Item ID Number	01710
Author	Carroll, Ray J.
Corporate Author	
Report/Article Title	Typescript: Report #4, September 1982
Journal/Book Title	
Year	0000
Month/Day	
Color	☐
Number of Images	24
Description Notes	Report consists of the following sections: The Control Population; Stratification of Battalions; Missing Battalion Exposure Data; An Exposure Index; Sampling Strategies

Report #4

R.J. Carroll, Ph.D.

September 1982

This report consists of the following sections:

The Control Population: A general discussion of some of the issues to be kept in mind when deciding upon the control (non-exposed) population;

Stratification of Battalions: To reduce the cluster effects investigated in Report #3, I suggest that battalions first be stratified into similar groups and then sampled.

Missing Battalion Exposure Data: If many battalion records are not usable for determining exposures, we have to consider some alternatives and their potential effects.

* An Exposure Index: I conclude that a priori construction of a single, one dimensional exposure index is probably not feasible and is not even necessary. Multidimensinal indices will be more appropriate. *

Sampling Strategies: I propose a framework for using multidimensional exposure indices to design the sample. Alternatives are considered and analyzed.

Because there are many unknowns relating to the feasibility of conducting an Agent Orange study, and because my function is an advisory one, this report cannot be considered a protocol. I believe this report raises many issues which should be resolved either prior to or by means of the anticipated pilot study. Of course, I continue to be very willing to work with and advise the VA, outside epidemiologists and the survey firm picked to develop the final protocol and sample.

The Control Population

It is crucial that any study of the health consequences of exposure to Agent Orange include an adequate control population of non-exposed troops. The control or non-exposed population must be so similar to the exposed population that we can be sure that differences in health status are due to exposure to Agent Orange and not due to differences in the make-up of the exposed and non-exposed populations. Conversely, the non-exposed population must be such that if no detrimental health consequences are discovered due to exposure, we must be sure that this result is not due to having the non-exposed group less healthy than anticipated.

For example, consider the situation of false negatives, i.e., some soldiers reported as not exposed actually had heavy exposure. My reports #1 and #2 (which were based on the assumption that those listed as heavily exposed actually were so exposed) make clear that if the false negative rate is high, any detrimental health effects of Agent Orange will tend to be diluted. We could thus conclude that Agent Orange exposure is not harmful to the ground troops when in fact it is.

On the other hand, suppose the non-exposed control group would be expected to be in better health than the exposed group, perhaps because they had higher socio-economic status or were much less involved in the field. In this instance, we will tend to exaggerate the effects of Agent Orange exposure. We could thus conclude that Agent Orange exposure is harmful to the ground troops when in fact it is not.

What are the characteristics of an adequate control group? First, it

should ideally consist of combat troops in Vietnam who were extremely unlikely to have been exposed to any potentially harmful chemicals. Secondly, the control groups should be selected to be as similar as possible to the exposed group on a set of variables (confounders) relating to health status, e.g., previous disease, MOS, socio-economic status, etc.

What happens if such a control group cannot be found? Of course, one can try substitution of another group, e.g., combat troops who were stationed in Korea at the same time or combat troops whose unit was readied for deployment in Vietnam but sent elsewhere. Whether such substitution will suffice is not clear to me.

One group I have not mentioned is non-combat troops who were stationed in Vietnam. For example, non-combat troops might take the place of the non-exposed Vietnam combat troops if the latter were not found feasible to identify. There are obvious confounding problems in that the non-combat troops are likely to have higher socio-economic status, although appropriate stratification can help. Further, a comparison between exposed combat troops and non-exposed non-combat troops mixes up effects: is it Agent Orange or merely combat itself (or both) that is to blame?

Even if an adequate control group of non-exposed combat troops is found to be available, it has been suggested that a third group also be established and surveyed: non-exposed non-combat troops in Vietnam. I do not see such a strategy as particularly appropriate if the goal of the VA study is to study the effects of Agent Orange. It may be nice to have three groups and try to simultaneously study the combat experience, but I worry that such a strategy

will lower the statistical power of what I have been led to believe is the most pressing question: effects of Agent Orange. A three group study (exposed combat, non-exposed combat, non-exposed non-combat, all in Vietnam) does not seem to me to be appropriate, given limited resources.

Stratification of Battalions

One eay to reduce the effects of sampling battalions (Report #3, the two-stage cluster sample effect) is by means of adequate stratification. For example, suppose that we have identified all those battalions in a group named "Likely Heavily Exposed." I am not necessarily advocating at this time that such identification be done, but for the moment suppose this is the line of attack decided upon. Having made this identification, it is probably foolhardy merely to randomly sample a fixed number of "Likely Heavily Exposed" battalions, because these battalions may still be very heterogeneous. What needs to be done is to further classify the "Likely Heavily Exposed" battalions into smaller but more homogeneous subgroups that may be important for types of exposure, areas of operation, degree of actual combat, etc. I am obviously in no position to designate these subgroups or strata, as such stratification ought to be done by someone more knowledgeable than I about the Vietnam era.

Having formed these strata of battalions, we might then randomly select a few battalions from each strata. Such a scheme will tend to be more efficient than the alternative of no stratification, and this improvement in efficiency can be very large.

We may also decide to sample from a group of battalions designated "Unlikely Exposed." Stratification can also be done here. It would appear to be sensible but may not be possible to make the strata identical for the "Likely Heavily Exposed" and "Unlikely Exposed" groups.

Further stratification of individuals within battalions will also be desirable to reduce the effects of confounders.

A point that should also be addressed in the pilot study is this: even in those battalions with adequate exposure data, are the data sufficient for every soldier? Individual exposures may be missing (company records lost, for example); if so we have obvious important difficulties, which are more in the framework of classical sampling theory.

An Exposure Index

There are really two uses for exposure indices, and I think it is vital to keep the distinction in mind. The first use is for choosing the sample. The second use is in analyzing the sample; in this case there is the flexibility to produce a number of indices and try to relate them to health status (while keeping in mind, of course, the multiple testing problem). This second use of exposure indices will not really concern me at this time.

The key question then is the desirability of developing a single exposure index to be used as a vital component of the sampling design. To answer this question one has to return to the purpose of the study. Stating the study's purpose is not easy for me to do, but suppose it can be reduced to

"Do combat troops in Vietnam who had a heavy exposure to Agent Orange now have poorer health than those troops not exposed to Agent Orange?"

If this reasonably captures the purpose of the VA study, then we must ask if a single exposure index will adequately distinguish between "heavily exposed" and "non-exposed" troops. In particular, this index must be agreed upon by all the major interested parties before embarking on the sampling, or else seven years hence we might read "VA's Agent Orange Study: were the 'heavily exposed' really exposed at all?".

In my discussions with Mr. Levois, I have boome concerned that there simply is not enough good information available to construct a single

exposure index on which to base the VA study. I think alternatives should be explored and their consequences studied (see the section of this report "Sampling Strategies"). It might be feasible to convene a panel of physicians and other experts from outside the VA with a charge to develop a concensus exposure index which is both medically and politically sound.

A different conceptual framework for using exposure indices in designing the sample should prove more fruitful. The basic idea is to construct a multidimensional index which measures various facets of exposure, and then base the sample on this multidimensional index. I know of one example of a study for which I served as a consultant and which used a multidimensional approach quite successfully. The SENIC study of the Centers for Disease Control (Dr. Robert Haley, Principal Investigator) was designed to see if programs for surveillance and control of nosocomial (hospital acquired) infections were at all successful. A panel of experts was convened and helped develop a two-dimensional index measuring the two aspects surveillance and control, and this index was used to choose the sample by stratification. SENIC was also politically sensitive as well as difficult scientifically, and it might serve as a potential guide for the VA (see the American Journal of Epidemiology, May 1980).

To give some idea of how a miltidimensional index might be used in the VA study, suppose that "exposure" consists of two conceptual facets, "Aborted Mission Exposures" and "Usual Mission Exposures." A panel of experts is convened and develops an "Aborted Mission Exposure" index (AME index) and a "Usual Mission Exposure" index (UME index). This panel also designates levels of the AME and UME indices which are called Low, Medium and High, forming a

matrix such as in Figure #1. Depending on the goals of the study, the VA could then sample from the cells of Figure #1. For example, all sampling could be done from the upper left and lower right corners (Low-Low versus High-High); such sampling would be sensitive to detecting the effects of Agent Orange.

It is my belief that a priori construction of a multidimensional exposure index such as outlined above is more feasible than constructing a single index and may be very useful in designing the study.

N.B. I will later deal with the distinction between low-medium-high and a "dose response" type relationship using indices.

Sampling Strategies

In this section I will propose and study a framework for a sampling plan which is based upon the idea of a two-dimensional exposure index. To keep the framework simple, I will pretend that simple random sampling is possible, thus ignoring for the moment the more complex issues of stratification, battalion clustering, misclassification and varying exposure levels of individuals within battalions (the first two of which I have previously discussed, although not in this context). After this framework has been thoroughly studied, I think we will be in a position to confront more of the complex issues. In particular, Report #5 will focus on misclassification as it relates to sampling strategies. For the moment, I want to try to make the basic framework clear and find out if it meets the needs of the VA.

As in the section on exposure indices, I will assume that it is possible to roughly cenceptualize exposure as either "Aborted Mission Exposure" (AME) or "Usual Mission Exposure" (UME); this is for me just a working hypothesis and should really be explicated by those more knowledgeable than I about Agent Orange. Having conceptualized exposure in this way, I see a panel of experts (including some from outside the VA) as developing indices ranging from 0-100 which measure AME and UME. I then see this panel as developing groups based on the exposure indices, say Low-Medium-High for each. Thus, I am envisioning that each soldier can be categorized into Low-Medium-High on both the AME and UME indices. This leads to something like Figure #1. (N.B.: I do not know if such a construction can actually be done).

To a fairly large extent, this is the basic scheme used in the SENIC study mentioned in the section on exposure indices. Already, however, many questions

arise:

Q1: IS IT NECESSARY TO CATEGORIZE THE EXPOSURE INDICES? Not really, although such categorization is convenient and fairly standard.

Q2: HOW ARE THE CATEGORIES TO BE CHOSEN FOR PICKING THE SAMPLE? Basically, one would hope on medical and not statistical grounds. I do not think it would be useful to define LOW AME as the 33rd percentile of the AME index. Rather, LOW AME should be medically meaningful.

Q3: IS IS NECESSARY TO HAVE EXACTLY THREE GROUPS LOW-MEDIUM-HIGH FOR EACH INDEX? No. In fact, one might well want more. SENIC used four, but I use three to make subsequent calculations more transparent.

Q4: CAN WE EVER USE THE AME AND UME INDICES THEMSELVES AND NOT JUST THE CATEGORIES? Yes, especially in the analysis. One might well want to develop a "dose-response" relationship based on the indices.

Q5: CAN WE USE MORE THAN TWO INDICES? Yes, but I think much more than two indices would become unwieldy.

Let us now suppose that the framework of Figure #1 has been accepted (actually, it will inevitably be modified and include various stratifications). How should the sample be picked? To a major extent, this depends on the purpose of the study. What I will now do is consider a few sampling strategies, and then discuss the purpose for which they are ideal, as well as their drawbacks.

Strategy #1 (Sample only from the LOW AME-LOW UME and HIGH AME-HIGH UME cells of Figure #2).

This strategy is ideal for the purpose of determining whether or not high Agent Orange exposures in ground combat troops are harmful to future health,

when compared to an appropriate control group. This strategy will have the highest statistical power for such a comparison and will probably result in the lowest number of misclassifications (see Report #5 to follow).

While Strategy #1 maximizes the statistical power of a comparison between no and heavy exposure, it is not a very good method for estimating a "dose-response" relationship between exposure indices and health status. In addition, Strategy #1 will tell us nothing about which of AME or UME is the most harmful, if indeed the High-High exposure group has worse health than the Low-Low exposure group. These may not be much of a drawback, but the principal investigators of the study ought to be the ones making the substantive decision as to the purpose of the study.

Strategy #1 is the method of choice for the specific question of whether high Agent Orange exposure is harmful relative to low exposure. It is not an acceptable method if one instead wants to know how risk depends on exposure indices, especially for the middle range of exposure.

Strategy #2 (Sample from all the cells of Figure #2, either equally or proportional to size).

This strategy has a more general purpose from Strategy #1, being oriented to estimating the relationship between health status and exposure indices, especially for differing amounts of exposure. Further, it will help identify whether an AME is more harmful than a UME. My understanding of the need to compare Low-Low versus High-High exposures and do the best study possible of the effect of Agent Orange suggests that Strategy #2 will not be appropriate.

Strategy #3 (Sample from all cells, but over-sample the LOW-LOW and HIGH-HIGH cells).

This is a compromise between the two earlier strategies. One way to think of this strategy is to consider it as a three group study, the groups being Low-Low , High-High and others. Such a view is probably misleading. I prefer to think of Strategy #3 as a way of comparing Low-Low and High-High exposures while at the same time enabling us to get some sort of "dose-response" relationship between exposure indices and health status and some understanding of the relative importance of AME and UME. There being no free lunches, we get both high versus low exposure comparisons and dose-response simultaneously with lower efficiency and higher misclassification errors; the key question is how much efficiency do we lose?

It is my understanding that the high versus low exposure comparison is the most important one, so I will take the view of asking how much can we sample from outside the Low-Low and High-High cells of Figure #1 (to get a "dose-response" relationship) before losing a significant amount of statistical power for comparing High-High and Low-Low cells. In a later report I will address the problem of additional misclassifications caused by using Strategy #3.

Suppose that the VA can afford to obtain the health status of a total of N_* individuals. We observe N_{LL} from the Low-Low exposure cell, N_{HH} from the High-High exposure cell, leaving us with $N_* - N_{LL} - N_{HH} = N_0$ to be chosen from the other seven cells. A reasonable strategy for this example but one which may have to be modified is to apportion the $N_* - N_{LL} - N_{HH} = N_0$ observations equally in the remaining seven cells. This is illustrated in Figure #2.

An example will help to illustrate the potential problems with sampling from cells other than the LOW-LOW cell and the HIGH-HIGH cell. Suppose that

$$p = \Pr(\text{disease in the LOW-LOW cell}) = .005 ,$$

and the relative risks for disease in all nine cells are as given in Figure #3. Note that I am trying to detect a doubling in risk when the soldier has HIGH UME-HIGH AME ($r_{HH} = 2.0$), so that in this example the disease rate for the HIGH-HIGH cell is $pr_{HH} = .01$. The other relative risks I have chosen arbitrarily but conservatively. For example, the relative risk for HIGH-HIGH-MEDIUM AME is taken in this example as only 1.4, meaning that the dinease rate for this cell is .007 and, in 1,000 soldiers, we only expect the HIGH-MEDIUM cell to have two more incidents of disease than the LOW-LOW cell (versus five more in comparing HIGH-HIGH and LOW-LOW). Further, suppose we agree to follow the UCLA protocol and insist a priori that an acceptable probability for concluding that Agent Orange exposure is harmful when it really is not harmful is $\alpha = .01$. Finally, suppose we are able to observe the health status of $N_* = 12,000$ individuals. Thus, if we follow Strategy #1 and sample $N_{LL} = 6,000$ from the Low-Low cell and $L_{HH} = 6,000$ from the High-High cell, the statistical power for detecting the hypothesized doubling of relative risk is 81% (see Report #2). On the other hand, suppose we decide $N_{LL} = 3,900$, $N_{HH} = 3,900$ and $N_0 = 4,200$, so we take 600 observations in each of the remaining cells. Then, if we simply compare the disease rate in the Low-Low cell to that of the High-High cell, the statistical power drops to 61%.

I will call Analysis #1 the simple comparison of the disease rates in the Low-Low and High-High cells. A more complex Analysis #2 compares the disease rates of the combined Low-Low, Low-Medium, Medium-Low cells to those in the High-High, Medium-High, High Medium cells; the Appendix gives the technical

details. For Analysis #1, we see that Strategy #1 had power 81% while Strategy #3 with 600 observations in each of the seven outside cells had power 61%; the corresponding results for Analysis #2 are 81% and 58%. This drop in power of 20% (Analysis #1) or 23% (Analysis #2) is a serious one and illustrates the following basic fact:

Even when there are no misclassification errors, there are choices of p and the table of relative risks for which too extensive sampling from outside the Low-Low and High-High cells causes a serious loss of statistical power in the main comparison between very low and very high exposures.

In the previous example, we had p = Pr{disease in Low-Low cell} = .005. If we next try p = .01 but keep the same table of relative risks (see Figure #4) then we get

Table #1

	Analysis #	Sampling Strategy #	Power
(p=.005)	1	1	81%
	1	3	61%
	2	1	81%
	2	3	58%
(p=.01)			
	1	1	99%
	1	3	91%
	2	1	99%
	2	3	89%

The drop from 99% statistical power to 89% - 91% power caused by sampling outside the Low-Low and High-High cells when p = .01 is still somewhat discouraging, although much less severe than that encountered when p = .005. This illustrates a second basic fact:

For the same table of relative risks, the loss of statistical power due to sampling outside the Low-Low and High-High cells becomes greater as the disease probability in the Low-Low cell becomes smaller.

We next will vary the particular version of Sampling Strategy #3. In the previous version, we took a total of N_0 = 4,200 observations outside the Low-Low and High-High cells. In this next strategy, we will take half as much (N_0 = ~~4,200~~ 2,100), thus allocating 300 rather than 600 observations to each of the other seven cells. Note that this second Sampling Strategy #3 will be much less informative about dose-response. The power results are as follows:

Table #2

(Second) Sampling

Analysis #	Strategy #	Power
(p=.005)		
1	1	81%
1	3	73%
2	1	81%
2	3	71%
(p=.01)		
1	1	99%
1	3	97%
2	1	99%
2	3	96%

These losses of power are much less dramatic, illustrating the following fact:

For small disease probabilities in the Low-Low cell, sampling outside the Low-Low and High-High cells can be done without much loss of statistical power if the other seven cells are lightly sampled.

From the preceeding analysis as well as many others I have done, I can make some fairly definite conclusions even if I ignore the misclassification problem. Using a stringent Type I error rate $\alpha = .01$ as in the UCLA protocol and using a conservative table of relative risks, it appears that for the rarer diseases ($p \leq .005$), the ability to detect a doubling of risk going from the Low-Low cell to the High-High cell can be compromised if the middle range of risk is oversampled. Thus, if rarer diseases are of major interest (such as $p = .005$), only relatively few observations (less than 20%), if any, should be taken from the middle range of risk. For more common diseases ($p \geq .02$) or if the goal is to discover a tripling of relative risk ($r_{HH} = 3.0$, not 2.0 as heretofore), then some sampling from the middle range of exposure will entail considerably less potential loss.

For diseases which are not so rare (e.g., $p = .01$), no definite conclusion can be made at this point, because I have not yet illustrated the effect of misclassification. It is useful to repeat that only sampling the Low-Low and High-High cells gives the highest statistical power and the lowest misclassification rate.

Additional Remarks

There are still many questions that need to be answered. Among these are the following:

(1) What are the available control groups of non-exposed troops?

(2) Will battalions of generally high exposure have troops who were definitely not exposed, or are these false negatives?

(3) What variables can be used for battalion and individual stratification?

(4) What is the extent of missing battalion exposure data?

(5) What features of exposure need to be considered in construction of a multidimensional exposure index?

(6) Can a multidimensional exposure index be constructed?

(7) For what alternatives (p and r_{HH} of the section on Sampling Strategies) should we be designing the study?

(8) What are the difficulties with follow-up to look at health status?

(9) How will the power calculations change when the effects of misclassification and battalion cluster sampling are also considered?

(10) Can we estimate the misclassification error rates?

Technical Appendix

The choice $\alpha = .01$ made by UCLA is a conservative one ($\alpha = .05$ is more usual) but still a good idea in setting sample sizes and discussing effects of different analysis and design strategies. In this appendix, I am assuming that all disease probabilities are fairly low and that, unrealistically, there are no misclassifications; the latter case will be dealt with in a future report.

For Analysis #1 of the section on sampling strategies, statistical power is computed as in Report #2, i.e., by treating $2 \text{ Arcsin } \sqrt{\hat{p}_{LL}}$ as normally distributed with mean $2 \text{ Arcsin } \sqrt{p_{LL}}$ and variance $1/N_{LL}$.

For Analysis #2, I am going to compare the weighted disease rate

$$w_{LL}p_{LL} + w_{LM}p_{LM} + w_{ML}p_{ML} = p(L)$$

against the disease rate

$$w_{HH}p_{HH} + w_{HM}p_{HM} + w_{MH}p_{MH} = p(H),$$

where

$$N(L) = N_{LL} + n_{LM} + N_{ML}$$

$$N(H) = N_{HH} + N_{HM} + N_{MH}$$

$$w_{LL} = N_{LL}/N(L) \ , \ w_{LM} = N_{LM}/N(L)$$

$$w_{ML} = N_{ML}/N(L) \ , \ w_{HH} = N_{HH}/N(H)$$

$$w_{HM} = N_{HM}/N(H) \ , \ w_{MH} = N_{MH}/N(H) \ .$$

Assuming that the true disease probabilities are all relatively small, it turns out that to a first approximation, which is sufficient for my purposes of illustrating effects on power of different sampling strategies, $2 \text{ Arcsin } \sqrt{\widehat{p(L)}}$ is normally distributed with mean $2 \text{ Arcsin } \sqrt{p(L)}$ and variance $1/N(L)$; a similar result holds for $2 \text{ Arcsin } \sqrt{\widehat{p(H)}}$.

This means that the approximate statistical power for Analysis #2 is

$$1 - \Phi(z_{\alpha} - 2\sqrt{1/N(L) + 1/N(H)}(\text{Arcsin}\sqrt{p(H)} - \text{Arcsin}\sqrt{p(H)}).$$

A more complex analysis based on additive or log-linear additive tables could also have been considered, but I do not believe the necessary additivity should be assumed when making these important power calculations.

If we assumed a less conservative configuration of relative risks by changing Figure 4 to the following (reading across rows) 1.0, 1.1, 1.25, 1.1, 1.25, 1.5, 1.75, 1.25, 1.75, 2.0, then the power figures in Table #1 would become 81%, 61%, 81%, 66%, 99%, 91%, 99%, 94%, while those in Table #2 would become 81%, 73%, 81%, 75%, 99%, 97%, 99%, 97%. If the interest was in detecting tripling of relative risk ($r_{HH} = 3.0$), all the analysis and sampling strategies gave power of over 99%.

Figure #1

CONCENSUS AME INDEX
(ABORTED MISSIONS)

Concensus UME Index (Usual Missions)	Low	Medium	High
Low			
Medium			
High	These troops had high UME exposure.		These troops have high exposure on both indices.

Figure #2

DISTRIBUTION OF OBSERVATIONS FOR A TOTAL SAMPLE OF SIZE N_*

AML CATEGORICAL INDEX

UME Categorized Index	Low	Medium	High
Low	N_{LL}	$N_0/7$	$N_0/7$
Medium	$N_0/7$	$N_0/7$	$N_0/7$
High	$N_0/7$	$N_0/7$	N_{HH}

$$N_0 = N_* - N_{LL} - N_{HH}$$

For example, if we can take 12,000 observations total, we might take

$$N_{LL} = 3,900 \qquad N_{HH} = 3,900 \qquad N_0 = 4,200 \ .$$

Figure #3

"ABORTED MISSION EXPOSURE INDEX"

"Usual Mission Exposure Index"	Low	Medium	High
Low	p	r_{LM}	r_{LH}
Medium	r_{LM}	r_{MM}	r_{MH}
High	r_{LH}	r_{HM}	r_{HH}

r_{ij} = relative risk of disease (relative to the Low-Low cell) for soldiers in r i, column jg, i.e.,

$$r_{ij} = \frac{\Pr(\text{disease, row } i, \text{ column } j)}{\Pr(\text{disease, Low on both indices})}$$

$$p = \Pr(\text{disease, Low on both indices})$$

Figure #4

A SPECIFIC EXAMPLE OF RELATIVE RISKS

p = Pr{disease in the Low-Low cell}

α = Type I error (probability of finding Agent Orange a health hazard when it really is not).

r_{III} = relative risk for disease in the High-High cell

RELATIVE RISKS

	Low	Medium	High
Low	$r_{LL} = 1.0$	$r_{LM} = 1.1$	$r_{LH} = 1.2$
Medium	$r_{ML} = 1.1$	$r_{MM} = 1.3$	$r_{MH} = 1.4$
High	$r_{HL} = 1.2$	$r_{HM} = 1.4$	$r_{HH} = 2.0$

For an illustration, take $r_{HH} = 3$.