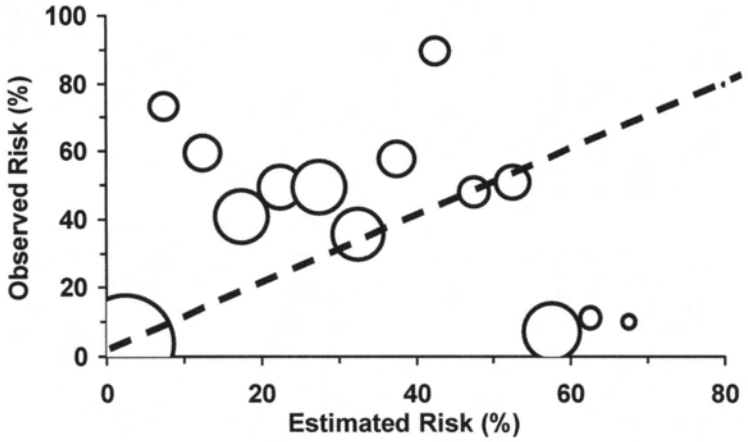


Figure 17. Graphical Comparisons: Observed *vs.* Estimated Group Incidences: Example Grouped by Quintiles (increments of 5% estimated incidence). The dashed line is the observed = estimated identity line. Bubble area is proportional to group size.

Notwithstanding difficulties arising from grouping effects, Parsons, *et al.* [22] have shown that the utility of these comparisons is limited once confidence limits on both the observations and the model-estimated incidences are considered.

7. Model Validation

All of the preceding work is descriptive. At best, it simply establishes that a model provides a good description of experience in a sample taken from a broader population. In order to use the model to manage risk in future exposures of individuals from that broader population, it must be shown that inferences about behavior in that population are validly made from model behavior in the sample. Such model "validation" is accomplished by using the above techniques to evaluate model goodness-of-fit to data other than that to which the model was fit, and by experimental trials, all involving samples from the same broad population of interest.

8. Acknowledgements

I am grateful to Dr. Edward D. Thalmann and Dr. R. S. Srinivasan for critically reviewing this manuscript. Their many suggestions for clarification and elaboration contributed materially to its final content and form.

9. Literature Cited

1. Akaike H. Information theory and an extension of the maximum likelihood principle. Petrov BN, Csaki F. 2nd International Symposium of Information Theory. Budapest: Akademiai Kiado, 1973: 267-81.

2. Collett D. Modelling Survival Data in Medical Research. London: Chapman and Hall, 1994.

3. Conkin J, Kumar KV, Powell MR, Foster PP, Waligora JM. A probabilistic model of hypobaric decompression sickness based on 66 chamber tests. Aviation, Space and Environmental Medicine 1996; 67:176-83.

4. Elandt-Johnson RC, Johnson NL. Survival Models and Data Analysis. New York, NY: Wiley, 1980.

5. Farewell VT. The use of mixture models for the analysis of survival data with long-term survivors. Biometrics 1982; 38:1041-6.

6. Farewell VT. Mixture models in survival analysis: Are they worth the risk? Canadian Journal of Statistics 1986; 14(3):257-62.

7. Finney DJ. Statistical Method in Biological Assay. Third edition. London: Charles Griffen and Company, Ltd., 1978.

8. Fishman GS. Monte Carlo: Concepts, Algorithms, and Applications. New York: Springer-Verlag New York, Inc., 1996.

9. Gault KA, Tikuisis P, Nishi R. Calibration of a bubble evolution model to observed bubble incidence in divers. Undersea and Hyperbaric Medicine 1995; 23(3):249-62.

10. Gerth WA. Decompression sickness during flying after diving: Motivation for mechanism. (This Workshop).

11. Gerth WA, Vann RD. Development of iso-DCS risk air and nitrox decompression tables using statistical bubble dynamics models.: National Oceanic and Atmospheric Administration, 1996; Final Report, NA46RU0505.

12. Gerth WA, Vann RD. Probabilistic gas and bubble dynamics models of DCS occurrence in air and N_2O_2 diving. Undersea and Hyperbaric Medicine 1997; 24(4):275-92.

13. Greenhouse JB, Silliman NP. Applications of a mixture survival model with covariates to the analysis of a depression prevention trial. Statistics in Medicine 1996; 15:2077-94.

14. Hilborn R, Mangel M. The Ecological Detective. Confronting Models with Data. Princeton, New Jersey: Princeton University Press, 1997.

15. Homer LD, Bailey RC. An analogy permitting maximum likelihood estimation by a simple modification of general least squares algorithms. Bethesda, MD, 1977; NMRI Tech Rep 77-55.

16. Hosmer DWJr, Lemeshow ST. Applied Logistic Regression. New York, NY: Wiley, 1989.

17. Kalbfleisch JD, Prentice RL. The Statistical Analysis of Failure Time Data. New York, NY: Wiley, 1980.

18. Ku HH. Notes on the use of propagation of error formulas. Journal of Research of the National Bureau of Standards - C. Engineering and Instrumentation 1966; 70C(4):263-73.

19. Lee ET. Statistical Methods for Survival Data Analysis. 2nd edition. New York, NY: John Wiley & Sons, Inc., 1992.

20. Marquardt DW. An algorithm for least-squares estimation of nonlinear parameters. Journal of the Society of Industrial Applied Mathematics 1963; 11:431-41.

21. Oreskes N, Shrader-Frechette K, Belitz K. Verification, validation, and confirmation of numerical models in the earth sciences. Science 1994; 263:641-6.

22. Parsons YJ, Weathersby PK, Survanshi SS, Flynn ET. Statistically Based Decompression Tables V: Haldane-Vann Models for Air Diving. Bethesda, MD, 1989; NMRI 89-34.

23. Thalmann ED, Parker EC, Survanshi SS, Weathersby PK. Improved probabilistic decompression model risk

predictions using linear-exponential kinetics. Undersea and Hyperbaric Medicine 1997; 24(4):255-74.

24. Van Liew HD, Burkard ME, Conkin J. Testing of hypotheses about altitude decompression sickness by statistical analyses. Undersea and Hyperbaric Medicine 1996; 23(4):225-33.

25. Weathersby PK. Meta-analysis of diver decompression data. (This Workshop).

Appendix A. Surviving Fraction and Improper Distributions

Some individuals in a population under study may not experience the event of interest regardless of how long they are observed. The existence of such a "surviving fraction" is accommodated by adoption of an improper density distribution. A surviving fraction in the population can also be accommodated explicitly by presuming that the population under study is a mixture of two subpopulations; one subject to failure according to a given failure time distribution, $f'(t)$, and the other not subject to such failure. In such a mixture model, there is a probability $P(\mathrm{S})$ that a given individual is a member of the susceptible subpopulation (in state S), and a probability $P(I)$ that the individual is a member of the other immune subpopulation (in state I) [5,6,13]. Thus,

$$P(\mathrm{S}) + P(I) = 1 . \qquad \text{(A.1)}$$

It is important to note that the presence of an immune subpopulation cannot be observed. It can only be inferred if many of the largest observations are right-censored. Under the supposition of *Eq.* (A.1), only individuals in the susceptible subpopulation can experience occurrence of the event. The probabilities in *Eq.* (1) thus become conditional on membership in the susceptible subpopulation:

$$P(0 \mid \mathrm{S}) + P(E \mid \mathrm{S}) = 1 , \qquad \text{(A.2)}$$

where $P(0|\mathrm{S})$ is the probability of no-event given membership in the susceptible subpopulation, and $P(E|\mathrm{S})$ is the probability of an event given membership in the susceptible subpopulation. These conditional probabilities also replace their unconditional counterparts in *Eqs.* (4)-(8), while $f(t)$ in *Eq.* (9) is replaced by $f'(t)$, the probability density distribution of $P(E|\mathrm{S})$. Note now that $f'(t)$ applies only to the susceptible subpopulation, all members of which will experience occurrence of the event if observed for a sufficiently long time. In general, the underlying density distribution $g(t)$ that determines $P(\mathrm{S})$ may be different from the distribution governing failure time, $f'(t)$, in the susceptible sub-population. Here, however, we assume that the processes governing occurrence of immunes are the same as those that govern occurrence of susceptibles in the population, and use the failure time density distribution as defined by *Eq.* (10) to determine $P(\mathrm{S})$ and $P(I)$ associated with the two sub-populations. Thus,

$$P(\mathrm{S}) = \int_0^\infty f_a(u)du , \qquad \text{(A.3)}$$

and

$$P(I) = \int_0^\infty f_b(u)du . \qquad \text{(A.4)}$$

which are constants for a given set of covariate values. Note that $g(t) = f_a(t)$ in this "degenerate " case.

We can now use the definition of conditional probability in *Eq.* (13) to show that $f'(t)$ in the mixture model is completely specified by $f_a(t)$ in the non-mixture model. We have for $P(E|\mathrm{S})$:

$$P(E \mid \mathrm{S}) = \frac{P(E \cap \mathrm{S})}{P(\mathrm{S})} = \frac{P(E)}{P(\mathrm{S})} , \qquad \text{(A.6)}$$

where we have noted that $P(E \cap S)=P(E)$ because the event can only occur in individuals that are members of the susceptible subpopulation. The unconditional probability $P(E)$ is obtained by rearranging *Eq.* (A.6):

$$P(E) = P(S) \cdot P(E \mid S) \tag{A.7}$$

The unconditional probability $P(0)$ is obtained by eliminating $P(E|S)$ and $P(E)$ in *Eq.* (A.7) using *Eqs.* (1) and (A.2):

$$1 - P(0) = P(S) \cdot [1 - P(0 \mid S)],$$

which rearranges to:

$$P(0) = P(S) \cdot P(0 \mid S) + [1 - P(S)]. \tag{A.8}$$

Finally, $P(E)$ and $P(0)$ must be the same for the mixture model as for the non-mixture model. Thus, for $t<T_r$, application of *Eqs.* (A.6) and (5) yields

$$P(E \mid S) = \int_0^t f'(x)dx = K \int_0^t f_a(u)du, \tag{A.9}$$

where $K = 1/P(S)$. For $t \geq T_r$, $P(E|S) = 1$ and $P(E) = P(S)$. The mixture model explicitly accounts for a "surviving fraction" of the population that never experiences occurrence of the event, and does so without having to divide the density function at an arbitrary T_r as in the non-mixture model. However, *Eq.* (A.9) shows that the density function, $f'(t)$, in the mixture model differs from the improper or "truncated" density function, $f_a(t)$, in the non-mixture model only by a proportionality factor K that is seen from *Eq.* (A.3) to be also completely specified by $f_a(t)$. Therefore, in the degenerate case with $g(t) = f_a(t)$, the mixture model provides no insights into the problem beyond those illuminated by the non-mixture model.

A mixture model may be useful when it is relatively certain that immune individuals actually exist in a population and that the processes governing the distribution of individuals between immune and susceptible groups differ from those governing failure in the susceptible individuals; i.e., when $g(t) \neq f_a(t)$. This is not generally the case in applications reviewed in this Workshop, where as a result, problems are handled using $f_a(t)$ in *Eq.* (10). The latter approach affords the added advantage that separate solutions for $P(S)$ need not be made as they must to use *Eqs.* (A.7) – (A.9) in mixture models.

Appendix B. Likelihood Construction

The likelihood for interval-censored failure times in Section 4.1.2.2. can be generalized to accommodate an arbitrary number of time intervals. As the intervals become infinitesimally small, the analysis becomes identical to one using *Eq.* (35) in which discrete failure times are used, illustrating that the latter is a special case of the interval-censored type of analysis.

We first stipulate that observations can be made on an individual only at discrete times; t_j, j=1,…, m; separated by intervals $\Delta t_j = t_j - t_{j-1}$, although the distribution that gives rise to the observations is presumed to be continuous. When there are both censored survival times and failures at a particular t_j, censorings are assumed to occur after failures in order to avoid ambiguity about which individuals remain subject to failure at t_j. We then note that an individual's progress through time in a survival study is a Markov process. As a result, the probability of an outcome in the j^{th} interval ending at t_j is dependent on the complete history of the individual up to time t_{j-1}, but only through the individual's status at time t_{j-1}, independent of his or her status at t_{j-2}, …, t_0. In the present context, the probability of an outcome in the j^{th} interval is thus conditional only on survival of the individual up to time t_{j-1}. The contribution of the i^{th} individual to the overall likelihood can thus be constructed as the product of the conditional probabilities of an outcome in each of the intervals [8,17]:

$$\mathrm{l}_i = P(t_0) \cdot \prod_{j=1}^{\varepsilon_i} \left[P\left(outcome_{ij} \middle| T_i \geq t_{j-1}\right)^{\alpha_{ij}} \right], \qquad \text{(B.1)}$$

where it is usually assumed that $P(t_0)$=1, and the censoring index α_{ij} is defined such that:

$\alpha_{ij} = 1$ if outcome$_{ij}$ is failure in the j^{th} interval, and;

$\alpha_{ij} = 0$ if outcome$_{ij}$ is survival through the j^{th} interval.

The upper index, ε_i, in *Eq.* (B.1) is the interval ($1 \leq \varepsilon_i \leq \mathrm{m}$) in which the individual is either observed to fail or, having survived through intervals 1 through ε_i, the interval in which the individual is last subject to failure while under study.

If the individual fails in the j^{th} interval, α_{ij}=1, and we have that:

$$P(outcome_{ij} \mid T_i \geq t_{j-1}) = P\left(t_{j-1} \leq T_i < t_j \middle| T_i \geq t_{j-1}\right), \qquad \alpha_{ij}=1 \qquad \text{(B.2)}$$

Note that this expression is the definition of the hazard at t_{j-1}, $h(t_{j-1})$, for a discrete distribution. Using the definition of conditional probability, *Eq.* (B.2) becomes:

$$P\left(t_{j-1} \leq T_i < t_j \middle| T_i \geq t_{j-1}\right) = \frac{P(t_{j-1} \leq T_i < t_j \cap T_i \geq t_{j-1})}{P(T_i \geq t_{j-1})}. \qquad \text{(B.3)}$$

Because failure in the j^{th} interval cannot occur without survival to time t_{j-1}, the numerator in *Eq.* (B.3) is the unconditional probability of failure in the j^{th} interval, $P(E_{ij})$, given by:

$$P\left(t_{j-1} \leq T_i < t_j \cap T_i \geq t_{j-1}\right) = P(t_{j-1} \leq T_i < t_j) = P(E_{ij}). \qquad \text{(B.4)}$$

Thus, using *Eq.* (42), we have:

$$P(E_{ij}) = S(t_{j-1}) - S(t_j). \qquad \text{(B.5)}$$

The denominator in *Eq.* (B.3) is also recognized as the probability of survival to t_{j-1} or longer, $S(t_{j-1})$. *Eq.* (B.3) can thus be rewritten:

$$P\left(t_{j-1} \leq T_{\mathrm{i}} < t_j \middle| T_{\mathrm{i}} \geq t_{j-1}\right) = \frac{S(t_{j-1}) - S(t_j)}{S(t_{j-1})} = 1 - \frac{S(t_j)}{S(t_{j-1})}. \qquad \text{(B.6)}$$

Recalling that the data have arisen from a continuous distribution, expressions for $S(t_{j-1})$ and $S(t_j)$ from *Eq.* (22) can be substituted to yield the following, after rearrangement:

$$P\left(t_{j-1} \leq T_i < t_j \middle| T_i \geq t_{j-1}\right) = 1 - \exp\left[-\int_{t_{j-1}}^{t_j} h(u)du \right]. \qquad \text{(B.7)}$$

If the individual survives through the j^{th} interval, α_{ij}=0, and we have that:

$$P\left(outcome_{ij}\middle|T_i \geq t_{j-1}\right) = P\left(T_i \geq t_j \middle| T_i \geq t_{j-1}\right) = \exp\left[-\int_{t_{j-1}}^{t_j} h(u)du\right]; \qquad \alpha_{ij} = 0. \qquad \text{(B.8)}$$

Using *Eqs.* (B.7) and (B.8), *Eq.* (B.1) is expressed:

$$l_i = \prod_{j=1}^{\varepsilon_i} l_{ij}, \qquad \text{(B.9.a)}$$

where

$$l_{ij} = \left\{\exp\left[-\int_{t_{j-1}}^{t_j} h(u)du\right]\right\}^{1-\alpha_{ij}} \cdot \left\{1-\exp\left[-\int_{t_{j-1}}^{t_j} h(u)du\right]\right\}^{\alpha_{ij}}. \qquad \text{(B.9.b)}$$

If the i^{th} individual fails in the x^{th} interval ($x \leq m$), the censoring index α_{ij} is zero for all intervals j=1, …, x-1, and unity for interval x (α_{ix}=1). Thus, $\varepsilon_i = x$, $\sum_{j=1}^{x} \alpha_{ij} = \delta_i = 1$, and *Eq.* (B.9.a) becomes:

$$l_i = \left\{\left\{\exp\left[-\int_{0}^{t_{x-1}} h(u)du\right]\right\} \cdot \left\{1-\exp\left[-\int_{t_{x-1}}^{t_x} h(u)du\right]\right\}\right\}^{\delta_i}; \qquad \delta_i = 1. \qquad \text{(B.10)}$$

However, if the i^{th} individual is right-censored in all intervals through the y^{th} interval ($y \leq m$) and is then removed from the study, the censoring index α_{ij} is zero for all intervals j=1, …, y. In these cases, $\varepsilon_i = y$, $\sum_{j=1}^{y} \alpha_{ij} = \delta_i = 0$, and *Eq.* (B.9.a) becomes:

$$l_i = \left\{\exp\left[-\int_{0}^{t_y} h(u)du\right]\right\}^{1-\delta_i}; \qquad \delta_i = 0. \qquad \text{(B.11)}$$

An expression essentially identical to *Eq.* (45) in Section 4.1.2.2. is obtained for the overall likelihood of N independent interval censored observations when *Eqs.* (B.10) and (B.11) are combined using *Eq.* (30).

It should be clear from *Eq.* (B.4) and *Eq.* (3.b) in Section 2.2.1 that as the intervals Δt_j are arbitrarily shortened with concomitant increases in the number of observation times, m, $t_{x-1} \rightarrow t_x$ and

$$P(E_{ix}) = \lim_{\Delta t \rightarrow 0} P\left(t_{x-1} \leq T_i < t_x\right) = P(T_i = t_x) = f(t_x), \quad \text{where } \Delta t = t_x - t_{x-1}.$$

Thus, referring to *Eqs.* (B.5) through (B.7), *Eq.* (B.10) approaches

$$l_i = \left\{S(t_x) \cdot \frac{P(E_{ix})}{S(t_x)}\right\}^{\delta_i} = f(t_x)^{\delta_i}. \qquad \text{(B.12a)}$$

If the Δt interval is sufficiently small so that $\int_{t_{x-1}}^{t_x} h(u)du \approx h(t_x)\Delta t << 1$ in the underlying continuous distribution, then by expansion, $1 - \exp\left[-\int_{t_{x-1}}^{t_x} h(u)du\right] \approx h(t_x)\Delta t$, which equals the dimensionless $h(t_x)$ in the discrete approximation of the continuous distribution. Therefore, *Eq.* (B.10) also approaches

$$l_i = [S(t_x) \cdot h(t_x)]^{\delta_i} . \qquad \text{(B.12b)}$$

(Note that the same result is obtained if we consider the interval from $t_{x-1} = t$ to $t_x = t+\Delta t$ as $\Delta t \to 0$. Also, the Δt factor remains with $h(t_x)$ in the strictly continuous case, but is omitted with no loss of rigor because an arbitrarily close discrete approximation can always be made.) Finally, *Eq.* (B.11) becomes

$$l_i = S(t_y)^{1-\delta_i} . \qquad \text{(B.13)}$$

Eqs. (B.12) and (B.13) are then combined using *Eq.* (30) to yield *Eq.* (35) of Section 4.1.2.1, the expression for the overall likelihood of N exact and right-censored observations. Discrete failure time data are thus seen to arise from a continuous distribution as a special case of interval censoring. Indeed, any set of observations arising from a continuous distribution will be discrete due to the limited precision of observation and yield a discrete distribution of survival times due to the inevitably finite size of the sample.

10. Glossary of Symbols

α	significance level or "p-value" in a significance test, equal to probability of committing a Type I error in the test
α'	transform of $\ln \tau_i$ in the logistic model for $P(E_i)$
$\boldsymbol{\alpha}$	vector of r parameters tested for significance in a likelihood ratio or approximate likelihood ratio test
$\tilde{\boldsymbol{\alpha}}$	Vector of null hypothesis values of parameters in $\boldsymbol{\alpha}$ tested in an approximate likelihood ratio test
α_{ij}	censoring index for the i^{th} individual at end of the j^{th} interval of a multiple interval-censored problem
AIC	Akaike Information Criterion
β_k	k^{th} parameter in the hazard function, $h(t)$
$\hat{\beta}_k$	maximum likelihood estimate of k^{th} parameter in the hazard function, $h(t)$
$\boldsymbol{\beta}$	vector of p parameters in the hazard function, $h(t)$
$\hat{\boldsymbol{\beta}}$	maximum likelihood estimators; vector of p maximum likelihood estimates of parameters in $\boldsymbol{\beta}$
$\mathbf{cov}(\hat{\boldsymbol{\beta}})$	variance-covariance matrix of $\hat{\boldsymbol{\beta}}$, the maximum likelihood estimators
$\mathbf{cov}(\hat{\boldsymbol{\beta}}_j, \hat{\boldsymbol{\beta}}_k)$	$(j,k)^{th}$ element in the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$
Δ	index for binary outcome class, 0 or 1
δ_i	censoring or outcome variable for individual i
ε_i	interval in which the i^{th} individual is either observed to fail or last subject to failure while under study
$f(t)$	probability density distribution of survival times

$f_j(t)$	partial probability density function for outcome j in a competing risks problem
$F(t)$	cumulative distribution function
$\hat{F}(t)$	maximum likelihood estimate of the cumulative distribution function
$\boldsymbol{\gamma}$	vector of parameters associated with $\mathbf{z}$
H_0	null hypothesis in a significance test
$\mathbf{H}(\hat{\boldsymbol{\beta}})$	Hessian matrix
$H(t)$	cumulative hazard function
$h(t)$	hazard function
$h(t;\mathbf{z})$	hazard function written to express explicit dependence on independent variables in vector $\mathbf{z}$
$h_j(t)$	partial hazard function for outcome j in a competing risks problem
$I_j(t)$	partial cumulative distribution function for outcome j in a competing risks problem
$\mathbf{I}(\hat{\boldsymbol{\beta}})$	observed information matrix
i	index for the i^{th} individual in the overall sample of N individuals
i_j	index for the i^{th} individual in group j
J	variable for outcome type in a competing risks problem, or variable for group in an incidence-only analysis (usage clear from context)
j	value of J for a particular outcome in a competing risks problem, or value of J for a particular group in an incidence-only analysis (usage clear from context)
j	index for possible failure time in discrete failure time or multiple interval-censored problems
k	index for parameters
l_i	likelihood of an observation made on individual i
l_{ij}	likelihood of an observed outcome on individual i in interval j of a multiple interval-censored problem
L	likelihood of N independent observations
LL_{max}	logarithm of the maximum likelihood of N independent observations
λ	constant hazard in the exponential distribution, or transform of μ in the log-logistic distribution
m	number of possible failure types in a competing risks problem
m	number of discrete times at which an observation can be made on an individual
μ	linear parameter in log-linear model of y = logarithm of the survival time
N	number of individuals in sample
n_j	number of individuals in group j
η_Δ	observed number of observations in a group with outcome Δ
$\eta_{e\Delta}$	expected number of observations in a group with outcome Δ
π	observed proportion of events in a binomial experiment

π_e	expected proportion of events in a binomial experiment
$\pi(\mathbf{z})$	probability of a response for the covariates in $\mathbf{z}$
p	exponent in the log-logistic distribution equal to the inverse of σ
p	number of elements in the β parameter vector
$P(0_i)$	probability of a right-censored observation (i.e., No-Event) on individual i
$P(E_i)$	probability of failure (i.e., of event E) in individual i
$P(A\|B)$	conditional probability of event A given occurrence of event B
PR_j	Pearson Residual for group j
r	number of parameters tested in a likelihood ratio or approximate likelihood ratio test
$S(t)$	survivor function
$\hat{S}(t)$	maximum likelihood estimate of the survivor function
σ	linear parameter in log-linear model of y = logarithm of the survival time
$\hat{\sigma}(t)$	standard error of $\hat{v}(t)$
t	particular value of the survival time, T
t_1	interval start time for single interval-censored survival time
t_2	interval end time for single interval-censored survival time
T	survival time variable
T_r	arbitrarily high value of the survival time, above which no individual under study will fail
τ	arbitrary time at which outcome is assessed in an incidence-only problem
$U(\beta_k)$	efficient score statistic for parameter β_k
u	dummy variable of integration
var[$g(x)$]	variance of the estimate of $g(x)$
$\hat{v}(t)$	log-log transform of $\hat{S}(t)$
w	variable for the error distribution of y
χ^2	Chi-square variate
x	interval in which an individual is observed to fail in a multiple interval-censored problem
y	last right-censored interval for observations on an individual in which the event of interest is never observed in a multiple interval-censored problem
y	transformed value of a response in a survival experiment
Z	standard normal variate
$Z_W(\hat{\beta}_k)$	Wald statistic for maximum likelihood estimate of parameter β_k
$\mathbf{z}$	vector of independent variables or covariates
$z_{\alpha/2}$	upper $\alpha/2$ point of the standard normal distribution

QUESTION: I have a question about meta-analysis. When data from different sources are used, you have to assume that the model describes the data sets in the same way. In other words, you must assume that only one model is valid for all data sets, is that true?

DR. GERTH: Yes.

QUESTION: Is there any way of discerning whether the separate data sets may really require their own separate models?

DR. GERTH: Yes, there is. One can put categorical dummy variables in the model that indicate to which subset of the overall fitted data a given model prediction applies. Any group- or site-specific dependence of the response is thus extracted into the coefficients of these dummy variables. The statistical significance of such dependence is then assessed by evaluating the significance of these coefficients.

The trouble with this approach is that the resultant model cannot then be used to make a general prediction except for a particular group or site.

But the question you raise is very important because decisions about data combinability under a single model are central to any meta-analytic undertaking. Inserting dummy variables to examine whether or not there are significant inter-group or inter-site differences is one way of examining whether or not the data are combinable under a single model.

DR. WEATHERSBY: There will be a presentation this afternoon that will address some of these issues in more detail.

Another question?

QUESTION: John Simms from the Submarine Research Lab. Could you briefly comment on the benefits, or the advantages and disadvantages, of using likelihood models versus other probabilistic models, such as Kaplan-Meier and Cox models?

DR. GERTH: Let me take up the models you mention in reverse order. In speaking of "Cox" models, I'll presume you mean the Cox proportional hazards model. This model is a so-called non-parametric model that can indeed be used to examine the influences of a large number of independent variables on the response.

A Cox model presumes some sort of baseline hazard that we don't always know. So, we don't have that in hand when we want to make predictions. A Cox proportional hazards model is great if your principal objectives are to describe data that you have in hand and identify important factors. However, after having finished the descriptive and factor identification aspects of the work, you don't emerge with a model that can be used to make predictions to prescribe future behavior. That requirement of our enterprise conditions a lot of what we do. It forces us to use parametric models.

Now to take up your other question about advantages or disadvantages of Kaplan-Meier survival curves and, I presume, why we do not use them. A Kaplan-Meier curve is another non-parametric approach to survival data that provides a way to view the response to only a single kind of exposure at a time. For example, we can take a population of individuals, hit them with one kind of exposure, and then record their subsequent survival experience. That experience can be viewed in terms of one Kaplan-Meier survival curve. If we then change the nature of the exposure by altering the values of the independent variables in some way, survival experience after this exposure would be viewed in terms of another Kaplan-Meier survival curve. The data in our work consists of survival experience after a large variety of different exposures. No single Kaplan-Meier curve can be used to describe all of this data.

In the approach we take, we can get values of the probability density function and derive the other functions, as I illustrated for the exponential distribution early on. The shape of any one of these distributions observed in a given data set can be illustrated and compared to the model-estimated shape for the same data set. In this fashion, illustrations of model behavior over a broad range of independent variable values can be obtained similar to those provided for only a single set of independent variable values by the Kaplan-Meier procedure.

QUESTION: Dr. Flook, Great Britain. Model parameter values determined using maximum likelihood techniques are sometimes quite different from those that have been measured in cases where we can measure the parameters. How does the model deal with that? Is it telling us that these parameters are unimportant?

DR. GERTH: The issue of whether a parameter is statistically important in a model; in other words, that its inclusion in the model provides a significant improvement of fit over the case when it is not included; is tested with a likelihood ratio test. The importance of a parameter in this sense, its statistical significance, is a separate issue from whether or not the value of the parameter makes sense in terms of the mechanistic or physiological model that might have motivated the form of the hazard function.

In a very real sense, likelihood maximization turns a mechanistic model into something else. The conformance of that "something else" to the original concept informs us about whether or not the original concept was correct. A parameter

may be statistically significant, but its value may differ considerably from that expected in the context of the model itself. For example, I might have a parameter for gas solubility in a mechanistic model that emerges from fitting to data with a value that does not make sense. No known substance may have a gas solubility of value even near the value we get from fitting. Such a result tells us that our concept of the hazard function is not correct. However, as nonsensical as the value might be, a finding that the parameter is statistically significant indicates that it is required in the hazard function -- with its nonsensical value -- in order for the function to fit the data well.

It is important to note that a fitted mechanistic model remains useful for making predictions even if it contains one or more parameters with values that do not make sense. The nonsensical parameter values simply indicate that the theory which motivated formulation of the hazard function needs to be further refined in order be a more complete and correct representation of the actual risk-governing processes.

Oh yeah. Rule Number 1 for modelers is that when the model doesn't work, it's never the model's fault. It must be some fault in the data.

NMRI Models of Human CNS Oxygen Toxicity

Paul Weathersby
Gales Ferry, CT

This presentation contains previously published material of Dr. Andrea Harabin of the Naval Medical Research Institute, NMRI. The organization of the presentation follows the list of questions circulated among the speakers before the Workshop.

Data

The full data set consists of 688 human exposures performed at the Navy Experimental Diving Unit. A number of smaller studies, mostly published by Butler and colleagues from 1979-86, were combined, since no single study was large enough for much of an analysis. Subjects were all immersed and exercising. That feature needs mentioning, since a prior analysis demonstrated that both immersion and exercise are significant risk factors for oxygen toxicity (1).

Only 23 different profiles were studied: about half were with a single continuous level of PO_2, while the others employed a sequence of 2 to 5 different levels. Subjects were breathing nearly 100% oxygen, so the independent variable used was actual ambient pressure.

What constitutes oxygen toxicity? The most dangerous - and unmistakable - effect of high-O_2 breathing is a grand mal seizure. Other less specific signs and symptoms occur, as shown in the following table:

Symptoms

1. Nausea
2. Irritability, dyspnea, sleepiness, dysphoria
3. Headache
4. Numbness, tingling
5. Dizziness, vertigo
6. Twitch
7. Hearing disturbance
8. Visual disturbance
9. Unconsciousness, aphasia
10. Convulsion

Various authors have attempted to define toxicity based on "any" symptom, or on "severe types of symptoms" - with less than fully satisfactory results. The present study used the sign and symptom criteria displayed real-time by the original investigators: exposure-stopping severity. There were 42 such cases among the 688 exposures.

Models

All models considered were hazard/risk rate formulations:

$$P\left(OxTox_{CNS}\right) = 1 - \exp\left(- \int_0^t risk(u)du \right)$$

As shown in Gerth's review in this volume, these risk functions have a number of excellent mathematical properties, and are especially useful when some amount of biological or other mechanistic information is available to guide the analysis. Three specific models will be described:

Model 0 applied a constant hazard. It has a single parameter, k, which ignores oxygen levels entirely. Its utility here is simply to provide a statistical reference: if we cannot find a another model that offers a substantial improvement in fitting success, we should quit.

$risk = k$

Model 1 is more descriptive, but not very physiological.

$$risk = \mathbf{a}\,(P_{O2} - \mathbf{Thr})^{\,b}$$

This formulation has 3 empirical parameters: a scale factor, ***a***, -roughly equivalent to the single parameter in Model 0; a threshold oxygen level, ***Thr***, below which no risk whatsoever is encountered; and an exponent, ***b***, allowing cumulative risk to build faster (or slower) than linear in oxygen.

Model 2 is considerably more complicated. We refer to it as an "autocatalytic model", since it embodies a positive-feedback feature. The model is conceived as following a putative toxic substance, ***X***:

$$risk = X(t) - Thr_X \qquad \text{,where}$$

$$dX/dt = a\,P_{O2} + k\,(P_{O2} - Pcrit)\,X(t)$$

The model has 4 estimated parameters: a scale factor, ***a;*** a baseline rate constant, ***k***; an oxygen level that overwhelms natural de-toxicification of ***X***, ***Pcrit***; and a "safety limit" on substance ***X***, $\mathbf{Thr}_X$. Note that once the exposure oxygen rises above the critical pressure, the rate constant for the toxic substance ***X*** actually reverses sign; so that the toxicant's level can rise quickly to extremely high values. This feature was necessary to describe a set of laboratory animal oxygen exposures designed to test different types of intermittent exposures (2). Simple first-order kinetics *do not* lead to a prediction that intermittency can provide any benefit .

Model Fitting

All models were fit using the Marquart algorithm to achieve a maximum likelihood solution. As with other complicated environmental modeling cases, multiple local maxima in the likelihood surface were found. Numerous sets of starting parameter values were needed to assure finding a global maximum. The final parameter covariance matrix was used to propagate error of all subsequent predictions, in keeping with standard methods.

Results

Parameters from all models are presented in the original paper (3). To keep the focus on methodology, only the log likelihood values are presented below:

0.	Constant hazard	1 parameter	-LL = 327.0
1.	$risk = \mathbf{a}\,(PO_2 - \mathbf{Thr})^{\,b}$	3 parameters	-LL = 301.4
2.	autocatalytic	4 parameters	-LL = 299.6

Both **Models 1** and **2** performed better than constant hazard rate (no oxygen dependence) by a likelihood ratio test ($p<.001$ for both). Improvement by 25 or so LL units with only a couple of parameters is a big deal in likelihood modeling. Therefore, both models have a noteworthy ability to fit the data. The apparent improvement of 1.8 LL units by the autocatalytic **Model 2** vs **Model 1** would be borderline for significance if the two models were in a subset relationship. However, they are not, and we have no direct statistically useful means of comparing the two successful models.

Both **Models 1** and **2** appeared to describe individual exposure profile safety. For example, on the 90-min exposure at 30 fsw depth, the observed outcome was 1 bad event in 40 subjects. That raw rate is 2.5%, with 95% binomial sampling confidence limits covering 0.1 to 5%. **Model 2's** prediction for this profile (propagated 95% confidence) was an incidence of 2 to 5%. **Model 1** had a very similar prediction range. Such an overlap of prediction and observed confidence limits occurred throughout the 23 profiles in the data. So there is little internal evidence of model failure.

Both models also performed with comparable success in separating exposures by level of risk, that is, in capturing the dose-response relationship. All predictions were ranked, then split into three groups by predicted risk. Predictions in each group were totaled, as were the total number of observed symptoms per group. The results for **Model 2** are presented here:

Risk level	*N total*	*N symptoms Obs.*	*N symptoms Pred.*	
0 - 5 %	272	5	7.2	
5.1 - 8 %	208	15	15.5	
8.3 - 15 %	208	22	20.8	Chi-square = 0.78 (p>.5)

Do we have any reason for choosing between the two "successful" models?

Application

Models 1 and **2** both appeared to describe the data well. Are they equal? We have a strong external reason for making a choice. The autocatalytic model fits an important animal data set better than the type of simple empirical function of **Model 1** (2). That animal experiment is important because it had carefully selected intermittent exposures that provided a major challenge in finding useful descriptive kinetics.

Models 1 and **2** differ in a specific testable way in exposures like those shown in Figure 1. In both of the exposures shown (solid lines in upper and lower panels) the "simple" oxygen exposure is the same. That is, both have the same total time at 50 fsw and at 20 fsw. But note in the top panel that the 50 fsw time is in a single block of time, while in the bottom panel the 50 fsw time is interrupted by a 20 fsw interval. (Parameter Pcrit for **Model 2** is between 20 and 50 fsw of pure O_2.) Note from the cumulative risk plots (dotted lines) that the interrupted profile only encounters half the total risk of the uninterrupted profile. On the other hand, similar calculations with **Model 1** predict an identical total risk for these two profiles.

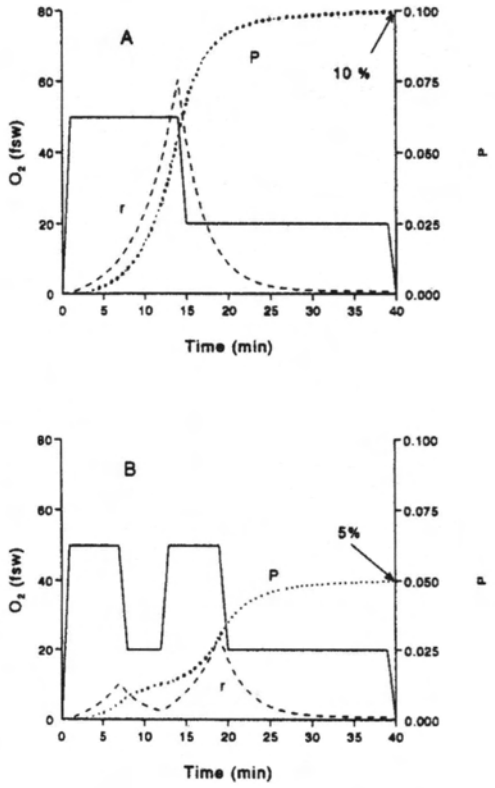


Figure 1. Effect of intermittent exposure on predicted CNS O_2 toxicity. (A) Predicted failure rate, r(t), and cumulative probability, P, for an exposure to 50 fsw for 14 min followed by 25 min at 20 fsw. (B) Predicted failure rate and cumulative probability for the same exposure as in A, but with the 50 fsw portion interrupted half-way by a 5 min excursion to 20 fsw. In each panel, r(t) was multiplied by 3 to appear on the same axis as P. [from ref (3)]

Conclusion

CNS oxygen toxicity can be modelled , successfully with survival functions. More than one formulation is capable of describing a large modern human data set. The autocatalytic model (**Model 2**) is recommended for further use, since it is capable of explaining the known benefit of intermittent exposure. Indeed, it could be used to optimize intermittency schedules – but they should be tested against actual data of intermittent human exposures.

References

1. Harabin AL, Survanshi SS, Homer LD. A model for predicting central nervous system oxygen toxicity from hyperbaric exposure in man: Effects of immersion, exercise, and old and new data. Bethesda MD: NMRI Technical Report 94-03, 1994.

2. Harabin AL, Survanshi SS, Weathersby PK, Hays JR, Homer LD. The modulation of oxygen toxicity by intermittent exposure. Toxicol Appl Pharmacol 93:298-311, 1988.

3. Harabin AL, S.S. Survanshi, L.D. Homer. A model for predicting central nervous system oxygen toxicity from hyperbaric oxygen exposures in humans. Toxicol Appl Pharmacol 132:19-26, 1995.

Modeling Diver Tolerance to Breathing Resistance

John Clarke
Navy Experimental Diving Unit
Panama City, FL

Human work performance is often limited by human tolerance to the impedances found in respiratory protective equipment such as fire-fighters' and miners' breathing apparatus and gas masks. Soldiers and sailors in a battle environment will inevitably face heavy exertion, potentially in a contaminated environment while wearing respiratory protective equipment. Unfortunately, any equipment that protects the respiratory system must also encumber it.

The primary concern of this presentation is the tolerance of divers to underwater breathing apparatus (UBA). Figure 1 (from reference 1) shows a diver attached to a chest-mounted, closed-circuit underwater breathing apparatus, otherwise known as a rebreather.

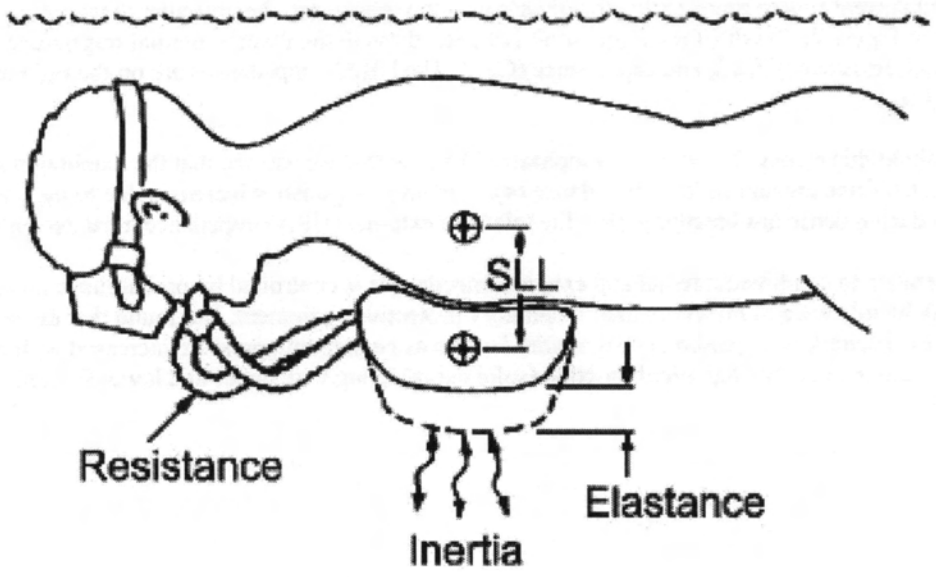


Figure 1. A closed-circuit underwater breathing apparatus provides various respiratory impedances to a swimming diver.

The combination of the diver and the UBA comprises, in an engineering sense, a system. The UBA or external portion of the system provides three respiratory impedances that the diver must overcome while breathing. Those impedances are resistance, inertance, and elastance. Static lung-loading (SLL) is a pressure off-set that strongly impacts elastance, and may affect resistance.

The combined man-UBA system can be modeled by an electrical circuit (Figure 2). Electrical analogs are frequently used in analyzing respiratory mechanics (2,3) because of the mathematical similarities between the various respiratory impedances and their electrical counterparts. Only the terminology and some symbols may differ. For instance, respiratory inertance (I) is analogous to electrical inductance (L), and respiratory compliance (C) is analogous to capacitance (C). To a first approximation, flow resistance (R) is identical to electrical resistance (R).

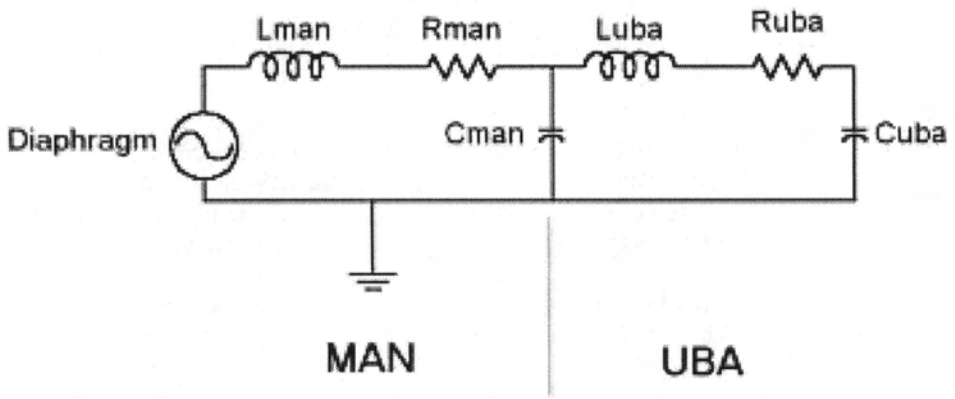


Figure 2. Electrical schematic of the combined system of diver and UBA.

An alternating current source representing the diver's gas flow-generator, the muscular diaphragm, is on the far left of the electrical diagram in Figure 2. The diaphragm forces air (current) through the diver's internal respiratory impedances; namely, inertance (L_{man}), resistance (R_{man}), and capacitance (C_{man}). The UBA's impedances are on the right half of the figure, labeled L_{uba}, R_{uba} and C_{uba}.

A key point about this analog that will be reemphasized later, is that we assume that the diaphragm and accessory musculature can tolerate a finite amount of total impedance (4). If internal impedance increases due to increases in respiratory resistance, as happens during dense gas breathing, then the tolerated external (UBA) impedance must decrease.

A diver's tolerance to combined internal and external impedances is controlled by probabilistic phenomena. In 1973, Bentley et al (5) tested the tolerance of miners to their respiratory protective equipment, and found that the probability of encountering respiratory discomfort increased in a sigmoidal fashion as peak mouth pressure increased with exercise. Figure 3 shows Bentley's best estimate of that sigmoidal function (solid curve) along with upper and lower 95 percent confidence limits (broken curves).

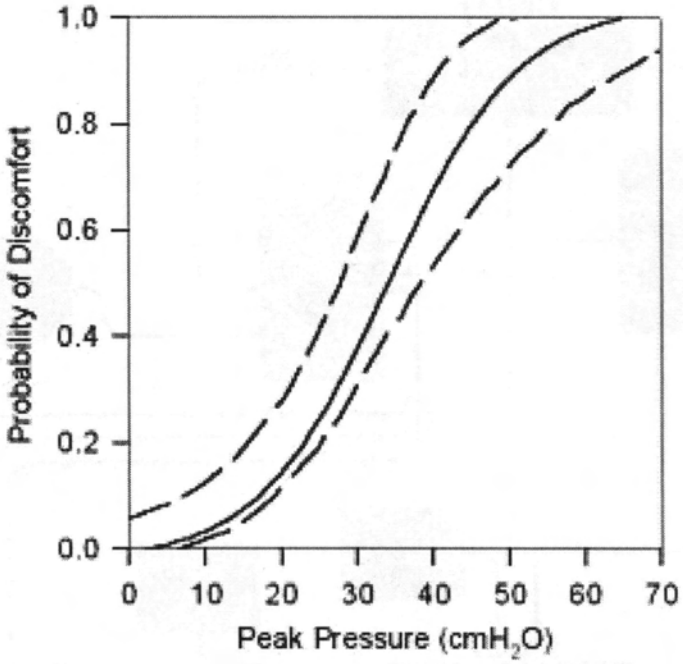


Figure 3. Probability of respiratory discomfort in use of mining respirators. From Bentley et al., 1973.

We have applied Bentley's approach to divers' tolerance to UBA. However, unlike Bentley, we focused on events that cause a diver to cease performing his mission. By definition, these events are binary, just as is decompression sickness. Either the diver stops work or he does not. The types of physiological events that can jeopardize a diver's safety, or cause a mission to be aborted, are shown in Figure 4, taken from reference (6). Various impedances (Z), are found on the bottom of the figure, just to the right of center. Ze is external impedance contributed by a UBA. Zi is internal impedance generated within the diver's airways and chest cage. Ztot is total impedance, internal plus external. Collectively, these impedances impede the diver's ability to ventilate (Ve).

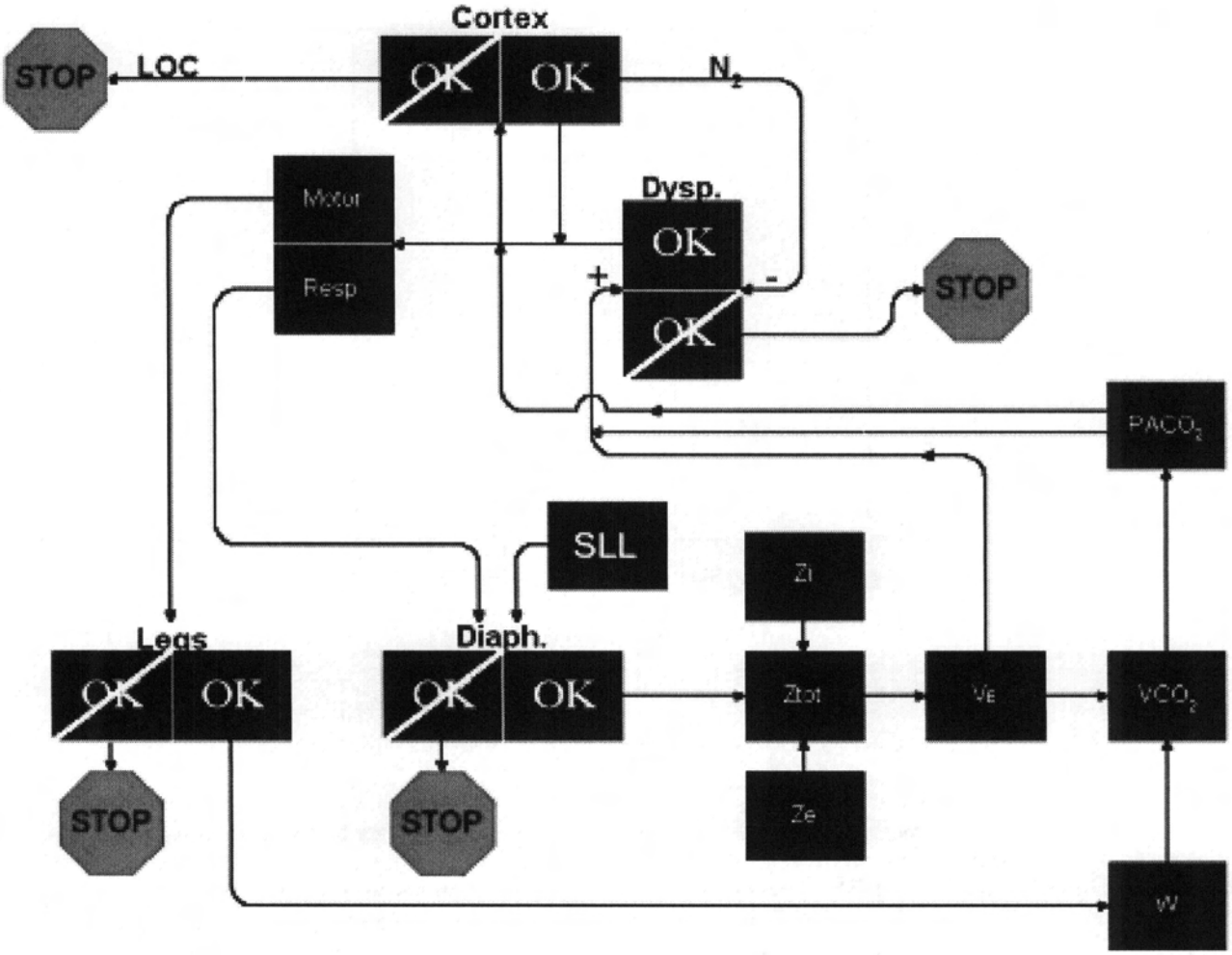


Figure 4. Sources of dive events (work stoppages). From the Lung at Depth, 1999.

Military and commercial divers go underwater to perform work (W). Work results in CO_2 production, which depending on the resulting ventilation, can modulate the arterial partial pressure of CO_2 ($PaCO_2$). $PaCO_2$ can in turn affect the central nervous system, the cerebral cortex and the portions of the central nervous system that generate the debilitating sensations of dyspnea or breathlessness. There is a stop sign associated with a level of dyspnea that is not "OK", meaning a diver who is feeling severely dyspneic or breathless will stop work. We classify the cessation of work due to dyspnea as an event; an untoward event.

There are a number of different event producers in Figure 4. If the diver does not stop work because of dyspnea, he may continue working with a reduced ventilation rate to minimize the sensations accompanying respiratory impedance. Such ventilatory depression can reach a point where arterial CO_2 reaches a very high level and produces CO_2 narcosis and loss of consciousness. That obviously would also be an event. If a diver neither becomes dyspneic nor loses consciousness, he could still fatigue his diaphragm. This would be a rare occasion, but would certainly be eventful since it would force the diver to stop work. Of course a diver's legs could fatigue, which would force work cessation. However, we do not consider leg fatigue to be a primary respiratory event.

We now apply our definition of a respiratory event to data generated at the Navy Experimental Diving Unit during the 1980s. The NEDU literature contains 240 man dives using a fixed exercise protocol. The data described in various NEDU

reports was comprised of dive depth, the type of gas being breathed, the exercise level in watts, values for peak-to-peak mouth pressure and gas density.

The particular exercise protocol used was graded exercise where a diver worked on a cycle ergometer for six minutes at 50 watts, and rested for four minutes, whereupon the exercise level increased by 50 watts. This work-rest cycle was repeated until 150 watts were successfully maintained for six minutes, or the diver stopped work (an event) due to respiratory insufficiency.

The details of the following analysis have been presented in references 7-9. I will present only a brief summary meant to illustrate the methods and challenges of parameter estimation in this application.

Table 1 shows the data organized for analysis. The first column contains a code for an event or non-event. A non-event is indicated by a zero, an event by the number one. The second column contains the differential peak-to-peak mouth pressures (ΔP) in cmH_2O. The third column was for gas density (ρ) in $gram \cdot liter^{-1}$. The fourth column identified the major inert gas constituent; either one for helium or zero for nitrogen.

Table 1. Individual dive data organized for analysis by the parameter estimating software

event	ΔP	ρ	gas
0.00000,	23.000,	3.2000,	1.000
0.00000,	22.000,	3.2000,	1.000
0.00000,	25.000,	3.2000,	1.000
0.00000,	23.000,	6.2000,	1.000
0.00000,	37.000,	6.2000,	1.000
0.00000,	19.000,	7.7000,	1.000
1.0000,	21.000,	7.7000,	1.000
1.0000,	24.000,	7.7000,	1.000
1.0000,	23.000,	7.7000,	1.000
1.0000,	29.000,	7.7000,	1.000
0.00000,	26.000,	7.7000,	1.000
1.0000,	23.000,	7.7000,	1.000
0.00000,	18.000,	7.7000,	1.000
1.0000,	18.000,	7.7000,	1.000
0.00000,	19.000,	7.7000,	1.000

Figure 5 shows summary statistics for the total data set plotting peak-to-peak mouth pressures against gas densities, showing the mean, with 1 standard deviation bars. For comparison, a gas density of 10 grams per liter is what would be expected at a temperature of 37°C for air at 255 feet, or a helium-oxygen gas mixture at 2,000 feet.

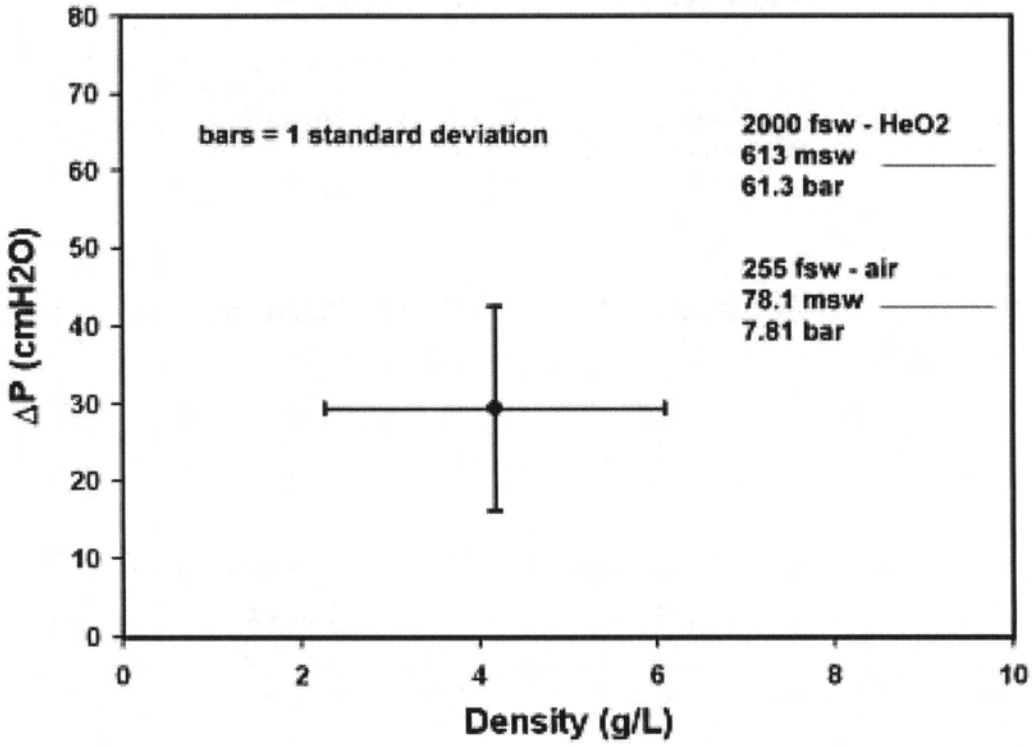


Figure 5. Summary statistics for the NEDU diver tolerance data set.

Figure 6 divides the data into those with nitrogen or helium as the inert gas. We see that the data were fairly evenly matched in terms of number of data points, mean ΔP and ρ.

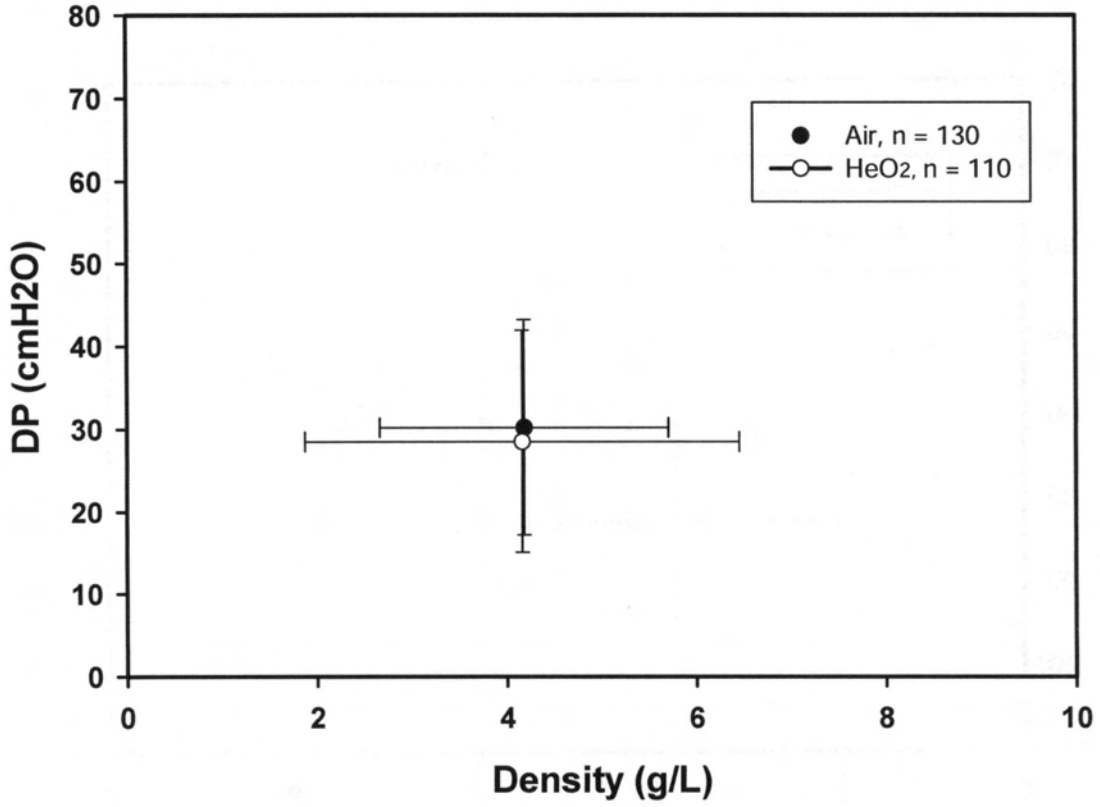


Figure 6. NEDU data set divided into nitrogen and helium based dives.

Figure 7 separates the eventful dives from the non-eventful dives. Eventful dives seem to correlate with somewhat higher gas densities and peak-to-peak mouth pressures.

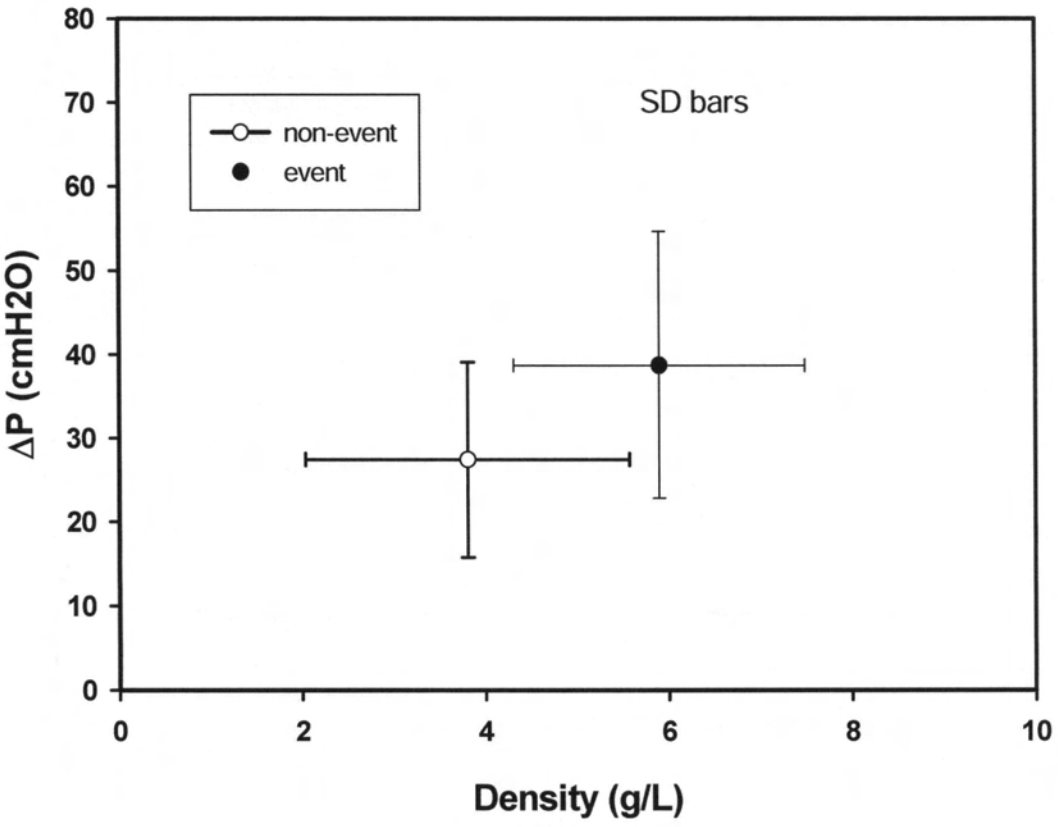


Figure 7. NEDU data set divided into eventful and non-eventful dives.

Our goal is to improve our statistical insight into diver tolerance of UBA by performing a more complete analysis using models of the data, and fitting the models to the data by the method of maximum likelihood.

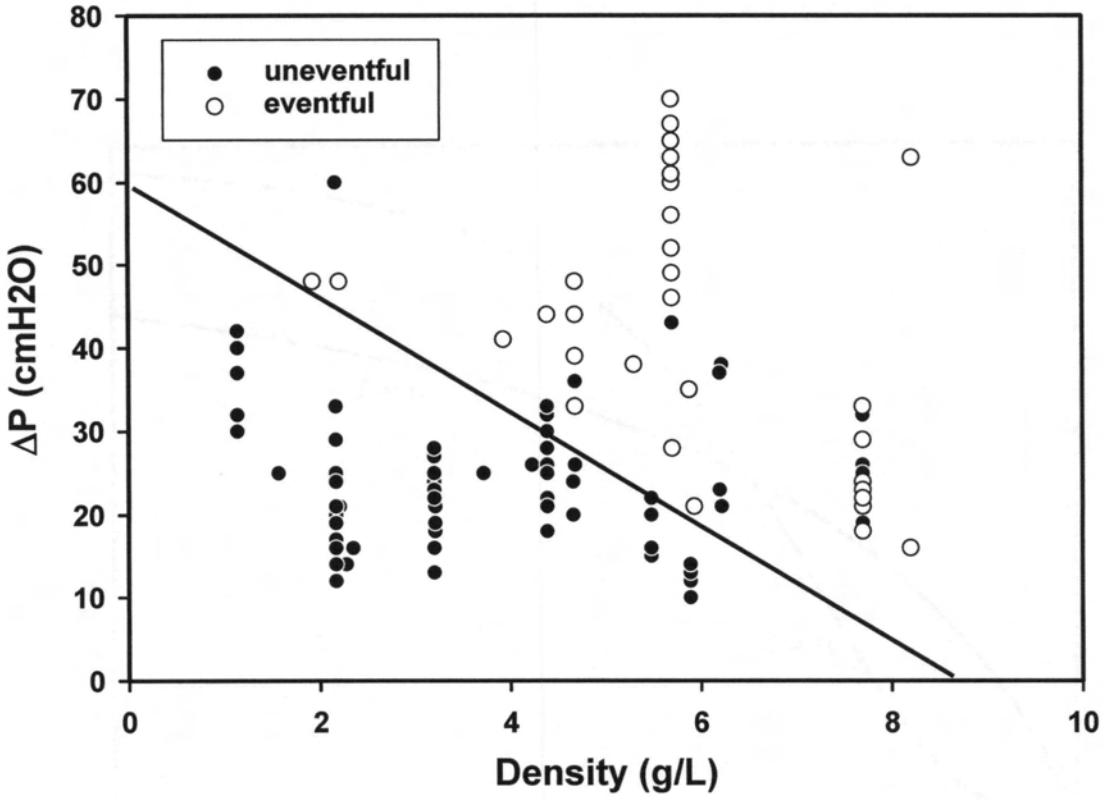


Figure 8. NEDU dives as a function of gas density and peak-to-peak mouth pressure. Black filled circles were uneventful, open circles were eventful.

Figure 8 shows one of the first analyses that suggested that maximum likelihood techniques might be useful. Gas density is on the X axis, with ΔP on the vertical axis. The dives marked by black filled circles were all non-eventful dives, while the dives with open circles were those where some respiratory event occurred. Multiple events on a dive could not be shown on this plot, but that data was nevertheless available for parameter estimation.

Figure 8 also shows a line fit by eye which approximately divides a region of eventful and uneventful dives. The eventful dives had either high gas density and low ΔPs or high ΔPs and low gas densities, or both. Our intent was to incorporate that threshold line in a model, and to use parameter estimation techniques based on maximum likelihood to find the best fit for that threshold line. The model would also define how event probability increased with distance above the threshold.

One of the best models was based on the Hill equation, a sigmoidal dose response curve commonly encountered in pharmacology (Figure 9). We are thus assuming that event probability is non-linearly related to respiratory impedance, and that impedance in general increases with ΔP and gas density.

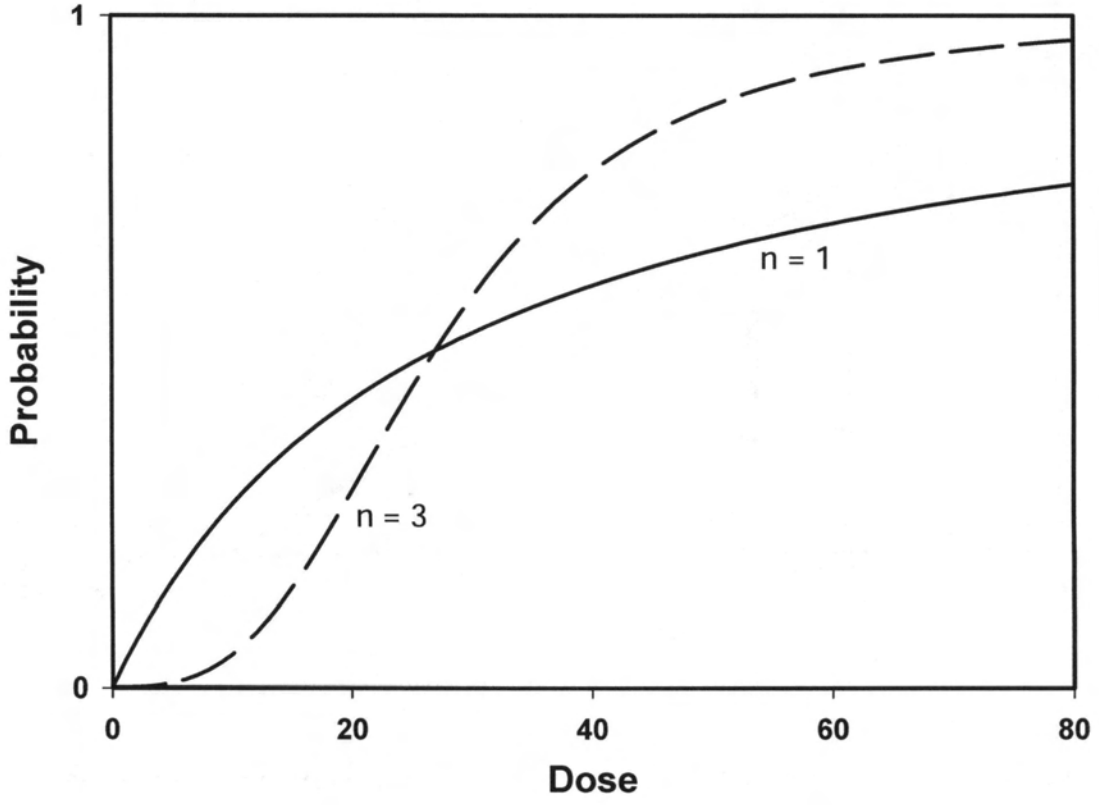


Figure 9. Two curves described by the Hill equation, differing only by the power of the power of the equation (or *n*).

If we define U as a respiratory "dose", then according to the Hill equation the probability of an event (P) is related to dose in the following manner:

$$P = \frac{U^n}{U^n + U_{50}{}^n} \tag{1}$$

We further define the dose (U) as follows:

$$U = \Delta P + B_1 \cdot \rho + B_2 \tag{2}$$

There are four parameters in this model that must be estimated. One is the power of the dose relationship (n), which determines how quickly event probability rises with increasing dose. Another is the mid-point of the sigmoidal curve (U_{50}). These two parameters are largely empirical. The last two parameters (B_1 and B_2) have a physiological source, where the respiratory dose, a function of respiratory impedance, is dependent upon ΔP and gas density (ρ). If we set U to zero and rearrange *Eq.* (2) for U, we obtain the following equation for the threshold line seen in Figure 9:

$$\Delta P = B_2 - B_1 \cdot \rho \tag{3}$$

where B_1 is the slope of that line, and B_2 is the Y-axis intercept.

The Hill equation is often considered to represent a sigmoidal response. Indeed, it will have a sigmoidal shape if "n", the power of U in the previous equation, is greater than one. It will not have a sigmoidal shape if n equals unity. The implications of this distinction will become apparent later.

Table 2 shows a few of the other models we examined. The first six are various forms of the Hill equation. Models 7 through 10 are based on the Gaussian probability distribution function:

$$PDF(U) = \left[\frac{1}{\sigma\sqrt{2\cdot\pi}}\right]\cdot\left[\left[WT\cdot e^{-\left[\frac{[U'-\mu_1]^2}{2\sigma}\right]}\right] + \left[[1-WT]\cdot e^{-\left[\frac{[U'-[\mu_1+60]]^2}{2.\sigma}\right]}\right]\right] \tag{4}$$

The simple Hill equation provided the best fit, and thus will be discussed below.

There were 42 events and 198 non-events out of a total of 240 man dives based on the NEDU data set. From this information we derived the so-called "null" model, a model that involves no mathematical model. To find the log-likelihood (LL) of the null model we found the fraction of events, which was 42/240 = 0.175. The fraction of non-events was 198/240 or 0.825.

$$LLnull = \ln(0.175^{42} \cdot 0.825^{198}) = -111.29$$

An alternative model should provide a log-likelihood significantly smaller than -111 to be considered a worthwhile improvement over the null model.

Table 2. Some of the models tested

Model No.	No. Estimated Parameters	Model
1	1	P_1 = constant Null model
2	3	$P_{He} = U'^n/(U'^n + U'_{50}{}^n)$, $U' = Pm + B(1)\cdot\rho$ $P_{N2} = P_{He}$
3	4	$P_{He} = U'^n/(U'^n + U'_{50}{}^n)$, $U' = Pm + B(1)\cdot\rho$ $P_{N2} = (1 + k) \cdot P_{He}$
4	5	$P_{He} = U'^n/(U'^n + U'_{50}{}^n)$, $U' = Pm + B(1)\cdot\rho + B(2)$ $P_{N2} = P_{He}$
5	5	$P_{He} = U'^n/(U'^n + U'_{50}{}^n)$, $U' = Pm + B(1)\cdot\rho + B(2)$ $P_{N2} = (1 + k) \cdot P_{He}$
6	5	$P_{He} = U'^n/(U'^n + U'_{50}{}^n)$, $U' = Pm + B(1)\cdot\rho + B(2)_{He}$ $P_{N2} = U'^n/(U'^n + U'_{50}{}^n)$, $U' = Pm + B(1)\cdot\rho + B(2)_{N2}$
7	3	$P_{He} = (1/\sigma)\cdot\int PDF\ dx$, PDF = Eq. 4, WT = 1 $P_{N2} = P_{He}$
8	4	$P_{He} = (1/\sigma)\cdot\int PDF\ dx$, PDF = Eq. 4, WT = 1 $P_{N2} = (1 + k) \cdot P_{He}$
9	4	$P_{He} = (1/\sigma)\cdot\int PDF\ dx$, PDF = Eq. 4 $P_{N2} = P_{He}$
10	5	$P_{He} = (1/\sigma)\cdot\int PDF\ dx$, PDF = Eq. 4 $P_{N2} = (1 + k) \cdot P_{He}$

The computer program used for the parameter estimation involved a modified Marquardt algorithm developed at the Naval Medical Research Institute, primarily by Bailey and Homer (10), and run on DEC machines until it was rewritten in 1989 using Microsoft Fortran on a PC. All of the following analyses were based on the 1989 PC version of the software.

Below are typical inputs (Table 3) and outputs (Table 4) from one of the parameter estimation runs. Out of 5 potential parameters, 2 were fixed, leaving us with a three-parameter model based on the Hill equation. The best estimate for the slope of the density dependence was 7.6, with a fairly small standard error (1.3) for the estimate. The estimate for the threshold was 59, with a standard error of 4.6. The slope of the mid-portion of the Hill equation was about 27, with a standard error of about 10. The first fixed parameter, the power of the Hill equation, was fixed at one, resulting in a negative log-likelihood of -69, a considerable improvement over the null model's log-likelihood of -111.

Table 3. Input to the Hill model, fixing two of the five parameters

```
FITTING PROGRAM = NLI.FOR, DATA FILE = UBATOTGA.DAT

NDAT NVAR NPRM ITER NFXD
 240    4    5   40    2

INITIAL PARAMETERS
  7.00000       0.00000       31.0000       1.0000       0.0000

FIXED PARAMETERS
   4    5

LAMD DELB EPS
 0.100000       0.100000E-05  0.250000E-02

FORMAT OF THE INPUT DATA FILE
    (10F20.0)

IPRT M2   IPV  ICP  IST  ILIK
   2    0    0    0    0    1

CONVERGENCE, PRINT LAST DELB
  .16E-01  -.93E-01   .13        .00        .00
```

Name of the data set

number of data "points", man dives

Power of the Hill eqn, and gas effect multiplier fixed

Table 4. Result of the parameter estimation

```
PARAMETERS AND STANDARD ERRORS

   7.6180        1.3066
   59.485        4.5928
   26.972        10.261
    .1000000E+01       PARAMETER FIXED
    .0000000E+00       PARAMETER FIXED

LOGLIK=     69.26724     LAMBDA=     2.94     DET=      .577E-01
```

slope of density dependence

threshold

slope of mid-portion of Hill equation

log-likelihood, a measure of fit

Figure 10 is a replot of Figure 8, but with the threshold line determined by parameter estimation rather than by eye. In addition, we have now added iso-probability lines. These delineate combinations of ΔP and gas density that yield equal probabilities of an event. The highest event probability shown (top most oblique line) is 50%.

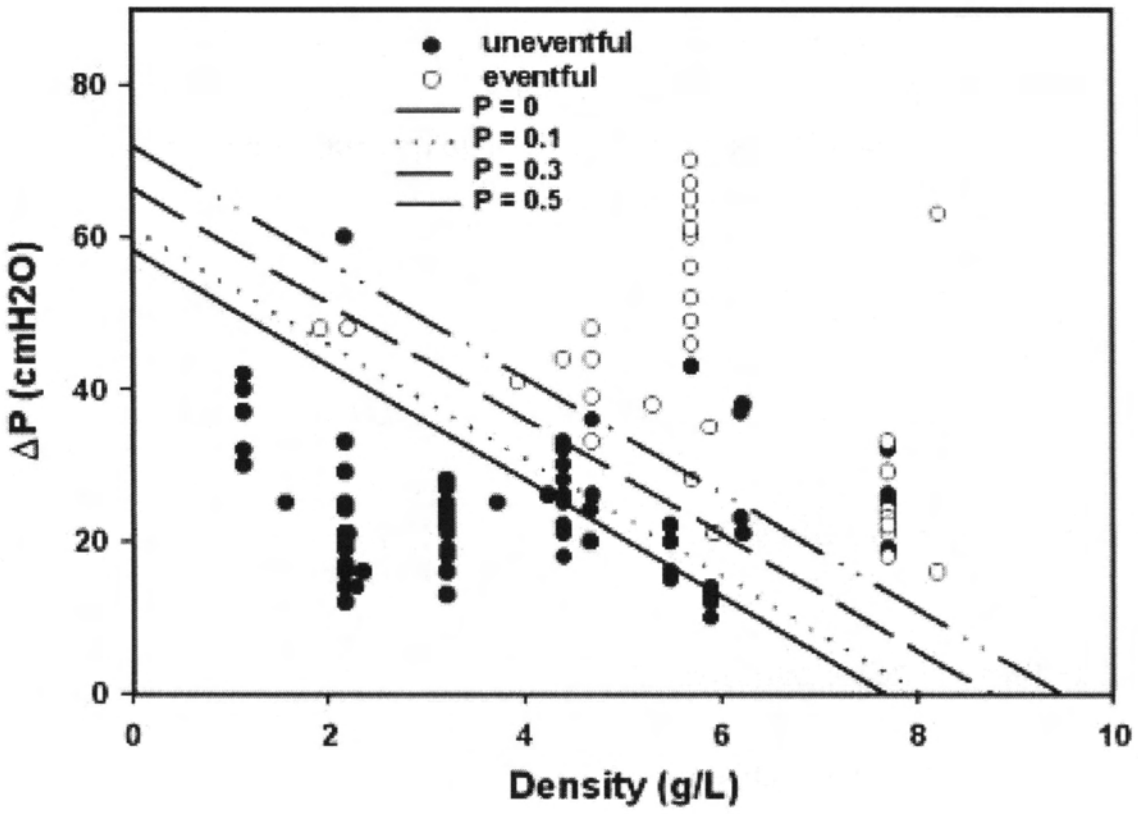


Figure 10. Replot of Figure 8 using a threshold line and isoprobability lines as determined by the maximum likelihood parameter estimation.

Next we added an additional parameter (Table 5), allowing us to test whether the background gas had an effect on event probability. The power of the Hill equation was again fixed at 1. This time the log-likelihood decreased from 69 down to 60 (Table 6). The additional parameter was -0.6 with a standard error of 0.1. That is, for a given gas density and ΔP, a nitrogen background gas reduced event probability compared to helium.

Table 5. Same Hill model, but adding one parameter (one less parameter fixed.)

```
FITTING PROGRAM = NLI.FOR, DATA FILE = UBATOTGA.DAT

NDAT NVAR NPRM ITER NFXD
 240    4    5   40    1   <-- 1 parameter fixed, power of the Hill equation

INITIAL PARAMETERS
  7.00000       0.00000       31.0000       1.0000       0.0000

FIXED PARAMETERS
   4

LAMD DELB EPS
 0.100000       0.100000E-05  0.250000E-02

FORMAT OF THE INPUT DATA FILE
    (10F20.0)

IPRT M2   IPV  ICP  IST  ILIK
   2    0    0    0    0    1

CONVERGENCE, PRINT LAST DELB
 -.13E-01    .28E-01  -.14        .00       -.98E-03
```

Table 6. Output of the 4 parameter estimation run

```
PARAMETERS AND STANDARD ERRORS

   7.2127         .66246
   60.777        2.0442
   8.3253        3.7123
    .1000000E+01       PARAMETER FIXED
  -.61957 <--    .10280   <-- gas effect multiplier

LOGLIK=     60.33887     LAMBDA=      .591E-03DET=      .849E-01
```

Improved log-likelihood

Are the differences between a log-likelihood of 69 and 60 significant enough to warrant the additional parameter? The log-likelihood ratio test measures the improvement in the likelihood of one model over another. The likelihood ratio (*LR*) is distributed as a Chi Square variable whose degrees of freedom (df) correspond to the numbers of constrained variables.

We compare a four-parameter model to a three-parameter model by the log-likelihood ratio test as follows:

$$LR = 2 \cdot (LL_{general} - LL_{specific})$$

$$LR = 2 \cdot [(-69.267) - (-60.339)]$$

$$LR = -17.857$$

For a difference of one degree of freedom (2 constrained parameters - 1 constrained parameter (Table 5)) we would only need a *LR* of -3.84 to accept the more specific, i.e. the four-parameter model. So, according to the log-likelihood ratio test, we should accept the improvement in model fit to the data based on four parameters. That is, the background gas does make a difference in event probability, all else being equal.

Let's see if we can estimate the fifth parameter, the exponent in the Hill equation. So far we have fixed it at unity. However, as Figure 9 shows, the typical sigmoidal shape of the Hill equation occurs only at higher values of "*n*". We let the parameter estimation algorithm search for this additional parameter by not constraining any of the parameters (Table 7.)

Table 7. Input for a 5 parameter estimation.

```
FITTING PROGRAM = NLI.FOR, DATA FILE = UBATOTGA.DAT

NDAT NVAR NPRM ITER NFXD
 240    4    5   50    0

INITIAL PARAMETERS
  7.00000       0.00000       31.0000       3.0000        0

LAMD DELB EPS
 0.100000       0.100000E-05  0.250000E-02

FORMAT OF THE INPUT DATA FILE
    (10F20.0)

IPRT M2   IPV  ICP  IST  ILIK
   2    0    0    0    0    1

CONVERGENCE, PRINT LAST DELB
  .11E-01    .98E-01    .15E-01    .82E-02    .15E-03
```

5 parameter model

Table 8. Output for the five parameter estimation

5 parameter model

```
PARAMETERS AND STANDARD ERRORS

   7.6896          1.1957
   59.700          7.2098
   13.024          12.465
   1.6322          1.6158      <-- Power of the Hill Equation
  -.62286          .10151

LOGLIK=     60.11195     LAMBDA=      .167     DET=      .293E-02
```

No improvement in fit over the 4 parameter model

As shown in Table 8, the additional parameter was estimated as 1.6, but the standard error of the estimate was large, also 1.6, and the log-likelihood was not improved over that of the four parameter model where n was fixed at unity. So, at least using the Hill equation, we cannot pull five parameters out of the data. We are limited to four.

In summary then, based on this particular model, the nitrogen background does seem to influence event probability, but the addition of a non-unity power to the Hill equation does not improve the fit.

Some of our newer work at NEDU has explored alternative parameter estimation methods. Logistic regression uses generalized linear models, fitted by maximum likelihood techniques (11). This analytical method focuses on the log-odds ratio or logit. The logit is sensitive to the probabilities of a binary event, or conversely, the probability is equivalent to the exponentials of the logit.

$$Logit = \ln\left[\frac{P}{1-P}\right]$$

$$P = \frac{e^{\log it}}{e^{\log it} + 1}$$

$$\ln\left[\frac{P}{1-P}\right] = B_1 + (B_2 \cdot \Delta P) + (B_3 \cdot \rho) + (B_4 \cdot \Delta P \cdot \rho) + \left(B_5 \cdot gas\right)$$

The last equation defines the logit as a function of ΔP, ρ, their interaction, and the type of diluent gas. There were once again 5 parameters to be estimated. The estimation routine was based on the logistic regression routines in the S-Plus 4.0 (Mathsoft, Inc.) statistical package.

The fit parameters (mean ± standard error) were as follows:

$B_1 = -24.11 \pm 4.57$
$B_2 = 0.44 \pm 0.09$
$B_3 = 3.67 \pm 0.73$
$B_4 = -0.06 \pm 0.01$
$B_5 = -3.78 \pm 0.87$

The standard errors were all relatively small compared to the best estimates. Figure 11 is a plot of predicted versus actual logits on the X-axis, versus pressure or the partial of V2 on the Y-axis. The prediction line runs right among the individual data points, and the distance of the data points from the line is a graphical representation of the residual fit. What is nice about this is that the data points and the residuals are centered around the line. There is no curvilinearity in the residuals, and the clustering of the data fit is relatively tight. So, at least at first blush, this looks like a reasonable logistic model.

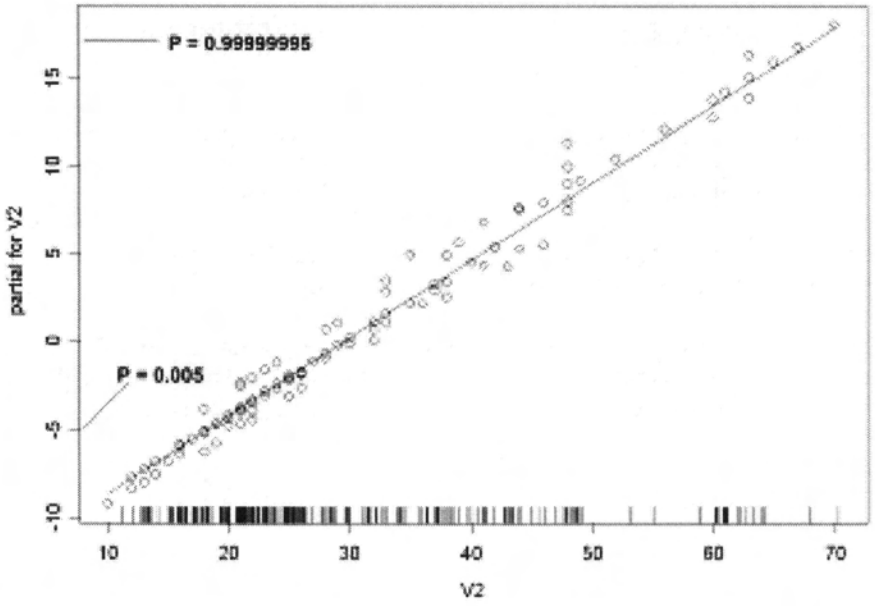


Figure 11. Plot of actual versus predicted fit of the logistic when applied to the NEDU data set.

A three-dimensional plot of the resulting fit is shown in Figure 12. It illustrates how event probability increases with both increasing ΔP and ρ. What is newly revealed by this approach is an apparent curvilinearity of our previously assumed linear threshold line marking the beginning of the rise in events.

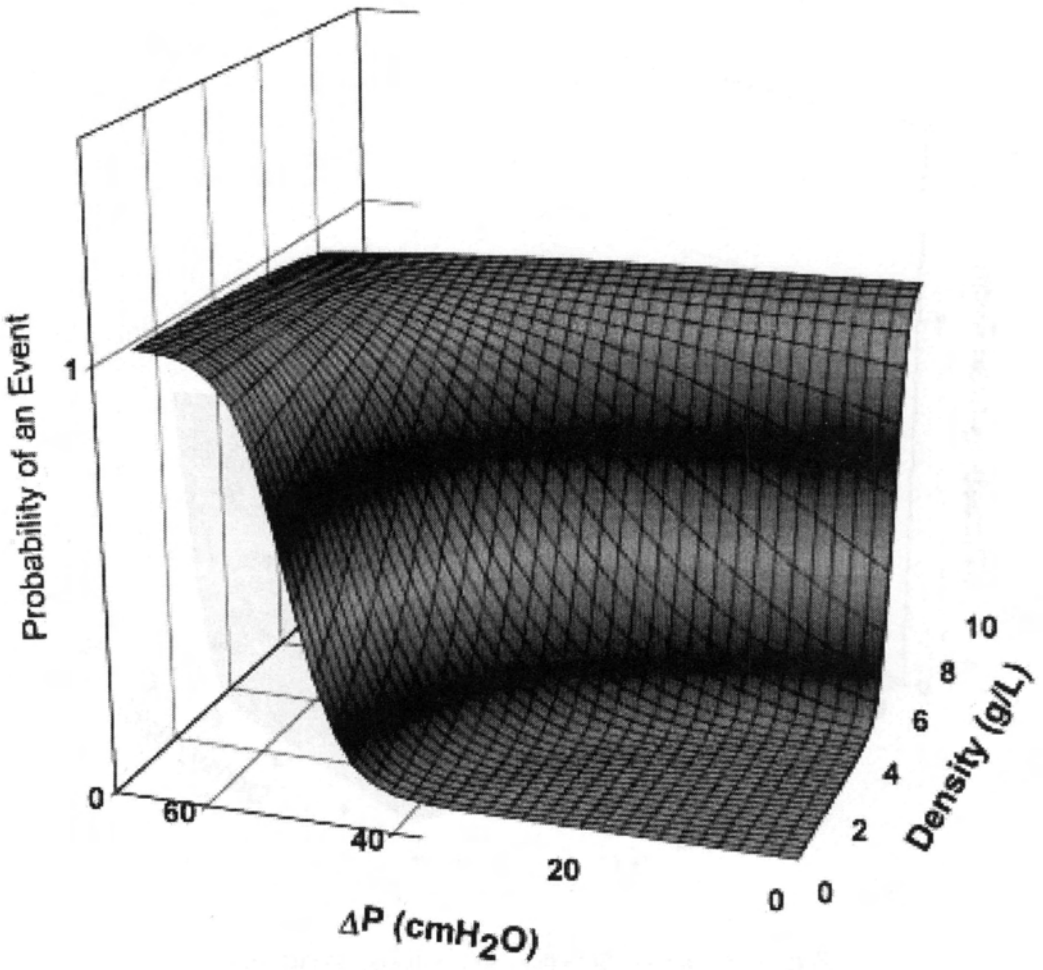


Figure 12. Three-dimensional plot of the logistic fit to the helium based NEDU data.

When plotting the effect of nitrogen atmospheres on event probabilities (Figure 13), we see an accentuation of the threshold curvilinearity, and displacement of the threshold to higher values of ΔP and ρ. The dip in event probability at both high ΔP and ρ is artifactual; it simply reflects the paucity of data obtained under those extreme conditions.

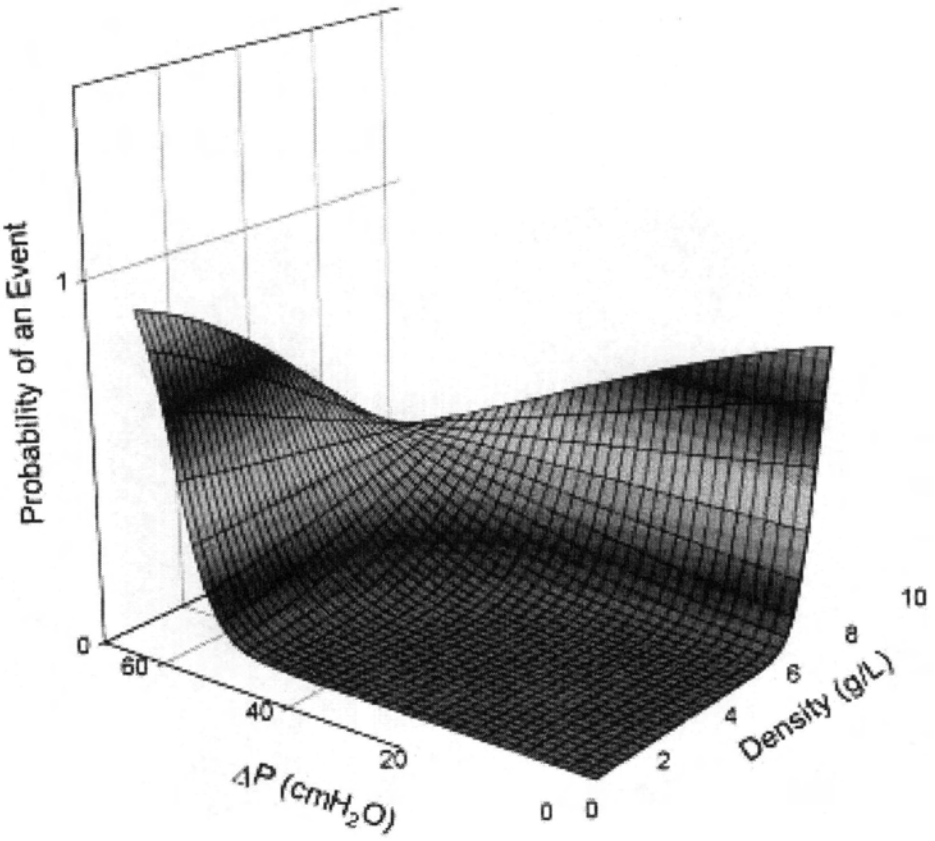


Figure 13. Logistic fit to the nitrogen based NEDU data.

Another alternative approach to parameter estimation is the use of neural networks to suggest more rigorous, albeit empirical models. I view it as an exploratory data visualization tool not dependent upon model preconceptions. Figure 14 illustrates a three layer, back propagating network that was successfully applied to the same diver tolerance data used for the previous maximum likelihood based parameter estimations.

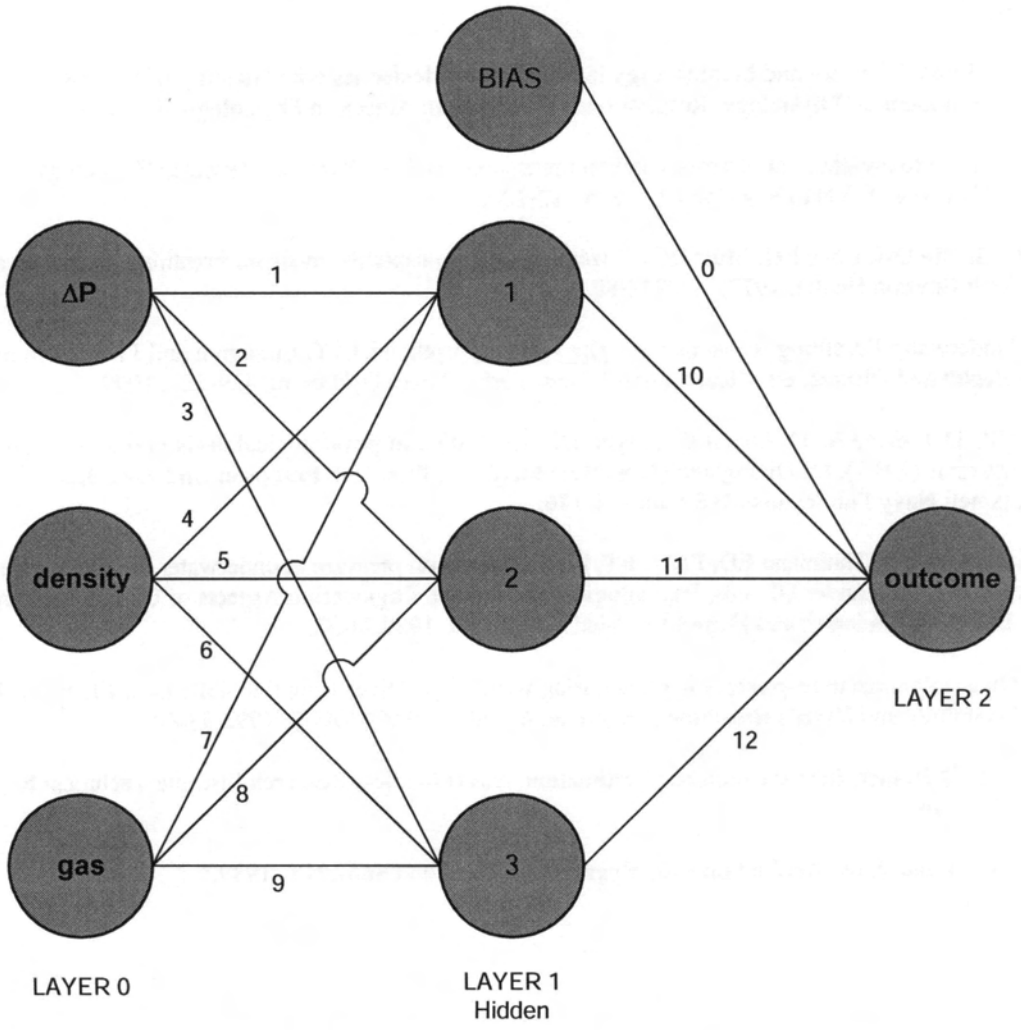


Figure 14. A three layer, back-propagating neural network applicable to the NEDU dive data.

The inputs to the neural net were, as usual, ΔP, ρ, and type of diluent gas. The output of the network was the probability of an event.

References

1. Warkander DE. Ergonomics of Breathing Apparatus, with Special Reference to Work of Breathing, Dead Space and Breathing Resistance. Ph.D. Dissertation, Chalmers University of Technology, Goteborg, Sweden.

2. DuBois AB, Brody AW, Lewis DH, Burgess BF Jr. Oscillation mechanics of lungs and chest in man. J Appl Physiol 1956; 8:587-594.

3. Mead J, Milic-Emili J. Theory and Methodology in Respiratory Mechanics with Glossary of Symbols. In: Fenn WO, Rahn H, eds. Handbook of Physiology. Respiration I, Washington: American Physiological Society, 1964: 363-376.

4. Mead J. Resistance to breathing at increased ambient pressures. In: Proc Symp Underwater Physiology, Goff LG, ed. Washington: Natl Acad Sci Natl Res Council, 1955:112-120.

5. Bentley RA, Griffin OG, Love RG, Muir DCF, Sweetland KF. Acceptable levels for breathing resistance of respiratory apparatus, Arch Environ Health, 1973, 27:273-280.

6. Clarke JR. Underwater Breathing Apparatus. In: The Lung at Depth, ed. CEG Lundgren and J Miller. In series, Lung Biology In Health and Disease, ed. Claude Enfant. New York, Marcel Dekker. pg.429-527, 1999.

7. Clarke JR, Vila D, Cabeca A, Thalmann ED, Flynn ET. The testing of physiological design criteria for underwater breathing apparatus (UBA). In: Diving and Hyperbaric Medicine, Proc 15th European Undersea Biomedical Society (E.U.B.S.), Israeli Navy Publications 1989; pp.171-176.

8. Clarke, JR, Survanshi S, Thalmann ED, Flynn ET. Limits for mouth pressure in underwater breathing apparatus (UBA). In: Lundgren CEG, Warkander DE, eds. Physiological and Human Engineering Aspects of Underwater Breathing Apparatus. Bethesda: Undersea and Hyperbaric Medical Society, 1989:21-32.

9. Clarke JR. Diver tolerance to respiratory loading during wet and dry dives from 0 to 450 m. In: Flook V, Brubakk AO, eds. Lung Physiology and Diver's Breathing Apparatus. Aberdeen: BPCC-AUP, 1992:33-44.

10. Bailey RC and LD Homer. Iterative parameter estimation. Naval Medical Research Institute Technical Report, 76-19, Bethesda, MD, 1976.

11. Hosmer D and S Lemeshow. Applied Logistic Regression. Wiley and Sons, NY, 1989.

A Log-Logistic Survival Model Applied to Hypobaric Decompression Sickness

Johnny Conkin, Ph.D.
National Space Biomedical Research Institute
One Baylor Plaza, NA-425
Houston, Texas 77030-3498

Dr. Johnny Conkin is an Assistant Professor at Baylor College of Medicine. Correspondence through: Life Sciences Research Laboratories, NASA / Johnson Space Center / SD3, Houston, Texas 77058.

ABSTRACT

Decompression sickness (DCS) is a complex multivariable problem. A mathematical description or model of the likelihood of DCS requires a large amount of quality research data, ideas on how to define decompression dose using physical and physiological variables, and an appropriate analytical approach. It is also proper to say that a high-performance computer with specialized software is required since thousands of exposure records with tens of variables are now available. Our DCS data from hypobaric decompressions of humans in altitude chambers come from published reports. Our decompression doses are variants of equilibrium expressions for evolved gas plus other explanatory variables. Finally, our analytical approach is survival analysis, where the time of DCS occurrence is modeled. A log logistic survival analysis is a powerful method to test competing hypotheses as well as to develop probability models about hypobaric DCS. Our conclusions are applicable to simple hypobaric decompressions, ascents from five to 30 min, and after mins to hrs of denitrogenation, called prebreathing. They are applicable to long or short exposures, and under conditions of rest or exercise at altitude. The ultimate goal is to apply our models to astronauts to reduce the risk of DCS during space walks, and explorations on the moon and on Mars.

INTRODUCTION

Scientists have been challenged to understand and prevent hypobaric decompression sickness (DCS) ever since the development of the jet engine, which took man high into the atmosphere. Decompression sickness in all its myriad forms and manifestations is fundamentally linked to evolved gas in the body. A fundamental axiom about DCS is that a transient gas supersaturation, also called over-pressure or pressure difference (ΔP), exists in a region of tissue. The sum of all gas partial pressures in that region is greater than the ambient pressure opposing the release of the gas. The metastable condition may resolve with a phase transition (in the presence of micronuclei), and some of the excess mass (moles) of gas in the form of bubbles may be accommodated by the tissue and cause no symptoms. The likelihood or probability of DCS increases as the evolved gas dose increases; this is a necessary but not sufficient condition in the mechanical view of DCS. All of the complex biophysical processes responsible for evolved gas in the tissue are not known. Even less is known about the linkage between evolved gas and subsequent signs or symptoms of DCS.

Because of complex and dynamic biophysical, biochemical, and physiological processes associated with living tissue, micronuclei and later bubbles may or may not form given the same experimental conditions. Even when bubbles grow, symptoms may or may not develop under the same experimental conditions. Therefore it is better (or appropriate) to consider DCS as a probabilistic rather than a deterministic event (1,21). By this I mean that the presence or absence of symptoms for the same individual under identical experimental conditions may or may not be observed from one day to the next. A quantitative description of DCS therefore requires a large number of quality research data (3), ideas on how to define a multivariable decompression dose, and analytical approaches that maximize the available information. A log logistic survival analysis provided us a powerful method to test competing hypotheses about DCS as well as provide DCS probability models (4-6,10).

METHODS

Selecting The Appropriate Hazard Function

Since the survival function S(t), cumulative distribution function (cdf) F(t), hazard function h(t), cumulative hazard function H(t), and probability density function (pdf) f(t) are different expressions of the same survival analysis, it is possible to derive all by just knowing one (2,10,13). The survival function is defined as $S(t) = 1 - F(t)$. Since the probability density function, $f(t) = dF(t) / dt$, is related to the hazard function, $h(t) = f(t) / S(t)$, the functional form of h(t) may be revealed given $F_n(t)$ from a plot of DCS data, where $F_n(t)$ is the empirical representation of F(t). An equivalent definition of h(t) is $dF(t) / dt / (1 - F(t))$. The mathematical relationship between h(t) and F(t) is clearer with this form. I will discuss our approach in terms of h(t) because an *a priori* rationale exists for determining h(t) for hypobaric decompression sickness.

The hazard function h(t) defines the instantaneous failure rate at a specific time, given that the subject survived to at least that specified time point without a response. It is expressed in hr^{-1} in our application. Lee (13) states, "h(t) gives the conditional failure rate; the probability of failure during a small time interval, assuming that the individual has survived to the beginning of the interval, or as the limit of the probability that an individual fails in a very short interval, t to $t + \Delta t$ per unit time, given that the individual has survived to time t". In our case, h(t) gives the probability of decompression sickness P(DCS) per unit time during the altitude exposure given that the individual has survived to time T while at altitude. The instantaneous failure rate for hypobaric DCS eventually goes to zero; some subjects never get DCS at a lower pressure, assuming the lower pressure is greater than about 2.5 psia since hypoxia and ebulism prevents humans from going to a vacuum. If they remain at the lower pressure long enough, say 48 hrs, then they will come into a new equilibrium with that environment and are not at risk for DCS unless they once again ascend to an even lower ambient pressure. This situation is different from the lifetime of light bulbs, for example. Eventually all light bulbs in a random sample will fail, so h(t) will never be zero for light bulbs. A new type of survival analysis called "cure models" may improve our current methods; these models properly address the reality that some subjects will never have DCS.

The function h(t) to describe DCS failure time might be selected based on a list of available functions, an understanding of the underlying failure process, a study of the cumulative distribution of the failure time $F_n(t)$, or combinations of all three. The function may increase, decrease, remain constant, or have a complex form due to an underlying complex process (13). Many variables interact to define the failure time (or survival time depending on your preference). The distribution of failure time for hypobaric DCS in a large data set from different tests is skewed to the right. Figure 1 shows 1574 cases of DCS in the Hypobaric Decompression Sickness Databank (HDSD, 3) partitioned into 0.2 hr intervals; it is a histogram representation of f(t), and the symbol $f_n(t)$ is used to signify the empirical representation of f(t). The solid curve is the histogram smoothed with the normal density function. The inset shows the same information replotted after a natural log transformation of failure time, and this distribution appears normal. There were some severe tests, and symptoms were reported prior to or immediately on arrival at the test altitude. The symptoms actually developed during ascent to altitude and these few cases were assigned a one min failure time in the HDSD since the convention was to start the exposure time upon arrival at the test altitude. This convention accounts for the few cases seen at the left of the otherwise normal log distribution. Figure 2 shows the cumulative DCS failure distribution of the 1574 cases of DCS described in Fig. 1. The inset shows an expanded view of the failure time over the first hr to better visualize the shape of F(t) near time = 0.

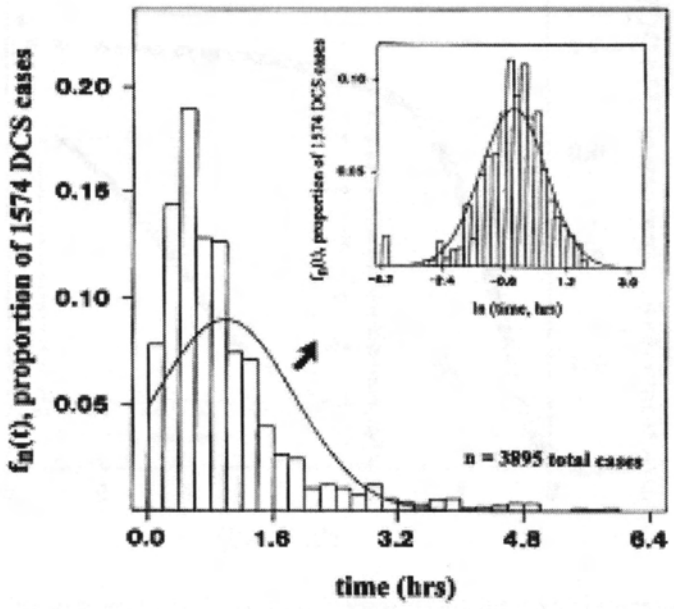


Figure 1. The histogram shows the proportion of 1574 cases of DCS as a function of time at altitude. The histogram is the empirical probability density function $f_n(t)$. The inset shows the natural log transformation of the skewed distribution into a normal distribution.

These data show that DCS under a variety of different test conditions is manifested early, within the first two hrs of exposure.

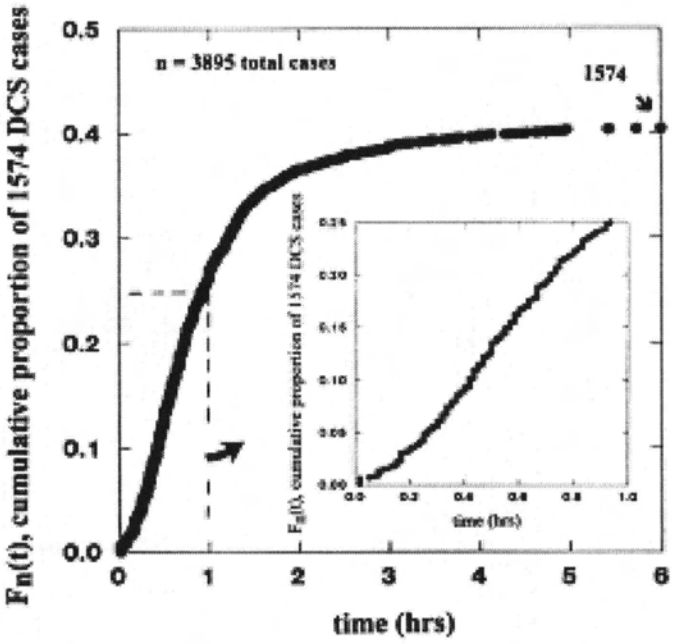


Figure 2. The empirical cumulative distribution $F_n(t)$ for 1574 cases of DCS out of 3895 exposures. F(t) is the cumulative distribution of failure time divided by the total number of records in the tests. The inset shows the same data but the time axis is limited to the first hr after reaching the test altitudes. The changing slope is easier to see on this expanded time scale, and this slope is important to select an appropriate survival model.

There are several observations about DCS that help to define an appropriate h(t). First, the rate at which DCS occurs is a function of time so the exponential distribution of failure time is not considered here. The exponential distribution defines h(t) as a constant so the time at altitude has no relation to the failure rate. If h(t) were constant, then the cumulative distribution of failure time, approximating the F(t), would be a increasing exponential defined as: $1 - \exp(- k * t)$, where k is a constant. The function S(t) would be a decreasing exponential defined as: $\exp (- k * t)$. The natural log transformation of S(t) yields $\ln S(t) = - k * t$, which is a linear function of time. It is easy to reject that the failure times come from an exponential distribution since a plot of ln S(t) against time in Fig. 2 is not a straight line, with the slope k being the constant hazard rate. Second, observations of failure times and symptom intensity also help to define h(t). The onset of a symptom is not instantaneous, and the risk of having a symptom increases with time. But it is unlikely that a person will get a symptom if he survives past some critical time since breathing 100% oxygen (O_2) (as is usually done at altitude) will ultimately reduce the nitrogen (N_2) pressure in the tissues. Also, some subjects with Type I (pain only) symptoms report that the intensity of pain reaches a peak, then subsides, and in some cases is completely gone before the end of a test. Third, observations about venous gas emboli (VGE) are helpful to define h(t) for DCS since evolved gas is fundamentally linked to a subsequent report of pain or other signs and symptoms (6,7). The two types of data share a common underlying etiology. Figure 3 shows the cumulative VGE failure distribution for 536 of 1401 records in the HDSD, not all tests produced VGE.

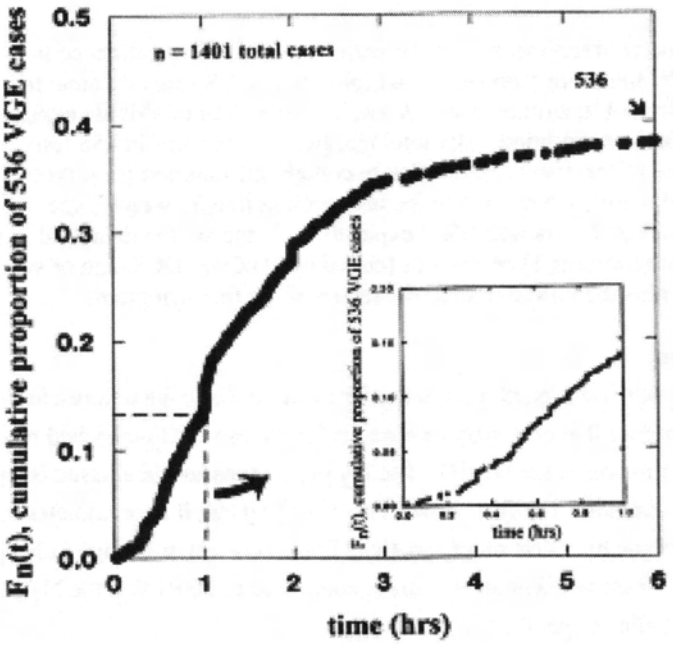


Figure 3. The empirical function $F_n(t)$ for 536 cases of VGE out of 1401 exposures. The inset shows the data up to one hour. Venous gas emboli are detected noninvasively with Doppler ultrasound technology. The pulmonary artery is insonated with the ultrasound beam and the presence of moving bubbles on the way to the pulmonary circulation is noted. Figure 2 has a similar shape and suggests that VGE and DCS share a common etiology.

Therefore in hypobaric decompressions, the instantaneous risk of DCS may increase with time, but only up to a certain point in time. The observed pattern of DCS and VGE failure time and intensity of symptoms leads us to conclude that the incidence of DCS from hypobaric decompressions would be described well with a h(t) that rises to a peak and then decreases with time. The log normal or log logistic survival models are good candidates, both providing for a non-monotonic h(t). Unfortunately, the functions F(t) and S(t) for both models may be "S" shaped. It is at the level of h(t) and f(t) that the two distributions are distinguishable. The log logistic model does not provide a slow increase of h(t) and f(t), but the log normal model does. The log normal is slightly better in most cases due in part to its ability to describe this "lag" component of h(t), but the log logistic model is easier to implement. Details about the log logistic survival model are documented in Appendix A.

DATA

The analyses presented here are based on results from documented hypobaric chamber tests and approaches (12) that account for failure and censored times. Investigators in the Navy have also exploited information about DCS failure time in divers (22). Failure time in our application is defined as the elapsed time from the beginning of a test after the decompression to the first report of a DCS symptom. Censored time is the elapsed time from the beginning of a test after the decompression to the scheduled end of the test, also called right censored time. We define h(t) in terms of several variables: $P1N_2$, P2, the presence or absence of exercise at P2, time at P2, presence or absence of VGE, etc., and use the notation: h(t; z) = f(time, P2, $P1N_2$, exercise, VGE, etc.) to denote the hazard function for a decompression dose model, where t is time and z represents various combinations of variables and constants. Appendix B lists some of the variables and their definitions in the HDSD used to model DCS.

The HDSD is a computerized repository of information about DCS experienced in hypobaric chambers that was reported in the literature (3). The literature represents a sample of the DCS research done from 1940 to present. The HDSD currently contains information from 456 altitude tests. A test is a collection of altitude exposures where one or more subjects were used to evaluate a particular test condition. The total number of exposures in 456 tests is 131,399. However 27 tests had 117,422 exposures, and none of the results reported here contain information from these 27 tests. A subset of the 456 tests provided detailed information for each subject in the test, such as height, weight, age, gender, failure time to first detection of VGE, etc. There were 211 tests with 3895 exposures. These are the data used in this report. The outcome or response variable is the presence (coded as 1) or absence (coded as 0) of any DCS sign or symptom, excluding paresthesia when it was the only symptom, plus the failure time to the report of the first symptom.

Management of O_2 prebreathe

Prebreathing 100% O_2 or O_2-enriched mixtures prior to a hypobaric decompression is an effective and often used technique to prevent DCS. Therefore it is necessary to account for the use of O_2-enriched mixtures prior to decompression in order to use the majority of information in the HDSD. The N_2 partial pressure in a tissue is an important variable in any mechanistic model about DCS. Equation 1 defines how $P1N_2$ is calculated; it approximates the more complex process of dissolved N_2 kinetics in living tissue by a first-order kinetics. Following a step-change in N_2 partial pressure in the breathing medium, such as during a switch from ambient air to a mask connected to 100% O_2, the N_2 partial pressure that is reached in a designated tissue compartment after a specific time is:

$$P1N_2 = P_0 + (P_a - P_0) * (1 - \exp^{-k * t}), \qquad \text{Eq. 1}$$

where $P1N_2$ = the N_2 partial pressure in the tissue after t mins, P_0 = initial N_2 partial pressure in the compartment, P_a = ambient N_2 partial pressure in breathing medium, exp = base of natural logarithm, and t = time at the new P_a in mins. The tissue rate constant k is related to the tissue N_2 half-time ($t_{1/2}$) for N_2 pressure in a compartment, and is equal to $0.693 / t_{1/2}$, where $t_{1/2}$ is the 360 min tissue N_2 partial pressure half-time, and 0.693 is the natural log of two. Half-time is the time taken for N_2 pressure to increase or decrease to one-half of the difference between the initial and final values. About 94% of this difference is achieved within four half-time periods. A half-time of 360 min is used because Type I altitude DCS and VGE have been shown to correlate well with long half-times, the use of 100% O_2 in altitude chamber flights eliminates faster compartments as potential contributors to DCS, and long half-times also govern the return of divers from saturation exposures. The initial, equilibrium N_2 pressure (P_0) in the tissue at sea level is taken as 11.6 psia instead of an average alveolar N_2 pressure of 11.0 psia. The use of dry-gas, ambient N_2 pressure as equilibrium tissue N_2 pressure (P_0), and as the N_2 pressure in the breathing mixture (P_a) makes the application of Eq. 1 simple. The ratio of $P1N_2$ to P2 is the Tissue Ratio (TR), where $P1N_2$ is the calculated N_2 pressure just prior to ascent to altitude and P2 is the ambient pressure after ascent. The importance and implication of TR as an expression of evolved gas is developed elsewhere (6,19).

I have described the logic that led us to select an appropriate h(t), briefly described our source of response and explanatory variables, and will now provide an example of the analytical steps that get us to a better understanding of hypobaric DCS.

THE ANALYTICAL PROCESS

The hazard function h(t) for the log logistic survival model (10) is:

$$h(t) = \lambda * (t^{\lambda - 1}) * \rho^{\lambda} / [1 + (t * \rho)^{\lambda}], \qquad \text{Eq. 2}$$

where λ and ρ are index (unitless) and scale (hr^{-1}) parameters to be estimated, respectively, and t is time in hrs in this application. When $\lambda > 1$, h(t) has a maximum, and resembles a bell shape.

The cumulative hazard function H(t) is obtained by integrating h(t). Thus:

$$H(t) = \int_0^t h(x)\,dx, \qquad \text{Eq. 3}$$

where x is the dummy variable of integration. Note that h(t) may not vary with time, as with the exponential model, but the integral of h(t) will give H(t) in terms of the starting and ending time at P2. A combination of Eq. 2 and Eq. 3 yields:

$$H(t) = \ln [1 + (t * \rho)^{\lambda}], \qquad \text{Eq. 4}$$

where ln is the natural logarithm. Since the survival function S(t) is also defined as:

$$S(t) = e^{-H(t)}, \qquad \text{Eq. 5}$$

We obtain the following expression for S(t) from Eq. 4 and Eq. 5 for the log logistic model:

$$S(t) = 1 / [1 + (t * \rho)^{\lambda}]. \qquad \text{Eq. 6}$$

The probability density function f(t) is:

$$f(t) = h(t) * e^{-H(t)}, \qquad \text{Eq. 7}$$

which may be expanded as follows from Eq. 2 and Eq. 4 for the log logistic model:

$$f(t) = \lambda * (t^{\lambda - 1}) * \rho^{\lambda} / [1 + (t * \rho)^{\lambda}]^2. \qquad \text{Eq. 8}$$

Now P(DCS) given failure time T ≤ the exposure time t becomes:

$$P(\text{DCS } T \le t) = 1 - e^{-H(t)}. \qquad \text{Eq. 9}$$

In order to account for variables other than time that influence P(DCS), we expand the hazard function h(t) but retain its functional form as given by Eq. 2. The gas phase contribution to h(t) could be as simple as 1 / P2, or as complex as $(((P1N_2 + c1) / P2) - 1)^{c2}$, but the exercise contribution is always in the form (1 + (c3 * exercise)), where exercise at P2 is one or zero, and c1, c2, and c3 are estimated parameters. The modified h(t; z) for the log logistic model that includes P2 and exercise is:

$$h(t; z) = \lambda * (1 / P2)^{c2} * [1 + (c3 * \text{exercise})] * (t^{\lambda - 1}) * \rho^{\lambda} / [1 + (1 / P2)^{c2} * [1 + (c3 * \text{exercise})] * (t * \rho)^{\lambda}]. \qquad \text{Eq. 10}$$

The function H(t; z) from Eq. 3 and Eq. 10 becomes an expression of decompression dose as a function of three variables associated with DCS plus the fitted parameters that maximize the agreement between dose and response:

$$\text{Dose} = H(t; z) = [\ln (1 + (1 / P2)^{c2} * [1 + (c3 * \text{exercise})] * (t * \rho)^{\lambda})], \qquad \text{Eq. 11}$$

and P(DCS) given failure time T based on P2, exercise, and time t at P2 becomes:

$$P(DCS\ T \leq t) = 1 - e^{-Dose}. \qquad \text{Eq. 12}$$

PARAMETER ESTIMATION BY MAXIMUM LIKELIHOOD

Maximum likelihood is the preferred method to optimize unknown parameters in a probability model where the response variable is dichotomous and the predicted value is a probability. The maximum likelihood method provides the probability that y = 1 (the response) given a value for "x" (the dose), and has been clearly explained by others (2,18,21). The likelihood function for a set of data containing (d + n) elements with some right censored times has two components, one for the failure times (subset d) and the other for the censored times (subset n). Denoting the failure times by t_i , i = 1, 2, ... , d, and the censored times by t_i , i = d + 1, d + 2, ... , n, the likelihood function (L) is (2):

$$L = \prod_{i=1}^{d} f(t_i) * \prod_{i=d+1}^{n} S(t_i). \qquad \text{Eq. 13}$$

A subject with DCS contributes a term $f(t_i)$ to the likelihood, the density of failure at t_i. The contribution from a subject whose survival time is censored at t_i is $S(t_i)$, the probability of survival beyond t_i.

The log likelihood (LL) is:

$$LL = \sum_{i=1}^{d} \ln f(t_i) + \sum_{i=d+1}^{n} \ln S(t_i). \qquad \text{Eq. 14}$$

The SYSTAT (ver. 5.03) Nonlin module (23) was used to estimate unknown parameters in the models. Estimation by maximum likelihood was accomplished by specifying the negative LL in the LOSS statement:

$$LOSS = - \ln (ESTIMATE), \qquad \text{Eq. 15}$$

where ESTIMATE is a number from one to zero from the LL function, as explained below. The LL function structured in SYSTAT for the log logistic model, as an example, is:

$$LL = [DCS * \lambda * (t^{\lambda - 1}) * \rho^{\lambda} / [1 + (t * \rho)^{\lambda}]^2] + [(1 - DCS) * 1 / (1 + (t * \rho)^{\lambda})]. \qquad \text{Eq. 16}$$

$f(t)$ or $f(t; z)$ $\qquad$ $S(t)$ or $S(t; z)$

The computer evaluates Eq. 16 for the first row of hundreds of rows of data. The first row contains values for the observed DCS (1 or 0), $P1N_2$ (psia), P2 (psia), exercise (1 or 0), and time (hrs): failure time when DCS = 1, or censored time when DCS = 0. When DCS is one, f(t; z) is evaluated, and when DCS is zero, S(t; z) is evaluated. The numerical result, between zero and one in each case, is called ESTIMATE, evaluated with initial values of the unknown parameters in the model, is used in Eq. 15. The LL calculation from Eq. 15 is repeated over all rows, and the LL is summed over all rows. The summed LL is then minimized using the Quasi-Newton algorithm (23). Iterations continue for parameters in the model until a predetermined convergence criterion is reached.

RESULTS

Table I is a compilation of a number of log logistic survival models for DCS, expressed as h(t; z), included in two of our reports (4,5). The table shows a progression from simple to more complex models. The complexity comes as we attempt to describe evolved gas with combinations of variables and constants associated with evolved gas, and with our notions of

how pain is perceived as tissues are deformed by evolved gas (see Appendix in ref. 6). Also, some information in the complex models is strictly correlative with DCS, such as the VGE information, which when added to the model improve the description of DCS failure time. Values and other details of the fitted constants are not reproduced here. Equation 17 identified prebreathe ($P1N_2$), the final altitude pressure (P2), the presence of exercise at altitude, and the length of the exposure as important variables to describe the DCS failure time in 1075 exposures. Figure 4 summarizes our main conclusions: for a given calculated N_2 pressure in the 360 min half-time compartment, DCS risk increases as P2 decreases (any vertical line through the curves), as time at P2 increases (two filled circles along the 4.3 psia curve), and if exercise is performed at P2 (two filled circles at 4 hrs exposure on the 4.3 psia solid and dashed curves).

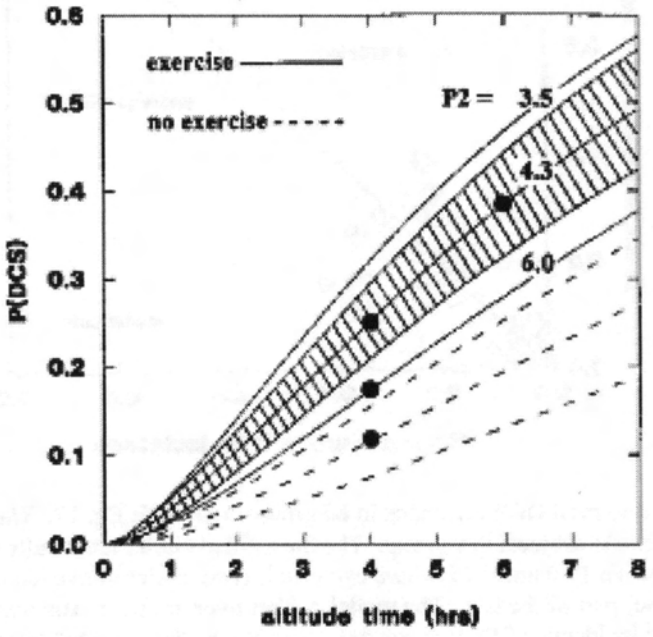


Figure 4. The P(DCS) at either 3.5, 4.3, or 6.0 psia with (solid line) or without (dashed line) exercise at a particular time after decompression. The ratio of $P1N_2$ to P2 (TR) in Eq. 17 was 1.65 for each curve, but notice the P(DCS) increases as P2 decreases at any particular time after decompression. The 95% confidence interval is provided for the curve specific to the 4.3 psia exposure that included exercise.

An important conclusion is that for the same TR, in this case 1.65, the risk of DCS is greater at a lower P2 for a given exposure time and exercise condition (two filled circles on the 4.3 psia and 6.0 psia curves at 4 hrs exposure). The fitted constant c1 in the numerator of Eq. 17 is responsible for this result, and other ways of accommodating the constant did not provide as good a fit of the model to the data. We suspect the importance of the constant is its linkage to metabolic gases in the evolved gas (11,19).

Once the best model from a family of models is determined, it is still not clear if there is a good fit of the best model to the data. The likelihood ratio test (8,13) defines when no further improvement is possible by adding more degrees of freedom (parameters to fit) to the model. However the test offers no absolute goodness-of-fit summary such as provided by the coefficient of determination (ρ^2) in least-squares regression. There are few available analytical tools, outside of a Statistics Department, to assess goodness-of-fit of a survival model. We use graphical approaches to "visually" assess goodness-of-fit. Figure 5 shows the predicted versus observed group incidence of DCS in 66 tests, the tests that provided the 1075 decompression records. A perfect description of the data by our model would require that all tests fall along the identity line. We have also validated this model in a set of data not used to optimize the model (4). We conclude that Eq. 17

(expressed through Eq. 12) describes reasonably well the DCS and no DCS cases in 1075 exposures, and could be used prospectively.

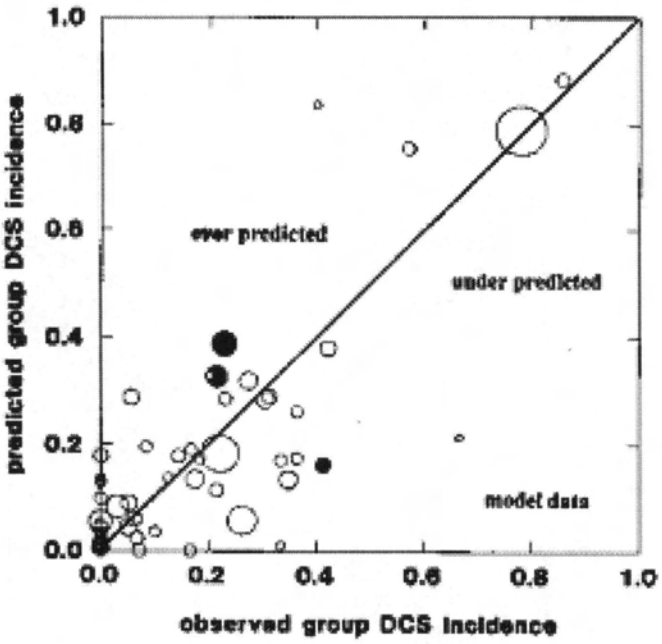


Figure 5. Predicted vs. observed DCS incidence in 66 groups used to fit Eq. 17. The area of a circle is proportional to the number of subjects in a group. The three filled circles are results from NASA tests at 4.3 psia with TRs between 1.60 and 1.65 where exercise is (two circles above identity line) and is not (circle below identity line) part of the test. The model neither over or under estimates the entire data set, but did over estimate the incidence of DCS in several small groups that reported no symptoms.

Figure 6 is a simulation based on Eq. 18 (expressed through Eq. 12) where data about VGE were available in 1322 records to improve the estimate of DCS failure time. The figure shows that the presence of Grade IV VGE increases the risk of DCS compared to all lesser grades. Additional information about the simulation is provided in the description of the figure. It can be argued that any information on VGE used to describe DCS is invalid since both DCS and VGE are responses to decompression. However the intensity and time course of VGE are information that relate (correlate) to a subsequent DCS symptom (7).

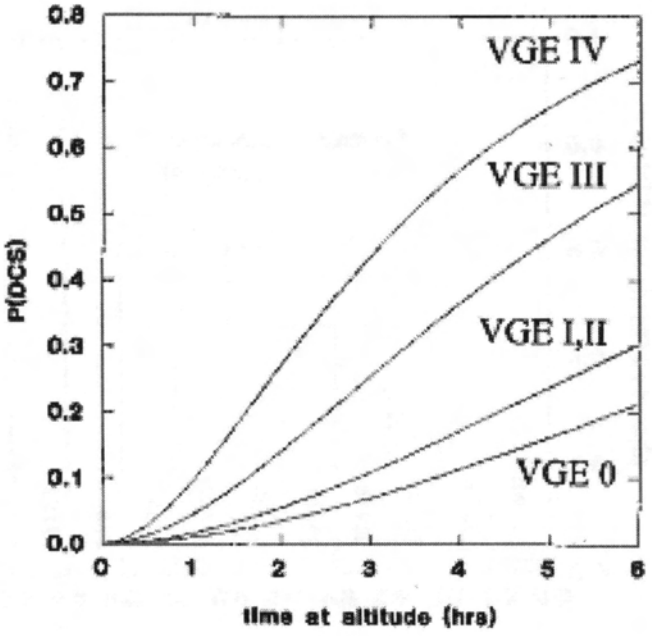


Figure 6. The P(DCS) versus time at altitude from Eq. 12, given by Eq. 18 in Table I, for a simulated decompression at a TR of 1.65 (7.1 $P1N_2$ / 4.3 P2), all with exercise, with VGETM of one hr, and with the presence of VGE at Grades I and II, III or IV, and the absence of VGE (Grade 0). Please review Appendix B for the definition of the variables used in this analysis.

We conclude that the inclusion of VGE information into our basic model (Eq. 17) was beneficial, and also improved the goodness-of-fit. Figure 7 is a visual representation of goodness-of-fit for Eq. 18. This presentation differs from Fig. 5 in that each subject in the 1322 exposures had a unique P(DCS) since no two subjects necessarily had identical VGE information. As before, we conclude that Eq. 18 describes reasonably well the DCS and no DCS cases in 1322 exposures.

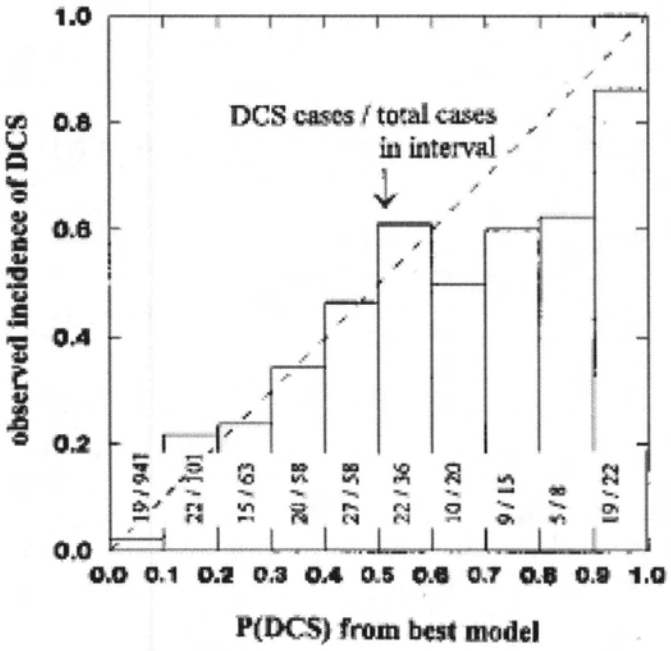


Figure 7. A bar graph to show the observed incidence of DCS in ten intervals compared to the predicted P(DCS) from Eq. 12, given by Eq. 18 in Table I. The 1322 records were first divided into ten probability intervals based on the P(DCS) from Eq. 12 for each record. The number of DCS cases in the interval were then divided by the total number of cases in the interval to give the incidence of DCS. Equation 12 did not systematically under or over predict the observed incidence. It did under predict the observed incidence in intervals from 0.60 to 0.90.

Equations 17 and 18 were attempts to develop useful hypobaric DCS probability models. Like others (20), we explored using survival analysis to test a specific hypothesis. We were curious about the linkage between evolved gas in a tissue and the report of a DCS symptom. Often elegant and complex models about bubble growth in tissue neglect this aspect of the problem. The published report (6) develops the rationale about how a power term fitted to our simple equations of evolved gas may link evolved gas to the P(DCS). Conceptually, as the intensity of a symptom increases (as a power) the P(DCS) increases to a certainty. Figures 8 and 9 show the dramatic improvement in describing the DCS failure times in 1085 exposures simply by including a power term in a simple expression (ΔP) of evolved gas.

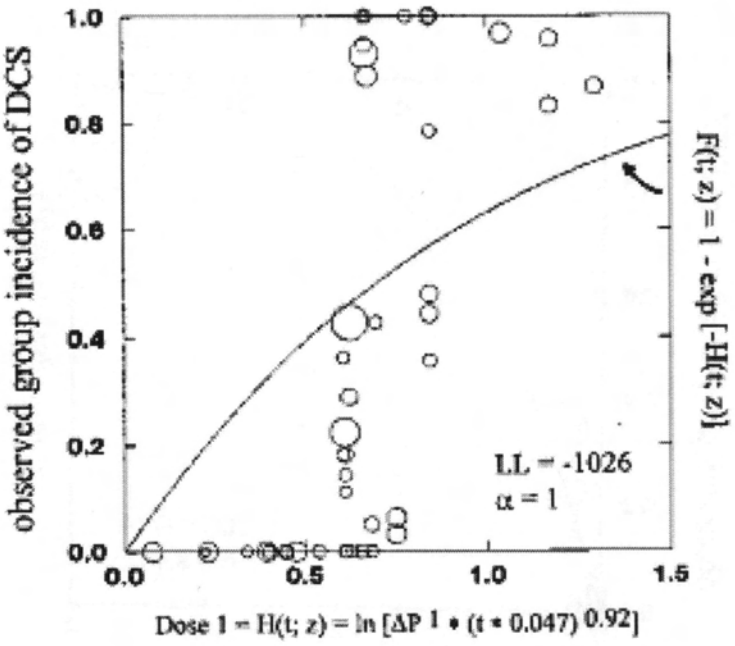


Figure 8. A scatter plot that shows the observed incidence of DCS in a group and the calculated decompression dose with Dose 1 = ln [1 + ($P1N_2$ - P2) $^{\alpha}$ * (t * 0.04728) $^{0.922}$], where α = 1, and $P1N_2$ is from the 360 min half-time, plus a curve from Eq. 12. The position of each circle along the vertical axis depends on the value of Dose 1 for each group. Superimposed on the circles is a solid curve from Eq. 12, given f(t; z) on the figure, that is the P(DCS) as a function of Dose 1. The area of a circle is proportional to the number of subjects in a group; the smallest circle represents a test with two subjects and the largest circle represents 77 subjects.

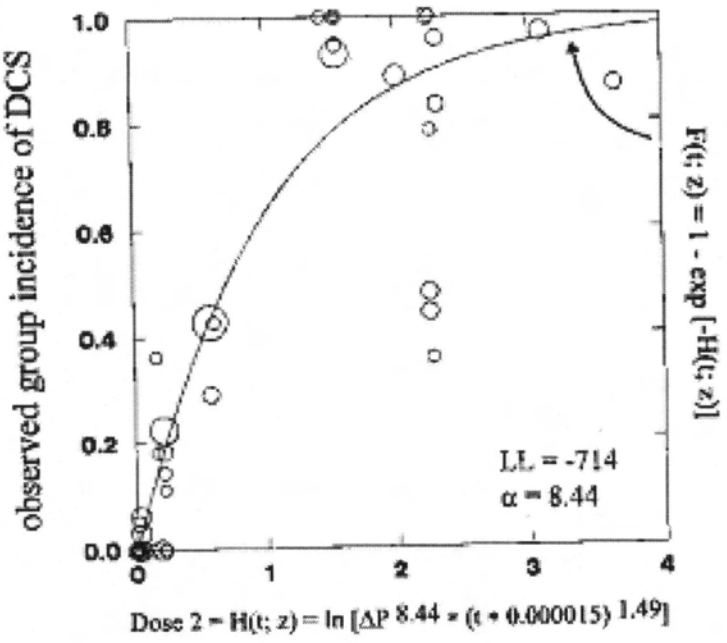


Figure 9. A scatter plot that shows the observed incidence of DCS in a group and the calculated decompression dose with Dose 2 = ln [1 + ($P1N_2$ - P2) $^{\alpha}$ * (t * 0.00001517) $^{1.491}$], where α = 8.44 , and $P1N_2$ is from the 91 min half-time, plus a curve from Eq. 12. The horizontal position of the circles are the same as in Fig. 8 but the vertical position has changed due to the recalculation of decompression dose. The goodness-of-fit was improved by estimating the half-time but the greatest improvement came from estimating α. The circles are positioned more symmetrically around the curve than in Fig. 8 and the LL improved from -1026 in Fig. 8 to -714 in this figure.

The solid curve on Fig. 8 from a model without a power term does not pass near the majority of group DCS incidence data as compared to the curve on Fig. 9. We were motivated to evaluate this concept based on an earlier analysis by Nims (15). Figure 10 shows that our survival model as a probability density function f(t; z) gave results similar to Nims, but our methods (statistical) were much different from those of Nims (deterministic).

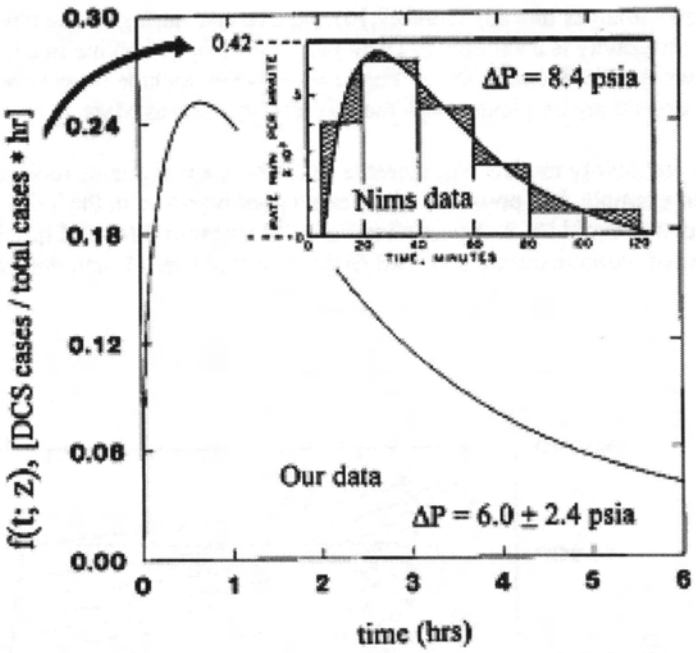


Figure 10. The resulting f(t; z) for the average ΔP of 6.0 psia from the f(t; z) equation on Fig. 9 where $\alpha = 8.44$, $\rho = 0.00001517$, $\lambda = 1.491$, and t1/2 = 91 mins. The f(t; z) resembles the shape of the curve from Nims (15, his Fig. 40) in a test with ΔP = 8.4 psia.

The shape, but not the magnitude, of the two curves are similar and yet Nims did not explicitly use a power term in the expression of DCS dose. Our observation that different methods lead to similar results reinforced our belief that conclusions from hypothesis testing with incomplete models should be verified with experimental data.

CONCLUSION

We have used survival analysis with maximum likelihood optimization as the basis to describe the failure time for DCS under a variety of decompression conditions tested in hypobaric chambers. Our first goal was to identify an appropriate hazard function. This was based on a survey of DCS and VGE data contained in a computerized databank and descriptions and observations on how DCS symptoms progress through time (Figs 1-3). The exponential survival model was clearly inappropriate, the log normal model was slightly better than the log logistic, but more difficult to implement. Other models for failure time distribution were also evaluated, but the log logistic model proved to be the best overall for our applications.

Our efforts over the past few years were directed toward developing probability models for DCS that accounted for major physical and physiological variables (Figs. 4-7). We have not completed the analysis of several variables known or suspected to influence the risk of DCS. Age and gender differences continue to be discussed as modifying factors for DCS. While it is difficult to include age and gender in a deterministic (theoretical) model of DCS, it is simple to include these variables in a statistical model. We are always surprised to find that one long half-time compartment (about six hrs) is adequate to describe the results from a variety of hypobaric tests at our disposal. We have brought empirical models into better agreement with bubble models by including a term to account for the presence and consequence of metabolic gases in total evolved gas.

Our second use of survival analysis was to test a hypothesis about the inclusion of a power term into simple expressions of evolved gas (Figs. 8-10). The goal here was to understand a mechanism about the perception of pain. An exciting area to explore with research and modeling is the biophysical linkage between evolved gas and perception of pain. The future for hypothesis testing and developing better predictive models for DCS is good because new and better data are

being collected and shared. New variables like adynamia (9,16) and exercise during prebreathe (14) are now being tested. Adynamia is a concept about how gravity is a variable in DCS, particularly how walking in a gravitational field influence micronuclei that in turn influence the likelihood of DCS. Future models that include these variables will have applications to astronauts during space walks, or walking on planets with reduced gravity such as Mars.

Applications for DCS probability models will increase since these are available tools and, if properly applied, can provide useful information. For example, it is possible to lose cabin pressurization in the T-38 aircraft. What is not known is if an emergency landing is needed to avoid DCS. We applied Eq. 17 (expressed through Eq. 12) under two scenarios for the T-38. The DCS risk for the loss of pressure during a normal flight is seen in Fig. 11, and during a high altitude flight seen in Fig. 12.

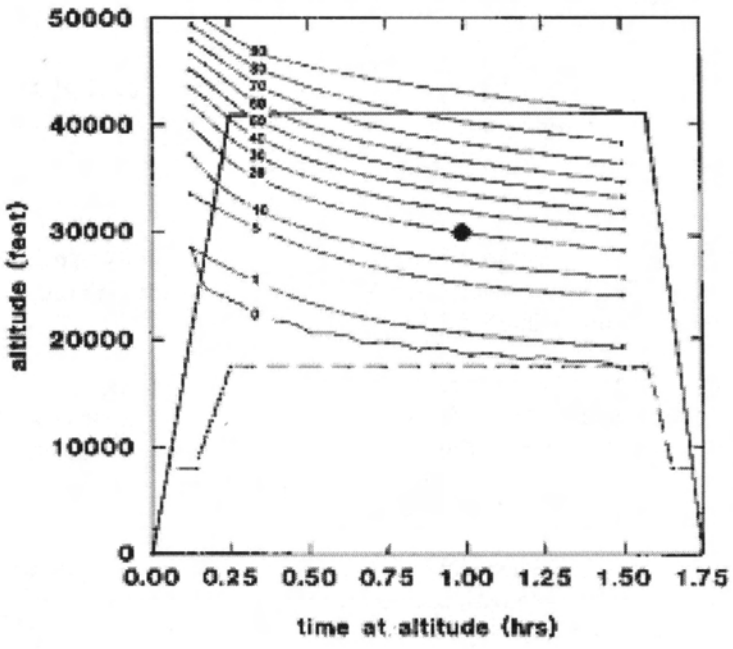


Figure 11. The approximate flight envelope (solid near vertical lines) of the T-38 and the resulting cabin pressure (dashed line) under nominal flight conditions. Transposed over the flight envelope are twelve DCS isoincidence isopleths for the condition where the crew is not physically active. The proper way to determine the risk is to select a time of exposure and the altitude of the exposure and then interpolate between the isopleths. There is no risk of DCS if cabin pressure is maintained. However a loss of cabin pressure for even brief periods of time can exposure the crew to ahigh risk of DCS. The likelihood of very serious DCS symptoms is greater as the risk of any DCS symptom increases.

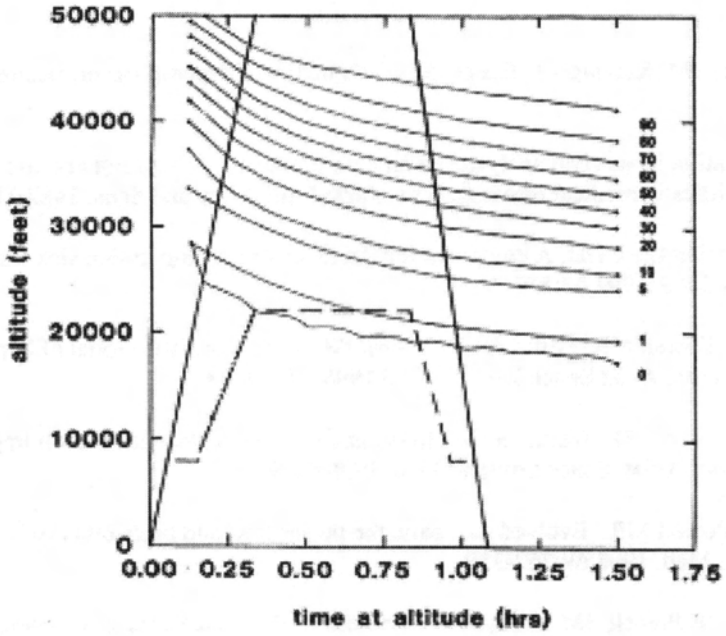


Figure 12. The flight and cabin pressure envelope under extreme flight conditions. Notice that even at the lowest cabin pressurization (22,000 feet) and 45 mins of exposure, the risk of any symptom of DCS is less than 5%. The majority of the risk is between zero and 1% under extreme flight conditions.

The aircraft can fly high but only for a shorter duration. Altitude, duration, prebreathe, and exercise at altitude are variables in Eq. 17. We assumed a limited use of O_2 during the flight (defined in ref. 17) and that the aviators were not physically active during the flight. Figure 11 shows the P(DCS) given that the aviator was exposed to a certain decompression for a certain time. Notice that below a normal cabin altitude of 18,000 feet, it is unlikely that DCS will occur. However a one hr exposure to 30,000 feet puts the aviator on the 20% DCS isopleth (solid point). During high altitude flight the cabin altitude can increase to 22,000 feet, but the flight time is limited to just over one hr. Figure 12 shows that the lowest cabin pressure (22,000 feet) with the T-38 at the highest operating altitude (50,000 feet) is associated with a risk of DCS between one and five percent. The information in Figs. 11 and 12 can help managers make flight rules that would prevent the loss of cabin pressure in a T-38 leading to the loss of an aircraft and crew.

ACKNOWLEDGMENT

This body of work, both published and presented at the 1998 Undersea and Hyperbaric Medical Society Workshop on Survival Analysis in Environmental Physiology, could not have been done were it not for the exchange of ideas between me and my colleagues: Michael R. Powell, Philip P. Foster, Alan H. Feiveson, Michael L. Gernhardt, Vasantha Kumar, James M. Waligora, Karin C. Loftin, Hugh D. Van Liew, Wayne A. Gerth, R. Srini Srinivasan, and Kallappa M. Koti. The National Aeronautics and Space Administration supported part of this work through the NASA Cooperative Agreement NCC 9-58 with the National Space Biomedical Research Institute.

REFERENCES

1. Berghage TE, Woolley JM, Keating LJ. The probabilistic nature of decompression sickness. Undersea Biomed. Res. 1974;1:189-96.

2. Brown BW, Jr. Estimation in survival analysis: parametric models, product-limit and life-table methods. In: Mike V, Stanley KE, eds. Statistics in medical research. New York: John Wiley and Sons, 1982:317-39.

3. Conkin J, Bedahl SR, Van Liew HD. A computerized databank of decompression sickness incidence in altitude chambers. Aviat. Space Environ. Med. 1992;63:819-24.

4. Conkin J, Kumar KV, Powell MR, Foster PP, Waligora JM. A probabilistic model of hypobaric decompression sickness based on 66 chamber tests. Aviat Space Environ Med 1996;67:176-83.

5. Conkin J, Powell MR, Foster PP, Waligora JM. Information about venous gas emboli improves prediction of hypobaric decompression sickness. Aviat. Space Environ. Med. 1998;69:8-16.

6. Conkin J, Foster PP, Powell MR. Evolved gas, pain, the power law, and probability of hypobaric decompression sickness. Aviat. Space Environ. Med. 1998;69:352-359.

7. Conkin J, PP Foster, MR Powell, JM Waligora. Relationship of the time course of venous gas bubbles to altitude decompression illness. Undersea Hyperbaric Med 1996; 23:141- 49.

8. Conkin J. Probabilistic modeling of hypobaric decompression sickness [Dissertation]. Buffalo, NY: State Univ. of New York at Buffalo, 1994.

9. Conkin J, Powell MR. Lower body adynamia reduces the risk of hypobaric decompression sickness. Aviat. Space Environ. Med. 1999; (in peer review).

10. Cox DR, Oakes D. Analysis of survival data. New York: Chapman and Hall, 1984:13-21.

11. Foster PP, J Conkin, JM Waligora, MR Powell, RS Chhikara. Role of metabolic gases in separated gas phase formation during hypobaric exposures. J. Appl. Physiol. 1998; 84:1088-95.

12. Kumar KV, Powell MR. Survivorship models for estimating the risk of decompression sickness. Aviat. Space Environ. Med. 1994;65:661-65.

13. Lee ET. Statistical methods for survival data analysis, 2nd ed. New York: John Wiley and Sons, 1992:8-18.

14. Loftin KC, J Conkin, MR Powell. Modeling the effects of exercise during 100% oxygen prebreathe on the risk of hypobaric decompression sickness. Aviat. Space Environ. Med 1997; 68:199-204.

15. Nims LF. Environmental factors affecting decompression sickness. Part I: A physical theory of decompression sickness. In: Fulton JF, ed. Decompression sickness, Philadelphia: WB Saunders, 1951:192-222.

16. Powell MR, Waligora JM, Norfleet WT, Kumar KV. Project ARGO - Gas phase formation in simulated microgravity. NASA Technical Memorandum 104762. Houston: Johnson Space Center, 1993.

17. Robinson RR, JP Dervay, J Conkin. An evidenced-based approach for estimating decompression sickness risk in aircraft operations. NASA Technical Memorandum TM-1999-209374, NASA Johnson Space Center, Houston, Tx. July, 1999.

18. Tikuisis P, Nishi RY, Weathersby PK. Use of the maximum likelihood method in the analysis of chamber air dives. Undersea Biomed. Res. 1988;15:301-13.

19. Van Liew HD, Burkard ME. Simulation of gas bubbles in hypobaric decompressions: roles of O2, CO2, and H2O. Aviat. Space Environ. Med. 1995;66:50-55.

20. Van Liew HD, Burkard ME, Conkin J. Testing of hypotheses about altitude decompression sickness by statistical analyses. Undersea Hyperbaric Med. 1996;23:225-33.

21. Weathersby PK, Homer LD, Flynn ET. On the likelihood of decompression sickness. J. Appl. Physiol. 1984;57:815-25.

22. Weathersby PK, Survanshi SS, Homer LD, Parker E, Thalmann ED. Predicting the time of occurrence of decompression sickness. J. Appl. Physiol. 1992;72:1541-48.

23. Wilkinson L. SYSTAT: the system for statistics. Evanston: SYSTAT Inc., 1990:342-87.

TABLE I. VARIOUS LOG LOGISTIC SURVIVAL MODELS FOR DCS

Model	Parameters	
log logistic survival model (null model)		
$h(t) = \lambda * (t^{\lambda - 1}) * \rho^{\lambda} / (1 + (t * \rho)^{\lambda})$	2 (λ, ρ)	
log logistic hazard function with additional variables and constants (accelerated model)		
$h(t; z) = [\lambda * z_n * (t^{\lambda - 1}) * \rho^{\lambda}] / [1 + z_n * (t * \rho)^{\lambda}]$		
$z_1 = 1 / P2$	2	
$z_2 = P1N2 / P2$	2	
$z_3 = (P1N2 / P2) - c$	3	
$z_4 = (P1N2 / (P2 + c1)) - 1.0$	3	
$z_5 = ((P1N2 + c1) / P2) - 1.0$	3	
$z_6 = (((P1N2 + c1) / P2) - 1.0) * (1 + (c3 * exercise))$	4	
$z_7 = ((P1N2 / (P2 + c1)) - 1.0)^{c2} * (1 + (c3 * exercise))$	5	
$z_0 = (((P1N2 + c1) / P2) - 1.0)^{c2} * (1 + (c3 * exercise))$	5	Eq. 17
$z_8 = z_0 * [1 + (c4 * vge)]$	6	
$z_9 = z_0 * [1 + (c4 * vge)] * \{1 + [c5 * (1 / vgetm)]\}$	7	
$z_{10} = z_0 * [1 + (c4 * mvge)] * \{1 + [c5 * (1 / vgetm)]\}$	7	
$z_{11} = z_0 * [1 + (mvge^{c4})] * \{1 + [c5 * (1 / vgetm)]\}$	7	
$z_{12} = z_0 * [1 + (c4 * vgeI,II)] * [1 + (c5 * vgeIII)] * [1 + (c6 * vgeIV)]$	8	
$z_{13} = z_0 * [1 + (c4 * vgeI)] * [1 + (c5 * vgeII)] * [1 + (c6 * vgeIII)] * [1 + (c7 * vgeIV)]$	9	
$z_{14} = z_0 * [1 + (c4 * vgeI,II)] * [1 + (c5 * vgeIII)] * [1 + (c6 * vgeIV)] * \{1 + [c7 * (1 / vgetm)]\}$	9	Eq. 18

Appendix A: Two Forms of the Log Logistic Survival Model

A common form of the log logistic survival function S(t) is:

$$S(t) = 1 - [\, 1 / (1 + e^{(-\omega)})], \qquad A1$$

where $\omega = [\ln(t) - \beta(2)] / \beta(1)$.

The distribution is specified as a two parameter distribution generalized to include the effects of covariates on survival times. The generalized log logistic is called an accelerated life model where the logarithm of survival time is a linear function of the covariates:

$$\omega = [\ln(t) - \beta(2) - \beta x1 * x1 - \ldots - \beta xn * xn\,] / \beta(1). \qquad A2$$

Other functional expressions of the model are:

$$h(t) = f(t) / S(t) \qquad A3$$

$$f(t) = e\, [-(\ln(t) - \beta(2)) / \beta(1)] / [(1 + e\, (-(\ln(t) - \beta(2)) / \beta(1)))^2 * \beta(1) * t] \qquad A4$$

$$h(t) = f(t) / [1 - (1 / (1 + e\, (-((\ln(t) - \beta(2)) / \beta(1)))))], \qquad A5$$

and the accelerated life model:

$$f(t; z) = e\, [-(\ln(t) - \beta(2) - \beta x1 * x1 - \ldots - \beta xn * xn\,) / \beta(1)] /$$

$$[(1 + e\, (-(\ln(t) - \beta(2) - \beta x1 * x1 - \ldots - \beta xn * xn\,) / \beta(1)))^2 * \beta(1) * t] \qquad A6$$

$$h(t; z) = f(t; z) / [1 - (1 / (1 + e\, (-((\ln(t) - \beta(2) - \beta x1 * x1 - \ldots - \beta xn * xn\,) / \beta(1)))))] \qquad A7$$

$\beta(1)$ = scale parameter
$\beta(2)$ = index or location parameter
βxn = parameter from regression for variable n
xn = value for the nth variable
t = time

An alternate form (10) of the log logistic survival model used in our analysis is:

$$S(t) = e\, [-\ln(1 + (t * \rho)^{\lambda})], \qquad A8$$

and expanded to include covariates as:

$$S(t; z) = e\, [-\ln(1 + (c1 * x1) * \ldots * (cn * xn) * (t * \rho)^{\lambda})]. \qquad A9$$

The h(t) expression of the log logistic model is:

$$h(t) = \lambda * (t^{\lambda - 1}) * \rho^{\lambda} / (1 + (t * \rho)^{\lambda}), \qquad A10$$

and the accelerated h(t) is:

$$h(t; z) = \lambda * (c1 * x1) * \ldots * (cn * xn) * (t^{\lambda - 1}) * \rho^{\lambda} /$$

$$[1 + (c1 * x1) * \ldots * (cn * xn) * (t * \rho)^{\lambda}] \qquad A11$$

ρ = scale parameter
λ = index or location parameter

cn = parameter from regression for variable n
xn = value for the nth variable
t = time

Appendix B: Some Variables in the HDSD

Dependent Variables

DCS: presence (1) or absence (0) of any sign or symptom of decompression sickness, excluding paresthesia when it was the only symptom.

DCSTM: failure time to the first sign or symptom of DCS or censored time to the end of the test in those without DCS (hrs).

Independent Variables

P1N2: calculated nitrogen pressure (psia) from Eq. 1 to account for all denitrogenation procedures.

P2: ambient pressure after ascent (psia).

EXERCISE: presence (1) or absence (0) of repetitive exercise planned for the test.

VGE: presence (1) or absence (0) of any Grade of VGE.

MVGE: maximum Grade of VGE (0 - 4) detected during the exposure.

VGEI: presence (1) or absence (0) of Grade I VGE as the maximum Grade of VGE recorded during a test.

VGEII: presence (1) or absence (0) of Grade II VGE as the maximum Grade of VGE recorded during a test.

VGEIII: presence (1) or absence (0) of Grade III VGE as the maximum Grade of VGE recorded during a test.

VGEIV: presence (1) or absence (0) of Grade IV VGE as the maximum Grade of VGE recorded during a test.

VGETM: failure time to the first VGE detected or censored time to the end of the test in those without VGE (hrs).

ALTTM: scheduled duration of the test or the time t at P2 in a simulation (hrs).

Testing of Hypotheses about Basic Mechanisms with Risk Functions

Hugh D. Van Liew, Ph.D
Navy Experimental Diving Unit
Panama City, Florida 32408

Statistics for new insights. How sound is current understanding of the basic mechanisms that give rise to decompression sickness (DCS)? In two recent publications (1,2), we have used the statistical modeling process to gain insight into the underlying mechanisms. Note that when we use the word "mechanistic," we do not mean to imply that we can provide a model that is based on fundamental principles; just the opposite -- we are trying to learn about the principles from the models. The main purpose of my talk here is to encourage others to compare a range of models; the presumption is that a model that gives a good fit to data is more in line with the basic mechanisms than models that give poorer fits.

One goal of a statistical analysis is simply to summarize and characterize the data. The more important matter is how the results of an analysis are used. Statistical analyses are often used to to provide instructions, such as the public health instructions about benefits of exercise and hazards of high blood cholesterol. The type of instruction that interests people at this workshop is decompression procedures to avoid DCS. However, my goal is different; my theme about developing insights as to basic mechanisms can be illustrated with a simple analogy, a scatter of points on an X-vs.-Y plot (Fig. 1).

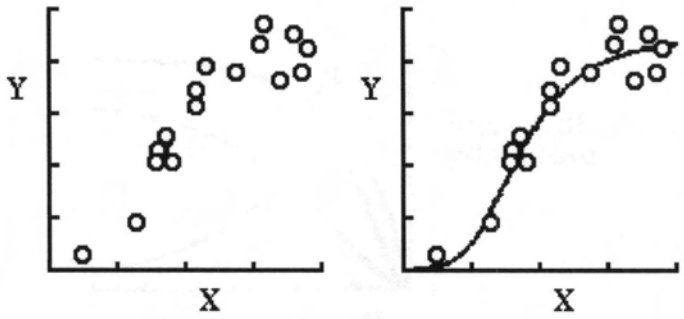


FIGURE 1. An investigator might be satisfied with fitting the points in the left-hand graph with a straight line, but a better fit is obtained with a sigmoid function (right).

The points in Fig. 1 fit well with a sigmoid function. The reader can envision other curves instead of the S-shaped one. The points would not fit as well to a straight line or some curve that does not have the inflection seen in the S-shaped curve. Suppose the data represent amount of a gas in a liquid (Y axis) as a function of the partial pressure of the gas in the liquid (X axis). An investigator who fit the points with a simple line would be missing two important insights: that the data may be generated by a chemical binding of the gas to a substance in the solution that is saturable (manifested by the tendency to level off at the top), and that the process involves interactions between binding sites (manifested by the sigmoid nature of the curve). The well-known carriage of oxygen by hemoglobin in blood has these characteristics.

In this simple illustration, the equation that is used to fit the data is "the model." It is convenient to think in terms of unexplained variability. With the points alone, all the scatter is unexplained. Fitting of a "model," a line or curve, provides an explanation for some of the variability. Even with the sigmoid fit of Fig. 1, some unexplained variability remains, seen as the deviations of many of the points from the curve. However, the unexplained variability is not as great as if the data had been fit with a less-optimal model, such as a straight line.

Is this remaining variability caused by random processes or is it because of orderly behavior of something that isn't accounted for by the model? The points in Fig. 1 may be influenced by a variable of secondary importance; call it variable Z. If so, a model that accounted for variable Z should improve the fit of the data and further reduce the unexplained variability.

The risk-function paradigm. In the paragraphs above, it was implied that 1) better models can be obtained by accounting for as many variables as possible, 2) better models come from actively seeking models (equations) that enter the variables in ways that give good fits to the data, and 3) the better models give better insights into the underlying phenomena that give rise to a set of data. I will now give an example of how these issues interacted together in analyses (2) of data from exposures of volunteers in altitude chambers (3). We purposely chose "forced descent" as our indicator of DCS. That is, we only counted a subject as positive for DCS when the exposure was terminated early for that subject; the supervisors of the exposure decided that the subject's condition was serious enough to warrant bringing him immediately to ground level. We believe that the forced descent criterion is more objective than simple reporting of symptoms by subjects.

Before proceeding, I will review the probability paradigm that we have used, which follows from early works about risk functions (4,5). The curve labeled r_i in Fig. 2 is instantaneous risk, which is related to the cases per time. We presume that r_i is a function of the pre-planned duration of the exposure to altitude. Use of pre-planned duration is admittedly a crude way of dealing with the timing of DCS occurrence, made necessary because of the information available. If we had data about the time that symptoms first occurred, it would be desirable to use survival analysis or the failure time analysis.

The shape of the r_i curve in Fig. 2 reflects one possibility for the relationship between risk and the preplanned exposure time. The rise-and-fall shape implies that, for a given exposure, DCS is apt to occur early rather than late. If so, short-duration exposures will have more cases per the total time than long exposures because the long exposures have innocuous time added to the high-risk time, whereas the high-risk time stands alone in short exposures.

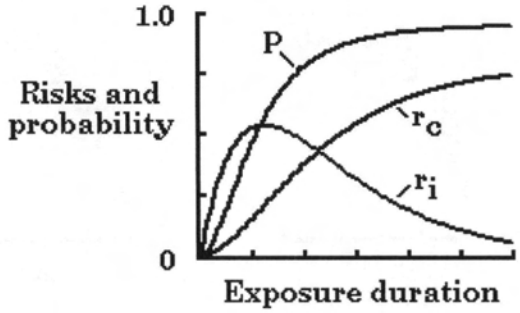


FIGURE 2. Instantaneous risk, r_i , is assumed to be a function of the pre planned duration of the exposure. The cumulative risk, r_c , is the integral of r_i. Probability is defined as an exponential function of cumulative risk. The value 1.0 on the Y axis pertains only to the probability curve; ordinate values for risk curves are not given.

The cumulative risk, r_c , is defined as the integral of the instantaneous risk, r_i. Our models (1,2) deal with two kinds of risk: the risk that is due to duration of altitude exposure, and other possible kinds of risk. We multiply the one by the other. An analogy for the second kind would be the risk of being injured in a damaging automobile accident; cumulative risk will increase as driving duration of the trip increases. However, other variables can impinge on the risk of being injured. If the brakes are bad, then for any duration, there would be, say, double the chance of injury. The shape of the r_i curve in Fig. 2 determines the shape of the r_c curve as a matter of definition but the r_c curve's height is affected by the other, time-independent sources of risk.

Finally, the probability is defined as follows: $P = 1 - e^{(-r_c)}$. The maximum-likelihood statistical analysis fits the probability curve to the data. To test the hypothesis that instantaneous risk rises and falls, one would devise a model containing an appropriate rise-and-fall mathematical formula, and one would compare that model with models having alternative formuli. The comparison we decided upon was that instantaneous risk simply rises as duration lengthens.

Analyses of altitude chamber tests. In our first analysis of over 7,000 exposures (1), the variables included in the analysis were atmospheric pressure at altitude, duration of altitude exposure, and duration of breathing of pure oxygen before the exposure (imposed to lower the nitrogen partial pressure in the tissue before the decompression). We found that a model

which utilized the rise-and-fall possibility was indistinguishable from a model which used a simple rise. Apparently the ability to distinguish between these two possibilities was masked by the magnitude of the unexplained variability.

Of the original 7,000 exposures, 4,000 had information on rate of ascent to altitude. In analysis (2) of this subset of data, we found, first of all, that the ascent rate was an important variable; addition of ascent rate to the previous model made a large improvement of log-likelihood. More interesting, when we retested the hypothesis about rise and fall of instantaneous risk with increasing duration of exposure, we found that the rise-and-fall option was highly significant this time (2). We believe that we unmasked the true situation by eliminating the portion of the unexplained variability ascribable to ascent rate, which is independent of the rise-and-fall variable, exposure duration.

Remarks about successful modeling. The data we have used (3) is clearly spotty; the altitude tests date from 1942 until the present, and purposes of different tests differed. For example, tests to find the limits of safety from DCS would involve exposures severe enough to elicit DCS occasionally, whereas training exposures would probably avoid risky exposures altogether.

Regarding the fit of models, we relied mostly on just the likelihood number and the likelihood ratio test. When the objective is development of insights into what is behind the data, we feel justified in relaxing statistical rigor.

One useful graph is a plot of DCS predicted by the statistical analysis versus DCS observed in the subjects. If the model were perfect, the points on Fig. 3 would all lie along the diagonal line; clearly there is much room for improvement of the model by "explaining" the unexplained variability. It is helpful to use the size of the points on such graphs to show information about one of the variables. For example, the point size on Fig. 3 is proportional to duration of prebreathing of oxygen before altitude exposure, a variable that should be inversely related to DCS incidence; as expected, the big points tend to fall at the lower left. However, Fig. 3 reveals that for small points (short prebreathing), the points are not well balanced along the line. Clearly, the model under-predicts DCS for short prebreathing durations.

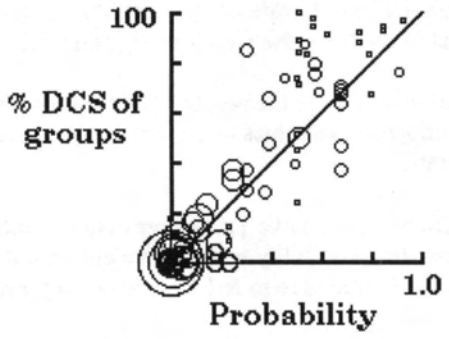


FIGURE 3. Plot of observed DCS in groups of subjects plotted against probability estimated by the statistical analysis. Size of the points is propor tional to duration of prebreathing; long prebreathing time is expected to be associated with low risk of DCS, in accord with the lay of the points.

Figure 4 continues the exploration of this problem. It shows isopleths of cumulative risk due to all variables except prebreathing of oxygen. We divided the data points into categories of "high other", "intermediate other," and "low other." When actual points were plotted on the Fig. 4 diagram, it was found that when "other" was high or low, the points fit reasonably well to the appropriate "other" curves. However, when "other" was intermediate and prebreathing time was low, the fit was terrible -- all points were above the curve, suggesting strongly that the poor fit of low-prebreathing, high-DCS

points in Fig. 3 were associated with intermediate "other." We were unable to devise a model that corrected this defect, but at least we had identified the problem for future attempts.

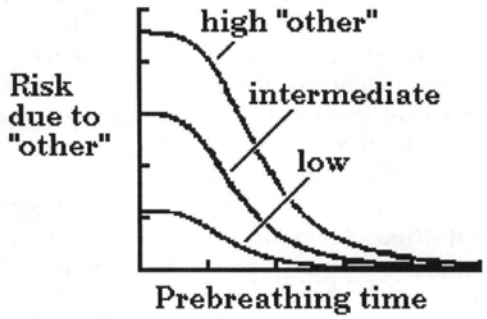


FIGURE 4. Predicted risk of DCS due to variables other than prebreathing time, categorized by intensity.

Summary

1) Models that yield significantly better fits to the data are probably closer to reality. To obtain insight about the nature of the processes that give rise to the data, one can compare alternative hypotheses, and the hypothesis which yields the better fit to the data is thus supported by the results of the statistical analysis. [It is probably best to avoid the notion that the hypothesis is proven by the analysis.]

2) Some so-called "unexplained variability" can be due to a sub-optimal model. One possibility for a sub-optimal model is that the model may be sub-optimal in the ways the variables are entered into the model [see 1) above].

3) A second possibility for suboptimal models is in the number of variables that are accounted for. A model of course improves if it accounts for as many of the important variables as possible. My example of this was the improvement of the altitude data fit when we added the ascent rate.

4) A model that accounts for more variables [3) above] may give better results in the hypothesis testing [1) above]. In our example, the advantage of the rise-and-fall possibility was not evident until the additional variable, ascent rate, was accounted for. Apparently the unexplained variability due to failure to account for rate of ascent was masking the truth in the original analysis.

References

1. Van Liew HD, Conkin J, Burkard ME. Probabilistic model of altitude decompression sickness based on mechanistic premises. J Appl Physiol 1994; 76:2726-2734.
2. Van Liew HD, Burkard ME, Conkin J. Testing of hypotheses about altitude decompression sickness by statistical analyses. Undersea Hyperbaric Med 1996; 23:225-233.
3. Conkin J, Bedahl S, Van Liew HD. A computerized databank of decompression sickness incidence in altitude chambers. Aviat Space Environ Med 1992; 63:819-824.
4. Weathersby PK, Homer LD, Flynn ET. On the likelihood of decompression sickness. J Appl Physiol 1984; 57:815-825.
5. Tikuisis P, Weathersby PK, Nishi RY. Maximum likelihood analysis of air and HeO2 dives. Aviat Space Environ Med 1991; 62:425-431.

Survival Models for Altitude Decompression Sickness

Nandini Kannan
Division of Mathematics and Statistics
University of Texas at San Antonio
San Antonio, TX 78249

Survival time is the time to onset of DCS symptoms. Again, you have certain functions which characterize the data. The survival function, S(t), is the probability that the individual survives or, in other words, does not experience DCS, beyond time t.

$$S(t) = P(T > t)$$

The cumulative distribution function, F(t), is the probability that the individual is symptomatic by time t.

$$F(t) = P(T \leq t) = 1 - S(t)$$

Finally, the risk (or hazard), r(t), is the instantaneous probability of DCS in a small interval, $[t,t+\Delta t]$, given that the individual has remained DCS free up to time t.

$$r(t) = -\frac{S'(t)}{S(t)}$$

where the prime denotes differentiation with respect to time.

A distinguishing characteristic of DCS data is that it is heavily censored. DCS is simply not observed in many individuals under study. While DCS might eventually develop in these individuals if the period of exposure were increased, the time to onset of DCS is unobservable for these individuals. The observed survival time on each of these individuals is said to be "censored". In the Air Force data that has been the basis of our work, the censoring is Type 1 or fixed. Every individual enters the experiment at a fixed time, and the time of exposure is pre-determined by the flight protocol.

Survival analysis consists of methods for identifying risk factors, also called "covariates", and assessing their effects on survival time. There are two fundamental approaches. In the parametric approach, the survival time T is assumed to have a known distribution with a particular functional form. Commonly used models are the exponential, Weibull, log-normal and log-logistic distributions. Physical and physiological models for DCS provide insights into the appropriate distributional model for T.

A feature of altitude DCS is that its incidence initially increases over time after decompression, but because of de-nitrogenation, this risk tends to level off and then decrease. Such behavior restricts our attention to so-called inverted bathtub models. Distributions that fall into this category include the log-normal, the log-logistic and the inverse Gaussian.

In the other non-parametric approach, no functional form for the underlying survival distribution is assumed. The most popular non-parametric approach entails use of the Cox Proportional Hazards model, in which the ratio of hazard functions of two individuals with covariate levels x_1 and x_2 is independent of time. The hazard in such cases is given by:

$$r(t \mid x) = r_0(t) \cdot \exp(x'\beta),$$

where $r_0(t)$ is the baseline hazard, x is the vector of covariates, and β is the vector of parameters to be estimated. Because the argument of the exponent on the right side of this expression is independent of time, $r(t)/r_0(t)$ is also independent of time. Subject and baseline hazards are consequently proportional. Specification of the baseline hazard can be a problem with this model. One approach is to use the Kaplan-Meyer estimate which is the non-parametric estimate of the survival function.

The advantages of the proportional hazards model include obviation of need to assume any particular form for the survival function. Additionally, risk factors that vary over time, or so-called "time-dependent co-variates", are readily incorporated, though the subject and baseline hazards are then no longer proportional. If the form of the survival function is known, however, parametric methods are always more powerful and hence preferred.

I won't go into too much detail on the background. You've heard most of the speakers. One of the earliest attempts at likelihood methods was -- were papers by Dr. Weathersby in '84 and '92, probabilistic models developed to predict the occurrence of DCS and likelihood methods used to find the estimates of the parameters. Then there were models based on bubble growth and mechanistic principles described by Van Liew et al. Kumar et al in a series of papers, one of the earliest attempts to actually use survival models to model the incidence of DCS. Most of the models were based on dichotomous response; you either observed or did not observe the symptoms, and different risk factors, such as tissue ratio, CMB, which is circulating micro-bubble stages, and the effects of exercise were all incorporated into these logistic models. Finally, Dr. Conkin has addressed this log-logistic model with tissue ratio and exercise, and also some models for bubble growth for the combination of the statistical and some mechanistic models for bubble growth.

The data set is from the US Air Force Armstrong Laboratory, consisting of records from 975 flight exposures to pressures below 314 mmHg. Data from additional exposures to pressure levels above 314 mmHg were omitted because the DCS incidence at these low altitudes was minimal. The data for each exposure included the altitude or pressure to which the individual was exposed (PRES), the planned time of the altitude exposure (TALT), the pre-oxygenation time (BR), and the level of exercise performed while at altitude. This exercise was classified into three categories; rest, mild and heavy; based on oxygen consumption. VGE data were also available for each exposure, collected at roughly 15-minute intervals. These were condensed into two variables; MAXB, which is the maximum VGE score observed during the entire exposure, and MAXD, which is the time at which the maximum VGE score was observed. Finally, the data for each exposure included the survival time or the time of DCS symptom onset.

Simple contingency tables were constructed in preliminary analyses to assess which factors should be considered and how such factors should be included in the models. The first of these is given in Table 1.

Table 1. Pre-Breathe Time (BR)

	CENSOR		
BR (min)	0 (%)	1 (%)	Total
0	90	10	535
15	35	65	23
60	47	53	610
75	44	56	127
90	62	38	13
135	56	44	116
240	100	0	2
Total			1426

The censor variable of zero (0) indicates no DCS, while the censor variable of one (1) indicates the individual had symptoms. We first looked at the O_2 pre-breathing time in minutes ranging from zero to 240, and we examined the DCS incidence in percent in the various categories. Conventional wisdom would indicate that the incidence of decompression should go down as the pre-breathe time increases, and that trend is evident in the tabulated results. However, there are some problems evident in these numbers. For zero minutes of pre-breathing, 90 percent of the individuals had no DCS, and 10 percent were symptomatic. The trend is not really increasing all the way through.

One of the reasons for this problem is people who pre-breathe tended to go to lower altitude, and the people who went up for longer, to higher altitudes, tended to have more pre-breathe time. So, that became a problem. You don't see the trend that you expect. How do you deal with this? One way is to make some kind of transformation of these risk factors.

Table 2. Maximum Observed Bubble Grade (MAXB)

	CENSOR		
MAXB	0 (%)	1 (%)	Total
0	84	16	608
1	76	24	71
2	54	46	96
3	56	44	209
4	39	61	439
Total			1425

We also examined the relationship between bubble grade and DCS incidence in these preliminary analyses. The maximum bubble grade is reported on the Spencer scale; zero for no bubbles and one through four for successively increasing levels of bubble profusion, and again percentage of DCS and no DCS for these different groups. We notice that for people with no circulating micro-bubbles, 84 percent were asymptomatic. However, if they had Grade 4, 39 percent of them were asymptomatic as well. So, the question is: What do bubble grades really tell us about DCS symptoms? This is a problem that we all have grappled with at some point or another.

The likelihood function for the parametric models we studied is given by:

$$L(\theta) = \prod_{i=1}^{M} f(t_i) \prod_{j=1}^{N-M} S(t_j)$$

where N is the total number of observations in the data set, θ is a vector of unknown parameters, M is the number of individuals with symptoms and for whom observed DCS onset times, t_i, are known, f(t) is the probability density function, and S(t) is the survival function for the N-M censored observations with no symptoms observed during the exposures. The density and survival functions are usually not of closed-form, and must be solved using iterative numerical techniques. So, you get the maximum likelihood estimates of the parameters, and once you have those, you can use the model with these predicted values to look at different profiles.

We used density and survival functions that give an inverted bathtub shape in their risk transforms. One such function was the log-normal distribution for the survival function, given by:

$$S(t) = 1 - \Phi\left(\ln(\lambda t/\sigma)\right); \qquad \lambda = \exp(-x'\beta)$$

where Φ is the cumulative distribution function (cdf) for the standard normal distribution, σ is the scale parameter for the normal, and β is a vector of unknown parameters associated with the covariate vector x.

We also used the log-logistic distribution for S(t):

$$S(t) = \frac{1}{1+(\lambda t)^{\gamma}}; \qquad \lambda = \exp(-x'\beta)$$

where γ is a scale parameter. The vector of risk factors enter into the model via λ in the exponential expression to the right.

We used three risk factors. The first, PRES, was simply the pressure in mmHg of the altitude exposure. The second variable, BRTALT, accommodated the effects of both O_2 pre-breathe time and time at altitude. BRTALT was defined as the ratio of one plus the pre-breathing time, BR, to time at altitude, TALT, or BRTALT=(1+BR)/TALT. This association of the

BR and TALT factors dealt with the problem of zero pre-breathes for low altitude exposures versus long pre-breathes for high altitude exposures. The last variable, EX, was a categorical variable with three levels to accommodate exercise effects.

We ran into some problems in the initial analysis. One of the problems arose from the large range of altitude exposure time; from 120 to 480 minutes; in the data. There was a lot of variation in the DCS onset time. So, we put weights into the likelihood function by breaking up the data into several categories. This is a relatively common way of dealing with large variations in data.

Table 3. Estimated Parameter Values for the Weighted Log-Logistic Model

Variable	DF	Estimate	Std.Err.	Chi-sq.	p-value
INT	1	-8.00	2.45	10.63	0.0011
PRES	1	2.53	0.44	32.57	0.0001
BRTALT	1	1.29	0.39	11.26	0.0008
EX	1	-0.53	0.14	13.68	0.0002
SCALE	1	0.60	0.03		

Max. Loglikelihood = -560.84

Table 3 gives the parameter estimates based on the maximum likelihood for this weighted model. The p-values for the three risk factors; pressure (PRES), the ratio of pre-breathe time to time at altitude (BRTALT), and exercise (EX); are all highly significant at p-values less than 0.001.

The maximum log-likelihood value is again a measure of how well a model fits the data. It's a negative 560.84. The fit of a log-logistic model can also be assessed by nesting it into the generalized F distribution. We found that the log-normal and log-logistic both provided satisfactory fits to the data, but we selected the log-logistic for further work because of its simpler form.

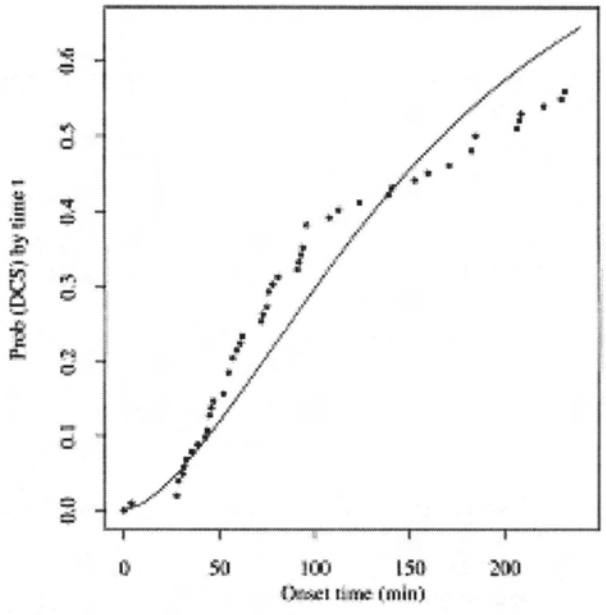


Figure 1. Cumulative DCS probability as observed (closed circles) and predicted by the weighted model (solid line) for a 240 min exposure with mild exercise to 231 mmHg after a 75 min O_2 pre-breathe.

DCS probabilities predicted by this model for several different flight profiles were compared with empirical data for those profiles. Figure 1 illustrates results for a 240 min exposure with mild exercise to 231 mmHg after a 75 min O_2 pre-breathe. The solid line is the predicted cumulative distribution function. The dots represent the empirical distribution or the actual data. As you can see, the predicted line tends to overlay the observed points rather well.

We ran a chi-squared test which is something you can do. It's not the best test for goodness of fit, but it does work. We broke up the data into several intervals and did an observed and expected test for this.

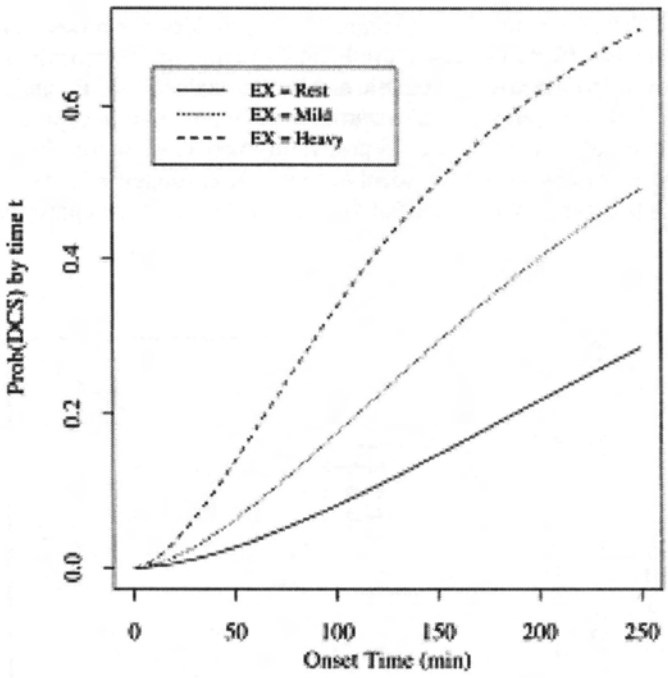


Figure 2. Effect of exercise on probability of DCS predicted by the weighted model during a 240 min exposure to 282 mmHg after a 60 min O_2 pre-breathe.

Figure 2 shows the effect of exercise on the probability of DCS predicted by the weighted model during a 240 min exposure to 282 mmHg after a 60 min O_2 pre-breathe. Three cumulative distribution functions are shown; one each for heavy exercise, mild exercise and rest. The probability of DCS at any exposure time is highest for heavy exercise and the lowest for rest, the sort of trend that is observed in the database.

In order to examine performance of a non-parametric model on these data, we also fit a proportional hazards model with the same three risk factors. Results are summarized in Table 4. Consistent with results from the parametric models, the p-values for all three risk factors are very significant. This model also provides an additional value for each covariate called the risk ratio, which has a useful practical interpretation. Considering the risk ratio for the EX parameter, for example, the risk of DCS for individuals who perform mild exercise is 100(2.066-1)% higher; i.e., almost double; the risk of individuals at rest, all other factors being the same.

Table 4. Estimated Parameters for the Proportional Hazards Model

Variable	Estimate	Std. Err.	Chi-sq	p-value	Risk ratio
PRES	-2.75	0.42	42.61	0.0001	0.064
BRTALT	-2.22	0.37	36.67	0.0001	0.109
EX	0.72	0.15	24.66	0.0001	2.066

Because the proportional hazards model is non-parametric, it provides a relatively unbiased description of the data. So, we used this model to examine the effect of different pre-breathing times on the probability of DCS in a particular profile for comparison to performance of the parametric log-logistic and log-normal models. Figure 3 shows the effect of pre-breathe time on DCS probabilities predicted by the proportional hazards model for a 240 min exposure with mild exercise to 282 mmHg. The model not only provides estimates of the DCS probability very close to the observed incidences, but yields cumulative distributions with shapes similar to those evident in Figure 2, in support of the theory that the inverted bathtub risk of the log-logistic model is appropriate. We concluded that the log-logistic is an appropriate model for DCS.

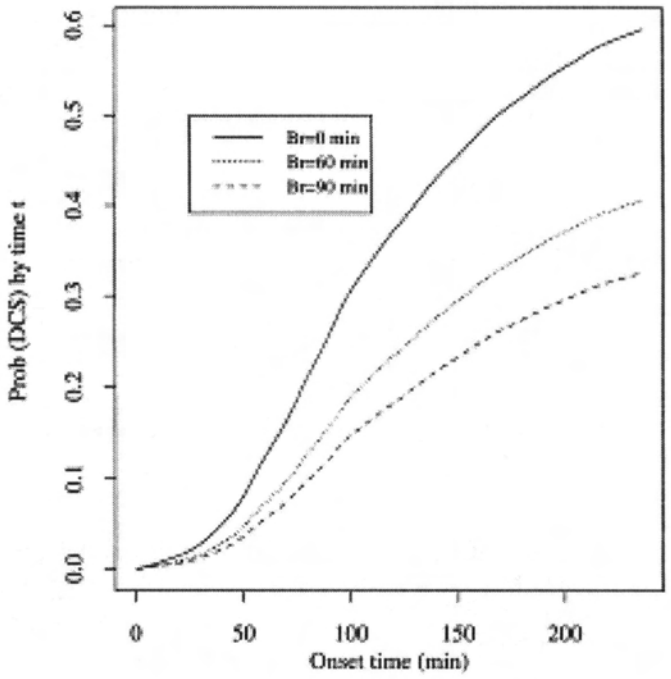


Figure 3. Effects of O_2 pre-breathe duration on DCS probabilities predicted by the proportional hazards model for a 240 min exposure with mild exercise to 282 mmHg.

Model performance is not sufficiently evaluated by comparing observed and predicted probabilities alone. Confidence bands for the cumulative distribution function must also be considered. This is not as easy as it sounds, however, because the models are heavily non-linear. You don't have closed-form solutions, and most commercially available software for survival analysis will give you confidence intervals for the DCS probability only at specified time points. In contrast, we wanted to get an overall band or envelope for the P(DCS), computation of which is more problematic. Most traditional confidence level bands are based on the Kolmogorov-Smirnoff test for the empirical distribution function. For censored data, we use the Kaplan-Meyer estimator, and again it becomes a huge problem. Not only can the bands fall below zero and

exceed unity, they also tend to be of constant, rather large width. Among the techniques that have been suggested to avoid these problems, the bootstrap technique is perhaps the better alternative.

The bootstrap technique was introduced relatively recently by Ephron in a paper in the Journal of the Royal Society. It's a resampling technique that allows you to use the data to generate a more efficient model or confidence band. Bickel and Kreeger in '89 showed that these techniques provide the right coverage probability, and for small sample data, they tend to also form the traditional methods based on the Kaplan-Meyer estimator. So while most techniques for calculating confidence bands work well for large samples because of their asymptotic properties, the bootstrap is particulary well suited for small samples. We decided to use the bootstrap to generate confidence bands. For a specific profile, we repeatedly generate separate samples with replacement from the original data. For each such bootstrap sample, we re-estimated the parameters by maximizing the likelihood function and then obtained the estimate of the cumulative distribution function (cdf) at several time points. Finally, the 5th and 95th percentiles of the cdf estimates at each time point were used as the 90% confidence bands. We processed 500 bootstrap samples for each profile in the dataset, and maximized the likelihood for each sample. Needless to say, this was a very compute-intensive procedure.

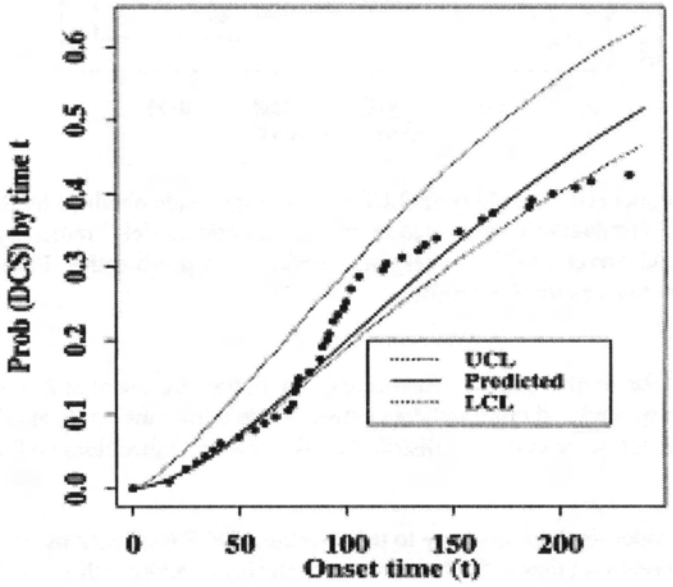


Figure 4. 90% upper (UCL) and lower (LCL) confidence bands obtained by the bootstrap technique for DCS probabilities predicted by the log-logistic model (Predicted) for a 240 min exposure with mild exercise to 231 mmHg after a 135 min O_2 pre-breathe. Filled circles are the observed DCS incidences for this profile.

Figure 4 illustrates results for one profile. The solid line is the predicted cumulative DCS probability, the points are the actual data points from the database, and the dotted lines are the upper and lower 90-percent confidence bands for the cumulative distribution function. Note that the confidence bands are not symmetric about the predicted line, a feature common in bootstrap-derived confidence bands. As you can see, the bands tend to be narrow at low exposure times and widen as exposure time increases. This behavior is expected as well, but it is clear that most of the observed DCS rates fall within the 90-percent band. These results support the idea that the model was working all right.

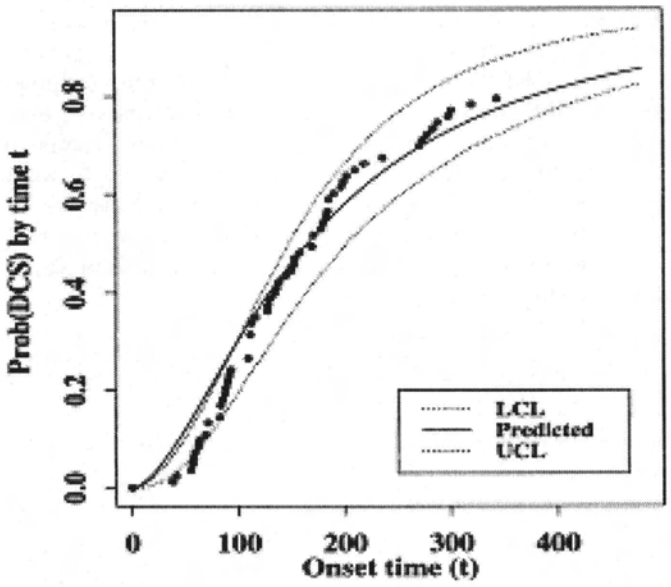


Figure 5. 90% upper (UCL) and lower (LCL) confidence bands obtained by the bootstrap technique for DCS probabilities predicted by the log-logistic model (Predicted) for a 480 min exposure with mild exercise to 253 mmHg after a 60 min O_2 pre-breathe. Filled circles are the observed DCS incidences for this profile.

Figure 5 illustrates results for another profile, consisting of an altitude exposure at 253 millimeters for 480 minutes, and mild exercise. As before, the predicted and observed data almost all lie within the upper and lower 90-percent bands. The model consequently captures nearly all of the empirical distribution function, and thus does well not just in goodness of fit but also in prediction.

One problem with these models is their tendency to under-predict DCS incidence by about 15 to 20 percent for profiles with zero or very low pre-breathing times. We have been attempting to address this problem by using VGE data in the form of the time of occurrence of the maximum observed bubble grade. We noticed that people with early observed Grade 4 VGE tended to have symptoms fairly early as well. We did a simple analysis using the time to observed maximum VGE, or MAXT, as another risk factor in our log-logistic model. Results are summarized in Table 5.

Table 5. Parameter Estimates for the Log-Logistic Model with MAXT

Variable	DF	Est.	Std. Err.	Chi-sq.	p-value
INT	1	-3.66	1.80	4.12	0.0424
PRES	1	1.34	0.31	17.37	0.0001
BRTALT	1	0.96	0.30	10.21	0.0014
MAXT	1	0.01	0.00	183.58	0.0001
SCALE	1	0.37	0.02		

Looking at the chi-squares, the MAXT variable tends to dominate all of the other risk factors. This is not surprising because MAXT is also affected by the exposure altitude, pre-breathing time and exercise level; i.e. MAXT covaries with these other factors. We hoped that including VGE data would alleviate model underprediction of DCS probability for profiles with low pre-breathing times.

MAXT can alternatively be defined as a deterministic variable obtained from solution of an appropriate bubble growth model for each exposure. One such model, for example, is being developed at the Armstrong Laboratory to predict the time at which bubbles reach their maximum radius. As with observed times to maximum VGE grade, using MAXT values from this bubble growth model also yields a dramatically improved model over that without a MAXT factor. However, the one concern I have is that the time to maximum bubble size is really random. Because it is not realistically considered as a deterministic factor, it should be incorporated as a time-dependent co-variate, which considerably complicates the analyses. But that's something to think about for the future.

In conclusion, log-logistic models seem to be most appropriate for DCS data. They have the right shape for the risk function, and predictions from these models agree closely with empirical data. The confidence bands contain observed DCS rates for most profiles. We hope the VGE data will improve the predictions for zero and low pre-breathe times, but that's going to be a difficult exercise, and how we incorporate it is still debatable. I believe that both statistical and mathematical models together, so we talk about mechanistic and non-mechanistic approaches, need to be used in conjunction to provide an overall better model to predict DCS. A validation study is currently underway at Brooks Air Force Base which will help us further fine-tune the model.

DR. GERTH: I think we do have one question. Dr. Southerland?

QUESTION: Dr. Southerland, NEDU. I thought that one of the requirements for boot-strapping was that the resampled data had to be independent and randomly selected from the population. The data that you are using, however, does not seem to be sampled that way. How does the kind of sampling you are doing affect your confidence in your confidences?

DR. KANNAN: Okay. I did the boot-strapping by sampling from all individuals or all the data points for each specific profile. For example, if I had 150 observations on a pressure of 282 and an exposure time of four hours, I used that as the initial data from which the bootstrap sample was taken. In that sense, I think our resampling procedure was okay. In contrast, if the bootstrap samples are taken from the entire database, the resultant confidence bands will be very far from correct.

Multinomial Bubble Score Model

Peter Tikuisis[1] and Keith A. Gault[2]
[1]Defence and Civil Institute of Environmental Medicine, Toronto, Canada M3M 3B9
[2]Navy Experimental Diving Unit, Panama City, FL

Model Demonstration

I would like to acknowledge Keith Gault , my co-author, as the person who conducted most of this research for his Master's degree.

It is generally accepted that bubbles are the precursors to decompression sickness (DCS) although the exact linkage between the two is not known. Bubble models are increasingly being used to predict DCS. We propose that the best test of a bubble model is to apply it against data that measure bubbles directly. This presentation will focus on the development of a model to predict the occurrence of bubbles as measured using Doppler ultrasound techniques.

Data

The data were obtained from measurement sites on the precordium, the left subclavian, and right subclavian of the diver's body at rest and exercise during and after a dive. Measurements were taken approximately every 30 to 40 min. Signals were graded according to frequency, duration, and amplitude, and converted into a single bubble grade, BG, ranging from zero indicating no bubble activity to four indicating maximum bubble activity using the Kisman-Masurel code (Nishi 1993). The resultant data are multinomal and categorical.

Both air and helium-oxygen (heliox) dives conducted at DCIEM were used in this study (see Table 1). The more than 2,000 man-dives in the data set included a variety of dives, single and repetitive, and some with oxygen decompression.

Table 1. Summary of dive data with BG recordings.

	Air	Heliox
Number of Trials	276	86
Number of man-dives	1,425	639
Bottom Depth (msw)	7.3 - 91	36 - 100
Bottom Time (min)	5.0 - 350	19 - 287
Dive Time (min)	9.5 - 300	19 - 287

The distribution of the bubble grade data are: BG = 0 (no bubbles detected) > 40%; BG = 1 and 2 (low bubble activity) < 30%; and BG = 3 and 4 (high bubble activity) < 30%. The advantage of these data is that they are fairly evenly distributed in contrast to the typically low occurrence of DCS. We purposely grouped bubble grades (i.e., 1 & 2, and 3 & 4) from the original 5 categories for modeling purposes, as explained below.

Model

We assume that the maximum predicted bubble size correlates with the maximum recorded bubble grade. The model does not take into account the times of occurrences of these maxima. The method of maximum likelihood estimation is used to fit the parameters using a modified Marquardt algorithm (Bailey and Homer 1977).

The bubble model is documented elsewhere (Tikuisis et al. 1994, Gault et al. 1995); hence, only key points will be presented here. The gas flux depends on the concentration difference between the gas in the bubble and that outside the

bubble in the tissue or fluid, and on the rate constant representing a diffusion barrier to gas transfer across the gas-bubble interface:

$$J_i = k_i \cdot \Delta C_i \tag{1}$$

where J is the gas flux, k is the rate constant, and ΔC is the gas concentration difference across the bubble interface for gas 'i'. The gas content of the bubble, N^g, is constrained by a mass balance and the amount of gas leaving or entering the bubble depends on the gas flux:

$$\frac{dN_i^g}{dt} = 4\pi R^2 \cdot J_i \tag{2}$$

Finally, tissue gas exchange with the blood is assumed to be perfusion-limited and is characterized by a time constant, τ, according to (Hills 1977):

$$\frac{dP_i}{dt} = \frac{P_i^{bl} - P_i}{\tau_i} \tag{3}$$

where P is the gas tension and the superscript 'bl' refers to its value in blood.

We now present the probability functions used to predict bubble grades. First, we consider the simpler binomial case. If our purpose was only to predict incidences of low versus high BG, the following expressions would suffice:

$$\begin{aligned} Pr_{lowBG} &= e^{-a \cdot R_{max}} \\ Pr_{highBG} &= 1 - e^{-a \cdot R_{max}} \end{aligned} \tag{4}$$

where Pr is the probability of a BG outcome, as illustrated in Fig. 1. The prediction of low BG prediction is a single parameter estimation involving R_{max} as an exponent. Clearly, predictions of a high incidence of low BG are associated with low values of R_{max}. Conversely, a high incidence of high BG is associated with large values of R_{max}. By definition, the two probabilities sum to unity. This binomial approach to BG prediction is analogous to the prediction of DCS.

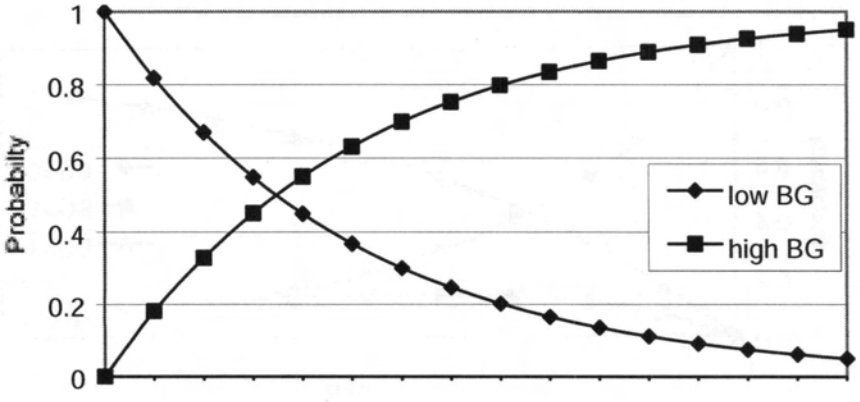


Figure 1. Illustration of the binomial probability function.

To provide a more elaborate prediction, an additional probability function was introduced to allow a trinomial outcome (based on advice from Dr. L.D. Homer). This allowed a separate prediction of BG = 0, as follows:

$$\begin{aligned} Pr_{BG=0} &= e^{\frac{-a \cdot R_{max}}{b}} \\ Pr_{BG=\{1,2\}} &= e^{\frac{-R_{max}}{b}} - e^{\frac{-a \cdot R_{max}}{b}} \\ Pr_{BG=\{3,4\}} &= 1 - e^{\frac{-R_{max}}{b}} \end{aligned} \quad (5)$$

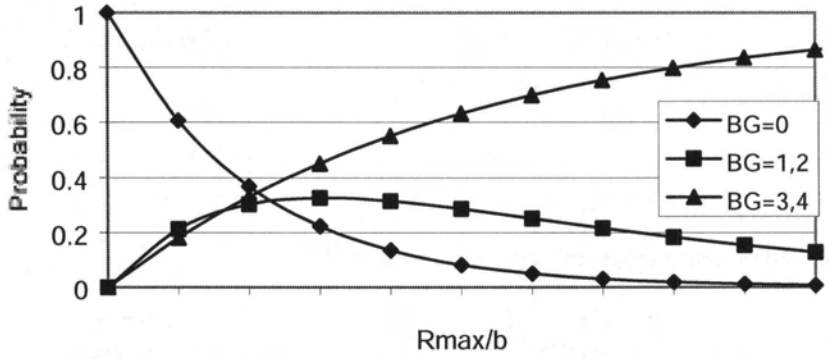


Figure 2. Illustration of the trinomial probability function (Eq. 5).

Close examination of the low BG (= 1, 2) prediction reveals that a second parameter, b, has been introduced, and the shape of this probability function is distinctly different from the other two. Yet, the summation of all probabilities to unity is preserved.

Figures 3 and 4 demonstrate the influence of parameter 'a' on the probability distributions. Low values of 'a' suppress the prediction of low BG. A higher value will boost the maximum probability of low BG allowing the possibility that the probabilities of the other two outcomes can be exceeded within a certain range of R_{max}.

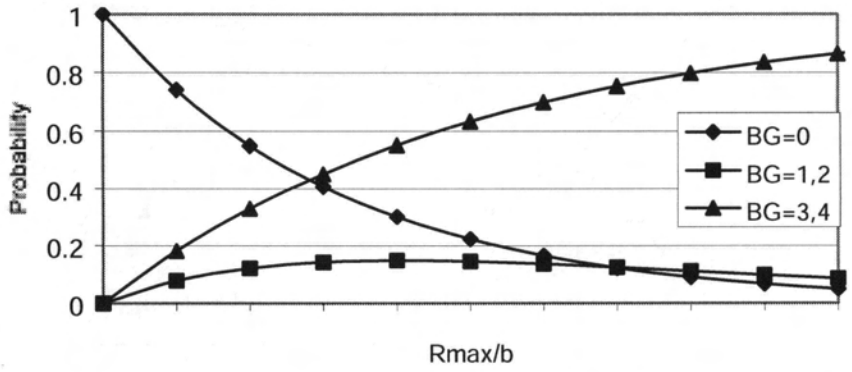


Figure 3. Illustration of the trinomial probability function with a low value of 'a'.

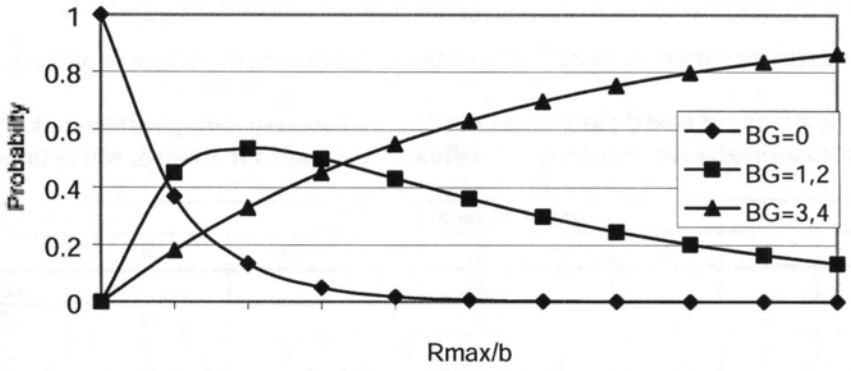


Figure 4. Illustration of the trinomial probability function with a high value of 'a'.

The likelihood function, L, is the product of the above probabilities, each raised to the power of the number of times, N, that the corresponding Doppler score was observed, i.e.:

$$L = Pr_{BG=0}^{N_0} \cdot Pr_{BG=\{1,2\}}^{N_{1,2}} \cdot Pr_{BG=\{3,4\}}^{N_{3,4}} \quad (6)$$

For computational convenience, the logarithm of L is used for parameter estimation:

$$LL = \sum N_0 \cdot \ln Pr_{BG=0} + N_{1,2} \cdot \ln Pr_{BG=\{1,2\}} + N_{3,4} \cdot \ln Pr_{BG=\{3,4\}} \quad (7)$$

Since all probabilities are less than or equal to zero, maximum likelihood is attained when LL has the lowest possible negative value.

The complete list of model parameters are the time constant of the tissue (τ), the rate constant of gas transferring in and out of the bubble (k), the gas solubility in tissue (δ), the volume of the tissue surrounding the bubble (ν), the surface tension of the bubble (γ), and the two scaling parameters (a, b). The expression that relates the gas solubilities and tissue volume is given in Tikuisis et al. (1994) and Gault et al. (1995), and primarily involves the minimum bubble size condition for bubble growth. The rationale for the other parameter choices are also presented in the cited references.

Results

Table 2 summarizes the variety of model configurations used to achieve the best fit of the data.

Table 2. Summary of model parameters used (√) and corresponding maximum log Likelihood values. Two estimates (for nitrogen and helium) were made for k and δ, where indicated by a √.

Parameters							LL
τ	k	δ	ν	γ	a	b	
							-2,229.7*
√	√					√	-2,189.1
√	√				√	√	-2,181.8
√	√	√				√	-2,158.9
√	√	√			√	√	-2,149.4
√	√	√	√	√	√	√	-2,146.4

**null model estimation*

The nine-parameter version (see Table 3) yielded the most significant improvement according to the log-likelihood ratio test. Note that the time constants of nitrogen and helium are markedly different, but their relative order is consistent with expectation. However, the large difference in gas solubilitities between the two gases suggest different tissue types. This is problematic but not surprising considering that the fit was performed on a single tissue (lack of data precluded the use of additional model tissues). The estimated value for the tissue volume is within the range used by other researchers. The very low estimated surface tension concurs with values reported by Paul Weathersby et al. (1982).

Table 3. Summary of model parameter estimates (± SE) of the best fit. The estimate of τ for helium was derived from the value for nitrogen.

Parameters	Nitrogen	Helium
τ (min)	27.9 ± 1.9	9.3 ± 8.3
k ($cm \cdot s^{-1}\ 10^{-6}$)	0.050 ± 0.013	55.5 ± 120
δ	0.0438 ± 0.0002	0.0096 ± 0.0079
ν ($cm^{-1}\ 10^{-4}$)	3.6 ± 0.9	
γ (dyne cm^{-1})	5.0 ± 2.2	
a	2.55 ± 0.09	
b ($cm\ 10^{-2}$)	1.39 ± 0.16	

Discussion

How well does the model fit bubble observations outside the calibration data set? We begin our examination of this question with a 45-meter seawater (msw) dive on air for 30 min. This was an experimental dive which involved sedentary divers half immersed in cold water, very unlike the dives in the data set used for the model calibration. As can be seen in Fig. 5, the model prediction is poor. This illustrates the model's lack of generality and risk of overextrapolation.

The next dive examined was similar, but involved working dives conducted on a semi-closed circuit breathing apparatus. Better agreement was obtained in this case. The next dive examined was a 45 msw for 30 min on heliox, and a similar level of agreement between the observed and predicted BG was attained as with the previous air dive. The last dive examined was a 15 msw for 4 days saturation dive where no bubbles were detected, in accordance with the model prediction of a zero bubble grade.

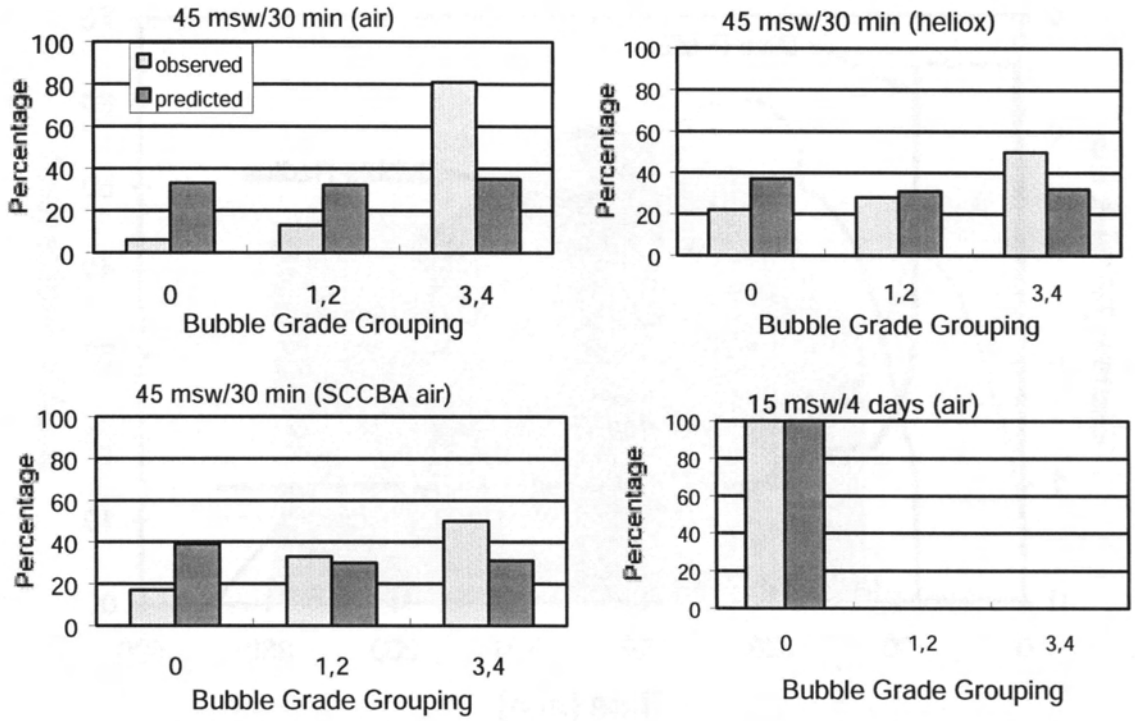


Figure 5. Comparison of observed and predicted bubble grades for various dives.

Although we have not used any time of occurrence information in the model estimation, it is informative to compare the evolution of predicted bubble sizes and observed bubble grades. The dive profile examined for this purpose in Fig. 6 is a 45 msw for 50 min dive on air. Superimposed on the plot of bubble radius are bubble grades for the six divers (identified by numbers 1 through 6) involved in this trial. For example, diver #1 had BGs of 0, 3, and 4 at about 70, 120, and between 150 and 220 min, respectively, and then BG began to decrease. Indeed, the history of all the divers' BGs tend to be described by the shape of the predicted bubble radius envelope. This reasonably strong chronological correlation is remarkable considering that the parameter estimations were based on the maximum bubble grade recorded, not on when it occurred or on the events preceding or following the maximum occurrence.

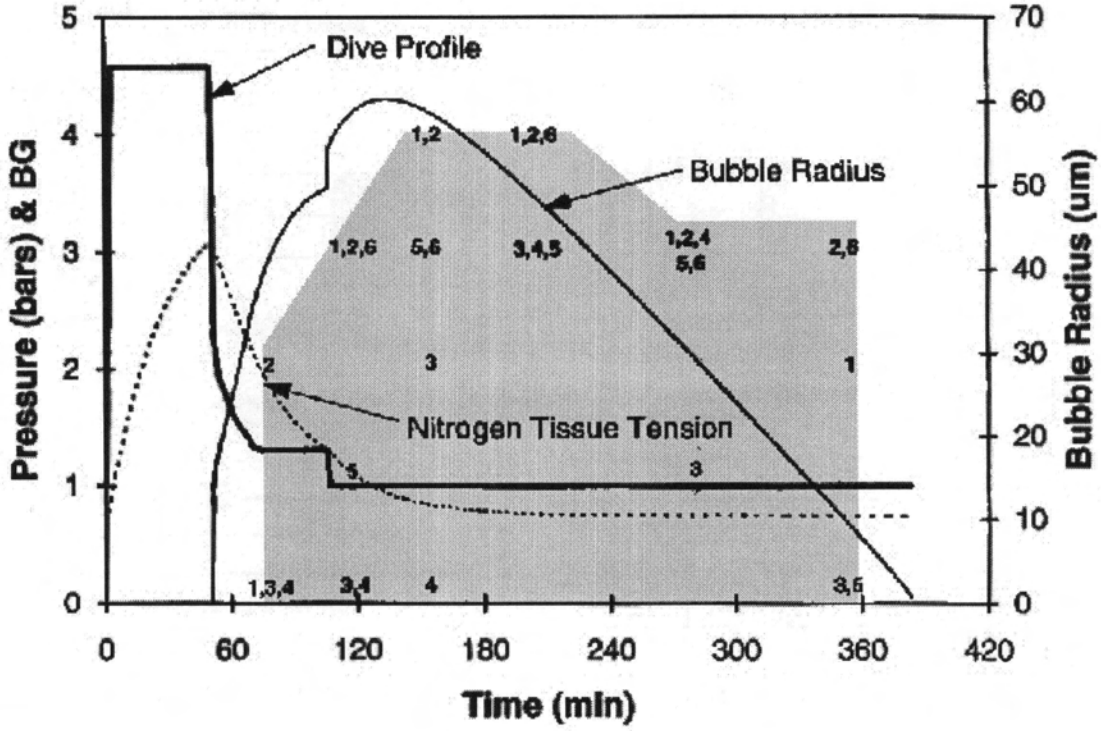


Figure 5. Comparison of observed and predicted bubble grades for a 45 msw for 50 min dive on air. The heavy solid line defines the dive profile; the dotted line shows the predicted tissue gas tension; the thin solid line shows the predicted bubble radius; and the numbers within the shaded region refer to the diver's identification and indicate the BGs recorded (scale on left axis).

Summary

The present model distinguishes three levels of intravascular bubble activity and represents a departure from the binomial outcome distribution usually applied in the prediction of DCS. It can be referred to as a 'competing risk' model according to the co-chair's (Dr. W.A. Gerth) overview.

Although we have some difficulties with the interpretation of some of the model parameter estimates, they fell within the range of acceptable human biological values. The strong chronological correlation between the maximum predicted bubble size and the maximum recorded bubble grade is important. This is because the time of these maxima lagged the occurrence of maximum gas super-saturation which generally occurs upon surfacing and has been considered representative of the maximum instantaneous risk of DCS. The bubble model prediction is consistent with the general observation that DCS symptoms usually occur after surfacing.

References

Bailey, R.C. and Homer, L.D. (1977). An analogy permitting maximum likelihood estimation by a simple modification of general least squares algorithms. Naval Medical Research Institute Report No. 86-51, Bethesda, MD.

Nishi, R.Y. (1993). Doppler and ultrasound bubble detection. In: Bennett, P.B. and Elliott, D.H. (eds.) The Physiology and Medicine of Diving. Saunders, London, pp. 433-453.

Tikuisis, P., Gault, K.A. and Nishi, R.Y. (1994). Prediction of decompression illness using bubble models. Undersea Hyperbaric Med. 21:129-143.

Gault, K.A., Tikuisis, P. and Nishi, R.Y. (1995). Calibration of a bubble evolution model to observed bubble incidence in divers. Undersea Hyperbaric Med. 23:249-262.

Hills, B.A. (1977). Decompression sickness, Vol 1: The biophysical basis of prevention and treatment. John Wiley & Sons, New York.

Weathersby, P.K., Homer, L.D. and Flynn, E.T. (1982). Homogeous nucleation of gas bubbles in vivo. J. Appl. Physiol. 53:940-946.

Probabilistic Models of Decompression Sickness During Flying After Diving: Motivation for Mechanism

Wayne A. Gerth
F.G. Hall Laboratory
Center for Hyperbaric Medicine and Environmental Physiology
Duke University Medical Center
Durham, NC 27710

Objectives

One of our goals in the F.G. Hall Laboratory is to develop a model under which decompression sickness (DCS) survival data from diving, altitude, and flying after diving can be combined and used to predict DCS risk during flying after diving. My objective in this presentation is to discuss issues that must be considered in construction of the hazard function to achieve this goal. These issues include the types of explanatory variables, or *covariates*, that must be accommodated, and the motivation for adoption of an unabashedly "mechanistic" approach to risk function design. We will see that this approach is the distinguishing feature between two very different classes of models that have been used to describe DCS occurrence.

Model Forms and Explanatory Variables

Any model is applicable only to populations for which the explanatory variables and factors in the model are known and in which no other confounding factors are active to influence the modeled outcome. In other words, any model is applicable only when all relevant heterogeneity in the population is accommodated in the covariates. Well-characterized parametric statistical distributions that meet this requirement are readily available. In this Workshop, for example, models for altitude DCS have been described based on the log-logistic function with *time-independent* covariates [Conkin, Kannan; This Workshop]. The hazard function for the i^{th} individual in these models is generally expressed:

$$h_i(t) = \frac{\lambda p \cdot g(\mathbf{z}_i, \boldsymbol{\beta})(\lambda t)^{p-1}}{1 + g(\mathbf{z}_i, \boldsymbol{\beta})(\lambda t)^{p}}, \qquad (1)$$

where $g(\mathbf{z}_i,\boldsymbol{\beta})$ is a function of a vector of parameters, $\boldsymbol{\beta}$, and the vector of explanatory variables for the individual, $\mathbf{z}_i$. These parametric models work well for modeling responses to exposures that have very simple time courses. For example, a pressure and respired gas profile for a contingency EVA (Extravehicular Activity) from Space Shuttle is illustrated in Figure 1. The profile consists of a switch to 100% O_2 breathing at the start of a 4 hr prebreathe, a 30 min decompression to space suit pressure of 4.3 psia (equivalent to the barometric pressure at an altitude of 30,300 ft in the US Standard Atmosphere), and a 6 hr EVA followed by return to cabin pressure and air breathing.

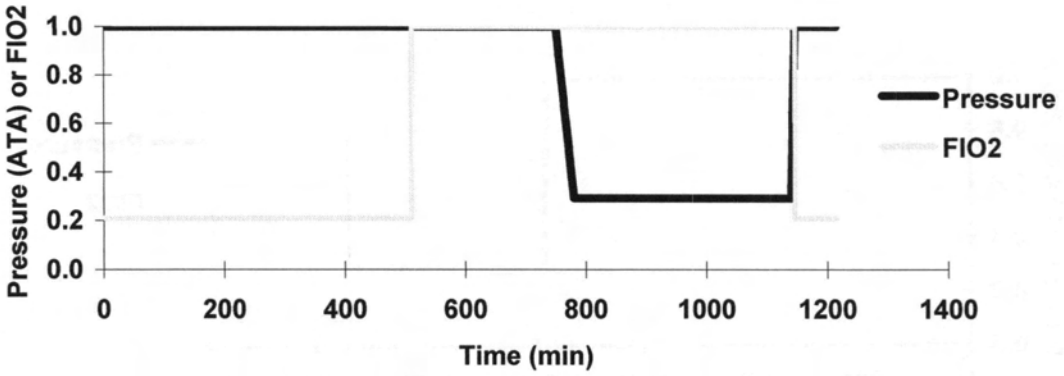


Figure 1. Shuttle contingency EVA pressure and respired gas profile consisting of a 4 hr 100% O_2 "prebreathe," a 30 min decompression to 4.3 psia Shuttle suit pressure, and a 6 hr EVA at suit pressure.

This exposure and other similar "square" decompressions can be rather easily described in terms of the ratio of the N_2 tension in a DCS-governing "tissue" at the time of decompression to the ambient pressure at altitude ("tissue ratio" or TR), the ambient pressure at altitude, and the time at altitude. Computation of the tissue ratio then requires a single additional parameter, the O_2 prebreathe time. Each of these covariates is a constant for any given exposure, allowing *Eq.* (1) to be used as a model of DCS hazard, with t=0 at the time of arrival at altitude.

If we complicate the exposure, however, assumptions must be made about potential confounding factors or such a model becomes inapplicable. For example, Figure 2 illustrates the pressure and respired gas profile for a nominal Space Shuttle EVA. The exposure includes decompression to a 12 hr stage at 10.2 psia before a second and final decompression to the EVA suit pressure of 4.3 psia. The profile begins and ends with air breathing, but 26.5% O_2 in N_2 is breathed during the 10.2 psia stage, and 100% O_2 is breathed starting 40 min before the final decompression and throughout the exposure at 4.3 psia. Figure 3 shows how this profile must be idealized to apply a simple log-logistic model of DCS that accommodates only a single decompression (no staging).

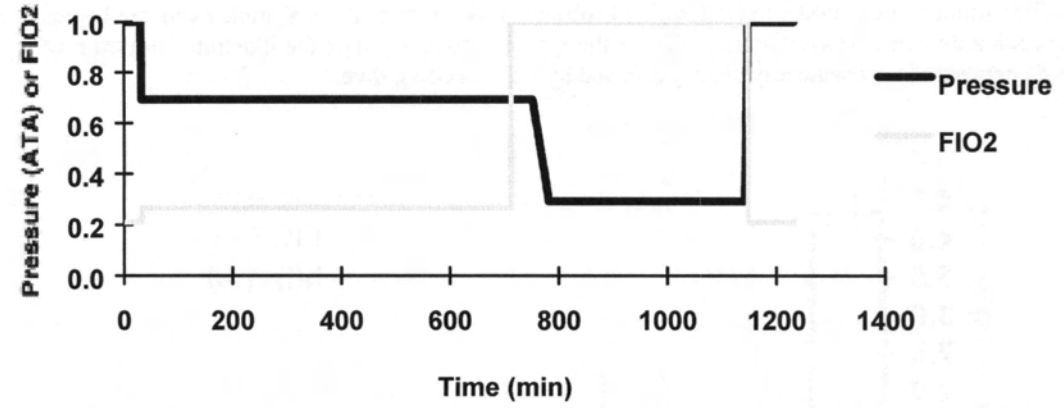


Figure 2. Shuttle EVA pressure and respired gas profile consisting of 12 hr residence at 10.2 psia breathing 26.5% O_2 in N_2 followed by 6 hr exposure at Shuttle suit pressure of 4.3 psia.

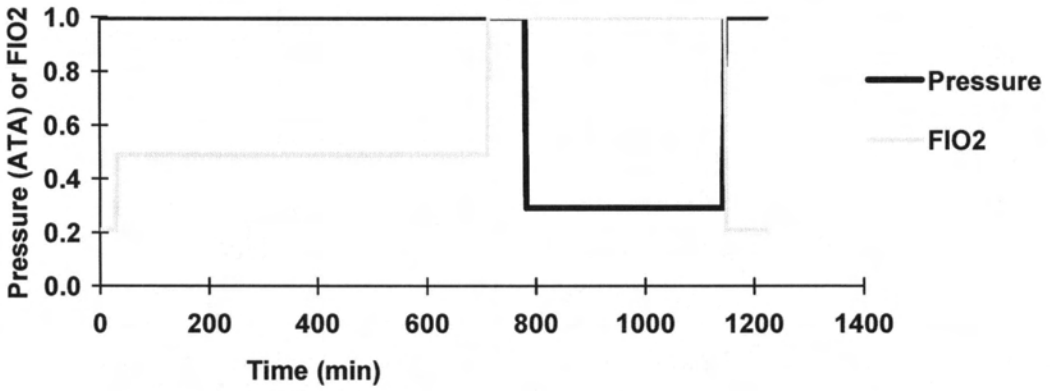


Figure 3. Idealization of profile in Figure 2 required to consider it in terms of a simple model applicable only to DCS occurrence after a single decompression.

As required by a simple model based on a single TR value for a given exposure, the profile in Figure 3 contains only a single decompression from 1.0 atm cabin pressure to 4.3 psia suit pressure. The 12 hr 10.2 psia stage between 60 and 780 min in Figure 2 is considered to be spent at sea level pressure, with the inspired oxygen fraction adjusted to yield the same alveolar PO_2 as that breathed during the 10.2 psia stage. The decompression is depicted as practically instantaneous, reflecting that effects on TR of N_2 washout during decompression are neglected. Such effects can be considered by adding another parameter, the ascent rate, to the covariate vector[13]. The profile in Figure 3 is equivalent to that in Figure 2 from the standpoint of DCS risk only if the change in pressure from 1 atm to 10.2 psia in the Figure 1 profile has no effect.

The presumption that a given decompression is unaffected by a preceding decompression is not always reasonably made. This is particularly true in flying-after-diving, where outcome during the altitude exposure is confounded non-negligibly by the diving history preceding it. A hypothetical flying after diving profile is illustrated in Figure 4, along with a corresponding DCS hazard profile estimated by a model to be described subsequently. Note that the altitude exposure begins before DCS risk from the preceding dive has decayed to zero. Under the model used to estimate the illustrated hazard profile, DCS risk during the altitude exposure is consequently clearly affected by the preceding dive.

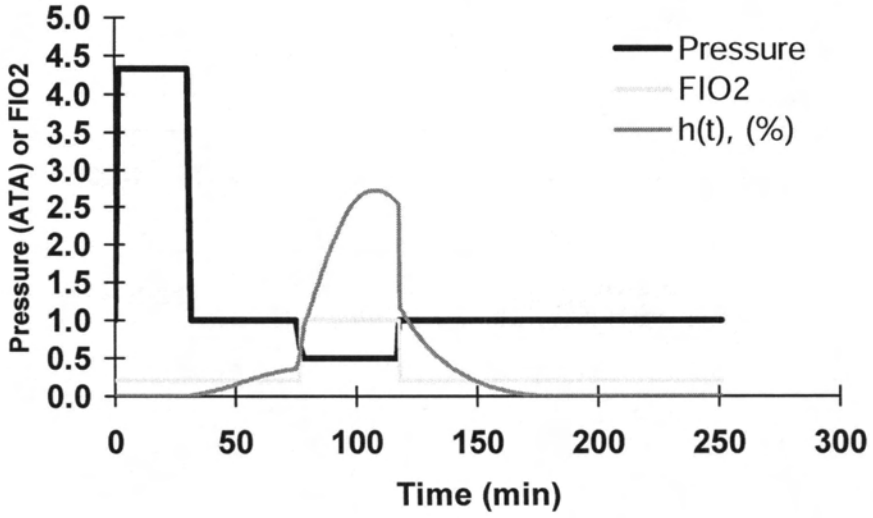


Figure 4. Flying after diving profile with corresponding model-estimated hazard profile.

The confounding effect of the dive history preceding the altitude exposure can be accommodated in simple models like that in *Eq.* 1 by addition of more elements to the **z** and **β** vectors of explanatory variables and parameters. In general, however, the number of time-independent covariates and parameters must increase in proportion to the complexity of the profiles to which the models will be applicable.

Time-Dependent Covariates

The proliferation of parameters with increasing profile complexity is overcome by using *time-dependent* covariates. The model then responds to a *covariate process* that is manifest in the temporal variations of each element in the covariate vector. These variations are not fixed in the model *per se*, but are accommodated by model structure fashioned to respond as these covariates change through time. Thus, the complexity of this structure does not need to increase as the complexity of the covariate process increases. Additionally, models based on such covariates allow computation of intermediate probabilities of failure based on covariate vectors that are complete only up to arbitrary times, so that such models can be used in real-time applications.

Time-Dependent Covariate Types

Time-dependent covariates fall into two classes; external and internal[6]. An external covariate is not directly involved in the failure mechanism, and can be *defined*; i.e., determined in advance for each individual under study although not constant through time; or *ancillary*; i.e., the output of a stochastic process external to the individual at risk. Thus a series of discrete pressure, time, and inspired oxygen values can be used to describe the time course of exposure in a planned laboratory decompression trial. The matrix formed by this series is an example of defined external time dependent covariates.

In comparison, an internal covariate is the output of a stochastic process that takes place in the individual under study. Because an internal covariate requires survival of the individual to generate the output, it carries intrinsic information about the survival history of the individual. VGE grade through time after decompression is a good example of a time-dependent internal covariate of interest in DCS studies.

Use of time dependent covariates requires recognition that the hazard at any time t is conditioned on the covariate process up to that time but no further. As a result, only values of such covariates before or concurrent with failure in a given individual can serve as predictors of failure.

Time-Dependent Covariates in an Accelerated Log-Logistic Hazard

Time-dependent covariates can be incorporated into models based on well-characterized parametric statistical distributions. For example, the hazard function for the i^{th} individual in log-logistic models for altitude DCS that use time-dependent covariates can be generally expressed:

$$h_i(t) = \frac{\lambda p \cdot g(\mathbf{z}_i(t), \boldsymbol{\beta})(\lambda t)^{p-1}}{1 + g(\mathbf{z}_i(t), \boldsymbol{\beta})(\lambda t)^{p}} \qquad (2)$$

where $g(\mathbf{z}_i(t), \boldsymbol{\beta})$ is a function of a vector of parameters, β, and a vector of time-dependent explanatory variables for the individual, $z_i(t)$. This type of model is called an *accelerated failure time model.* Note that in comparison with z_i in *Eq.* (1), $z_i(t)$ in *Eq.* (2) is the covariate vector for the i^{th} individual at time t. Formulation of the $g(\mathbf{z}_i(t), \boldsymbol{\beta})$ function is especially challenging because no particular form is suggested on purely mathematical or statistical grounds. However, in a DCS occurrence model, any formulation of this function must produce an instantaneous risk function, $h_i(t)$, that behaves in conformance with known or hypothetical influences of pressure, breathing gas and time in the $z_i(t)$ matrix. Thus, our only

guidance for specifying $g(\mathbf{z}_i(t),\boldsymbol{\beta})$ comes from our etiological understanding of DCS. It is then far simpler to specify $h_i(t)$ directly -- in terms of $z_i(t)$ and β -- than to specify $g(\mathbf{z}_i(t),\boldsymbol{\beta})$.

This can be illustrated with the following example. Let us stipulate existence of a hypothetical pressure profile that, in accord with known influences of pressure, breathing gas and time, would be thought to produce a pure, constant amplitude sinusoidal profile of DCS hazard. This sinusoidal form is easily expressed, but if used as $g(\mathbf{z}(t),\boldsymbol{\beta})$ in *Eq.* (2), yields the curve labeled "h(t), accel" in Figure 5.

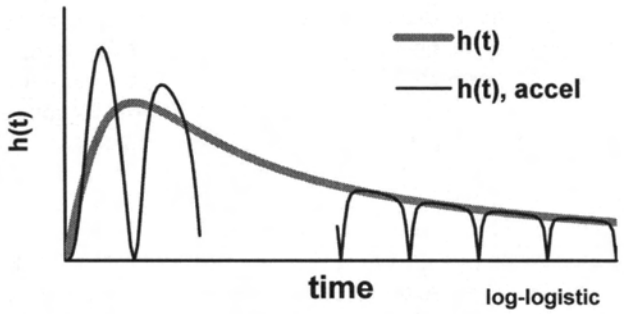


Figure 5. Plot of Eq. (2) with arbitrary values of λ and p and $g(\mathbf{z}(t),\boldsymbol{\beta})=1$ [curve labeld h(t)] and $g(\mathbf{z}(t),\boldsymbol{\beta})=1+\sin(\omega t+k)$ [curve labeled h(t), accel]. The sinusoid in the latter form of $g(\mathbf{z}(t),\boldsymbol{\beta})$ is superimposed on the underlying shape of the log-logistic distribution, h(t).

Features of the sinusoid are clearly evident, but with increasing time on the abscissa, its amplitude attenuates and its downward excursions become more acute under influence of the underlying log-logistic function. In this case, the sinusoid $g(\mathbf{z}(t),\boldsymbol{\beta})$ is the hazard h(t) and added influence of another distribution, regardless of its particular form, is undesirable.

As a result of such considerations, we have been compelled to eschew the strengths of well-characterized statistical distributions in favor of "mechanistic" forms for the hazard function that are defined wholly in terms of one or more explicitly-modeled processes as they evolve under the influence of time-dependent covariates. Under this "mechanistic" formalism, the theoretical processes and their descriptions completely prescribe the form of the arithmetic and logical relationships between the independent variables, parameters, and hazard in any given model. As with models based on time-independent covariates, mechanistic models can be "empirical," without claim that the modeled process(es) actually describe the real physiologic processes involved, or "physiologic", attempting to describe those processes as completely and accurately as possible.

Mechanistic Models of DCS Occurrence

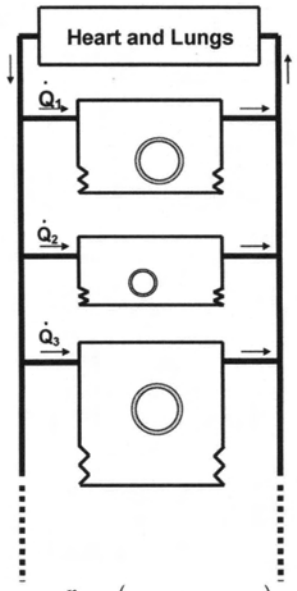
We have been developing physiologic mechanistic models, schematized in Figure 6, to describe DCS occurrence survival data. In these models, the body is envisioned as a collection of parallel-perfused compartments that each exchange gas with the atmosphere via the circulation and lungs. If this exchange lags behind a decrease in ambient pressure so that gas-supersaturation is produced in a compartment, one or more bubbles may nucleate and grow to relieve the gas-supersaturation and produce risk of DCS. These models provide the flexibility to accommodate both increasing or decreasing risk after decompression and a mechanistic foundation that conforms to the well-accepted idea that DCS is initiated by *in vivo* bubble formation and growth [5]. The different gas and bubble dynamics equations used in these models are described elsewhere [2, 3, 5].

These models require a continuous description of the pressure and respired gas profile. This is provided by encoding each profile as a sequence of nodes that give the pressure or depth and the inspired O_2 fraction at particular times in the profile. An unbroken description of the profile is then obtained by linear interpolation in the time domain between pressures

and respired O_2 fractions at successive nodes. Each node thus gives the conditions prevailing at the end of a profile stage that is either a travel stage (compression or decompression), an isobaric stage, a breathing gas switch stage, or a combination travel and breathing gas switch stage. The model is then exercised on the profile by sequentially processing these stages, preserving the model state at the end of each stage as the initial state for the next stage. A detailed format (the Augmented NMRI Standard Format) for listing profile nodes with outcome information and compiling these lists into machine-readable data sets of one or more profiles has been developed as described in Appendix A.

At the outset of our attempts to model DCS occurrence during flying after diving, two different gas and bubble dynamics models had been developed to model DCS incidence and time of occurrence in either diving or altitude exposures. In the diving DCS model [3, 5], schematized in Figure 6.A, only a single bubble is allowed to nucleate and grow in the each of three modeled compartments to produce DCS risk. This restriction is relaxed in the single compartment of the altitude DCS model [2], schematized in Figure 6.B, where multiple bubbles are allowed to nucleate and grow subject to a nucleation model described by Yount [17].

A



$$h(t) = \sum_{i=1}^{n} G_i \left(V_{B,i}(t) - V_{B,i}^{o} \right)$$

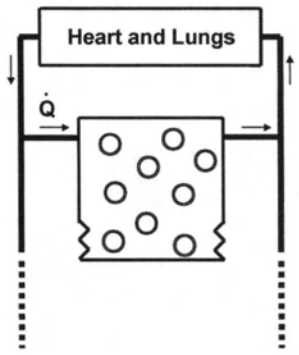
where

G_i = gain, compartment i

$V_{B,i}(t)$ = prevailing bubble volume, compartment i

$V_{B,i}^{o}$ = nuclear (or initial) bubble volume, compartment i

B



$$h(t) = \frac{\left(V_B(t) \quad V_B^{o} \right)}{V_t}$$

where

G = gain

$V_B(t)$ = prevailing bubble volume

V_B^{o} = nuclear (or initial) bubble volume

$N(t)$ = number of bubbles in compartment at time t

V_t = compartment volume

Figure 6. Schematics of: (A) three-compartment BVM(3) model for diving DCS, and; (B) single-compartment model for altitude DCS.

Each of these models describes observed DCS incidences and times of occurrence rather well in its own domain, but seriously underestimates these properties for exposures in the other's domain. Diving and altitude DCS data could thus not be combined under either of these models. As illustrated in Figure 7, however, observed DCS occurrence density distributions from these data appear sufficiently similar to motivate continued search for a single model that can accommodate all these data.

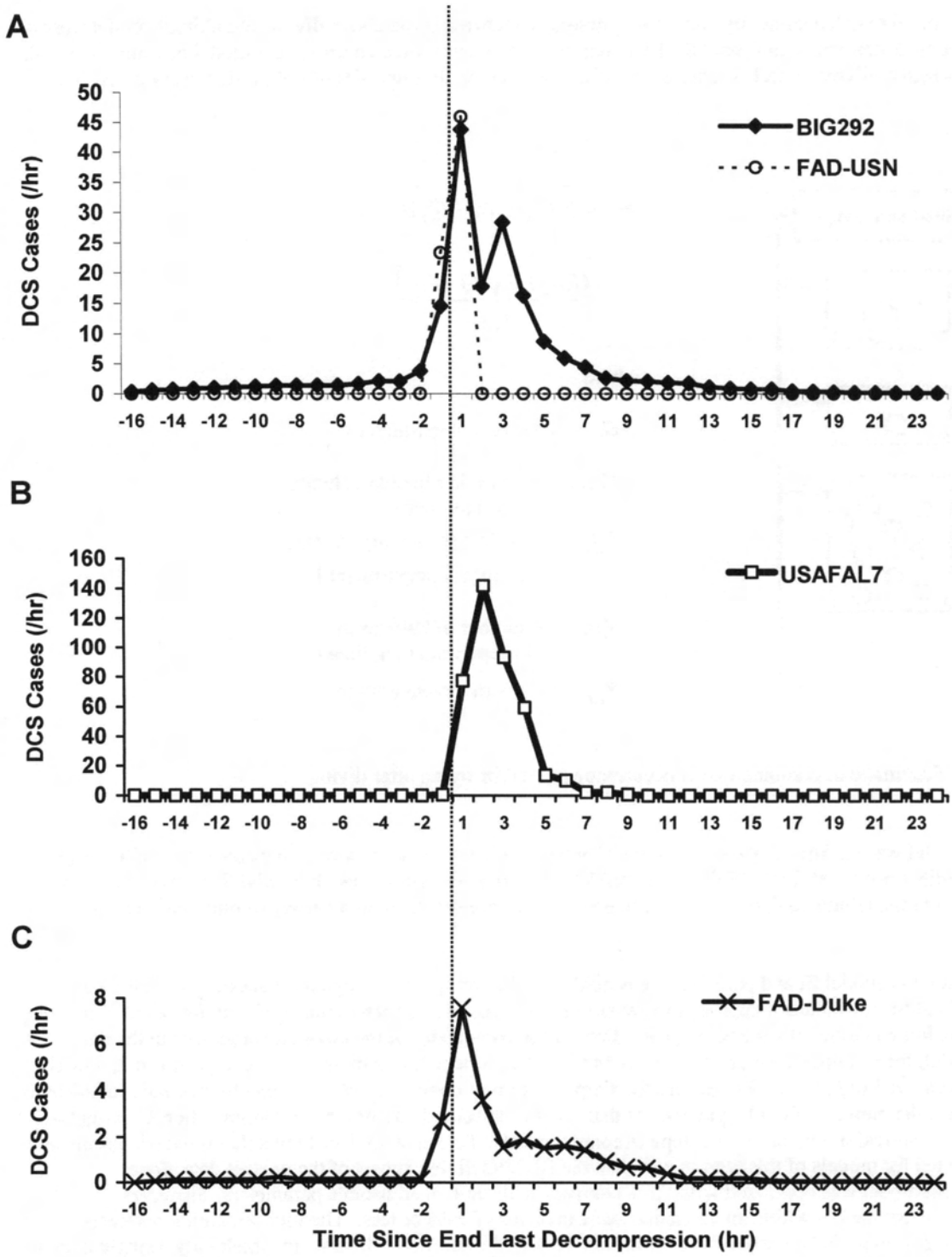
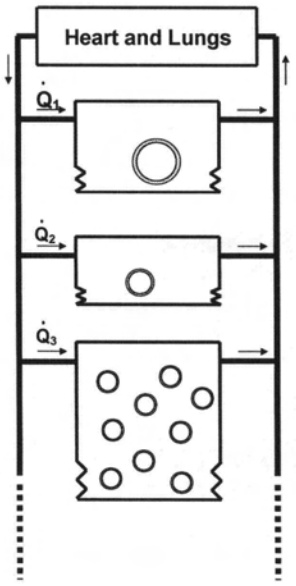


Figure 7. Observed DCS occurrence density distributions from four large data sets of diving (BIG292), altitude (USAFAL7) and flying after diving (FAD-USN, FAD-Duke) exposures. Lines between indicated data points in each panel are drawn for clarity only. The vertical dotted line is drawn through zero time since last decompression in the three panels. Other important summary information about the data sets is given in Table 1.

We began our search for such a model by combining important features of the above diving and altitude models into a single "combined" model schematized in Figure 8. This combined model is a three compartment model in which one of the compartments in the original diving model is replaced by a compartment with properties of the single compartment in the altitude model.



$$h(t) = \sum_{i=1}^{2} G_i\left(V_{B,i}(t) - V^o_{B,i}\right) + \frac{G_3 N_3(t)\left(V_{B,3}(t) - V^o_{B,3}\right)}{V_{t,3}}$$

where

G_i = gain, compartment i

$V_{B,i}(t)$ = prevailing bubble volume, compartment i

$V^o_{B,i}$ = nuclear (or initial) bubble volume, compartment i

$N_i(t)$ = number of bubbles in compartment i at time t

$V_{t,i}$ = volume, compartment i

Figure 8. Schematic of combined DCS occurrence model for flying after diving.

This combined model was optimized about a data set of 5663 exposures and known outcomes on diving, altitude and flying-after-diving profiles compiled from USN, USAF, NASA and Duke data resources. Essential features of the data are given in Table 1. Each of the tabulated data sets has been described in greater detail in a variety of other publications [3, 7-11, 14, 15].

A detailed evaluation of model fit and performance is neither within our present scope nor necessary to illustrate advantages and pitfalls of the mechanistic approach to hazard function design. A disadvantage of such functions is that they can require a relatively large number of parameters, regardless of the complexity of the covariate processes in the data. The present combined model, for example, contains 8 parameters for compartment 1, 8 parameters for compartment 2, and 10 parameters for compartment 3 (*c.f.*, Figure 8). Except for simplification of compartment 3 into one functionally identical to the other compartments, the number of model parameters that must be specified, either through optimization or assignment, can be reduced only by removal of one or more complete compartments. Earlier work had shown that three compartments are statistically warranted for models of this type to correlate the BIG292 diving subset of the present data alone. Accordingly, the present model was optimized with three compartments and 26 adjustable parameters. However, only 22 of these parameters were determined to within an estimated standard error of 30% or less. The four remaining parameters were associated with the compartment that emerged from optimization with the shortest half-time, in which only a single bubble could nucleate and grow after decompression. Each of these had an estimated standard error that exceeded the parameter value by a factor of 30 or more, and was thus poorly constrained by the data. Nevertheless, model structure requires nonzero values for these parameters, though the optimization could almost certainly have been completed after fixing these four parameters at reasonable arbitrary values.

Table 1. Training Data Summary Description

	Exposure Type	Data Set	# Exposures	# DCS Incidents	# DCS Marginals	Total DCS*
NMRI BIG292 (Diving)	Single Air	EDU885A	483	30	0	30.0
		DC4W	244	8	4	8.4
		SUBX87	58	2	0	2.0
		NMRNSW	91	5	5	5.5
		NSM6HR	57	3	2	3.2
	Single Air, Decompression	PASA	72	5	2	5.2
	Repetitive Air	EDU885AR	182	11	0	11.0
		DC4WR	12	3	0	3.0
		PARA	135	7	3	7.3
	Multi-Level Air	PAMLA	236	13	12	14.2
	Single Non-Air	NMR8697	477	11	18	12.8
		EDU885M	81	4	0	4.0
		EDU1180S	120	10	0	10.0
	Repetitive Non-Air	EDU184	239	11	0	11.0
		EDU885S	94	4	0	4.0
	Multi-Level,SDV; PO2=0.7 Decompressions	PAMLAOD	134	6	0	6.0
	Multi-Level,SDV; PO2=0.7 Transits	PAMLAOS	140	5	3	5.3
	Air Saturation	ASATEDU	120	13	27	15.7
		ASATNSM	132	18	21	20.1
		ASATNMR	50	1	0	1.0
	Non-Air Saturation	ASATARE	165	20	13	21.3
	Subtotals		3322	190	110	201.0
USN_FAD (Flying after Diving)	Misc. flying after diving	FAD-NAVY	128	69	13	70.3
USAF (Altitude)	Armstrong Laboratory, Brooks AFB altitude	USAFAL7	1194	401	0	401.0
DUKE_High O2 (Diving)	Single Air; O2 Decompression	MOONVQ	69	3	3	3.3
DUKE_SIO2 (Diving)	Repetitive Air, Surface O2	SIO2-85	197	4	1	4.1
		SIO2-93	38	1	0	1.0
	Subtotals		235	5	1	5.1
DUKE_NOAA (Diving)	Repetitive Non-Air, Surface O2	NOAA	94	1	0	1.0
DUKE_FAD (Flying after Diving)	40/120-SI-SimComFl	FAD40-1	51	1	1	1.1
	60/55-SI-SimComFl	FAD60-1	122	5	10	6.0
	60/55-1hr-60/20-SI-SimComFl	FAD60-2	126	6	3	6.3
	60/55-1hr-60/20-1hr-60/20-SI-SimComFl	FAD60-3	100	2	0	2.0
	100/20-SI-SimComFl	FAD100-1	109	7	2	7.2
	100/15-1hr-60/35-SI-SimComFl	FAD10060	113	5	3	5.3
	Subtotals		621	26	19	27.9
	Grand Totals		5663	695	146	709.6

** Each marginal outcome counted as 0.1 DCS*

A great advantage of the mechanistic approach is illustrated in Figure 9, where performance of the combined model on a flying after diving profile from the Duke FAD60-3 data set is shown. The complexity of this profile compared with those in Figures 1 and 3 is immediately apparent. The profile consists of three dives to 60 fsw separated by two 1 hr surface intervals, a 17 hr preflight surface interval, and a 4 hr exposure to 8,000 ft altitude. Model use of time-dependent covariates allows it to accommodate this profile, as well as other profiles of arbitrary complexity, without modification.

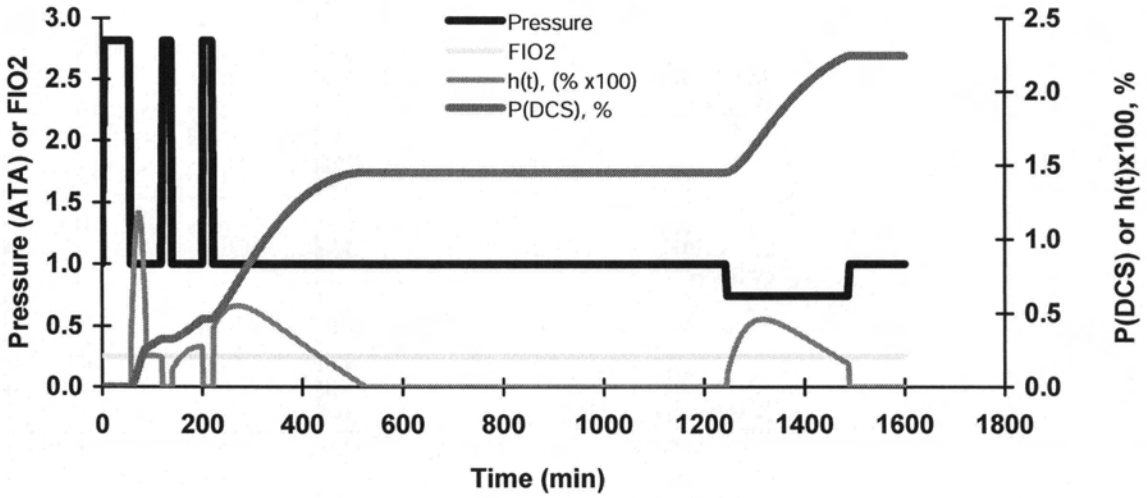


Figure 9. Flying after diving profile with corresponding model-estimated hazard and cumulative DCS probability profiles.

Evaluation of physiologic mechanistic models is also naturally extended beyond the usual assessments of goodness-of-fit by examining the hazard function as a formal statement of the hypothesis that the observed outcomes were in fact products of the processes that the hazard function was designed to represent. In the present case, DCS risk is formally expressed as a function of simulated bubble volumes and number densities in a collection of hypothetical, parallel-perfused gas exchange compartments. Model parameters are consequently associated with biophysical quantities that are independently measurable at least in principle, such as gas solubilities, diffusivities, blood flows, etc. If the model equations and logic provide complete and correct representations of the processes that actually gave rise to the DCS cases in the data, optimized parameter values should be within range of their corresponding measured or "physiological" values. Earlier examinations of this kind lead to the conclusion that the diving DCS model schematized in Figure 6.A was incomplete as a mechanistic description of DCS etiology, even though it provides a good empirical correlation of a large body of diving DCS experience [3, 5].

The present model, however, runs into trouble before consideration of its optimized parameter values. The altitude portion of the profile in Figure 9 is typical of all Duke-FAD profiles. The model result of zero hazard after return to ground at the end of the 4 hr altitude exposure is also typical of model performance on these profiles. Comparison to panel C in Figure 5 shows that the observed DCS hazard persists long after bubbles in the model, and hence DCS risk, have resolved with return to ground level. The model consequently fails to account for a relatively large number of DCS cases that were observed to occur after completion of the altitude portion of the profiles. We have not yet been able to account for this observed hazard in terms of prevailing bubble volume.

This model difficulty illustrates another challenge of the mechanistic approach. The behavior of mechanistic models is intrinsically constrained by model structure, so that elaboration of the model – and addition of parameters – within that structure may not improve model fit to a given data set. Model improvements may instead require expansion of model scope itself to accommodate additional mechanistic processes. In the present case, the indication is that such elaboration must include representation of processes that are initiated by bubble formation but persist to cause DCS risk after bubbles have resolved.

A final motivation for the physiologic mechanistic approach is belief that as a model embodies more complete and accurate representations of the processes that give rise to the outcome of interest, the model will provide more accurate extrapolations from the model calibration data. This has been demonstrated in one case where a gas and bubble dynamics model of DCS occurrence provides more accurate predictions of DCS incidence than a more empirical model in application to dive profiles in which high oxygen partial pressures are breathed [4]. Both models were calibrated about the same data, from which such profiles were purposefully excluded.

Conclusions

The complexity of the covariate processes that govern DCS outcomes under conditions of usual practical interest, such as during flying after diving, compels adoption of a "mechanistic" design of the hazard function in models of DCS occurrence. However, functions engineered to express DCS hazard in terms of the prevailing bubble number density and volume achieve only limited success in application to combined diving, altitude and flying-after-diving data. Yet more complex function(s) are required to account for etiologic processes that are initiated by bubble formation but persist to cause DCS after the bubbles have resolved.

Literature Cited

1. American Society for Testing and Materials. Standard Practice for Use of the International System of Units (SI). Philadelphia, PA, 1989; Document E380-89a.

2. Gerth WA, Vann RD. Statistical bubble dynamics algorithms for assessment of altitude decomopression sickness incidence. Brooks AFB, TX: USAF Armstrong Laboratory, 1995; Technical Report, AL/CF-TR-1995-0037.

3. Gerth WA, Vann RD. Development of iso-DCS risk air and nitrox decompression tables using statistical bubble dynamics models. National Oceanic and Atmospheric Administration, 1996; Final Report, NA46RU0505.

4. Gerth WA, Vann RD. A probabilistic gas and bubble dynamics model provides improved estimates of DCS risk during oxygen decompression from air dives. Undersea and Hyperbaric Medicine 1997; 24(S):29.

5. Gerth WA, Vann RD. Probabilistic gas and bubble dynamics models of DCS occurrence in air and N_2O_2 diving. Undersea and Hyperbaric Medicine 1997; 24(4):275-92.

6. Kalbfleisch JD, Prentice RL. The Statistical Analysis of Failure Time Data. New York, NY: Wiley, 1980.

7. Kiessling RJ, Duffner GJ. The development of a test to determine the adequacy of decompression following a dive. 1960; NEDU Research Report 2-60.

8. Kiessling RJ, Wood WB. The development of a test to determine the adequacy of decompression following a dive. Phase II. 1961; NEDU Research Report 3-61.

9. Logan JA. An evaluation of the equivalent air depth theory. 1961; NEDU Research Report 1-61.

10. Temple DJ, Ball R, Weathersby PK, Parker EC, Survanshi SS. The Dive Profiles and Manifestations of Decompression Sickness Cases After Air and Nitrogen-Oxygen Dives. Volume I: Data Set Summaries, Manifestation Descriptions, and Key Files. Washington, D.C.: Bureau of Medicine and Surgery, Department of the Navy, 1999; NMRC 99-02(Vol. I).

11. Temple DJ, Ball R, Weathersby PK, Parker EC, Survanshi SS. The Dive Profiles and Manifestations of Decompression Sickness Cases After Air and Nitrogen-Oxygen Dives. Volume II: Complete Profiles and Graphic

Representations for DCS Events. Washington, D.C.: Bureau of Medicine and Surgery, Department of the Navy, 1999; NMRC 99-02(Vol. II).

12. United States Committee on Extension to the Standard Atmosphere. U.S. Standard Atmosphere, 1976. Washington, D.C.: Supt. of Docs., U.S. Govt. Print. Off. (Stock No. 003-017-00323-0); National Oceanic and Atmospheric Administration (NOAA-S/T 76-15672), 1976.

13. Van Liew HD, Burkard ME, Conkin J. Testing of hypotheses about altitude decompression sickness by statistical analyses. Undersea and Hyperbaric Medicine 1996; 23(4):225-33.

14. Vann RD. Flying After Diving: A Database. Paul J. Sheffield. Flying After Diving: Proceedings of the Thirty-Ninth Undersea and Hyperbaric Medical Society Workshop. Bethesda, MD: Undersea and Hyperbaric Medical Society, 1989: 179-222.

15. Vann RD, Gerth WA, Denoble PJ, Sitzes CR, Smith LR. A Comparison of Recent Flying after Diving Experiments with Published Flying after Diving Guidelines. Undersea and Hyperbaric Medicine 1996; 23(S):36.

16. Weathersby PK, Survanshi S.S., Nishi RY, Thalmann ED. Statistically Based Decompression Tables VII: Selection and Treatment of Primary Air and N2O2 Data. Bethesda, MD, 1992; NMRI Technical Report 92-85.

17. Yount DE. Skins of varying permeability: A stabilization mechanism for cavitation nuclei. Journal of the Acoustical Society of America 1979; 65:1429-39.

Appendix A. Augmented NMRI Standard Format for DCS Data.

The Augmented NMRI Standard data file format is an elaboration of the NMRI Standard data file format described by Weathersby, *et al.* [16] for encoding dive profiles and their outcomes. The Augmented NMRI Standard data file format is backward-compatible with the NMRI Standard data file format. Thus, an interpreter for Augmented NMRI Standard data files can read files that conform to the older NMRI Standard format, but Augmented NMRI Standard data files may NOT be read using a strict NMRI Standard format interpreter.

A pressure exposure or profile can be described as a sequence of nodes that each provides point-in-time information about the prevailing ambient pressure, inspired gas and exercise. An unbroken description of the exposure is then obtained by connecting the pressures, inspired gas fractions and exercise levels at successive nodes with straight lines in the time domain. Each node can consequently be considered to describe the conditions prevailing at the end of a profile stage that may have been either a travel (compression or decompression) stage, an isobaric stage, a breathing gas switch stage, or a combination travel and breathing gas switch stage. The NMRI Standard and Augmented NMRI Standard formats use such a node-by-node convention with modifications to simplify recording of breathing gas switches and exercise periods.

The left-most fields for each line or node of a profile are strictly defined in this description. However, the format can be further augmented to accommodate additional information in new fields to the right of those defined in this description, and to the left of any optional COMMENT field (see below), as individual users may require. Profiles coded with such additional information, but otherwise conforming to the present format description, should be machine-readable by an Augmented NMRI Standard format interpreter without modification. The format is consequently flexible while providing a mechanism for ready exchange of essential information about diving, flying and flying after diving exposures.

Each profile is entered separately using standard ASCII characters. A profile is defined as a unique history of pressure, gas breathed, and outcome including symptom times. If more than one subject has the same profile, the replication can be noted in one profile and not entered again. The format of a profile is (items in square brackets are not always required):

Line 1: Identification data (free character labelling) [$P##] [$*##] [$TD]
Line 2: Originating gas, [No. of exposures, Outcome,] [T1,T2] [!comment]
Line 3: Time (min), Pressure (depth, fswg), [New gas, Switching time] [, Exercise Code] [!comment]
Line 4: ... Same as line 3
. " "
. " "
Last line: -9999.0 (or -0000.0)

Entries need not be column aligned and different levels of precision can be used for Pressure and Time entries. Separating commas between entries are required as shown. Each comma can optionally be followed by one or more spaces. Default units of time are elapsed minutes and default units of pressure are feet sea water gauge. These can be changed for a given profile using the $P and $TD commands in Line 1 as described below. A Time-Pressure node entry is required only when a change in the slope of the pressure or breathing gas profile occurs, such as at the start of compression, or when the breathing gas is changed. Changes in pressure or breathing gas are assumed to be time-linear between successive nodes. The maximum number of nodes in a profile is limited by software, not the format itself.

"Originating gas" is the Gas Code (see below) for the breathing gas on which the subject(s) were saturated at profile start (i.e., at 0.0 elapsed time). Such saturation is always assumed at reference "Surface Pressure," which by default is sea level pressure (0 fswg, 1.0 ATA). Saturation on "Originating gas" at a pressure other than the Surface Pressure is effected by entering a first node with Time=0.0 and Pressure equal to the desired saturation pressure. The reference Surface Pressure is NOT reset.

NOTE: The Duke interpreter of the NMRI Standard and Augmented NMRI Standard formats assigns saturation on "Originating gas" at 1.0 ATA Surface Pressure as the default initial condition for all profiles. A Time=0.0 first node is required only if the initial saturation pressure is not 1.0 ATA or the breathing gas is switched at profile start. Other interpreters may require a Time=0.0, Pressure=Surface Pressure (e.g.; Depth=0.0) first node to signal initial saturation on

"Originating gas" at Surface Pressure even when no breathing gas switch occurs at profile start. See DEFAULTS AND OPTIONS below for methods to code other initial conditions.

"No. of exposures" is an integer value indicating the number of occurrences of the profile that produced the indicated "Outcome", including any indicated T1 and T2 times. If the "No. of exposures" field is blank or zero, the profile is hypothetical, or one for which outcome data are not available. Such profiles are omitted from model optimizations, but can be included when the data set is used in other model exercises.

"Outcome" is a floating number of value 0.0 (no DCS) or 1.0 (DCS). Note that at least two profiles must be entered to record a dive in which DCS occurred in x of N participants ($0<x<N$); one profile to record the (N-x) exposures completed DCS-free, and the other to record the otherwise identical x exposures that resulted in x incidents of DCS. A fractional value to specify a "marginal" outcome is also allowed. In order to avoid ambiguity arising from this latter option, at least two profiles must be entered to record x incidents of DCS and y incidents of "marginal" DCS in N participants in any given dive ($0<x+y\leq N$); one profile to record the x incidents of DCS and the other to record the otherwise identical y exposures that resulted in y incidents of "marginal" DCS.

T1 is the "time last known DCS-free" and T2 is the "time at which definite DCS was first present" in total minutes elapsed for each of the *No. of exposures* subjects with the indicated *Outcome* ($0.0<Outcome\leq 1.0$). A separate profile must be entered for each set of DCS occurrences with unique *Outcome*, T1 and T2 combinations during any given dive in which multiple individuals participated.

"Originating gas" and "New gas" entries are specified using one of the Gas Codes listed in Table A.1.

TABLE A.1.Gas Codes in the Augmented NMRI Standard Format[a,b]

Code	FO2 or PO2	FN2 or PN2	FHe or PHe	Comments
Air and Nitrox codes				
1.mn	21% of Pamb	Balance	0.0	Air; mn ignored
2.mn	mn.0% of Pamb	Balance	0.0	Constant FIO2 = 0.mn
3.mn	0.mn ATA	Balance (Pamb-PH_2O-PO2)	0.0	Constant PIO2 = 0.mn ATA
13.mnopqr	m.nopqr ATA	Balance (Pamb-PH_2O-PO2)	0.0	High Constant PIO2: *c.f.*, 3.mn
Heliox (He-O2) codes				
4.mn	mn.0% of Pamb	0.0	Balance	Constant FIO2 = 0.mn
5.mn	0.mn ATA	0.0	Balance (Pamb-PH_2O-PO2)	Constant PIO2 = 0.mn ATA
15.mnopqr	m.nopqr ATA	0.0	Balance (Pamb-PH_2O-PO2)	High Constant PIO2: *c.f.*, 5.mn
Tri-mix (He-N2-O2) codes				
6.mnopqr	mn.o% of Pamb	pq.r% of (Pamb-PO2)	Balance	Constant FIO2
7.mnopqr	mn.o% of Pamb	p.qr ATA	Balance	Constant FIO2, Constant PIN2
8.mnopqr	m.no ATA	pq.r% of (Pamb-PO2)	Balance (Pamb-PH_2O-PO2-PN2)	Constant PIO2
9.mnopqr	m.no ATA	p.qr ATA	Balance (Pamb-PH_2O-PO2-PN2)	Constant PIO2, Constant PIN2
10.xxxxxx	PREVIOUS PO2, ATA	PREVIOUS PN2, ATA	Balance (Pamb-PH_2O-PO2-PN2)	(add He)
11.xxxxxx	PREVIOUS % of Pamb	PREVIOUS % of Pamb	Balance	(vent)
12.mnopqr	m.no ATA	pq.r% of Pamb	Balance (Pamb-PH_2O-PO2-PN2)	Constant PIO2, Constant FIN2
14.xxxxxx	PREVIOUS PO2, ATA	Balance (Pamb-PH_2O-PO2-PHe)	PREVIOUS PHe	(add N2)
16.mnopqr	mn.o% of (Pamb-PH_20)	pq.r% of (Pamb-PH_20)	Balance	Constant FIN2 and FIHe, wet gas

[a] In accord with the original NMRI Standard format, the Augmented NMRI Standard prescribes that constant PIO_2 gas codes be interpreted as wet-gas descriptions; FIO_2=PIO_2/(Pamb-PH_2O), where Pamb is the ambient hydrostatic pressure; because such gases are usually delivered from re-breathers. Other codes describe gases as from source, dry, NOT adjusted for water vapor.

[b] When using constant PIO2 codes (e.g., 13.mnopqr or 15.mnopqr), care should be taken to ensure that the PIO2 never exceeds the ambient pressure, Pamb. Depending on the interpreter, such instances either may cause a profile READ error or be accommodated through default assignment of FIO2=1.0 until the erroneous condition resolves with an increase in ambient pressure or a breathing gas change.

When a "New gas" is specified in a line, the change in alveolar oxygen partial pressure from the old gas is assumed to occur linearly over a period of time equal to the indicated "Switching time" beginning at the time and pressure specified in the line. Note that this convention obviates need for a separate node entry to signal completion of a breathing gas switch. The obviated node, which is implied by the "New gas" node, is inserted by the interpreter. Switching times for successive breathing gas switches cannot overlap: If a given gas switch has not completed before another is specified, a profile READ error is signaled.

The Duke interpreter allows a blank line to be inserted between successive profiles in a given file; i.e., before the first comment line of the second and following profile(s). Additionally, any comma preceded by another comma with no or only blank intervening characters is interpreted as a skipped or null entry in the line. A trailing comma after the last data entry in a line is not required. Finally, a comment field can optionally appear in any line to the right of required data, provided that its first series of non-blank characters contains at least one alpha or other non-numeric character. These provisions may not be supported by other interpreters.

Defaults and Options

A) A variety of different pressure units can be used. Two different pressure units, primary and alternate, can also be used in any single profile to facilitate entry of both hyperbaric and hypobaric pressures in the same profile.

Primary pressure units, which are fswg by default, can be changed on a profile-by-profile basis by including the string "$P##" anywhere in the first free-form comment line (record) of a profile, where ## is the pressure unit index for the desired primary pressure units. The unit index portion of the field can be either one or two digits ($P# or $P##). Supported pressure indices are given in Table A.2.

TABLE A.2. Pressure Unit Indices in the Augmented NMRI Standard Format

Index Number	Pressure Unit
1	FSWG, feet seawater, gauge
2	MSWG, meters seawater, gauge
3	FFWG, feet freshwater, gauge
4	MFWA, meters freshwater absolute
5	PSIA, pounds per square inch absolute
6	ThFt, Altitude*: Feet/10^3
7	FL, Altitude*: FL or Flight Level, Feet/10^2
8	ThM, Altitude*: Meters/10^3
9	ATA, atmospheres absolute
10	kPa, kilopascal absolute
11	MPa, megapascal absolute
12	bar, bar absolute
13	kgsc, kg/cm^2 absolute

**Altitude units; ThFt, FL and ThM; are always from the U.S Standard Atmosphere, 1976, expressed with respect to sea-level (1.0 ATA), regardless of the prevailing reference Surface Pressure [c.f., (C) and Pressure Conversions section below].*

Alternate pressure units are ThFt altitude by default and can be changed on a profile-by-profile basis by including the string "$*##" anywhere in the first free-form comment line of a profile, where ## is the pressure unit index for the desired alternate pressure units. The unit index portion of the field can be either one or two digits ($*# or $*##).

Each pressure entry in the ensuing profile description that is intended to be interpreted in alternate pressure units must be immediately followed (i.e.; no comma, tab, space or other delimiter) by the "*" character. Entries lacking this flag are interpreted in primary pressure units.

B) Time entries can be made in either elapsed time format (default) or in delta time format. In the latter, each time entry for a profile node gives the delta time in minutes since the preceding node. Delta time format is invoked on a profile-by-profile basis by including the string "$TD" anywhere in the first free-form comment line of a profile. Profiles with first lines lacking this string are interpreted in elapsed time format.

C) The reference Surface Pressure can be changed on a profile-by-profile basis to a value less than sea-level pressure to effect appropriate interpretation of gauge pressure units for diving at altitude. Saturation on "Originating gas" at a reference Surface Pressure other than the default sea-level pressure is specified by a first node that contains a -1.0 Time entry with a Pressure entry equal to the desired Surface Pressure. The Pressure entry can be in either primary pressure units or alternate pressure units with the "*" flag, in conformance with the conventions described in (A) above. A gauge pressure in this first Time=-1.0 node is interpreted with respect to the default sea-level Surface Pressure, while all subsequent gauge pressures are interpreted with respect to the new Surface Pressure. A "New gas" entry in a Time=-1.0 node initiates a gas switch at Time=0.0 from saturation on "Originating gas" at the new Surface Pressure. Note that an

initial Time=-1.0 node is similar to an initial Time=0.0 node in the NMRI Standard Format, but also causes the Surface Pressure to be reset with corresponding effects on interpretation of 'gauge' pressure units.

Two starting nodes must be used to code a profile that occurs with reference to a surface pressure other than default and that begins with the subject(s) at initial saturation on "Originating gas" at a pressure other than the new surface pressure. A first node with Time=-1.0 must be used to code the desired surface pressure. A second node with Time=0.0 then resets the initial saturation conditions at the indicated pressure on "Originating gas". Note that the surface pressure reset node, with Time=-1.0, causes the initial saturation conditions to be set equal to saturation on "Originating gas" at the new surface pressure. The subsequent Time=0.0 node then resets the initial saturation pressure (again), without effecting the surface pressure setting. If the nodes are reversed, with the Time=0.0 node entered before the Time=-1.0 node, the reset of the initial saturation conditions that accompanies the latter will override the effect of the first Time=0.0 node.

D) A profile ended with a "-9999.0" TIME entry is terminated at the TIME of the preceding node. A profile ended with a "-0000.0" TIME entry signals the presence of an end-stage of indefinite duration at the last entered PRESSURE and breathing GAS CODE. The latter allows a suitably coded DCS model to run until all DCS risk decays to zero, without requiring explicit specification of an arbitrarily long post-decompression stage.

E) Supports entry of an Exercise Code in any node to indicate performance of exercise in the period beginning at the node time and ending at the elapsed time of the next node. (This convention is similar to that for the "New gas" entry which, when present, indicates that breathing of the New gas starts at the node time.) Nodes that are included only to indicate start or finish of an exercise period must include the time, in appropriate format, followed by four commas before the Exercise Code to signal absence of pressure and gas switch information. The Exercise Code is a floating value as follows, where the square brackets indicate optional information:

0[.opq] (or absent)	REST or END EXERCISE [.opq] ignored]
mn[.opq]	Begin exercise of type mn [at intensity opq].

If [.opq] is absent or opq is 0 with nonzero mn, an appropriate unit response; e.g., 1 vice 0 in a simple binary EXERCISE or REST code; is assumed.

Definitions of various exercise type and intensity codes are user or site specific and, for proper interpretation, must be included as separate accompaniments to the data set(s) in which they are used. For example, the opq intensity field can be used to indicate an (opq x 10)% increase in whole-body O2 consumption above resting baseline:

mn.005	Begin exercise of type mn increasing O2 consumption by 50%
mn.010	Begin exercise of type mn increasing O2 consumption by 100%
mn.120	Begin exercise of type mn increasing O2 consumption by factor of 12

Pressure Conversions

Units of pressure are converted according to the following primary definitions [1]:

1 atm = 760.000 torr
1 bar = 100,000 Pa
1 psi = 6,894.76 Pa
1 torr = 133.322 Pa

Units of pressure expressed as water depth below sea-level are converted using the following additional standard definitions as adopted by the Undersea and Hyperbaric Medical Society:

1 bar = 32.6457 fsw (assumes seawater density = 1.02480 gm/cc)
1 msw = 10.0000 kPa (assumes seawater density = 1.01972 gm/cc)

1 bar = 33.4702 ffw (assumes freshwater density = 0.999552 gm/cc)
1 mfw = 9.80229 kPa (assumes freshwater density = 0.999552 gm/cc)

Units of pressure expressed in terms of geometric altitude above sea-level are converted using defining equations for the *U.S. Standard Atmosphere, 1976* [12]. These equations give pressure P in atmospheres absolute (atm abs) as functions of geometric altitude above seal-level A in kilometers (km):

$$P = \left[\frac{288.15}{288.15 - 6.5A}\right]^{-5.25588} \quad ; A < 11 \text{ km} \tag{A.1}$$

$$P = 0.22336 \cdot \exp\left[0.15769 \cdot (11 - A)\right] \quad ; 20\text{km} > A \geq 11 \text{ km}. \tag{A.2}$$

These equations are inverted to obtain the following expressions for geometric altitude A in kilometers (km) as functions of pressure P in atmospheres absolute (atm abs):

$$A = \left\{\frac{288.15 - \exp\left[\ln(288.15) + \frac{\ln(P)}{5.25588}\right]}{6.5}\right\} \quad ; P > 0.22336 \text{ atm abs} \tag{A.3}$$

$$A = 11 - \left\{\frac{\ln\left(\frac{P}{0.22336}\right)}{0.15769}\right\} \quad ; 0.05403 \text{ atm abs} < P \leq 0.22336 \text{ atm abs} \tag{A.4}$$

The above expressions cover the relationship between geometric altitude and atmospheric pressure over the entire physiological range; from below sea-level to above the Armstrong line at 62,800 ft (19.14 km), where atmospheric pressure equals the vapor pressure of water at 37°C (47 mmHg). In this physiological region, the *U.S. Standard Atmosphere, 1976*, of the United States Committee on Extension to the Standard Atmosphere (COESA) is the same as COESA's *"U.S. Standard Atmosphere, 1962,"* and is identical with the International Civil Aviation Organization (ICAO) "Manual of the ICAO Standard Atmosphere," as revised in 1964. The definition of the Standard in this region was also adopted in the *ISO Standard Atmosphere* (ISO 1973) by the International Standards Organization (ISO) in 1973. (c.f., [12])

Improving on a "Good" Model

Erich C. Parker[1], Shalini S. Survanshi[1], Paul K. Weathersby[2]
[1]Naval Medical Research Institute
Bethesda, MD
[2]Gales Ferry, CT

This paper presents an example of a process by which an existing model, considered to be a good model of decompression sickness (DCS) risk, can be improved. Along the way, some questions as to what constitutes a good model, what conditions require an improvement and what constitutes an improvement will be raised.

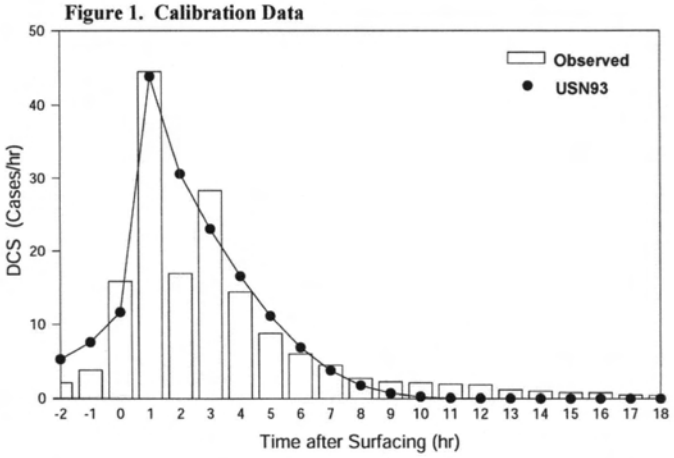
In general, the process works as follows: The decompression modeling program at the Naval Medical Research Institute (NMRI) developed an empirical model (as opposed to a mechanistic model) and calibrated it to high-quality observed data. This model was successful in describing outcomes in a large, diverse set of diving data [1] and provided several useful products [1,2,3,4]. Later we identified an area where the model did not perform well enough to provide a useful answer. At this point we had two options to improve the model's performance: 1) Add data to specifically address the area of poor performance, and 2) Modify the model itself. After each step, we have to ask two important questions: 1) Have we fixed the problem that we identified? 2) How do we know we fixed it?

The model used as the example is known as USN93 (the acronym identifies the sponsor and the vintage of the work) [1]. This is an empirical probabilistic model of decompression sickness risk, optimized by fitting to data using the maximum likelihood method. The data used to calibrate the model consist of the time, depth and the gases breathed by the diver and the outcome of the dive (DCS or not-DCS). The outcomes of those exposures are coded in a binomial sense as zero for safe outcome and one for DCS outcome. A third category, referred to as a marginal outcome, is considered somewhat worse than completely safe, but not as bad as a DCS outcome. Such marginal cases are coded as one-tenth of a full DCS case. The DCS and marginal outcomes are further modified by the time of the event occurrence. A form of time of event occurrence is used which represents the interval during which the case develops. This interval begins with the last time that the subject was known to be safe and ends at the time when the subject was considered to have acquired DCS [5].

The USN93 model accumulates DCS risk as a time integral of instantaneous risk and this risk, or hazard, function is proportional to over-pressure. This over-pressure consists of the model's calculated nitrogen tissue pressure when it is in excess of the ambient pressure. Over-pressures, or instantaneous risks, are summed over multiple independent parallel compartments. There are typically one to three compartments, depending on the size and complexity of the calibration data set. USN93 was calibrated with 3,322 dive exposures which included 190 DCS and 110 marginal events.

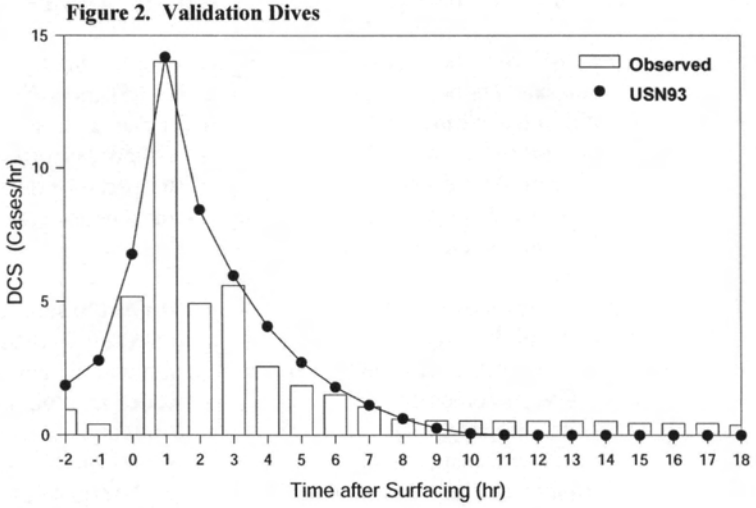
How does the model perform after calibration? One way of answering that is to look at how the calibrated model predicts outcomes in the calibration data itself. Figure 1 shows time after surfacing on the x-axis in increments of hours, and on the y-axis, the number of DCS cases. The vertical bars represent the observed DCS rate (number of DCS cases per hour observed). The solid line represents the USN93 model's prediction of DCS cases for each time category. The model performs well at describing the features of this distribution: A success in that measure of its predictive ability.

Figure 1. Calibration Data



Another way the goodness-of-fit question can be asked is to look at how the model predicts DCS risk in the data that was not included in the calibration data set, sometimes referred to as validation data. In this case, there were roughly 1,500 dive exposures not included in the calibration data set for various reasons. One question to ask, of course, is why was this data not included in the calibration data set? There are a number of reasons. For example, one of the original criteria for inclusion in the calibration data set was that the divers be immersed during the study. Some of these dives included divers who were dry, in a compression chamber, during exposure. Other parts of these data simply arrived too late to be included in the calibration, so were used here as validation data. Figure 2 shows that USN93 provides essentially the same quality of fit to the data not included in the calibration as it did for the calibration data of Figure 1.

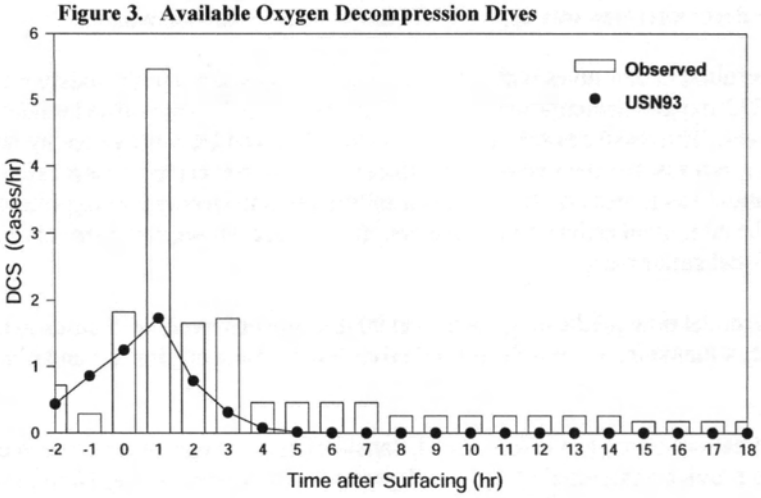
Figure 2. Validation Dives



One of the criteria for a model to be considered 'good' is that it provide a useful product. One such product that USN93 has led to has come to be known as the U.S. Navy Dive Planner [2]. This is an implementation of the model in which the user can input depth, bottom-time and gas mix to plan a dive of arbitrary complexity. Upon reaching the end of bottom time, the dive planner will provide a decompression profile to follow. This dive planner has been approved by the U.S. Navy for special operations dives and is extensively in use today.

USN93 performs well in these goodness-of-fit measures and has proven to be of practical value, so it can be considered to be a 'good' model. Why does it need improvement? There are areas in which its performance is less than optimal. One such area occurs in dives which use 100% oxygen during decompression [6]. This is an area of diving operations which, during model development, was not considered to be of primary importance, but has since become a critical question. USN93 does a poor job of predicting the DCS outcomes in O_2-decompression dives. From Figure 3, it is clear that USN93 dramatically under-predicts the number of observed cases.

Figure 3. Available Oxygen Decompression Dives



What is happening inside the model to cause this under-prediction? The risk of DCS in USN93 is proportional to the N_2 over-pressure, and wash-out during decompression is controlled by both the kinetic time constant and the relative levels of ambient and tissue nitrogen. When 100% O_2 is breathed, N_2 wash-out is enhanced due to the absence of a nitrogen component of ambient pressure. Since the risk that the model accumulates is proportional to this over-pressure, a diver breathing 100% O_2 has substantially less accumulated risk than a diver breathing air. Observations indicate that oxygen during decompression should be beneficial, but USN93 over-estimates that benefit. USN93 under-predicts the observed DCS incidence in O_2 decompression dives by 60 percent.

There were 729 O_2-decompression dives withheld from the original calibration data set. One approach to improve the model is to simply add these dives to the calibration data and recalibrate. When this is done, the model under-predicts the DCS incidence in O_2 dives by only 30 percent: Certainly an improvement but not enough to declare a success.

We then postulated two modifications to the model. The first was to introduce the idea of oxygen acting as a circulatory drug with nitrogen wash-in/wash-out kinetics slowed based on the PO_2 present in the diver's breathing gas. This idea was derived from a publication by Anderson *et al* [7] from the hyperbaric group at SUNY-Buffalo, in which substantial slowing of nitrogen wash-out with a dependence on PO_2 was observed.

The second modification involved oxygen acting in part as an inert gas; that is, some part of the PO_2, the pressure of oxygen breathed by the diver, contributes to the risk producing over-pressure as though it were another inert gas.

These two modifications appeared promising with the available data set (original 3322 dives plus the 729 O_2 dives) but did not provide statistically significant improvement to the log-likelihood fit; that is, not enough of an improvement in the log-likelihood to justify the added parameters necessary to implement these modifications. These models still provide a 15-to-20-percent under-prediction of the observed DCS incidence.

A prospective oxygen decompression dive trial [4] provided the next necessary component. In this trial, the Dive Planner provided an oxygen decompression schedule following an air dive. These were dry dives (divers not immersed) with divers breathing air during the bottom time, followed by 100% oxygen during the decompression, at either 60 or 40 feet. This dive trial was conducted in two phases.

The initial phase can be summarized as having asked the following question: Can the dive planner provide an oxygen decompression schedule that works? The short answer is: No, it did not. The dive planner failed to provide an adequate oxygen decompression. The second phase of the dive trial asked a different question: What percentage of the dive planner's recommended *air* decompression time, spent breathing 100% oxygen, provides an adequate decompression? At the end of a dive's bottom time, the dive planner recommends a certain total air decompression time. Perhaps some proportion of that total air time could be taken at 40 feet breathing oxygen to provide an adequate decompression. Twenty percent was found to be a reliable proportion. For example, a diver conducting an air dive, which resulted in a dive planner-recommended total decompression time on air of 100 minutes, could take 20 minutes at 40 feet breathing 100% oxygen to

provide an adequate emergency decompression. This procedure of taking 20 percent of the dive planner's air decompression as an emergency oxygen decompression was approved by the U.S. Navy for use with the dive planner.

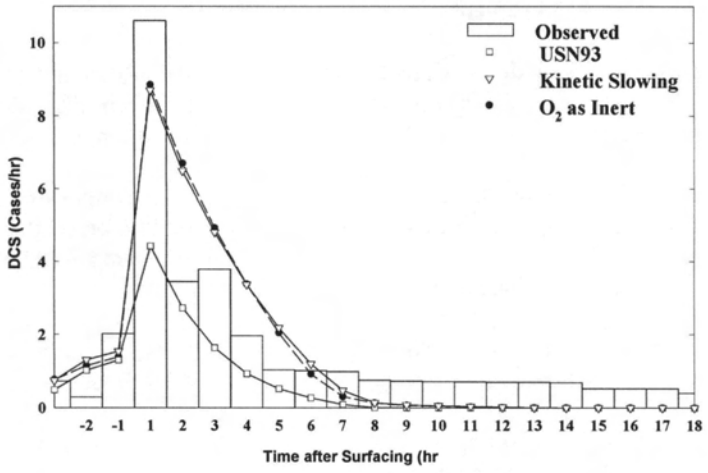
This dive trial resulted in 284 dives with 17 DCS cases. Adding these to the existing calibration data set gives over 4300 dives, including 1,013 oxygen decompression dives. Recalibrating the unmodified model with this expanded data set yields a small improvement. This model now under-predicts the observed DCS incidence by only 25 percent, still not enough to be considered a success, but the two oxygen effect modifications can now be recalibrated as well. The nitrogen kinetic slowing modification was a success. It provided a significant improvement in log-likelihood, enough of an improvement to justify the additional estimated parameters. In this case, an improvement of about 12 likelihood units was achieved with two additional parameters.

In addition, this model now predicted greater than 90% of the observed DCS cases in the oxygen decompression dives. Equally important, it maintains the excellent predictive ability in the original air and nitrox dives that successful USN93 exhibited.

The other modification, or oxygen effect model, considering some part of the oxygen as an inert gas, was also a success. This model also provided essentially the same degree of improvement in log-likelihood and predicts 90+ percent of the incidence in the oxygen dives, while maintaining the original quality of predictive ability.

Figure 4 shows the DCS incidence in the 1,013 oxygen dives as predicted by these models. Each bar represents the observed DCS incidence in one hour. The original USN93 model badly under-predicts DCS occurrence. The two curves that almost overlay each other are the predictions of the two modified models. They both provide adequate representations of the observed time distribution of DCS incidence in these dives.

Figure 4. Updated Oxygen Decompression Data



After adding the new data, re-calibrating the models and finding the two modifications to be successful, the remaining question is: What constitutes goodness-of-fit? The log-likelihood is our primary measure of fit to the data. One advantage log-likelihood has is that it is an objective measure, and direct comparisons can be made within model groups. Another advantage of likelihood as a measure of fit is that it allows the formality of the likelihood ratio test for testing the significance of adding or removing parameters from a model. One disadvantage of likelihood is that in models of this complexity, it can be highly prone to finding local minima, or local 'best-fits'. In 10-to-12-dimensional parameter spaces, the shape of the likelihood surface can provide many opportunities to stop searching before finding a global best-fit. A large number of different sets of starting parameters are needed in order to thoroughly explore a model of this complexity. For the models presented here, on the order of thousands of different starting parameters sets, rather than tens or hundreds, were needed.

The other general category of goodness of fit measure that we apply is the model's ability to predict outcomes, as shown in the figures. One quantification of this predictive ability is the chi-squared test. A note of caution in the use of chi-squared tests is in order, due to their extreme sensitivity to the necessary categorization of the data.

The examples shown here are categorization of the DCS outcomes by time after surfacing, but the data can also be categorized in other ways. For example, dive data can be categorized by groups of increasing depth or groups of increasing risk level or by type of dive. For example: single dives, repetitive dives, saturation dives, etc. Chi-squared tests can lead to contradictory results depending on how those categorizations are chosen. For example, grouping the data in this study by type of dive, opposite conclusions regarding the significance of fit were achieved by two equally valid categorizations. If the data were grouped as follows: air dives, non-air dives, saturation dives and oxygen dives, the chi-squared test indicates that the model is a good predictor of the outcomes. Grouping these same dives as single dives, repetitive dives, saturation dives, and oxygen dives yields a chi-squared indication of poor predictive ability. The same data and same model with equally valid groupings yields a contradictory conclusion.

This categorization issue has surfaced in other studies as well. For example, in a different study using only no-decompression dives [8] and categorizing dives in terms of ascent rate (seemingly an important factor in no-decompression dives) a strong sensitivity of chi-squared tests was found with regard to where the boundaries of those categories of ascent rate occur. Slight changes in boundaries, which push an important sub-set of data from one group to another, result in substantial chi-squared test differences. These tests also tend to be highly sensitive to extreme cases. In this instance, two high-speed ascents, which can unduly sway the result to indicate significance of predictive ability when in fact there is none in the absence of these outlying cases.

Application and interpretation of such chi-squared tests can be misleading. Practitioners are encouraged to present their results in as much detail as possible so that summary measures like chi-squared tests are not the sole gauges of success.

References

1. Statistically based decompression tables XII: Repetitive decompression tables for air and constant 0.7 ata PO_2 in N_2 using a probabilistic model, S.S. Survanshi, E.C. Parker, E.D. Thalmann, P.K. Weathersby, Naval Medical Research Institute Technical Report 97-36 (v.1), Bethesda, MD; 1997.

2. Statistically based decompression tables X: Real-time decompression algorithm using a probabilistic model, S.S. Survanshi, P.K. Weathersby, E.D. Thalmann, Naval Medical Research Institute Technical Report 96-06, Bethesda, MD; 1996.

3. Calculating decompression in Naval Special Warfare SEAL Delivery Vehicle diving operations utilizing the Real-Time Dive Planner, D.J. Valaik, E.C. Parker, S.S. Survanshi, Naval Medical Research Institute Technical Report 96-54, Bethesda, MD; 1996.

4. Dry decompression procedure using oxygen for Naval Special Warfare, S.S. Survanshi, E.D. Thalmann, E.C. Parker, D.D. Gummin, A.P. Isakov, L.D. Homer, Naval Medical Research Institute Technical Report 97-03, Bethesda, MD; 1997.

5. Predicting the time of occurrence of decompression sickness, P.K. Weathersby,. S.S. Survanshi, L.D. Homer, E.C. Parker and E.D. Thalmann, J. Appl. Physiol. 72(4): 1541-1548, 1992.

6. Probabilistic models of the role of oxygen in human decompression sickness, E.C. Parker, S.S. Survanshi, P.B. Massell and P.K. Weathersby; J. Appl. Physiol. 84(3): 1096-1102, 1998.

7. Oxygen pressures between 0.12 and 2.5 atm abs, circulatory function and N_2 elimination, D. Anderson, G. Nagasawa, W. Norfleet, A. Olszowka, and C. Lundgren, Undersea and Biomed. Res. 18: 279-292, 1991.

8. Does the dive profile affect the manifestations of decompression sickness? R. Ball, D. Temple, S.S. Survanshi, E.C. Parker, P.K. Weathersby, Undersea Biomed. Res. 24 (Supplement), 1997.

Meta Analysis of Diver Decompression Data

Paul K. Weathersby, Diana J. Temple, Erich C. Parker
Naval Medical Research Institute
Bethesda, MD

As in all scientific investigations, the success of modeling the occurrence of decompression sickness (DCS) relies upon the quality of the data used. First consider the ideal, or *natural,* data set.

A *natural* data set
- Has the same dive type (rig, immersion, temperature...)
- Was obtained at a single research center
- By a single investigator, and
- With a single subject pool (and maintaining a consistent attitude about what subjects should expect), and
- Conducted and scored by a single detailed protocol

In the history of decompression sickness research, there are numerous examples of studies missing one or more of these qualities. Moreover, the better studies tend to be "small" by modeling standards. There are **zero** *natural* data sets with N > 1000, and only 6 *natural* data sets with N > 300 in the NMRI data collection. (7)

For a decompression model to allow predictions over an operationally useful range, we cannot use *natural* data alone. We must combine studies. *The process of combining data sets is referred to as* ***meta analysis***. In non-hyperbaric medicine, meta analysis has achieved "fad" status, and engendered controversies for many of the same reasons we relate below. In the wrong hands, meta analysis becomes a targeted search for studies that, taken together, support a pre-conceived belief that a "P<.05 effect" exists, but could not be demonstrated in the *natural* data cited. One infamous example is the EPA assessment of cancer risk from second-hand smoke (4).

Several approaches can be followed to assemble the *unnatural* data set. At different times, our group has followed four different paths described below:

a) Ignore all differences among the *natural* data sets
b) Select "similar" data to combine
c) Test for data combinability (under a specified model)
d) Revise prior data for similarity

a) Ignore Differences

This approach offers a definite advantage: it is the easiest. Naturally, ease is associated with a major disadvantage: uncritical combinations of data can seriously bias results.

For simplicity consider the following contrived example:

Study A used deep dives with low DCS incidence
Study B used shallow dives with high DCS incidence

The result of an uncritical meta analysis (say with simple logistic regression on the combined data from A and B) will be the conclusion:
-- The risk of DCS decreases with deeper depths !

To decrease the possibility of such a bias, this easy path must be avoided.

Each *natural* data set must be scrutinized to find important differences in the way the data was obtained. For example, we have been struck by the huge variation in reporting of post-dive manifestations (7). Likewise, the diagnostic criteria seem to have evolved since the 1940's. Two reported actual cases are:

Study A
" No complaints upon surfacing. That evening, diver noticed slight swelling of left wrist lasting 2 hours. No pain in this joint, but it felt stiff. Probably a sprain as this does not resemble bends." Diagnosis: Not DCS
Study B

"No complaints following the dive. About 24 hours post surfacing, diver noticed a dull 2/10 ache in his right hip, which resolved spontaneously after about 13 minutes." Diagnosis: DCS.

Even the most generous accommodation of physician diagnostic variability could not ignore this demonstration of differing interpretations on just what is DCS. Other recent reviews of cases – by experienced diving medical officers - have also noted unsatisfactory differences in diagnostic standards (2, 3, 5).

Less subjective than the diagnostic quandary are problems arising from differences among data sets in such factors as temperature, immersion, etc. Our analysis of DCS risk in dry/sedentary vs. immersed/exercising subjects (see below) demonstrated less than a 30% effect (14), but we still resist automatic combinations of data with those different attributes.

Even tabulated raw results must be questioned. A detailed reading of an earlier study, followed by examination of institutional archives, led us to change tabulated depth by several fsw (1). We believe the report used a convention of deepest depth (the feet of standing subjects) rather then the mid-chest convention adopted for our other data files.

The first successful model of air diving combined data from different studies without a comprehensive attempt to ensure similarities (12). Indeed, many serious data questions were simply noted in passing (as an initial proof of concept).

b) Select Data which is Similar

Sometimes, studies appear to have so many similarities, that it might be almost *natural* to simply combine them. The advantage is to increase the amount of data, but the disadvantage is that unrecognized biases can be introduced.

By the mid 1980's, the value of having a collection of data sets with an achievable degree of similarity was recognized. In a laborious iterative process, some 23 studies were accumulated which possessed the following specifications (17):

- Done in military laboratories in 1972 or later
- Depth/time records available to 1 fsw/30 sec
- Regular pre- and post-dive checks by military Diving Medical Officers
- Original investigator available to assist in resolving discrepancies

Each of the above specifications evolved from the sequence of quality control issues uncovered during data assembly and review. The resulting data sets were later referred to as Primary Data for modeling and made freely available to investigators (17). Included in the documentation were details on immersion, temperature, acclimatization and other possibly relevant information.

The labor involved in generating and scrutinizing Primary Data has led to repeated temptations to use "easier" data. Some discussion of the problems found in commercial or military occupational diving records, and why those sources will seldom produce Primary Data, have been reported (15).

c) Statistical Test for Data Combinability

Sometimes a formal statistical test can answer the question "Are these data sets compatible?" The advantages are transparency and objectivity (although definitions of "compatibility" can be problematic). The disadvantage is that the answer may be irrelevant to the study at hand.

c1. Overall model compatibility

One test of "compatibility" asks the question:
Does this mathematical model describe a combined set of data "about as well" as the model describes the component data sets, examined individually?
(The "about as well" is defined in the Likelihood Ratio Test – see Appendix B).

Any answer in this test depends upon the model making sense in the first place. If you don't think much of the model, the test is not worth performing. And the test can sometimes be uselessly "weak", if there are insufficient cases of DCS in the component data sets.

On the other hand, the test may be uselessly strong. A statistical answer that a discernable difference in data can be demonstrated, might be less important than production of a model and parameters that has some "acceptable" ability to describe the full range in the data.

c2 Parametric tests

It may be possible to more precisely test a suspected data difference. In (14), we examined some pressure exposures that were dry/sedentary, while others used immersed exercising subjects. A risk model was constructed with parameters defined for dry conditions, but with a single immersion-difference parameter assigned to displace risk for the wet divers. When that difference was shown to be essentially zero, we concluded that immersion was not a strong risk factor - and that dry dives could therefore be combined with wet ones for modeling. Other parametric tests examined oxygen effects (13).

Sometimes a single observation can dominate a data set and models calibrated from it. This occurred (16) when a short duration dive was followed by DCS late after surfacing. For the exponential gas exchange to maintain a calculated overpressure that late, the allowable range in time constants was only a few minutes. Failure of this dive to combine easily with other data would not be surprising.

d) Revise Prior Data for Similarity

In the evolution of Primary Data standards (17), many older studies were uncovered which had thorough details on the diving profiles, and various notes as to medical outcome. In examining the DCS events and their descriptions of the signs and symptoms, it became apparent that historical and contemporary definitions of DCS were different. (The "back when men were men" phenomenon.) The advantage of using the older studies was a great increase in available data. The expected disadvantage was that the older reports might have ignored – or - "underdiagnosed" some cases that would be diagnosed and treated as DCS today.

One of us (PW) moderated a panel of military Diving Medical Officers (E. Flynn of NEDU-NMRI, C. Harvey of NSMRL, K. Sawatsky of DCIEM, H. Schwartz of NEDU, and E. Thalmann of NEDU-INM-NMRI) who had participated in modern diving trials and were still active in the diagnosis of DCS in diving trials. During a one-day meeting in September 1988, examples of DCS manifestations were reviewed from published military trials extending over 40 years (9, 10, 11, 13, 18). A consensus document on how to approach retrospective re-diagnosis was drafted, then reviewed by all participants. The revised standards are presented in Appendix A.

Although a document was produced that could serve as a diagnostic tool, the participants were not particularly comfortable with it. As physicians, they were naturally hesitant about second guessing one of their own - especially through a written document that lay-people might be tempted to apply. Many of the fragmentary case reports reviewed were insufficient to convince panel members of the presence of DCS and whether the "case" would receive recompression treatment today. Problems were most prevalent in cases labeled as "marginal" in modern studies, and as "minor bends", "moderate pain", etc in older studies.

The system of re-diagnosis in the Appendix was not actually applied for several years. It was recently resurrected for two reasons:

- desired expansion of data with manifestations to allow epidemiology, and
- re-emphasis on high risk profiles no longer ethically feasible for study.

The re-diagnostic criteria have been applied to 9 older studies dating from 1945 to 1957 from NEDU (3 of which had been partially used in (12)). They were also applied to 2 recent NMRI studies (6, 8) after issues arose in the consistency of Undersea Medical Officer staffing. Compilation of the older studies, several modern trials not included in (17), and the existing Primary Data - complete with manifestations - has been released in a 2-volume Technical Report (7). The compilation encompasses 7400+ exposures with 800+ manifestation reports from 36+ original studies.

Conclusions

Data quality will continue to be an issue in a field where the scientific use of data is so recent. The NMRI Technical Reports that have recovered, reviewed, and codified the results of older studies are intended to provide a stable platform for anyone intending to model decompression sickness.

There is no single "right" answer to whether data should be combined. Different objectives will lead to different decisions. The responsibility, as always, rests with the analyst to justify, or at a minimum clearly document, the source of any data.

References

1. des Granges M. Standard air decompression table. Washington DC: NEDU Report 5-57, December 1956.

2. Flynn ET, EC Parker, R Ball. Risk of decompression sickness in shallow no-stop diving: An analysis of Naval Safety Center data 1990-1994. Bethesda MD: NMRI Technical Report 98-08, May 1988

3. Francis TJR, DJ Smith, JJW Sykes. The prevention and management of diving accidents. Alverstoke UK: Institute of Naval Medicine Report R93002, Feb 1993.

4. Jinot J, D Barry. Environmental tobacco smoke: science vs. rhetoric. Risk Analysis 15: 91-96, 1995, and follow-on letter, 16:303-304, 1996.

5. Kelleher PC, TJR Francis. INM diving accident database: analysis of 225 cases of decompression illness. Alverstoke UK: Institute of Naval Medicine Report R93048, 1993.

6. Survanshi SS, ED Thalmann, EC Parker, DD Gummin, AP Isakov, LD Homer. Dry decompression procedure using oxygen for Naval Special Warfare. Bethesda MD: NMRI Technical Report 97-03, Apr 1997

7. Temple DJ, R Ball, PK Weathersby, EC Parker, SS Survanshi. The dive profiles and manifestations of decompression sickness cases after air and nitrogen-oxygen dives. Vol. I: Data set summaries, manifestation descriptions and key files, Vol II: Complete profiles and graphic representations for DCS events. Bethesda MD: NMRC Technical Report 99-03, 1999.

8. Thalmann ED and RW Hamilton. Multiday air saturation at 20-24 fsw: Data report. Naval Medical Research Center Technical Report, *in preparation.*

9. Van Der Aue OE. Surface decompression, derivation and testing of decompression tables with safety limits for certain depths and exposures. Washington DC: NEDU Report 5-45, 1 Jan. 1945

10. Van der Aue OE, Kellar RJ, Brinton ES. The effect of exercise during decompression from increased barometric pressures on the incidence of decompression sickness in man. Washington DC: NEDU Report 8-49, Mar 1949

11. Van der Aue OE, Kellar RJ, Brinton ES, Barron G., Gilliam HD, Jones RJ. Calculation and testing of decompression tables for air dives employing the procedure of surface decompression and the use of oxygen. Washington DC: NEDU Report 13-51, November 1951.

12. Weathersby PK, SS Survanshi, LD Homer, BL Hart, RY Nishi, ET Flynn, and ME Bradley. Statistically based decompression tables. I. Analysis of standard air dives: 1950-1970. Technical Report of the Naval Medical Research Institute, Bethesda, MD: NMRI 85-16, 62 pp. March 1985.

13. Weathersby PK, BL Hart, ET Flynn, and WF Walker. Human decompression trial in nitrogen-oxygen diving. Technical Report of the Naval Medical Research Institute, Bethesda, MD: NMRI 86-97, 44 pp. Aug 1986.

14. Weathersby PK, SS Survanshi, and RY Nishi. Relative decompression risk of dry and wet chamber air dives. Undersea Biomedical Research, 17:333-352, 1990.

15. Weathersby PK and SS Survanshi. Data quality for decompression modeling. In: Operational dive and decompression data: collection and analysis, Proceedings of the European Undersea Biomedical Society Workshop, W Sterk and RL Hamilton eds. EUBS, Amsterdam, pp. 94-99, 1991.

16. Weathersby PK, SS Survanshi, LD Homer, EC Parker, and ED Thalmann. Predicting the time of occurrence of decompression sickness Journal of Applied Physiology, 72:1541-1549, 1992.

17. Weathersby PK, SS Survanshi, R.Y Nishi, and ED Thalmann. Statistically based decompression tables. VII. Selection and treatment of primary air and N2O2 data. Joint Technical Report of the Naval Submarine Medical Research Laboratory, Groton, CT and the Naval Medical Research Institute, Bethesda, MD: NMRI 92-85 and NSMRL No. 1182, 112 pp. Sept 1992.

18. Workman RD. Surface decompression from air dives. Washington DC: NEDU Report 10-57, 1957.

Appendix A

Diagnosis Criteria: from 22 NOV 1988

First step: Separate outcome into 3 categories:

Cat A. Definite DCS (Symptom within 24 hour unless saturation dive)
Cat B. Unknown Outcome
Cat C. Not DCS

Second Step: Separate **Cat A** further:

Cat A-1 Definite DCS. Requiring recompression therapy by 1988 standards
Cat A-2 Definite DCS. NOT requiring recompression by 1988 standards. Difference between A-1 and A-2 is 1988 perception of whether lack of treatment will lead to morbidity in the diver.

Specific Description:

Cat A-1. Definite DCS requiring recompression.

- Any suspicious symptoms leading to and relieved by recompression
- Joint pain persisting as tabulated below whether recompressed or not:

		One Joint	Multiple Joints
Pain	Mild	60 min +	30 min +
	Moderate	30 min +	15 min +
	Severe	15 min +	8 min +

- Dyspnea, unless clearly from barotrauma or anxiety hyperventilation syndrome
- Any spinal neurologic symptoms, supported by signs, regardless of duration
- Any brain symptoms, such as visual blurring, "mental sluggishness", regardless of duration
- Any inner ear symptom, such as unsteadiness, vertigo, hearing loss, unless clearly from barotrauma

Cat A-2. Definite DCS not requiring recompression.

- Joint pain not persisting as long as tabulated under A-1
- Fatigue, moderate or severe
- Skin itch in immersed air or N_2-O_2 divers (Itch in dry chamber dives and HeO_2 dives is probably due to local skin mechanisms that would confuse modeling of primary symptoms)
- Skin rash or mottling, if only symptoms (When combined with non-persistent (A-2) joint pain, becomes A-1).
- Default diagnosis: Symptoms reported as "Mild bends, not requiring recompression" which do not fit other categorization criteria

Cat B. Unknown outcome; data insufficient for 1988 diagnosis

- Headache, typical and common for this diver
- Vague abdominal pain, not related to trauma or barotrauma
- Vague chest pain, not related to trauma or barotrauma
- Vague symptoms of any kind NOT responding to recompression or oxygen therapy attempted soon after dive (say 18 hours for non-saturation dive).

Cat C. Not DCS

- No post-dive symptoms reported
- Joint pain or fatigue, mild and consistent with recent exercise
- Sharp pain, consistent with joint sprain or impact injury
- Vague symptoms similar to Cat A-2 NOT responding to recompression therapy attempted not soon after dive
- Skin itch in dry chamber dives and He-O_2 dives.

Appendix B

Likelihood Ratio Tests of Combining Data

In the Likelihood Ratio test, the procedure is to fit each data set alone, and then the data combined. The total of the maximum likelihoods from the component sets will be better than from the large combined set, but perhaps only by an amount likely to be due to chance when the number of estimated parameters are properly accounted for. The test statistic is the ratio of likelihoods (or difference in log-likelihoods), which is Chi-square distributed under the null hypothesis of indistinguishable differences.

Take a specific example. Suppose we are using a 2 compartment model, with a time constant and scale/gain coefficient as parameters for each (4 total parameters).

4 parameter model fit to data set A $-LL_A = 42$
4 parameter model fit to data set B $-LL_B = 38$
(so 8 parameters give a combined -LL = 80)

4 parameter model fit to combined data set A+B $-LL_{AB} = 84$

Likelihood Ratio = $2 * [LL_{AB} - LL_A - LL_B] = 8$
Chi-square with 4 df is 7.8 (at $p < .1$) and 9.5 (at $p < .05$)
So the data sets A and B appear somewhat - but not strikingly - different.

Cold Exposure Survival Model

Peter Tikuisis, Ph.D.
Defence and Civil Institute of Environmental Medicine
North York, Ontario, Canada M3M 3B9

Model Demonstration

This presentation will begin with a demonstration of the model that we have developed for predicting survival times during cold exposure (Tikuisis 1995, 1997). Hereafter, we will refer to this model as CESM (Cold Exposure Survival Model), and following the demonstration, we'll discus the challenges of making such predictions.

The user interface (see Fig. 1) of the model accepts inputs according to three separate categories. The first category pertains to the characteristics of the subject such as age, gender, weight, height, body fatness, and fatigue, plus the level of water immersion. The second category consists of environmental factors (the air exposure factors of temperature, humidity, and wind speed are not shown in Fig. 1 because of the example chosen below) and the last category concerns the clothing protection on the individual.

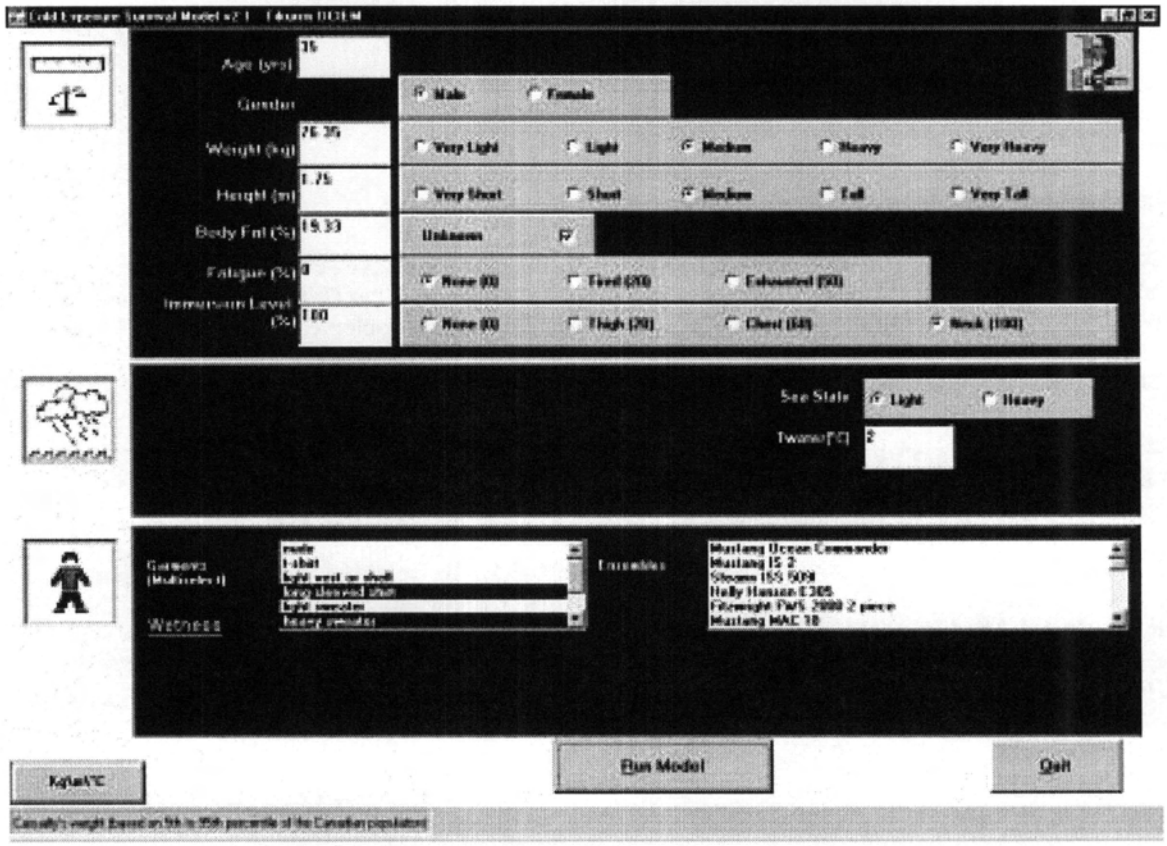
For example, let us choose a 35 yr old male, as shown in Fig. 1. The weight of the individual can be entered directly if known, otherwise it must be estimated. To accommodate the latter, CESM provides a menu from which the individual's weight can be selected. The 'very light' category refers to the 5^{th} percentile of the population, 'light' refers to the 25^{th} percentile, etc. up to 95^{th} percentile for the 'very heavy' category. Height is similarly selected. Body fat (BF) is an important determinate of survival time, however, its value is rarely known. In this case, CESM determines the %BF according to a regression formula based on age, gender, weight and height. In the present example, we have chosen the 50^{th} percentile for weight, and height leading to a BF of 19.3%.

CESM can be applied to situations involving cold air exposure and/or cold water immersion. For this demonstration, we will assume conditions that an individual might have faced in the water after the sinking of the Titanic. We'll assume that the individual is not fatigued and is immersed to the neck-level, thus only the water parameters apply. In this example, we select a light sea state and a water temperature of 2°C. Clothing can be selected in any combination of different garments by making appropriate selections from the clothing menu shown on the lower left of the input screen in Fig.1. Alternatively, actual clothing ensembles can be selected from the adjacent menu on the right. Among these are coveralls, survival suits, etc. Let us suppose that our unfortunate individual is wearing a long-sleeved shirt and a heavy sweater. Clothing is specified for the torso only since the other regions of the body are assumed to be clothed to the same level of protection.

To recap, we have selected a 50^{th} percentile male of 35 years of age, neck-immersed in light seas at 2°C, and wearing medium-weight clothing. The model predicts times to two different stages of body cooling on the basis of these inputs; a functional time of 1.4 h and a survival time of 2.8 h (see output screen of Fig. 1). The functional time is the predicted time for the individual's deep body temperature (T_{db}) to decrease to 34°C at which point the individual would suffer motor and cognitive impairments. The survival time is the predicted time for T_{db} to reach 28°C at which point unconsciousness is likely to occur. If we change the value of one of the input factors, say the weight, then CESM predicts shorter times for a lighter individual and vice-versa. In this case, %BF changes automatically to correspond to the changes in the weight, or any of the other individual characteristics.

This brief demonstration covered only the body cooling portion of the model. An additional calculation pertaining only to neck-level immersion in water provides the probability of finding the individual alive at the predicted functional time. That is, if CESM predicts that 1.4 h elapses before the individual's T_{db} reaches 34°C, then what is the chance of finding that person alive at that time? Causes of death other than hypothermia are considered here. This calculation will be explained later, but suffice it for now that in the present example, there is a 86% chance of finding the individual alive at 1.4 h if flotation is worn and 57% if not (see output screen of Fig. 1).

In summary, CESM predicts times to specific body temperatures corresponding to functional and survival times, and secondly, it provides a prediction of finding an immersed individual alive at the predicted functional time.



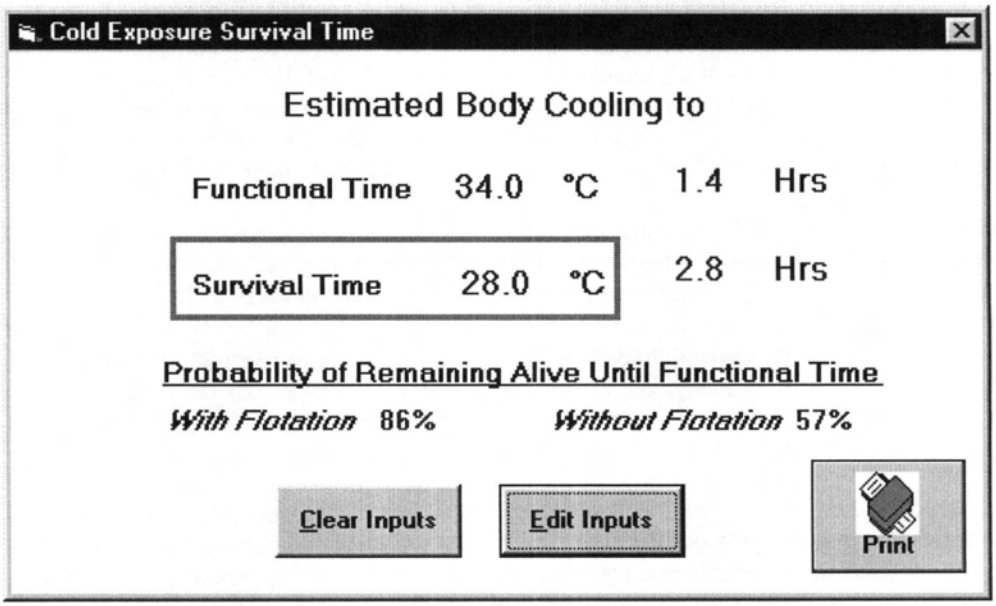


Figure 1. Displays of the input and output screens of CESM (see text for explanation).

Model Calibration/Validation

To calibrate and check CESM predictions, we require information on the characteristics of the individual, exposure conditions, and the level of clothing protection. We also need to know the individual's T_{db} at the end of the exposure, and the individual's state of cognition or consciousness associated with their deep body temperature.

This represents the desirable data. Data with this level of detail are usually only available from controlled studies and are limited to conditions that do not exceed a mild hypothermic state. Typically, the deep body temperature is not allowed to go decrease below 35°C in laboratory experiments, and thus much of the controlled T_{db} data reside between 35 and 37°C. There are many accidental cases involving severe hypothermia, but of these, only very few are documented to the level of detail required to calibrate or validate CESM. As a result, predicting the time course of body core cooling to 28°C is extremely extrapolative.

On the other hand, data are available for the statistics of survival during cold water immersion and the probability of finding someone alive as a function of time during such exposures. We will now consider in more detail the deterministic prediction of the rate of body cooling and the probabilistic prediction of survival outcome for water immersion.

Body Cooling Prediction

An important assumption in CESM is that the individual is considered sedentary. That is, the only source of body heat in addition to the resting metabolism is shivering. Any activity beyond this would contribute to internal heat production, which cannot be predicted unless the actual activity is known. The sedentary assumption is a reasonable one for accidental exposures to cold and it represents a worst case scenario. We also assume a normal physiological response to cold. Relevant information that can be obtained from laboratory experiments on individual responses to cold is coded into the model. The possibility of death due to causes other than hypothermia are not considered, at least in the model prediction of the rate of body cooling.

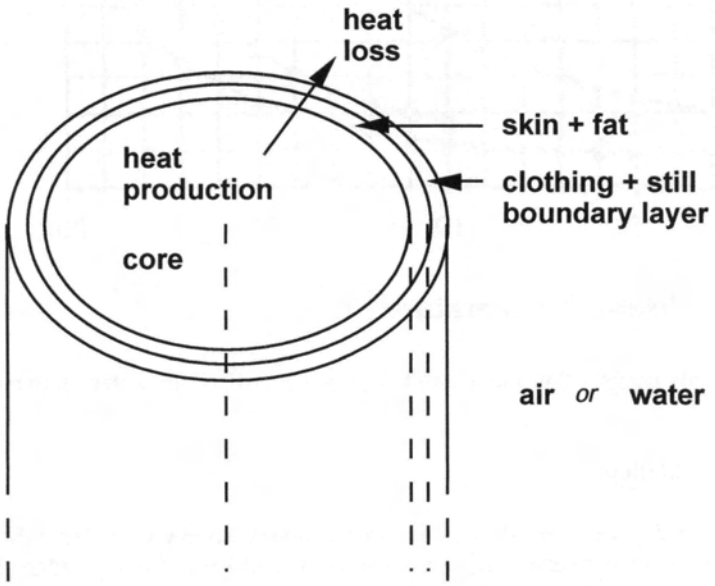


Figure 2. Schematic of the body cooling portion of the model.

The model is schematized in Fig. 2 and is essentially based on a core-shell configuration. If heat is produced within the core of the cylinder, that heat passes through two layers of insulation. The internal layer of insulation is represented by the skin and fat of the individual, and the external insulation layer is represented by clothing and the still boundary layer. As indicated earlier, CESM can be applied to problems involving cold exposure in both air and water environments.

An important feature of the model is its discrimination of different body types. During the model demonstration, the impact of different body size conditions was discussed. Typical model predictions for three body fat conditions are shown in Fig. 3 with survival time plotted against water temperature. Lean individuals have the fastest body cooling rates and hence

the shortest predicted survival times while fat individuals should last longer, with other factors being the same. Similar survival curves can be constructed with variations in other factors, such as sea state, level of clothing protection, etc.

We conclude our consideration of the body cooling portion of the model by reiterating that this prediction is deterministic. It predicts not "if", but "when" lethal hypothermia will occur. Although this portion of the model is based on physical principles of heat conduction and physiological responses to cold stress, its predictions of body cooling to temperatures below 34°C are extrapolative. Little is known about what happens to the body's response to cold when its deep body temperature drops below this level. This is particularly challenging for model development since we are unable to acquire controlled data for these situations. Instead, we must rely on case histories, yet the documentation is rarely as good as required, and there are simply too few detailed cases to support a statistical approach at this time.

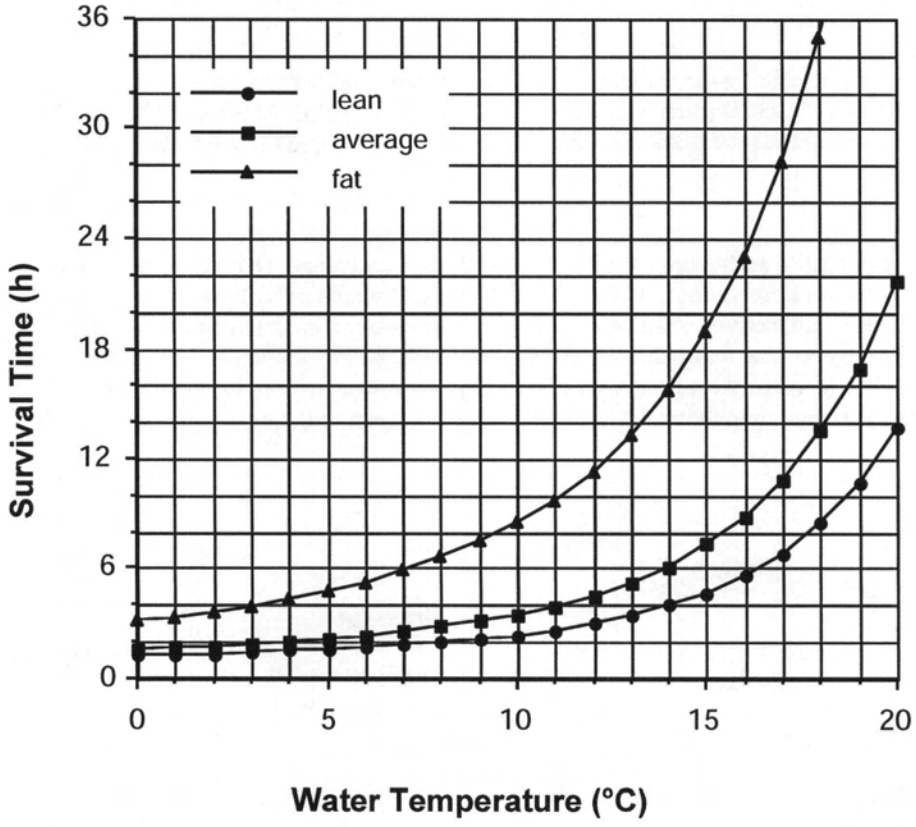


Figure 3. Model predictions of body cooling times to a core temperature to 28°C for water-immersed individuals of different body fatnesses.

Immersion Survival Outcome Prediction

The model prediction of finding someone alive during immersion is based on data from the U.K. National Immersion Incidence Survey. These data were analyzed and modelled by Oakley and Pethybridge (1997). The U.K. survey covered 930 incidents in which there were 66 deaths. The factors that were considered in the model included immersion time, water temperature, and whether or not the individual wore a buoyancy device. The model was based on the following logistic form:

$$Pr = \frac{\theta}{1+\theta} \qquad (1)$$

where Pr is the probability of finding the immersed individual alive and the quantity θ is defined in terms of three parameters; α, β and γ; and two independent variables; time of immersion (t_{imm}) and water temperature (T_w):

$$\theta = \exp[\alpha + \beta \cdot \ln(t_{imm}) + \gamma \cdot T_w] \qquad (2)$$

The data were segregated into three groups according to whether a buoyancy device was worn, not worn, or if unknown. The model was then fitted to each group using maximum likelihood. Resultant parameter values for each of the groups are given in Table 1 and the percent survival rate is shown in Fig. 4.

Table 1. Immersion survival outcome model parameter values (see Eq. 2).

Buoyancy Device	*Parameters*		
	α	β	γ
Yes	5.55	- 0.888	0.121
No	3.99	- 0.888	0.120
?	2.52	- 0.888	0.291

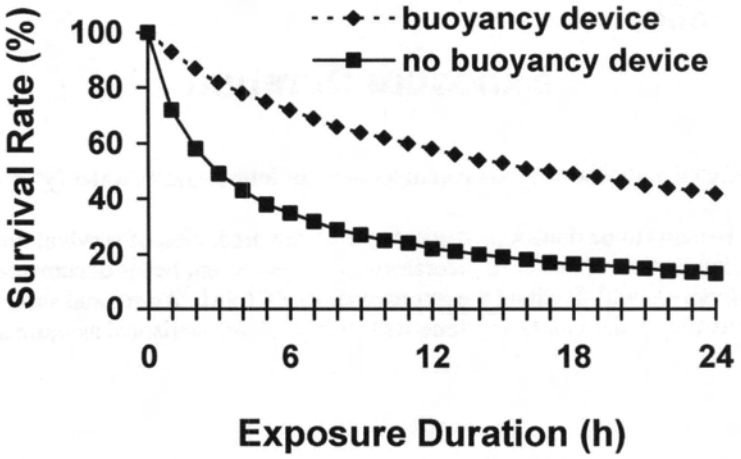


Figure 4. Model-predicted influence of personal flotation on survival outcome of individuals immersed in 5°C water.

The survival outcome does not take into account any of the factors used in the body cooling portion of the model; that is to say, it doesn't consider individual characteristics, sea state, or clothing protection. Figure 5 illustrates how well the U.K. model of survival outcomes compare to the observed data used to calibrate the model. The data were categorized according to immersion times beginning with consecutive 15-min periods, and ending with time periods that cover 1 to 2 h, 2 to 4 h, and > 4 h, respectively. Although the agreement between predicted and observed rates appears good, the chi square is significant only at the 0.5 level.

The prediction of the immersion survival outcome is a probabilistic approach in which the factors are immersion time, water temperature, and whether or not a buoyancy device is worn. Model parameters, however, are biased since the statistics neglect immersions of very short duration. Accidents in which individuals are rescued within a few minutes are usually not reported, nor are incidents in which bodies are not found.

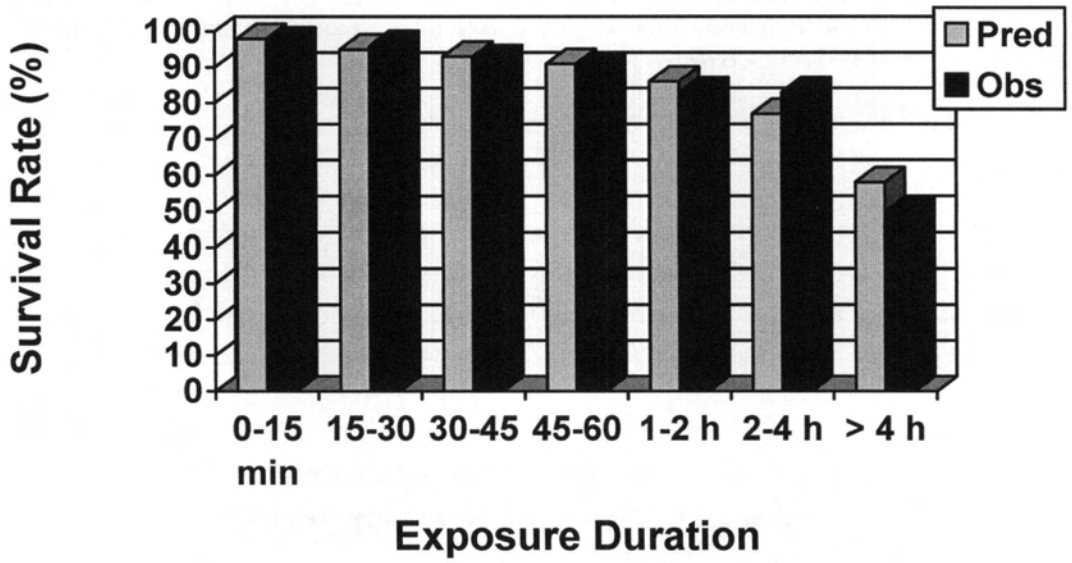


Figure 5. Comparison of predicted and observed survival outcomes for immersion in water ($\chi^2 = 4.46$; df = 5; p >0.5).

Clearly, much work remains to be done with respect to both the prediction of survival times due to hypothermia and the survival outcomes during immersion. Continued laboratory experiments and better documentation of the status of accidentally cold-exposed individuals will facilitate the improvement of CESM. Additional surveys of accidental immersions with particular attention to the immersion time will strengthen the statistical assessment of immersion survival outcome.

References

Oakley, E.H.N. and Pethybridge, R.J. (1997). The prediction of survival during cold immersion: results from the UK national immersion incident survey. Institute of Naval Medicine Report No. 97011, Alverstoke, UK.

Tikuisis, P. (1995). Predicting survival time for cold exposure. Int. J. Biometeorol. 39: 94-102.

Tikuisis, P. (1997). Prediction of survival time at sea based on observed body cooling rates. Aviat. Space Environ. Med. 68: 441-448.

Critique of Methodology

Frank E. Harrell, Louis D. Homer
(Editors' note: This section is arranged in order of the presentations.)

Overview of Survival Functions and Methodology.
(Wayne Gerth)

DR. HARRELL: I think there were many nice things that Wayne laid out. One of them is how different pieces of the data enter into the likelihood. Another is it how the fit of a model to the data is influenced by the different event times or event time intervals.

There were just a couple of things that I might pick on Wayne for. He mentioned that the Cox model cannot predict outcomes, but that is not really true. While the Cox model is designed to predict relative effects, like hazard ratios, it can also predict absolute effects, such as survival probability. It may not do so quite as accurately as a parametric model, but it is almost the same.

Then Wayne made a point about which my comment is going to seem very subtle: that there are formal ways to test for unnecessary parameters in a model, and that deletion of those parameters will ameliorate over-fitting. That really is not the case, because we know from simulation studies that use of your data to discern which parameters should be in the model is just another kind of over-fitting. The only time we see gains from looking at multiple models and thinking that we are reducing problems with over-fitting is when we consider the final model as if it were pre-specified and ignore how we got there. So, the value of tests Wayne described is really a mirage.

DR. HOMER: I, too, enjoyed Wayne's presentation. A couple of things came up right off the bat. One was a question asked about what should be done with a data set that the model does not quite fit in a meta-analytic combination of data sets; whether it was all right to define a new category. Paul gave us a very nice summary of such situations.

One of the things that was not mentioned explicitly can sometimes be done with two different data sets that are not compatible under a single model: Generalize the model with just a few new parameters. Instead of trying to fit the different data sets with wholly separate models, keep some of the parameters in common between the data sets under a single model and specify other parameters to accommodate differences between the two data sets. I think Erich's talk gave us something very close to what I have in mind as an example for that.

Another thing came up with Wayne's talk. What do you do with parameter values that are unexpected; that is, parameters that do not make physical sense in a mechanistic model? One of the answers that both Peter and Wayne addressed is that we just have to face the fact that the model sometimes does not fit. Another thing that was not mentioned is that occasionally the standard errors on those non-fitting parameters are so large that it is inconclusive whether they are revealing much. I think Peter in fact was guilty on one occasion of bad-mouthing one of his own models that appeared to me to have been perfectly all right.

NMRI Models of CNS Oxygen Toxicity.
(Paul Weathersby)

DR. HARRELL: Paul described three models for the probability of oxygen toxicity: a constant hazard model, a power function model, and a four-parameter model. The last two really gave about the same log-likelihood. Paul went on to give us really nice descriptions of the different shapes these two models can take, even though the differences did not matter so much for the calibration data set he used. However, he gave a nice derivation of how these differences affected model performance on other data sets.

The general comment I had about Paul's presentation, which relates to many of them, is that there is a real problem with categorizing continuous predictions into intervals when assessing the accuracy of model predictions. The intervals can be very arbitrary, and also it is not a very powerful way to assess predictive accuracy. In Paul's particular case, he divided outcomes into three risk intervals, and got a Chi-squared with three degrees of freedom of 0.78. That sounds pretty good, but such a validation of model accuracy has fairly low resolution.

Another suggestion that will also come up in discussion of other talks is the possibility of rank ordering the severity of multiple symptoms rather than just using the presence or absence of any one symptom. There might be some information

to be gained by making a judgment about the relative severities of the different symptoms.

DR. HOMER: I am also very unhappy with the Chi-squared test as a way of trying to judge how well I have done. I would say that generally speaking, if it tells me that I have a bad situation, I believe it. If it tells me that I have an acceptable situation, the test is a piece of trash.

Another thing that I try to do is categorize the results into bins containing nearly equal numbers of events. The reason is that Chi-squared is notorious for being poor if the cell sizes get small. We all are often in such situations, and it is not unusual for us to worry about a bin that is empty. So as I prepare results for publication in a chi-square table, I know that somebody is going to want to see the table with equal numbers. I will usually do a calculation in which I have near equal number of events in the bins and try to keep those numbers each larger than five.

Paul also brought up something about global maxima and local maxima, which Erich echoed in a reminiscence of some of his tortured experiences. I could sense the pain in Erich's voice as he was telling us about how he has searched and searched and searched, thought he had the answer, and then made the mistake of running one more run, which then went on for another two or three weeks to achieve a new higher likelihood. I think you have to do it that way: There is no substitute for trying lots of different parameter starting values. When working with multiple parameters, you can be easily deceived into believing that you have achieved a global likelihood maximum, which is revealed only with continued work to have been a local maximum.

Modeling Diver Tolerance to Breathing Resistance.
(John Clarke)

DR. HOMER: John's talk sparked something new in me that I had not noticed before: When he went to his logit model, he included interaction terms. Now the inclusion of interaction terms in these complicated situations is very, very important. I just got through working on a paper trying to predict renal complications for diabetics, in which I chanced to start by including all the parameters at the beginning, and then progressively eliminated parameters that seemed to be less important. I ended up with a model that had fair predictive capabilities, but with none of the linear terms remaining, only quadratic terms. So, when you are getting ready to do some of this modeling, if it is appropriate, I would urge you to follow John's example.

DR. HARRELL: I think John presented a really nice description and graphic lay-out of the different events that would cause you to stop working; dyspnea, unconsciousness and fatigue of the diaphragm. I suggested when talking about Paul Weathersby's presentation that we might want to look at the severity of the symptoms, and in John Clarke's presentation, we see that there is actually a complication in doing that for his set-up. He said that a subject can get dyspnea and stop working; or might not get dyspnea, work awhile longer, and then lose consciousness. You could argue that it is a good thing for a subject to get dyspnea, because it provides a warning not to continue working, not to proceed and be put at risk for a worse outcome. So, when you are trying to rank the order of different events, you could possibly run into problems, as occurrence of a minor event acts as a premonition of a worse event and thereby precludes occurrence of the worse event.

DR. HOMER: Do you want to launch onto a larger treatment of having more than binary events, because having only binary events is sometimes very confining?

DR. HARRELL: Right. I think that is going to be a common theme here for several things. But I wanted to come back to your statement about binning, and specifying Chi-square intervals based on the number of events rather than on the number of subjects. I really do not like any of the binning approaches. I just think they give you an accuracy curve with too large a variance. I will be talking later about using smoothing techniques for getting accuracy curve estimates.

A Log-Logistic Survival Model Applied to Hypobaric Decompression Sickness.
(Johnny Conkin)

DR. HARRELL: Johnny Conkin used a log-logistic model for looking at altitude decompression. The dependent variable was intervals of time until DCS. There was one model that also looked at time until VGE, the grade of VGE and its location. I think there is a real problem with this particular model in that when you write down the model that includes a time-dependent covariate, you have to be very careful to make it so that the time dependency is done in the framework of the instantaneous hazard. The time dependency must be expressed explicitly as a function of time. In other words, if you construct a variable that is time until VGE, the value of that variable is unknown at time zero, and at time one minute. Specification of the variable value at those times requires you to look into the future. Putting such a variable in the model

consequently gives you the wrong likelihood. It yields a model that is very difficult to interpret, particularly if for subjects who do not get VGE, you put in the censoring time for that particular observation. If you put in the maximum observation time as the time until VGE for these cases, that is a fairly arbitrary type of variable to include in a model.

I am very dubious of using a "looking into the future" type of variable. When you are going to have time-dependent covariates, they need to be included in a very special way that causes ongoing modification of the hazard, and not by using a summary value such as time until the intervening event. This is something that we will revisit because it also applies to at least one other talk given today.

Johnny had several models that got fairly complex as he added variables into the particular formulation that he had. By the time a certain complexity is reached, it no longer helps so much to have a model that looks mechanistic. The model tries to make sense out of the underlying physiology, but as it gets so complex, you might as well have a completely empirical model. It would be like an ordinary regression model that has the right product terms for interactions, and the right square or higher-order polynomial terms. You will find that such a model will fit just as well as having the fairly complex ratios with different things in numerators and denominators.

DR. HOMER: Johnny got me thinking about why the astronauts do not have more cases of the bends. I was not paying as close attention to some of the rest of it. I still want to come back, if we have time later, and hear his thoughts on that. I will move on to the next paper.

Testing of Hypotheses About Basic Mechanisms with Risk Functions.
(Hugh Van Liew)

DR. HOMER: Dr. Van Liew gave me a very useful and interesting perspective that shifted away from using models to make predictions towards using models to understand how things work. The one thing that he mentioned over and over again, that really rang a cord with me, was plotting data. I am glad I do not have to do it with pencil and paper any more, but I really think that there is no substitute for visualizing, first your original data, and then how well the model works or does not work in numerous different ways. I noticed also that Dr. Van Liew treated us to multiple variable views of how well the data was doing. I think all of this is absolutely essential to making progress.

DR. HARRELL: I really liked the way Hugh talked about including the ascent rate in the model, and how that uncovered the shape of the hazard function. Once you adjust for that variable, you got much closer to the truth and the underlying function of time.

He talked about unexplained variation, and I think there is just one other source of unexplained variation he could maybe stress a little bit more. That source is simple randomness, because at a certain level that we cannot measure, unexplainable randomness will always be present. You might include another variable and explain some of the variation, but there is always going to be some noise we cannot deal with.

Survival Models for Altitude Decompression Sickness.
(Nandini Kannan)

DR. HARRELL: Nandini Kannan covered a real nice overview of the big field of survival analysis. I was glad that she talked about the Cox proportional hazards model since, that is, by far and away, the number one survival analysis model used.

She mentioned that a parametric model is more powerful, and that is the one thing I would take issue with. The Cox model has the same power to assess the impact of certain measurements or risk factors as a parametric model. What she might have been alluding to, and I think I mentioned this when discussing Wayne's talk, is that parametric models would be slightly more precise in getting survival predictions - but not by much.

Now, one interesting thing that Nandini did was to fit an improper time-dependent covariate. I think she did it on purpose to show why you should not do that. The issue was fitting the time of occurrence of the maximum bubble grade, and that has the problem that I talked about earlier, where inclusion of a "looking into the future variable" gives you the improper likelihood. It is not the correct likelihood for time-dependent covariates. It is pretending that time until maximum bubble grade could be known at time zero. She pointed out that you get this huge Chi-squared for that variable, which is, you know, not surprising because that is sort of pre-destined.

She had some nice examinations of goodness of fit, and looked at the empirical distribution function. There was one minor problem in saying that something fits if the cumulative distribution function is within the confidence bands of the

bootstrap, because you really have to consider that the empirical distribution function also has its own confidence bands. There are consequently two sources of error: one from the variance in the estimates; say from a Cox model; and the other from variance in the observed survival distribution. The observed survival distribution is also just an estimate, not the gold standard.

DR. HOMER: On this VGE business, I think it does not help us in writing diving tables or making predictions. But suppose someone was to say to you: "I put it in to try and understand something about mechanism, and it really did help us to predict." Would you feel the same way?

DR. HARRELL: Yes. I would, unless you enter it correctly, not as a baseline variable. It has to be entered in a fashion that updates the value of a time sequence type of variable.

DR. HOMER: Well, it becomes a mixed model, but if it is a factor that is improving the description, even though it is not consistent, and even though it is useless as a predictor, would you feel you might have learned something about the mechanism?

DR. HARRELL: Yes. But, the way you would have to do it is to put the bubble grade in as an instantaneous measurement at each time. You cannot simply use the time until the maximum bubble grade.

DR. HOMER: Not to have a consistent model. Yes.

DR. HARRELL: Nor even to learn about it.

DR. HOMER: I was also very impressed with the use of the bootstrap. I have very consistently used propagation of error formulas, but those really are only very good if you are fortunate enough to have a well-behaved likelihood surface, and you are close to normality. In fact, very often we are not. Here is a simple way to look at how far away you might be. Take one of your parameters, and its estimated standard deviation. Run two standard deviations up, and two standard deviations down, and check the likelihood surface to see if the likelihood ratio test agrees with the approximate t-test. What you will often find is that the likelihood surface gives you much larger confidence regions on one side than the other. Such situations violate the usual assumption that you have a normal surface in the region of the estimate.

DR. HARRELL: Dr. Kannan also had real nice description of the confounding due to the dependence of oxygen pre-breathing time on the intended altitude. She also had a really nice description of the classical accelerated failure time formulation. We have seen other log-logistic formulations talked about today, but the one she described is the more common one; the accelerated failure time model in which you write about how your risk factors affect the median time until an event.

Now, one thing I remain unclear about is why a large variation in onset times would motivate us to use a weighted likelihood. So, I would need to talk to her about that to understand it.

There was a really nice use of the Cox model in estimating an underlying hazard function, and in looking at what shape you actually get. So, you are estimating this hazard function non-parametrically, essentially empirically, and then you can check to see whether it looks like a log-logistic hazard. That is a good way to justify the use of a log-logistic.

There was one additional issue in Dr. Kannan's use of a time-dependent covariate because, if I understood her correctly, the model under-predicted for low pre-breathing time. That signals to me that maybe there is a main effect that was not modeled correctly. It was not that you needed to do anything really fancy, but there may have been a linearity assumption that was used for the effect of pre-breathing time that might have been relaxed, that might have made the model fit without having to add some other variable.

Multinomial Bubble Score Model
(Peter Tikuisis)

DR. HOMER: I enjoyed Peter's talk. I think Peter and I have talked a lot over the years about trying to get those parameters to end up agreeing with solubility, if it is supposed to be solubility, and with the real diffusion coefficient if it is to be one. I see finally that you got the time constant down. That was very satisfying to me. Those 360-minute time constants have given me trouble for years. I have no idea what they mean physiologically.

DR. HARRELL: I got a little bit worried when Peter began by saying that bubble grade was a multinomial variable.

That sort of sets up a worry flag for me that maybe he was not going to use all the information in the bubble grade because a multinomial analysis would treat the outcome like Chevrolet, Chrysler or Ford, where you do not have any way to order those cars. Bubble grade is something that is naturally ordered. So, I would have called it an ordinal response variable. You would call it a continuous response variable if you could measure it more accurately than to five values. But Peter grouped his five initial bubble grades into three categories: 0; I and II, and; III and IV. I think there is a risk that Grade I is not the same as Grade II, and that Grade III is not the same as Grade IV. Such categorization may consequently translate into little losses of power and precision.

In analyzing data that really is naturally ordered, I try to keep from making ties, at least any more ties in the data than there already are when the data are measured. So, I would try to deal with that as an ordinal variable.

Now, since the model was an exponential in the probabilities for different categories, I think it does end up utilizing the variable as ordinal, really not multinomial, although it was doing so for only the three values allowed under the grouping scheme used. So, what I would attempt to use there is a classical ordinal regression model that allows you to have as many levels as you want, and that always makes sure that the ordering of those levels is used. These models are really generalizations of the Cox or Spearman correlation approaches.

DR. HOMER: Did you try more than those levels, Peter?

DR. TIKUISIS: No, we did not.

DR. HOMER: It does raise an interesting issue, though. We have talked about considering different kinds of decompression sickness, like Type I and Type II. There are some areas in which we perhaps should be interested in multinomial models, though indeed if one has continuous data, generally it should stay that way.

Probabilistic Models of DCS During Flying After Diving: Motivation for Mechanism.
(Wayne Gerth)

DR. HOMER: I was curious when you described the failure of the model on Duke flying after diving data. Did I read it correctly that there were something like two cases of DCS occurring late, and it is on those observations that you were basing this assessment?

DR. GERTH: Yes. But only two cases occurred early, while about eight or so occurred late. The overall number of hits in those data is still very low.

DR. HOMER: So, it was not a large number of failures. Now let me ask you, when you treated those cases, did they respond?

DR. GERTH: Yes.

DR. HOMER: So, what do you think? Was it a bubble?

DR. GERTH: We have to be careful about trying to infer "truth" from any aspect of model performance. The point I was trying to make is that I cannot model DCS risk with a bubble during the time frames of those DCS incidents, no matter what I do with it.

DR. HOMER: No, I understand that. I mean you might have to give up the bubble, but I am curious to know whether you went one step further. I presume that you treated those cases with hyperbaric therapy and observed that they responded.

DR. GERTH: Clinically, they responded.

DR. HOMER: Okay. So, do they have a bubble or not?

DR. GERTH: I do not think that in those cases, where we had very late onset, that we were treating a bubble any longer. We were treating something that a bubble caused.

DR. HOMER: Okay. So, you are saying that we can successfully treat something that is not a bubble by recompressing it, is that right?

DR. GERTH: Remember that recompression also entailed administration of oxygen. So, we might have been treating an ischemic condition.

DR. HOMER: Boy, it is awfully hard to get you to admit that it is not the bubble that is doing it, isn't it?

DR. GERTH: Ed is going to give a presentation later this week in which he illustrates that there is not much we can do to bubble dynamics models to account for the persistence of a bubble under those conditions for much more than a few hours.

DR. HOMER: Well, you could give them up and go after something else.

DR. GERTH: I will back off and say we do not know whether the model is true or not. However, I cannot make the model I described fit those late-onset cases if you hold it to what it is supposed to represent, i.e. that DCS risk arises from presence of a bubble. I cannot make a bubble last long enough in that model.

DR. HOMER: I think that looking at your data that way and saying, "No, the model will not account for this," is really in the finest traditions of what you ought to be doing. This takes me all the way back to Kaplan. So, thank you for a nice observation.

DR. HARRELL: Wayne incorporated time-dependent covariates in his model through their effects on the hazard function, not through their effect on the time axis, on the failure time variable, if I understand the model correctly. That would mean that, unlike the way we saw Dr. Kannan present the accelerated failure time family, this was not an accelerated failure time model; and that is fine. It is just that sometimes it is useful to model in terms of how you accelerate or delay a failure. How do you move the time axis rather than how do you multiply the hazard function? This other approach, using classical log-logistic or log-normal accelerated failure time models, is often interesting to entertain.

Wayne had a good discussion of confounding factors and an excellent lay-out of the shapes of profiles and where the time origin is. He had a wonderful description of the covariate process and made the point that when you are using time-dependent covariates, you do not need to know the future values of them to look at how they modify the risk at a certain point in time. You just look at the current values of the measurements.

One thing I did not understand was his sinusoidal example where he said, "Here is an underlying hazard function if we do not know about bubble formation or how other variables contribute to the hazard." He showed an overall hazard function that you would have if you did not know the values of some time-varying covariate, like bubble formation, and then a fast-moving sinusoidal shape that would be the hazard if you knew the values of that covariate. The maximum points of the sinusoid sort of followed the simpler hazard. I really was not motivated about that. It seems to me that the more you know, the less the hazard function is going to stay smooth, and the more it is going to vary up and down. If you get in a car, for example, and start driving at a hundred miles an hour, you know that your hazard function takes an instantaneous increase. But you also know that when you stop the car the hazard will come back down. So, I need to talk to him some other time to understand why the sinusoidal shape hazard would not be realistic.

DR. HOMER: I would add that I enjoyed your remark just now, about how use of different models provides different views of what is working and what is not to prompt a different sort of thinking. I think I am one of those people who is partial to the hazard models, as Wayne is, but using the other models sometimes does give you another way of looking at things.

DR. HARRELL: I might add that the accelerated failure time models are especially nice if you want a simple calculation of, say, median failure time. They are very simple in form.

Improving on a "Good" Model.
(Erich Parker)

DR. HOMER: I want to reiterate my sympathy for Erich's convergence problems. The other note I had for Erich was on goodness of fit assessment. We have covered both topics in several different directions.

DR. HARRELL: I want to dwell on goodness of fit assessment for one second more, particularly when lack of fit is viewed interval-by-interval, or when there has been some sort of binning going on. Erich made an excellent point how the

goodness of fit depends drastically on how you categorize a continuous variable.

I think when looking at the problem with the initial model under-predicting, you still probably have to do some more smoothing before you really make a firm conclusion about exactly how much underfitting it is doing. If you find that the model just does not seem to predict well for one little interval, I would look at the surrounding intervals or use a continuous type of moving average, something to get a smooth estimate.

The bottom line is: we know there are dangers in tweaking a model. You will eventually make it fit the data that you have, but it will then fit future subjects less well. When you are making modifications to the model, you want to really be sure to do it in a very patterned way that is not specific to one small interval.

So, I would tend to use a lot of smoothing even for that purpose.

DR. HOMER: Do you happen to have a reference with you on that? Because this business of binning is very important to all of us. The journals definitely expect us to be doing something like that, and I have not talked with anybody who is happy about doing it. So, if you happen to have a reference with you, that would help.

DR. HARRELL: Yes. I will show one when I get to my little talk with an example of using the smoothing method. We had a paper in Statistics in Medicine in '96 that shows how to do that.

The most often cited method for looking at calibration accuracy is probably the Hosmer and Lemeshow paper from years ago. That was for looking at accuracy of logistic models by binning the data using the percentage of events instead of the average predicted number of events. They actually had a newer paper out in Statistics in Medicine a couple of years ago showing that their prior method had serious problems. With the one they proposed about 15 years ago, they found that just using different definitions of the bins (how you calculate a decile, for example) can really change your Chi-square value. So, the people that really invented that Chi-squared goodness of fit test now have serious reservations about using a test that requires that sort of categorization.

Meta-Analysis of Diver Decompression Data.
(Paul Weathersby)

DR. HOMER: I liked Paul's first solution. Just go ahead and ignore it all because no available solution appears completely satisfactory. Thinking of all the attendant problems as he was talking, I do not know what else one can do.

I have watched Erich do test after test to see whether different data sets were compatible. The problem in those endeavors is that you have a multiple comparison difficulty. In comparisons of eight or 10 different data sets, some of them have to come out looking not as good as the others. So, I do not even know that formal statistical tests of data combinability are completely satisfactory.

DR. HARRELL: Yes. A related issue is that the test for combinability really does not have much power. If you are trying to feel comfortable that you can combine two data sets because the test for whether they were combinable was very insignificant, you cannot really feel that comfortable. It is very difficult to know what sort of P value cut-off to use. You certainly would not use 0.05, but people are very confused right now about whether to use 0.1 or 0.2. It needs to be relatively high because of the low power of that particular test.

I think sometimes, if you have the right kind of data documentation, you can rescore outcomes using uniform criteria. You may have had some experience with that, but I could see where that would have some advantages if you have the right narrative descriptions accompanying each dive that can be scored by impartial objective reviewers.

DR. HOMER: Of course, you do not choose those sets randomly either, do you? One sort of bright spot is that consideration of this problem invariably starts you thinking about why data sets may be different. Occasionally, you may come up with a new idea. So, the problem is still worth worrying about.

DR. HARRELL: Yes. I want to commend Paul, too, for coming up with a list of criteria for judging similarity of studies. I think that is a very necessary step.

Cold Exposure Survival Model
(Peter Tikuisis)

DR. HOMER: I thought the way you coupled the deterministic model into a probabilistic problem was interesting. One of the things that I saw in one of your papers, again coming back to subject matter rather than methodology, was that at about the time when you would like to really know what is going on, the metabolic rate of the person takes a nosedive. Is that the major deficiency in the predictive model that you described for us?

DR. TIKUISIS: Yes. That is one of the great uncertainties.

DR. HOMER: If it is exceedingly variable, the whole prediction must just come to pieces about that point.

DR. HARRELL: Peter had a really beautiful interface for using the predictive instrument. What programming system was that using?

DR. TIKUISIS: Visual Basic.

DR. HARRELL: That was really nice.

One of the things Peter addressed in that predictive instrument, that few people are willing to talk about, is that you use a number of predictive variables; and then invariably there is one variable you cannot obtain. You cannot measure it, or you are in a hurry. He had a built-in sort of imputation. I guess the percent fat was one of them, and there might have been others. So, he had some logic going on behind the scenes that was really nice for imputing missing predictor values.

I, too, really liked the combination of deterministic and probabilistic models. There was just one little point I would take issue with. Peter quoted a goodness of fit test. Again, we have the problem with the categorization. But I think his was a Chi-squared test that was calculated on the training data, and that is not really informative because you know the training data is always going to fit. Training data will always be on the 45-degree line. I would really use a bootstrap for getting an assessment of goodness of fit. That really adjusts for over-fitting as much as possible.

Promising Approaches to Experimental Design

Louis D. Homer

A remarkable thing about my talk is that I heard it earlier today at lunch, from Dick Vann, and I liked his so much that I decided I had to tell the rest of you about it.

Dick was much more eloquent than my outline slides. We talked about planning experiments, that is, experimental design. One view of diving research is to think of a particular diving series as being a binomial problem, in which you either develop the bends or you do not. That view has a lot of literature behind it on how you plan the size of the experiment. Another approach is to use a model. But if you are going to use models, and you are going to estimate parameters (rather than the probability of decompression sickness directly) then there are a number of things that you should begin thinking about.

One point is that you really do not have to be exactly on the mark with respect to the probability of bends that you are going to have. Suppose you are trying to predict a one-percent profile. You may be perfectly all right in doing a five-percent calibration trial for your model, and extrapolating to one percent using the model. That is a relief, but then you still have the responsibility for deciding where you should do that trial. Should you do it at a one percent incidence? Should you do it at five percent? Should you do it somewhere else?

You need a design that represents the plan for the dive. The design will specify the kind of dive, and results (data) which is to be used with the model to estimate the parameters of the model. A good design should minimize costs in some sense, and should optimize some function, F, of the parameters. For example it might minimize the number of dives and at the same time minimize the variance of the estimated probability of decompression sickness. The function F will depend on the parameters to be estimated, β, and the design, D.

$$F(\beta, D)$$

In the simplest case, F might just be the probability of bends.

$$F = P(bends)$$

One might choose to minimize (optimize) the variance of a parameter of the model, or some more complex function of the parameters and the variance covariance matrix. If we choose the variance of F, V, we can estimate that variance with a propagation of error formula:

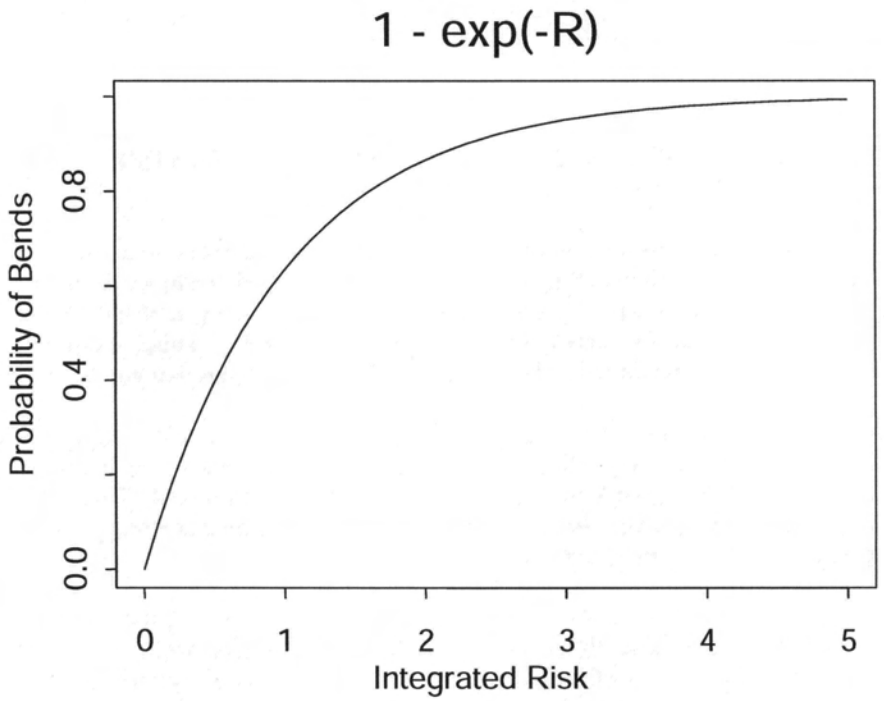
$$V(F) = \left(\frac{\partial F}{\partial \beta}\right)^2 V(\beta)$$

If F is not a parameter itself, the variance of F still is a function of the parameters and the variance-covariance matrix. Generally, we will have in mind some function that we are trying to optimize. Again, it might be as simple as the variance of the parameter itself. Then we have one or more designs, call them Design-1, Design-2, Design-3; choices that we have in mind for how to run the dives. We can simply try out each design with computer simulations and choose the one which minimizes the variance of F.

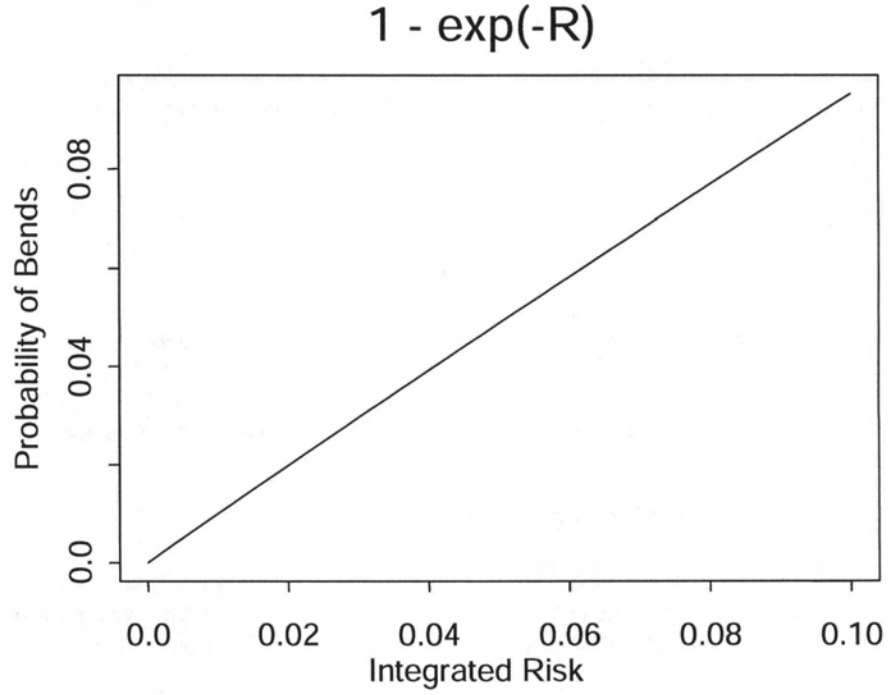
A "SIMPLE" PROBLEM TO OPTIMIZE EXPERIMENTAL DESIGN

Now I am going to treat you to a simple example. First I want to convince you that in some settings, it really does make a difference to think about planning the design. I want to stimulate ways of thinking about what you need to do (in considering possible designs) that you might not have done, if you did not go to the trouble of thinking about the variance of F.

The particular example that I am going to use is simple. You go down to a depth, stay eight minutes, and come back. The trial is trying to understand how depth is related to the risk of decompression sickness. (I am going to end up telling you that the best design will be to bend as many people as you can). Now the figure below is like a lot of the curves that you've seen before.



The Y axis is probability of bends, and the X axis is integrated risk. The reason I put it here, is to remind you that most of the time we are working way down in the lower left corner. If I expand that corner and just show that to you, this is what it looks like.



The following few lines of math explains analytically why the initial curve seems to have become a line. If you expand the infinite series for the exponential term in the first term for probability of bends, you end up with the conclusion, that in this narrow area, the probability of bends must be close to that integrated risk.

$$P(bends) = 1 - 1 + R - \frac{R^2}{2} + \ldots$$

$$P(bends) \cong R$$

Now I am going to take you on another very big leap. I am going to tell you that for this simple example, the integrated risk is going to be proportional to the depth:

$$R = \beta d$$

This assumption allows me to give you a nice simple formula for the variance. Referring back to the propagation of errors formula, the variance of the probability of bends is going to be the square of the depth times the variance of that proportionality constant, beta. Remember that the variance of the estimated (binomial) probability is going to be probability, P, times one minus P and divided by the number of dives, N. I then make the substitutions, and I end up with the same expression but now including the proportionality parameter, beta, the depth, and the number of dives. In algebra, we have:

$$P(bends) \cong \beta d$$

$$V(P) \cong V(\beta) d^2$$

$$V(\hat{P}) = \frac{P(1-P)}{N}$$

$$V(\hat{\beta}) d^2 = \frac{\hat{\beta} d (1 - \hat{\beta} d)}{N}$$

$$V(\hat{\beta}) = \frac{\hat{\beta}}{dN} - \frac{\hat{\beta}^2}{N} \cong \frac{\hat{\beta}}{dN}$$

The variance of the proportionality factor, beta, is approximately equal to the proportionality factor itself, divided by the depth and the number of dives. So, to make the variance small, you make these terms in the denominator (depth and N) as large as you can.

Other ways of saying this, are that in order to have a fixed variance in the parameter of interest, beta, you simply need to bend a fixed number of people, and you have two ways of doing it. Either you can take a lot of low-risk dives or you can take a few higher-risk dives. But to have the same variance, you have to bend exactly the same number of people. Or from another view, if you want to double the depth; that allows cutting the number of dives by half, but you have the same variance and the same number of DCS events.

Now, that was a little parlor game. Reality is sometimes a little more complex. First of all, P is not necessarily small in some problems. I think I saw one slide, where for the flying people, they were advertising 80 or 90 percent incidence. That is marvelous. I wish we had had stuff like that in the Navy to deal with.

Occasionally the time of the event is known. It was not in the example case that I gave you above. If the time of the event is known, you would be foolish not to use that in your model.

Also, usually you have many parameters. Six, eight, 10, a dozen, are not unusual, and usually the design and the choice of design is fairly complicated.

In Wayne Gerth's model presented today, he wants to dive you, bring you to the surface, let you go have a drink, and then go flying. So, you have to be able to take care of all of that. (I did not see the drink step in his model.)

OPTIMIZING MORE COMPLEX DESIGNS

My perception of the more complex problem, is that you still are going to have some function that you want to optimize. For our purposes here, we will talk about simply minimizing the variance of that function.

The function could be even simpler than the example above, something as simple as the probability of bends. It might be like the example above; and in Dick's talk that he gave at lunchtime, he was interested in the slope of something. So, it could be dependent on one parameter; or it could be some other function of several of the parameters. But you need to develop an expression for variance. I have suggested using propagation of errors, because I do not know how to do any better. But, I must say I am going to start looking for better ways after some of what I have heard today. For example, one could use Monte Carlo simulations to obtain estimates of the variance of F.

We will assume for now that we are using a propagation of error estimate of the variance of F, obtain a variance covariance matrix from estimation performed with data from a simulated dive, and then we will apply the variance matrix to the calculation of the variance of the function to be optimized.. The first step of this is to look at the propagation of error formula:

$$F_i = \left(\frac{\partial F}{\partial \beta_i} \right)$$

$$V(F) = F_{11}V_{11} + 2F_{12}V_{12} + F_{22}V_{22} + \cdots$$

where the F_i are partial derivatives of the quantity of interest, F, with respect to each of the parameters, the V terms are from from the parameter variance-covariance matrix (the V_{11} element is the variance of parameter 1), and F_{11} represents the square of the partial of F with respect to the first parameter, and so on. In this manner you develop your expression for the variance of F.

To estimate the variance you do some Monte Carlo simulations. You start with the best values that you now have of the collection of parameters, the betas; and a projected design, Design-1. The simulation produces a sample data set. From that data set, you estimate the simulation betas; and, in the process of maximum likelihood estimation, you get the variance-covariance matrix for that simulation. With this variance-covariance matrix, one can use the propagation of error formula to estimate the variance of F. You repeat the simulation over again with Design-2, and with Design-3. Get your new variances, calculate the variance of F for each, and simply ask yourself which design is going to get you where you want to go (mininum variance on F). In order to be reliable, the Monte Carlo procedures must be repeated, so that you obtain a collection of estimates of β and a collection of estimates of F and its variance for each design. Then you would choose the design providing (usually) the smallest estimates of variance for F. If you can afford the computer time, it would be possible to estimate F many times, and calculate a variance for F from the repeated simulations rather than from the propagation of error formula.

Directions in Statistical Methodology for Multivariable Predictive Modeling

Frank E Harrell, Jr
Division of Biostatistics & Epidemiology
Department of Health Evaluation Sciences
University of Virginia School of Medicine
Health Sciences Center Box 600
Charlottesville, Virginia 22908

There are several elements of developing reliable statistical models. Some of them are

1. Choosing between theoretical and empirical models
2. Selecting the model structure or model family
3. Modeling the shapes of the effects of independent variables, or how to best transform them to fit model assumptions
4. Diagnosing the fit of the model
5. Quantifying the precision of parameter estimates and the overall accuracy and predictive power of the model
6. Drawing inferences about associations
7. Validating the model so as to get an idea of its likely performance in the future
8. Presenting the model graphically to non-statisticians

Readers may want to see tutorials in *Statistics in Medicine* in 1996 and 1998 for detailed case studies of the development, validation, and graphical presentation of multivariable empirical models, for survival and ordinal response data, respectively.

In terms of measuring the accuracy of predicted values from fitted statistical models, many authors categorize predictions so that simple summaries can be derived. For example, it is common to stratify predicted risks into deciles and to plot the proportion of events in each decile vs. the mean predicted risk in that decile. It is easy to show that the assessment one obtains from such a procedure is very dependent on how intervals of predicted risk are selected. There are many advantages to using statistical measures of predictive accuracy that do not require grouping the data. One popular summary index for is the area under the receiver operating characteristic (ROC) curve. This is a measure not of absolute accuracy but of strictly discrimination accuracy. R^2 is another measure that primarily assesses discrimination. One of the more common indexes that combines absolute or calibration accuracy with discrimination accuracy is the Brier quadratic probability accuracy score. This is a mean-squared error measure used by the U.S. Weather Service for judging accuracy of rain forecasts.

An excellent method for assessing absolute predictive accuracy when the response is binary is the smooth calibration plot based on the lowess nonparametric regression smoother (Cleveland, 1979). An example of a non-parametric calibration is illustrated in Figure 1. This example is from one of the must studied of all risk prediction problems - assessing risk of individual patients undergoing open heart surgery. The predicted probability of operative death is shown on the X-axis. The average probability is around 0.03 across the spectrum of patients. The actual probability, or our best estimate of it, is shown on the Y axis. We don't know the actual risk, but we can estimate that by a non-parametric regression on (predicted risk, 0/1 binary outcome). What we're trying to show is agreement of predicted results with the 45-degree line, which is the line of perfect prediction. This curve is just a fancy sort of moving average between the predicted risk and the observed zero-one, or binary, outcome.

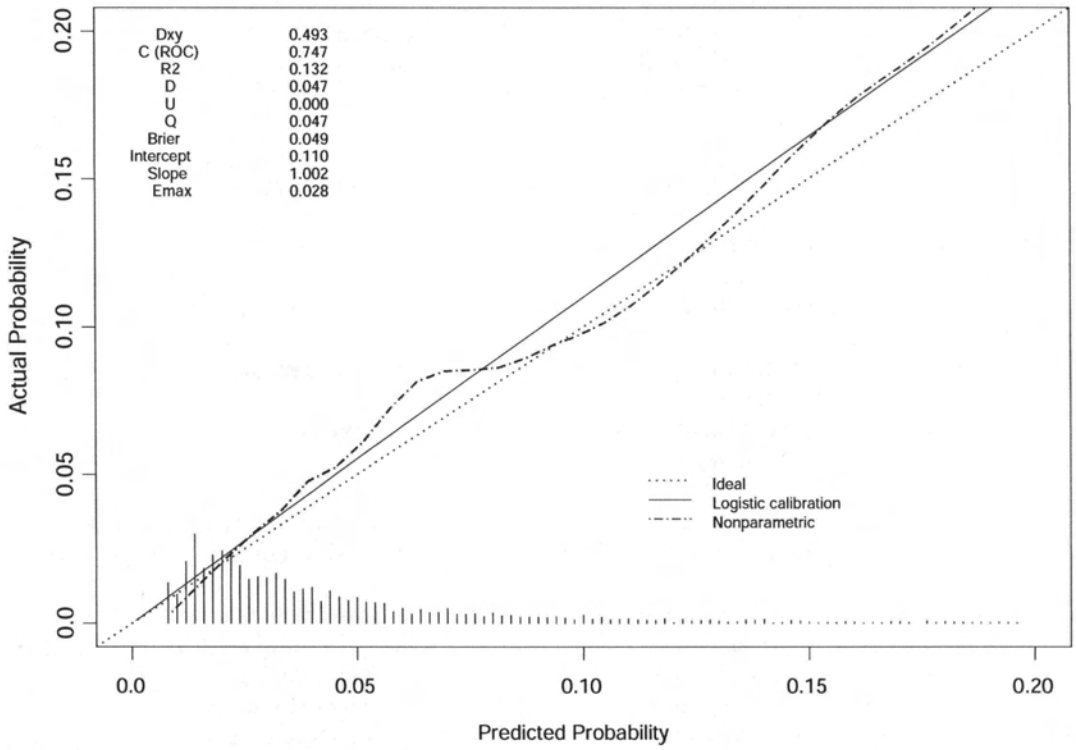


Figure 1: A nonparametric calibration curve for assessing absolute predictive accuracy.

We can see that we are very close to the identity 45-degree line, especially where we have a lot of data at risks less than 0.05. The histogram at the bottom of the graph depicts the distribution of actual predicted risks. There are many other very well developed and justified indices that quantify the fit of predicted to observed binary outcomes in various ways, but limited space precludes covering them here.

For assessing the predictive accuracy of survival models there is a rank correlation index that generalizes the ROC area to quantify discrimination ability. This assesses our ability to discriminate individuals having early failures from those having late failures. Assessment of calibration accuracy is not as well developed for survival models other than using arbitrarily stratified Kaplan-Meier estimates.

Modeling Tools

There are many modeling tools that become more helpful as the number of variables increases. We have just seen one use of nonparametric regression such as lowess. Instead of using predicted risk as the independent variable, we could use pressure or depth. Nonparametric regression methods are not only available for a single predictor or X variable, but there are well developed generalized additive nonparametric models for multiple X variables (Hastie & Tibshirani, 1990). There are also models for optimally transforming the X and Y sides of the equation simultaneously for a continuous response variable Y (Tibshirani, 1988). Piecewise polynomials (spline functions) that are almost non-parametric can also be used for modeling shapes of effects.

I note in passing that empirical regression methods can be used to do formal tests of adequacy of biomathematical models. Suppose one had a pre-specified model, say $y = 1 - \exp(x^2/h)$, and we want to test whether that model is adequate. We can embed that model inside a more general one such as $y = 1 - \exp(x^2/h + \text{spline}(x))$, where spline(x) represents a piecewise cubic polynomial whose coefficients are to be estimated. By jointly testing the regression parameters of the spline function for significance we are testing the adequacy of the x^2/h equation.

Model uncertainty is a big problem that statisticians like to keep hidden in the closet. If one tests many models and picks the one that fits best, that model will never predict as well on future data. There is another problem in picking the best model on the basis of data: standard errors and P-values are no longer appropriate (Faraway, 1992). This is a well-kept secret because it takes much more analyst time do correct inferences that take into account model uncertainty. Currently the best way to do that is to use a bootstrap process as Faraway described. The bootstrap is a very valuable tool for getting standard errors and confidence limits and other things, because no statistical theory for computing those for non-pre-specified models is available.

Even when the model form is pre-specified, the non-parametric bootstrap is a good way to estimate standard errors of individual parameters without even assuming that the model is correct, and the bootstrap does not assume normality of estimates. The bootstrap can also be used to get standard errors of predicted values and confidence bands for them. This involves sampling with replacement from the original data and studying how the model changes as it is fit to each of many such samples.

The bootstrap is also very valuable for validating models because its inventor, Brad Efron, has also developed a different kind of bootstrap procedure for estimating the optimism in a measure of predictive accuracy. One can validate summary indices, get calibration plots corrected for over-fitting, and get an estimate of how well a model performs without waiting for that new data set. This is not an external validation that would validate how you collect data or how you enroll subjects in the study, but it is an internal validation that correctly penalizes the R^2 and other measures for over-fitting.

Another modeling procedure gaining popularity is model averaging. When there are competing models, better predictions will often be obtained by averaging their predictions rather than choosing a single model. Bayesian modeling is the fastest growing area in statistics, and researchers owe it to themselves to look into Bayesian methods. A frequentist method that parallels some aspects of Bayesian modeling is penalized maximum likelihood estimation (discussed in Harrell *et al.*, 1998). This technique is useful in problems that involve too many parameters and not enough subjects, where parameter estimates need to be discounted to make the model more conservative and not over-fit.

There is a problem particular to decompression research that traditional statistical models do not account for – lack of independence of observations. In decompression research it is not uncommon to have some divers participate in multiple dives within a given study. If these divers have some similarity with themselves across multiple dives, or if they tire over the course of the study, for example, that generates a dependence or partial redundancy in the data. Simply speaking, one should not get as much credit for one diver making multiple dives as for separate divers each making one dive. This may not greatly affect parameter estimates, but it really affects confidence intervals (they will be too narrow) and standard errors. There's a need in many of applications for correcting variances using some sort of cluster sampling approach or using the cluster bootstrap (an especially easy solution; see Feng, McLerran, and Grizzle, 1996).

It may also be worthwhile to include a subject's track record up to the date as a predictor. According to David Southerland, if a diver has had multiple bends, she might be more prone to bends than another diver, and that might be a powerful predictor.

Time-dependent covariables is another useful aspect of modeling. These can provide powerful and interpretable analyses when the covariables pertain to experimentally-controlled conditions. If covariables are "internal" or out of control of the investigator, it is very difficult to interpret model parameters. But such covariables can still be useful for understanding patterns of risk. For example, one might understand how to use a VGE grade profile by seeing how changes in VGE relate to the instantaneous risk, and that might help one to score or grade the severity of VGE. Internal time-dependent covariables are not very handy for prospective use.

There are models that are under-utilized such as a family of logistic models for ordinal response that uses just the ordering of the response variable. These models do not assume any spacing between the levels of the response and do not require any grouping of, say, bubble sizes.

Another consideration is whether the time until a symptom develops is more important to predict than the severity of the symptom when it does develop. There are main types of failure time models, such as the accelerated failure time family, for focusing on the time until the event. A model that has not yet seen very much use, but that might come into play here has been developed by Berridge and Whitehead (1991). Their model combines an ordinal logistic model for severity of an event with a Cox model for the time until the event. Their application was a clinical trial for preventing migraine headaches in which the goal was to delay as long as possible the next headache or, when one gets a headache, to minimize its severity.

In summary, there is a real explosion of statistical methodology for model development, diagnosing fits, and validating models. Many of the new techniques allow us to completely avoid categorization of variables by using nonparametric smoothers or flexible parametric modeling. Software such as S-Plus (see Harrell, 2000 for example), is starting to keep up with the latest statistical developments, opening new opportunities for merging empirical and biomathematical models.

References

1. Berridge DM. Whitehead J. Analysis of failure time data with ordinal categories of response. Statistics in Med 10:1703-10, 1991.
2. Cleveland WS. Robust locally weighted regression and smoothing scatterplots. J Am Statistical Assoc. 74:829-836, 1979.
3. Faraway JJ. The cost of data analysis. J Computational and Graphical Statistics 1:213-229, 1992.
4. Feng Z, McLerran Dale, Grizzle J. A comparison of statistical methods for clustered data analysis with Gaussian error. Statistics in Medicine 15:1793-1806, 1996.
5. Harrell FE, Lee KL and Mark DB. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statististics in Med 15:361-387, 1996.
6. Harrell FE, Margolis PA, Gove S, Mason KE, Mulholland EK, Lehmann D, Muhe L, Gatchalian S and Eichenwald HF and WHO/ARI Young Infant Multicentre Study Group. Development of a clinical prediction model for an ordinal outcome: The World Health Organization Multicentre Study of Clinical Signs and Etiological Agents of Pneumonia, Sepsis and Meningitis in Young Infants. Statististics in Med 17, 909-944 , 1998.
7. Harrell FE. `Design`: S-Plus functions for biostatistical/epidemiologic modeling, testing, estimation, validation, graphics, prediction, and typesetting by storing enhanced model design attributes in the fit. See hesweb1.med.virginia.edu/biostat/s/Design.html.
8. Hastie T and Tibshirani R. *Generalized Additive Models*. London: Chapman & Hall, 1990.
9. Tibshirani R. Estimating transformations for regression via additivity and variance stabilization. J Am Statist Assoc 83:394-405, 1988.

General Discussion

Editors' Note: The transcript of this session did not allow identification of all of the speakers. We apologize for any omission, or even incorrect assignment, of people's remarks.

R. VANN (Duke): I am interested in life after death. Here I do not mean this in a theological sense. But people die, light bulbs burn out, and survival analysis addresses these processes pretty well. A difficulty is that divers are cured of decompression sickness and live to dive another day.

My question is: what is the future of statistical procedures that will accommodate either a cure or a resurrection? Did I get a hint from one of your last slides, Frank, that similar problems exist elsewhere, such as the time to the next migraine, or time to the next incident of decompression sickness? Is the [reference to Berridge and Whitehead in the bibliography of F. Harrell's paper] approach something that we might use to address the problem of repetitive diving better than we are doing now with survival analysis?

F. HARRELL: The state of survival analysis, now, is incredibly well developed for looking at multiple events per subject. The classic examples would be multiple heart attacks, or multiple strokes. There are some fantastic techniques for dealing with those, and then there are techniques for dealing with multiple different *kinds* of events. But for your particular question, the closest thing I know of that's fairly well developed is if you look at death as the end point, and you have an intervening event, such as a heart attack. You can follow a subject who after some period of time suffers a heart attack, at which time his or her outcome goes from zero to one. Then you can say that if the subject survives the heart attack a certain length of time, the heart attack loses its impact. The outcome can come up (and be a one) and then it can wane. So, if you survive the heart attack a few months, it is almost like you didn't have one.

That still does not address your question, though, because you are asking about, say, a heart attack being your main end point, and then the end point resolving. Since you gave me a heads-up on this question at our break earlier, I had toyed with the idea of using a two-stage model. You model the time until the event, and then given that you had DCS, you model the time until resolution of symptoms. I think that approach has some promise, but it will probably add more parameters to the overall model. You won't get something for nothing.

V. FLOOK (Britain): This is more of a comment addressed to Erich Parker. One of the things that happens when bubbles form is that the bubbles grab the gas and make it very much more difficult for the inert gas to be washed out of the body. I would hazard a guess that this is perhaps the single biggest factor that was missing in your model when you looked at the effect of oxygen breathing post-decompression.

It obviously is, to some extent, a random effect. In some people, the bubbles will have formed before you start the oxygen breathing, while in other people, they will not have formed, or at least you think they will not have formed. Bubble formation will slow down the wash-out and very much reduce the effect of oxygen breathing. If you run our bubble model, this comes out automatically. As soon as bubbles form, you see the tissue partial pressure drop right back down almost as low as arterial partial pressure.

There's a second issue from that. I suspect that perhaps this is one of the reasons, Wayne, why your bubbles were not lasting so long. The effect was being taken into account.

W. GERTH: The bubble model is the risk function in that model, in that hazard function. So, all of those things are built in.

V. FLOOK: I am running the Van Liew bubble model, and I get bubbles lasting for many, many hours. As an example, there is a paper in last year's UPS proceedings where after a two and a half minute submarine escape exposure, the model predicts the bubbles to last for six hours.

W. GERTH: Yes. We get bubble lifetimes that long, too, under certain conditions. But not after return to ground level following an altitude exposure of the kind that I showed. It is under those conditions that I cannot do anything to our bubble model that allows the bubble to persist beyond the point of return to ground, and remain consistent with the way the bubble must behave to account for risk after the dives that precede the altitude exposure.

E. THALMANN (Duke): You are generating a controversy. I remember sitting at NMRI with Dr. Homer one day scratching my head over how we could validate these models. I said, why don't we just take our sample of dives and randomly select different data sets and compute parameter values and see how they agree with one another? When we were

finished, he had convinced me that such a procedure would not be very useful. I just learned that this is something called bootstrapping, which Dr. Harrell thinks is useful. Lou, how do you view bootstrapping as a way to validate the parameters that you compute, given that your data set already exists, and you may not get any more?

L. HOMER: I would not use bootstrapping as a validation of parameter point estimates, but as a means of judging the variance of the parameters. I think that, on the average, if you use resampling from the same data set, you are going to get the same point estimates.

F. HARRELL: It might be worth elaborating on just how the bootstrap works for validating accuracy. You take samples with replacement from all of your subjects. You refit the model, say, a hundred times, taking 100 of those samples, and you find out how much that model falls apart when you use it to predict all of your original data.

So, you are studying how some measure, say an R-squared, reduces when you go from how well it seemed to work in the bootstrap sample compared to how well it works in the original sample. It's really a kind of a backwards idea of what we usually mean by training and test sampling.

W. GERTH: So, if I understand that right, you sample your data many times, and you compute parameter values. Then you turn around and use those parameter values to look at the whole data set. If you have a good model and a reasonable data set, you would not expect to see much difference between parameter estimates from the different samples?

F. HARRELL: Right.

W. GERTH: But if you see a large variability, in other words, if your parameter values become wildly dependent on the specific sample that you take, then how do you know what the problem is? Is it the data? In other words, do I have inhomogeneous data, which cannot be combined, or do I have an unsatisfactory model? How do you decide?

F. HARRELL: Well, you can look at multiple indexes of fit. I tend to look at two types. One is the discrimination ability, and one is the calibration, or absolute, accuracy. You have certain characteristics that you see, such as a shrinkage of your predictions. The 45-degree line, instead of being 45, it tilts toward a flat line. That would be a symptom of over-fitting, or not having enough information to estimate the parameters that you tried to estimate. It would tell you that you somehow need to either make the model more conservative, or wait until you get more subjects.

W. GERTH: I was impressed that you recommended we use the C index as a measure of goodness of fit. No one here today showed that we do in fact use it. We have shied away from it, perhaps because we do not understand it well enough.

My particular question with regard to the C index is: how do we apply it in altitude decompression sickness problems? Because there, failure during an altitude exposure terminates the exposure for the person that fails, while those people that survive remain at altitude to accumulate more risk. So calculated risk for the survivors is always higher than calculated risk for people that failed.

F. HARRELL: You are saying it is right-censored?

W. GERTH: Yes. So, do you have to do the C index evaluation at each failure time that you have?

F. HARRELL: What Wayne is talking about is a measure of concordance between your predicted and observed responses. What you do is: look at all pairs of subjects, and ask how often did the subject that had a higher predicted risk actually have the earlier failure in that pair. You look at all possible pairs. That's why you need a computer.

But if you have a subject who was censored before the other subject failed, that pair is ignored. That is how you take censoring into account. This is not a measure of goodness of fit; it is a measure of discrimination ability.

W. GERTH: The reason I asked the question is because you distinguished between two sorts of tests, discrimination ability being one of them, and that's an assessment that we tend to give little weight to at the moment.

F. HARRELL: Yes. That is just an overall single number summary of how you can separate low and high risk. I think it is worth doing.

A. BRUBAKK (Norway): I am very impressed with lots of the statistical things that have been done. But by the criterion that the model should give added insight into what is actually going on, I feel somewhat at a loss. Many of the models do not tell me much about the actual process. One of the problems might be that what you are only trying to model gas dynamics and bubble formation, when your actual end point is clinical symptoms. Although we know that there is a relationship, we do not know exactly what that relationship is. I would venture to say that if there is a lot of gas, the primary bubbles are perhaps very important. But, if there are few bubbles, as there probably are when diving many modern tables, the biochemical effects would probably play a significant role. Thus, all your trying to adapt your model would be an exercise. You may adapt your model, but it does not actually tell you what is going on.

One interesting thing that came out of the study done in the North Sea, where they did a lot of diving on many different profiles, was that the only dive time and depth were correlated to DCS incidence, nothing else. The kind of profile did not matter. In a way, such a result says that it does not really matter *what* kind of procedure you are using, as long as you use *some* kind of procedure. That governs your rate of ascent, and it doesn't really matter if you stop for two minutes or 10 minutes perhaps.

I have a different comment that relates to use of bubble grades. For instance, the way that Peter displayed a continuous line on the Y axis, with Grades 1, 2, 3 and 4. That gives a totally wrong impression because if you have a Grade 3 and 4, that covers something like hundreds, at least according to our studies. A Grade 3 can be, say one, but the Grade 4 can be also hundred bubbles. There is a very large volume here. It is quite obvious that if you do not find a correlation between VGE, the number of bubbles, and the incidence of decompression sickness; it might be partly because of the way you use these.

So, I think if we are going to proceed on that, we have to get the linear scale on those bubbles. If you use that, you can actually see that some models are able to predict how much gas is produced in the vascular system.

QUESTION: Could you transform the bubble rate to make them non-linear? I think you mentioned something like that.

E. PARKER: Well, I may be able to make a comment that helps. Dr. Brubakk's insight into what goes into bubble scores is always appreciated, but is somewhat outside the modeling methodology focus that we're on today. Dr. Tikuisis' approach combined categories for reasons that made sense under a data-limited condition, and then Dr. Harrell pointed out that there are additional techniques available to us that can preserve the full ordering of the events. This does not require that there be a linear relationship between one point and the next on the scale. So, we may be able to have our cake and eat it, too.

QUESTION: I fully appreciate that, but in many of the presentations made here, you actually did try to match the bubble grades to your models. I am just making the comment that in some cases, I think you may have used [bubble grade] in a way that did not take account of the extreme nonlinearity of this scale.

L. HOMER: The general point is about trying to use ordinal values, if you have them., rather than treating them as categories. It is still an important piece of advice, whether one were to conclude that it did or did not apply to bubble grades.

R. VANN: I have a question concerning repeated measures. I have a data set now that has Doppler bubble measurements during open water dives, and a recorded dive profile. Some of these subjects were repeated. Now, I understand that there is a new technique that will let you handle that somehow.

Would it be safe to say that an approach to determining the importance of considering these repeated measures would be to do it both ways? Consider all subjects to be the same, and then consider them to be different. Then see how significant that was in how the models fit the data; according to whether repeated measures was used, or all were assumed the same?

F. HARRELL: Yes. I think the approach is the so-called sandwich variance estimator. The basic idea is that you fit the data, ignoring the fact that some subjects are repeated. You get the ordinary parameter estimates which are valid unless the within-subject correlations are fairly strong. But then, the variances are not valid. You use this new variance estimator, or you can use the bootstrap to get valid estimates of the variance. Then if the variance increases substantially you have evidence of intra-cluster correlation so the cluster correction is needed.

L. HOMER: In that case, though, even the bootstrap is not a complete protection. The problem is that your variance is small because you really do not have that many subjects. It is very hard to cure that, but I am not so sure that the [within subject] correlation is all that good.

We have tried in some instances to decide whether there were divers who were prone to this or that. It is not an easy thing to demonstrate. I think you would say physiologically that it is probably true based on some studies, but it is not a strong effect.

QUESTION: What is your recommendation then when you analyze these data? See if there is a difference?

L. HOMER: No harm in that.

E. PARKER: Maybe, unless you pick the way that proves your point.

COMMENT: That's the principle of maximum satisfaction.

W. GERTH: I am not going to resolve the following issue here, but I want to make a comment. There is a real difference between identifying a factor as important, as statistically significant in improving a model's fit to data, and ascribing to that significance any sort of important or significant insight into a mechanism. We, in this field, need to come to grips with those two things. I will ask this question of the bunch: when is it that we can claim to have learned something about mechanism when we show that a factor required in a mechanism is statistically important? How much can we learn about mechanism by model fitting?

F. HARRELL: I think Erich's paper is a good summary of that. He may not have the complete handle on exactly what oxygen is doing, but it is not innocuous. Oxygen is an important factor.

W. GERTH: What is the mechanism?

F. HARRELL: Well, you are not going to unravel the mechanism until you are first convinced that it is a factor.

W. GERTH: So, we have established the importance of this "thing" as a factor. Now what does that tell us about the mechanism that we posited to manifest that factor in the model? We will just have to leave that open, but I do think that is yet an unanswered issue in the approach that we take here, and what the kind of importance we ascribe to our results.

H. VAN LIEW (Panama City): It does not prove it, but you start thinking along the line that the data leads you, and you look for other evidence of it.

L HOMER: I think if you believe in a factor, put it into a logistic model, and it comes out to be important, that does not say why it is important. It just says it is. I think what Wayne is saying is just because it is important, how can you determine why? My sense from much of what was said today is that you really cannot tell very much. For example, there are many ways to include an "ascent rate" in a model besides just plopping it in. In principle, you could take a completely different model where your "ascent rate" is factored into bubble size, or something else, to get as good a fit. You are faced with determining whether this simple logistic model over here is as good or better than this more complicated bubble model, even though they both work as well. Our take has been that you need to demonstrate your mechanisms independently of the fitted data.

In other words, do experiments to show that in fact the "ascent rate" does impact bubble size, which definitely does cause bends. Just by simply fitting it to a data set where all you know is the outcome, all you learn is that your mechanism is consistent with the data. Such work does not tell you that your particular mechanism for including a factor is any better than the next guy's. It all really comes out just in the data fit.

F. HARRELL: The only thing I could think to add to those nice statements is that you can postulate an unmeasured variable that might explain away the effect of, say, ascent rate. Then you can do a sensitivity analysis to find out how likely it is that some other variable could explain the effect that you are attributing to the one variable that is currently in your model. In terms of causal inference, I think we will start to see this sort of sensitivity to unmeasured variables being a part of our arsenal.

V FLOOK: I think the answer really is much more fun, because having identified something that's important to you, you then put up three or four models of how that might be acting, and spend the rest of your life playing with them.

W. GERTH: Amen.

Closing Remarks

Wayne A. Gerth

The task has fallen on me to wrap this up. I will not endeavor here to summarize everything that has been covered today or develop any sort of conclusions. Instead we promise to provide a collection of today's presentations in a Proceedings to be published by the Undersea and Hyperbaric Medical Society.

In 1984, Drs. Weathersby, Homer and Flynn published a paper that set in motion a sea change in the way that we have come to reconcile theory with data in environmental physiology. Their formalization of a way to make one conform to the other was a seminal contribution in the area of decompression studies. The approach they outlined has since blossomed into a variety of different papers with applications of the technique to problems beyond decompression sickness. The value of their contribution is represented by the interest that motivated everybody's participation here today.

One of the principal purposes of this Workshop has been to provide a snapshot of the state-of-the-art of this work. We will hope that the Proceedings will provide a point of reference from which future work can be launched, and from which people can look to the past to find papers relevant to their particular interest.

On behalf of us all sitting in front you today, let me thank you very much for joining us. Without any further ado, we'll close.

Thank you.

(Applause)
(Whereupon, the meeting was concluded.)

List of Participants

John Clarke
Navy Experimental Diving Unit
321 Bullfinch Road
Panama City, FL 32408

Johnny Conkin
USRA
NASA, Johnson Space Center
Mail Stop SD3
Houston, Texas 77058

Wayne A. Gerth
Duke University Medical Center
Box 3823
Durham, NC 27710

Frank E. Harrell
Medicine-Health Evaluation Sciences
University of Virginia School fof Medicine
Health Sciences Center Box 600
Charlottesville, Virginia 22908

Louis D. Homer
Legacy Research
Mount Hood Medical Center
2801 N. Gantenbein Ave.
Portland, Oregon 97227

Nandini Kannan
Division of Mathematics and Statistics
University of Texas, San Antonio
6900 North Loop 1604 West
San Antonio, Texas 78249

Erich C. Parker
Naval Medical Research Institute
8901 Wisconsin Avenue
Bethesda, MD 20889

Edward D. Thalmann
Duke University Medical Center
Box 3823
Durham, NC 27710

Peter Tikuisis
Defence and Civil Institute of Environmental Medicine
1133 Sheppard Ave. West
P.O. Box 2000
North York, Ontario M3M 3B9
Canada

Hugh D. Van Liew
Navy Experimental Diving Unit
321 Bullfinch Road
Panama City, FL 32408

Paul K. Weathersby
5 Ferry View Drive
Gales Ferry, CT 06355