TECHNICAL REPORT
RERF TR 9-84
業績報告書

# LIFE TABLE ANALYSIS WITH SMALL NUMBERS OF CASES: AN EXAMPLE – MULTIPLE MYELOMA IN HIROSHIMA AND NAGASAKI

# 少数例を用いた生命表解析の1例：広島・長崎における多発性骨髄腫

DAVID G. HOEL, Ph.D.
ROBERT I. JENNRICH, Ph.D.

放影研
RERF

RADIATION EFFECTS RESEARCH FOUNDATION
財団法人　放射線影響研究所
A Cooperative Japan – United States Research Organization
日米共同研究機関

**RERF TECHNICAL REPORT SERIES**
**放影研業績報告書集**

The RERF Technical Reports provide the official bilingual statements required to meet the needs of Japanese and American staff members, consultants, and advisory groups. The Technical Report Series is not intended to supplant regular journal publication.

放影研業績報告書は，日米専門職員，顧問，諮問機関の要求に応えるための日英両語による公式報告記録である．業績報告書は通例の誌上発表論文に代わるものではない．

*The Radiation Effects Research Foundation (formerly ABCC) was established in April 1975 as a private nonprofit Japanese Foundation, supported equally by the Government of Japan through the Ministry of Health and Welfare, and the Government of the United States through the National Academy of Sciences under contract with the Department of Energy.*

放射線影響研究所（元ABCC）は，昭和50年4月1日に公益法人として発足したもので，その経費は日米両政府の平等分担により，日本は厚生省の補助金，米国はエネルギー省との契約に基づく米国学士院の補助金とをもって運営されている．

放影研 RERF RADIATION EFFECTS RESEARCH FOUNDATION
財団法人 放射線影響研究所

# LIFE TABLE ANALYSIS WITH SMALL NUMBERS OF CASES: AN EXAMPLE – MULTIPLE MYELOMA IN HIROSHIMA AND NAGASAKI

# 少数例を用いた生命表解析の1例：広島・長崎における多発性骨髄腫

DAVID G. HOEL, Ph.D.[1]; ROBERT I. JENNRICH, Ph.D.[2]

*RERF Director, on leave from National Institute of Environmental Health Sciences[1]; and National Institute of Environmental Health Sciences, now at UCLA[2]*

放影研理事，米国環境保健科学研究所から休暇中[1]；米国環境保健科学研究所，現在の所属：California 大学 Los Angeles 校[2]

## SUMMARY

Life table analysis techniques in epidemiology depend upon the asymptotic properties of the statistical test methods employed. In some instances, the statistical procedures indicate highly significant results which are, in reality, unjustified. This phenomenon may occur when the asymptotic methods are applied in situations where the cases of interest are few in number. This situation is illustrated by the 20 multiple myeloma deaths observed in the RERF Life Span Study cohort. A permutation test is applied to the life table data, although the test requires the false assumption that the censoring distribution is independent of the radiation dose. A simulation test is developed which does not require equal censoring, which has the same asymptotics as the usual test methods, and which is less likely to overestimate significance in small samples. It is found that both of these small-sample tests provide reasonable numerical solutions. In addition, the simulation test is recommended in general for analyzing life table data with unequal censoring. Finally, by using the small-sample tests, the frequency of death from multiple myeloma is shown to be positively associated with radiation dose ($P<0.01$).

## 要　約

疫学における生命表技法は使用する統計検定の漸近的性質によって異なる．ときには統計的手法が，実際には理に合わない結果を，有意性が非常に高いものとして示すこともある．このような現象は対象症例数の少ない場合に漸近法を用いたときに起こり得る．放影研の寿命調査対象者に見られた20例の多発性骨髄腫はこうした例である．置換検定(permutation test)を生命表資料に適用したが，それには censoring 分布が放射線量に依存しないという偽の仮説が必要である．等しい censoring が不要で，通常の検定法と同じ漸近性を有し，少数の対象例でも有意性の過大評価を起こしにくいシミュレーション検定を開発した．これらの少数対象例検定法はいずれも妥当な数値結果を示すことが分かった．更に，シミュレーション検定は一般に等しくない censoring をもつ生命表解析に用いるとよい．最後に，少数対象例検定を用いると，多発性骨髄腫の死亡率は放射線量と正の関係を示す($P < 0.01$)．

Approved 承認 16 May 1984

Printed 印刷 October 1984

## INTRODUCTION

In the course of life table analyses of epidemiologic data, one may be confronted with a small number of deaths from a relatively rare disease. The problem to be faced is then one of determining the reliability of the statistical methods of analysis used. The application of asymptotic statistical tests to small-sample situations comes into question. The unsuspecting investigator may not be aware of the potential difficulty in this particular situation. Sometimes it becomes apparent when a calculated statistical significance level runs contrary to the investigator's common sense. In this paper one such example will be examined, recommending a small-sample approach which, hopefully, will aid in reducing errors caused by the misapplication of asymptotic statistical methods to epidemiologic data.

## 緒　言

疫学データの生命表解析において，比較的まれな疾病による少数死亡例を扱う場合がある．その場合直面する問題は，使用した統計的分析法の信頼性を決定することである．漸近的統計検定を少数対象例に応用することが問題となるのである．疑い深くない調査者であれば，この特定の状況における潜在的難点に気付かないかもしれない．時折，計算した統計的有意レベルが調査者の常識に反することが明らかになる．本報告では，そのような1例を取り上げ，漸近的統計方法の疫学データへの不適切な応用によって起こる誤差を少なくすると期待される少数対象例検定法を検討する．

## MATERIALS AND METHODS

Since 1950, ABCC/RERF has followed prospectively over 100,000 atomic bomb survivors of Hiroshima and Nagasaki.[1] This cohort has provided some of the best dose-response data for the effects of ionizing radiation on humans. The findings with regard to cancer have been particularly noteworthy. As the cohort ages and the person-years at risk increase, the precision of the risk estimates improves and new endpoints often appear. One example is the recent appearance of multiple myeloma. Currently, there seems to be a sufficient number of these cases to analyze vis-à-vis a possible association with radiation dose.

Traditionally, the RERF Life Span Study (LSS) cohort has been divided into 20 subcohorts in order to study the possible association of any given disease with radiation dose. This subdivision can be visualized as a four by five matrix where four city/sex and five age-at-the-time-of-the-bomb (ATB) categories are used. Within each of these 20 subcohorts, a life table analysis can be carried out and a trend test performed in order to assess any possible radiation dose-response relationships. The subcohorts can then be combined statistically in order to give an overall statistical assessment of a possible radiation dose-response relationship.

## 材料及び方法

1950年以来，ABCC/放影研は，広島・長崎の原爆被爆者10万人以上の前向き調査を実施してきた．[1] この対象集団により，電離放射線のヒトに対する影響に関する最良の線量反応データが提供されてきた．癌に関する所見は特に注目に値する．対象集団の年齢及びリスク人年が増加するにつれ，リスクの推定が正確になり，新しい結果が度々現れる．最近の多発性骨髄腫の出現はその一例である．現在，これらの症例は，放射線量との関連について解析するのに十分な数に達しているようである．

特定の疾病と放射線量の関連を研究するために，以前から放影研寿命調査対象集団は20の副対象集団に分割されている．この分割は4群の都市/性及び5群の原爆時年齢を使用する4×5の行列とみなすことができる．何らかの放射線量反応関係がないかを評価するために，これら20の副集団各々について，生命表解析及び傾向検定を実施することが可能である．その後，副集団を統計的に結合し，ある可能な放射線量反応関係の全体的な解析を統計的に実施することができる．

## RESULTS AND DISCUSSION

### Overestimation by Asymptotic Test

In Table 1, the number of deaths from multiple myeloma in each dose group within each

## 結果及び考察

### 漸近的検定による過大評価

表1は，各副集団内の各線量群における多発性骨髄腫による死亡数を示す．線量の傾向を表す漸近的

subcohort is presented. The asymptotic log-rank $\chi_1^2$ value (1 df) for trend in dose is given.[2] The overall $\chi_1^2$ of 11.5 for trend in dose appears to be highly significant, as do the values for some of the individual subcohorts. Of particular interest is the Hiroshima female aged 20-34 ATB subcohort, which gives a $\chi_1^2$ value of 26.5. Although the group contained 8,887 subjects, only two multiple myelomas were recorded, making the significance level of the trend test value suspect. In all probability, what is being observed is a failure of large-sample theory in a small-sample setting, resulting in an overvaluation of the significance of the observed results. The trend test used was based on the log-rank test of Mantel[3] and Cox.[4] A trend test based on the modified Wilcoxon test of Gehan[5] and Breslow[6] gave similar results, as did the log-rank and modified Wilcoxon homogeneity tests for equality in all eight dose groups.

対数階数$\chi_1^2$値（自由度1）を示す.[2] 幾つかの副集団の値と同様に，全体の線量の傾向を表す$\chi_1^2$値11.5は非常に有意と考えられる．特に興味深いのは，$\chi_1^2$値26.5を示す広島・女性・原爆時年齢20～34歳の群である．その群には8,887人の対象者が含まれているにもかかわらず，多発性骨髄腫は2例しか記録されておらず，傾向検定値の有意レベルを疑わしいものにしている．多分，ここで見られる現象は，対象例が少数である場合に多数対象例の論理を適用すると，観察結果の有意性を過大評価する結果となることから生ずるのであろう．使用した傾向検定はMantel[3]及びCox[4]の対数階数検定に基づく．Gehan[5]及びBreslow[6]の修正Wilcoxon検定に基づく傾向テストも，8線量群すべてにおける均一性に関する対数階数及び修正Wilcoxon等質性検定と同じく，同様の結果を示した．

TABLE 1 MULTIPLE MYELOMA DEATHS IN THE LSS COHORT, 1950-78
表1 寿命調査対象集団中の多発性骨髄腫による死亡，1950－78年

| Age ATB in Years | Hiroshima | | Nagasaki | | Pooled $\chi_1^2$ |
|---|---|---|---|---|---|
| | Male | Female | Male | Female | |
| 0-9 | - | - | - | - | - |
| 10-19 | 0.10*<br>group 0-1** | - | - | - | 0.10 |
| 20-34 | - | 26.5<br>group 2-1<br>group 7-1 | 3.05<br>group 5-1 | - | 22.8 |
| 35-49 | 0.17<br>group 1-2 | 1.98<br>group 0-2<br>group 1-1<br>group 5-1 | 6.53<br>group 0-1<br>group 3-1<br>group 7-1 | 0.07<br>group 0-1<br>group 1-1<br>group 3-1 | 3.7 |
| 50+ | - | 0.26<br>group 0-1<br>group 1-2 | - | 0.02<br>group 2-1 | 0.22 |
| Pooled $\chi_1^2$ | 0.26 | 13.2 | 9.6 | 0.09 | |
| | 7.2 | | 4.8 | | 11.5 |

*Value of $\chi_1^2$ for log-rank trend test (1 df).
対数階数傾向検定の$\chi_1^2$値（自由度1）

**One multiple myeloma in exposure group 0. There are eight exposure groups 0: 0 rad, 1: 1-9 rad, 2: 10-49 rad, 3: 50-99 rad, 4: 100-199 rad, 5: 200-299 rad, 6: 300-399 rad, 7: 400+ rad. The doses used for the trend tests were 0, 3.7, 21.8, 70.4, 141.2, 242.2, 343.7, and 524.7 rad.
0被曝群で多発性骨髄腫1例．8群の被曝群，0：0 rad，1：1～9 rad，2：10～49 rad，3：50～99 rad，4：100～199 rad，5：200～299 rad，6：300～399 rad，7：400＋ rad がある．傾向検定に使用した線量は0，3.7，21.8，70.4，141.2，242.2，343.7，及び524.7 rad である．

The data in Table 2, which led to the $\chi_1^2$ value of 26.5, suggest where the problem might lie. The second multiple myeloma is observed in the last dose group, which contains only a small fraction (less than 1%) of the subjects at risk at the time of the tumor. To understand how such a large $\chi^2$ value might arise, consider only the second tumor, and only the first and last dose groups. The numbers at risk in these groups are 3,904 and 92, respectively, with a tumor observed in the last dose group. In this simplified situation there is no difference between trend and homogeneity tests, and none between the log-rank and modified Wilcoxon tests. They are all computed by the formula

$\chi_1^2$値26.5を導いた表2のデータは問題がどこに存在するかを示唆する．第2番目の多発性骨髄腫は最終線量群に観察されるが，その群には腫瘍時の観察対象者のわずか一部分(1%以下)しか含まれていない．このように大きな$\chi^2$値がどのようにして現れたかを理解するために，第2番目の腫瘍のみ，並びに最初の線量群及び最終線量群のみを考慮する．これらの線量群の観察対象例数は各々3,904及び92であり，腫瘍は最終線量群に観察された．この単純化した状況においては，傾向検定と等質性の検定の間，及び対数階数と修正 Wilcoxon 検定の間に差異はない．それらはすべて

$$\chi_1^2 = (s-p)^2/pq \qquad (1)$$

where p = 92/3904 = 0.02 is the proportion of subjects at risk in the last dose group, s=1 is the number of tumors observed in that group, and q=1−p. For this simplified problem, $\chi_1^2 = 41.4$, a value even greater than the $\chi_1^2 = 26.5$ value cited earlier.

という公式によって計算される．ただし p=92/3,904 =0.02は最終線量群における観察対象例の割合，s=1はその群で観察された腫瘍例数であり，q=1−pである．この単純化した問題においては，$\chi_1^2 = 41.4$であり，前に引用した$\chi_1^2 = 26.5$という値より更に高くなる．

TABLE 2 SUBJECTS AT RISK AT TIME OF MULTIPLE MYELOMA DEATH IN THE LSS SUBCOHORT, HIROSHIMA FEMALES AGE 20-34 ATB, BY DOSE GROUP
表2 寿命調査副集団中の多発性骨髄腫による死亡時における観察対象者：線量群別，原爆時年齢20〜34歳の女性，広島

| Dose Group | | | | | | | | Tumor Group* |
|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
| 3972 | 2303 | 1636 | 475 | 250 | 97 | 59 | 95 | 2 |
| 3904 | 2270 | 1610 | 469 | 247 | 95 | 58 | 92 | 7 |

*The first tumor was in Dose Group 2 and the second tumor was in Dose Group 7.
第1番目の腫瘍は線量群2で，第2番目の腫瘍は線量群7で発生した．

With regard to the $\chi_1^2 = 26.5$ value, one might elect to end the statistical analysis at this point by ignoring it with the observation that large sample results can greatly overestimate statistical significance in a small-sample setting. Moreover, the $\chi_1^2 = 26.5$ value is but one of 20 such values associated with the subcohorts. As mentioned

$\chi_1^2 = 26.5$という値に関しては，人によっては，多数対象例用の結果が少数対象例の場合には有意性の大幅な過大評価を起こし得るということを認め，この値を無視し，その時点で統計的解析を終えてしまうかもしれない．更に，$\chi_1^2 = 26.5$という値は，副集団に関連する20の同様の値の一つにすぎないのである．

previously, when all 20 data sets are statistically combined, one finds a total of 20 tumors and an overall log-rank trend test value of $\chi_1^2 = 11.5$. This result may not be as highly significant as it seems. The data in Table 3 shows that among the 20 tumors, 3 are notable in that the size of the dose groups in which these tumors occur, combined with the sizes of all groups with higher doses, is less than 2% of the total number of subjects at risk at the time of the tumor. It is less valid to disregard these data on the grounds that one is dealing with extremely small samples. Moreover, a $\chi_1^2 = 11.5$ value for the entire data set is not likely to be ignored by those summarizing results from the study. However, this $\chi^2$ value corresponds to a two-tailed P value of 0.0007, which probably overstates the case.

先に述べたとおり，20のデータ・セットを統計的に結合した場合，計20例の腫瘍が観察され，全体的な対数階数傾向検定値は$\chi_1^2 = 11.5$になる．この結果は実際にはそれほど有意でないのかもしれない．表3のデータは，腫瘍が発生した線量群にそれ以上の線量群を結合しても，腫瘍時の観察対象例数の2％以下にしかならないという点において，20例の腫瘍中3例は注目に値することを示している．扱っている症例が非常に少数であるという理由から，これらのデータを無視することはより妥当性に欠ける．その上，データ・セット全体の$\chi_1^2 = 11.5$という値は，研究結果を要約しようとする者にとって無視し難い．しかし，この$\chi^2$値は両側のP値0.0007に対応し，事実を誇張するものであろう．

TABLE 3 SUBJECTS AT RISK AT TIME OF MULTIPLE MYELOMA DEATH BY LSS SUBCOHORTS

表3　寿命調査副集団別，多発性骨髄腫による死亡時における観察対象者

| City/Sex/Age ATB* | | | Dose Group | | | | | | | | Tumor Group |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
| 1 | 1 | 2 | 2162 | 1546 | 687 | 154 | 134 | 60 | 36 | 60 | 0 |
| 1 | 1 | 4 | 2063 | 1206 | 816 | 200 | 186 | 51 | 19 | 47 | 1 |
| | | | 1805 | 1073 | 742 | 186 | 166 | 44 | 15 | 39 | 1 |
| 1 | 2 | 3 | 3972 | 2303 | 1636 | 475 | 250 | 97 | 59 | 95 | 2 |
| | | | 3904 | 2270 | 1610 | 469 | 247 | 95 | 58 | 92 | 7 |
| 1 | 2 | 4 | 3416 | 1801 | 1516 | 384 | 222 | 84 | 40 | 51 | 0 |
| | | | 3015 | 1622 | 1343 | 338 | 190 | 68 | 32 | 41 | 0 |
| | | | 3014 | 1622 | 1343 | 338 | 190 | 68 | 32 | 41 | 1 |
| | | | 2807 | 1510 | 1238 | 312 | 178 | 63 | 29 | 35 | 5 |
| 1 | 2 | 5 | 1569 | 804 | 654 | 160 | 65 | 30 | 12 | 11 | 1 |
| | | | 745 | 391 | 298 | 72 | 30 | 14 | 4 | 4 | 0 |
| | | | 542 | 290 | 227 | 57 | 23 | 10 | 3 | 3 | 1 |
| 2 | 1 | 3 | 231 | 244 | 139 | 96 | 82 | 51 | 23 | 18 | 5 |
| 2 | 1 | 4 | 282 | 334 | 196 | 100 | 88 | 42 | 22 | 23 | 3 |
| | | | 263 | 306 | 184 | 89 | 80 | 40 | 20 | 22 | 7 |
| | | | 184 | 220 | 132 | 71 | 59 | 27 | 14 | 15 | 0 |
| 2 | 2 | 4 | 331 | 754 | 440 | 114 | 89 | 52 | 21 | 32 | 1 |
| | | | 265 | 564 | 335 | 85 | 72 | 39 | 17 | 25 | 0 |
| | | | 243 | 517 | 310 | 81 | 64 | 37 | 15 | 23 | 3 |
| 2 | 2 | 5 | 103 | 304 | 175 | 51 | 30 | 15 | 7 | 4 | 2 |

*City/Sex/Age ATB: Hiroshima=1, Nagasaki=2; male=1, female=2; age ATB: 0-9=1, 10-19=2, 20-34=3, 35-54=4, and 55+=5.

都市/性/原爆時年齢：広島＝1，長崎＝2；男性＝1，女性＝2；原爆時年齢：0～9＝1，10～19＝2，20～34＝3，35～54＝4，55+＝5．

In this situation, an exact test with reasonable power is desired. If the censoring intensity were equal in the dose groups, either the log-rank or modified Wilcoxon permutation test would provide such a test. The censoring in the LSS data is dose related, although not strongly so. Thus, the permutation test must, as usual, be considered an approximate test.

このような状況においては，かなり有効で確実性の高い検定が望まれる．線量群における censoring 強度が等しいのであれば，対数階数検定又は修正 Wilcoxon permutation 検定がその役割を果たすであろう．寿命調査データの censoring は，強度にではないが，線量に関係している．このように，permutation 検定は通常どおり近似検定と考えなければならない．

**The Permutation Test**

The score statistics for the log-rank and modified Wilcoxon tests can be written as linear rank statistics of the form

**Permutation 検定**

対数階数検定及び Wilcoxon 検定のスコア統計は，下記の形式の一次階数統計として書き表すことができる．

$$U = \sum_{i=1}^{n} a_i z_i \; , \qquad (2)$$

where $a_i$ is a score attached to the i-th time-ordered observation, and $z_i$ is a covariate representing some property of the observation, such as its dose group membership. For log-rank tests, if the i-th time-ordered sample is the r-th tumor, then $a_i = c_r$, where

ただし $a_i$ は第 i 番目の経時的観察に付属するスコア，$z_i$ は線量群への所属など，観察のある特性を表す covariate である．対数階数検定に関しては，経時的に第 i 番目の症例が第 r 番目の腫瘍であるとすれば，$a_i = c_r$ が成立する．ただし

$$c_r = 1 - \sum_{s=1}^{r} n_s^{-1} \; , \qquad (3)$$

and $n_s$ is the number of subjects at risk at the time of the s-th tumor. If the i-th time-ordered observation is censored at or after the time of the r-th tumor, but before the (r+1)-th tumor (if any), then $a_i = C_r$, where

であり，$n_s$ は第s 番目の腫瘍時の観察対象者数である．第 i 番目の経時的観察を第r 番目の腫瘍時，又はその後に，しかも（もし存在するならば）第 r+1 番目の腫瘍以前に censor するのであれば，$a_i = C_r$ が成立する．ただし

$$C_r = -\sum_{s=1}^{r} n_s^{-1} \; . \qquad (4)$$

For a trend test, $z_i = d_j$, where $d_j$ is the dose received by the i-th time-ordered subject. The permutation log-rank trend test proceeds by considering the values of the U statistic (2) obtained from all possible permutations of the $z_i$ or, as is usually done and will be done here, as a sample of these permutations.

である．傾向検定に関しては，$z_i = d_j$ が成立する．ただし $d_j$ は経時的に第 i 番目の対象者が受けた線量である．Permutation 対数階数傾向検定は，$z_i$ のすべての可能な permutation により得たところの，又は通常，及び本報告においても行うように，これらの permutation の標本として得たU統計量（2）の値を考慮することにより実施する．

Let $\widetilde{U}_1, \ldots, \widetilde{U}_{1000}$ denote the values of the U statistic (2) computed from the data and 999 random permutations. Call

$\widetilde{U}_1, \ldots, \widetilde{U}_{1000}$ が，データ及び 999 の無作為置換から計算したU統計量（2）の値を示すとする．

$$PP = \text{number}\,(\widetilde{U}_i \geqslant U)/1000 \; , \qquad (5)$$

the permutation P (PP) value for the log-rank trend test. Attention will be restricted here to this particular case, since the modified Wilcoxon trend test and the homogeneity test differ only in detail and not in substance.

を対数階数傾向検定の permutation P(PP)値とする．修正 Wilcoxon 傾向検定及び等質性の検定は，細部が異なるだけで，本質的な差異はないので，この特定の場合のみを考慮する．

Because there are nearly 9,000 subjects in Table 2, the direct application of (2) would be expensive, and because its application to the complete data set given in Table 3 would cost even more, so (2) was not used directly. For Table 2, the $z_i$ in (2) take only the dose values, $d_1, \ldots, d_8$, and the $a_i$ take only the four values, $c_1$, $C_1$, $c_2$, and $C_2$. Thus (2) can be written in the form

表2中の対象者数は9,000人に近く，（2)を直接適用するには多大の労力を要し，表3の完全なデータ・セットへの適用には更に大きな労力を要するので，(2)を直接には使用しなかった．表2では(2)の $z_i$ は $d_1$，…，$d_8$ の線量値のみをとり，$a_i$ は4個の値 $c_1$，$C_1$，$c_2$，$C_2$ のみをとる．したがって下記の形式で(2)を書き表すことができる．

$$U = \sum_{i=1}^{m} \sum_{j=1}^{g} \tilde{a}_i d_j N_{ij} \,, \qquad (6)$$

where $\tilde{a}_1, \ldots, \tilde{a}_m$ are the distinct values of the $a_i$, and $N_{ij}$ is the number of subjects in group j which have the value $\tilde{a}_i$. Expression (6) is less expensive to evaluate than (2), but that is not its primary advantage. Table 4 displays the $N_{ij}$ matrix for the data in Table 2, together with row and column totals, $N_{i+}$ and $N_{+j}$. In this context, the problem of computing random permutations of the $z_i$ in (2) is replaced by that of finding random tables $N_{ij}$ with specified marginals $N_{i+}$ and $N_{+j}$. One may randomly compute the first three rows of Table 4 and obtain the last row by subtraction. In this manner only 143 random numbers are dealt with explicitly, rather than randomly permuting a vector of length 8,887. This process reduces the computing costs by a factor of 60.

ただし $\tilde{a}_1$，…，$\tilde{a}_m$ は $a_i$ の個々の値であり，$N_{ij}$ は $\tilde{a}_i$ 値をもつ j 群の対象者数である．式(6)の数値計算には式(2)ほど労力を要しないが，それがこの式の主な利点ではない．表4は表2のデータの $N_{ij}$ 行列及び，行と列の各合計 $N_{i+}$ と $N_{+j}$ を示す．このような状況においては，式(2)の $z_i$ の無作為置換を計算する問題に代わり，特定の限界 $N_{i+}$ 及び $N_{+j}$ をもつ無作為表 $N_{ij}$ を計算する問題が起こってくる．表4の最初の3行を無作為に計算し，最終行を引き算によって求めることもできる．この方法では，8,887の長さのベクトルを無作為に並べかえるのでなく，143の無作為数のみを陽に扱う．この方法は計算に必要な労力を1／60に減少させる．

TABLE 4 DISPLAY OF THE $N_{ij}$ MATRIX FOR THE DATA IN TABLE 2

表4　表2のデータの $N_{ij}$ 行列表示

| | $N_{ij}$ | | | | | | | | $N_{i+}$ |
|---|---|---|---|---|---|---|---|---|---|
| | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| | 68 | 32 | 26 | 6 | 3 | 2 | 1 | 3 | 141 |
| | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| | 3904 | 2270 | 1610 | 469 | 247 | 95 | 58 | 91 | 8744 |
| $N_{+j}$ | 3972 | 2303 | 1636 | 475 | 250 | 97 | 59 | 95 | 8887 |

The P value for Table 2 obtained from this calculation was PP=0.003. The corresponding $\chi_1^2 = 7.26$ is considerably less than the asymptotic value $\chi_1^2 = 26.5$ cited earlier.

この計算により求めた表2のP値は，PP＝0.003である．これに対応する$\chi_1^2 = 7.26$は先に引用した漸近値$\chi_1^2 = 26.5$をかなり下回る．

To compute a PP value for the data in Table 3, some method of combining city/sex/age blocks is required. Following Cochran and Mantel, the pooled score statistic is used:

表3のデータのPP値を計算するためには，都市/性/年齢ブロックをある方法で結合することが必要である．Cochran 及び Mantel の方法によると，結合したスコア統計量を使用する．

$$U_+ = U_1 + \ldots + U_g \ , \qquad (7)$$

where $U_1, \ldots, U_g$ are the score statistics (2) for the nine city/sex/age blocks that contained multiple myelomas; the others being zero by default. Comparable random values of $U_+$

ただし$U_1, \ldots, U_g$は，多発性骨髄腫を含む9個の都市/性/年齢ブロックのスコア統計量(2)である．その他は省略時解釈によって0である．匹敵する無作為値$U_+$

$$\widetilde{U}_{i+} = \widetilde{U}_{i1} + \ldots + \widetilde{U}_{ig} \ \ ; i=1, \ldots, 1000$$

can be obtained by permuting within blocks. For each b, the values $\widetilde{U}_{1,b}, \ldots, \widetilde{U}_{1000,b}$ were obtained from the data in the b-th block in Table 3 plus 999 random permutations using the methodology above.

はブロック内の並べかえを行うことにより求めることができる．各bについて，上記の方法論に従い，表3中第b番目のブロックのデータ及び999の無作為置換から値$\widetilde{U}_{1,b}, \ldots, \widetilde{U}_{1000,b}$を求めた．

The P value for Table 3 obtained from this calculation was PP=0.009 with a corresponding $\chi_1^2 = 6.81$. While this is considerably less than $\chi_1^2 = 11.5$ value for the pooled asymptotic log-rank trend test, the permutation test clearly indicates a statistically significant dose-response relationship of multiple myeloma with radiation.

この計算から得た表3中のP値は PP＝0.009であり，$\chi_1^2 = 6.81$が対応する．この値は，pool した漸近的対数階数傾向検定の値$\chi_1^2 = 11.5$をかなり下回り，permutation 検定が多発性骨髄腫と放射線の統計的に有意な線量反応関係を示すことは明らかである．

**The Simulation Test**

Motivated by a popular heuristic justification for the asymptotic log-rank and modified Wilcoxon tests, a simple alternative test is suggested which has the same asymptotics, but is less likely to overestimate statistical significance in small samples. Unlike the permutation test, this test is not exact under the assumption of equal censoring, but it is expected to be quite accurate when the censoring is heavy, as in the case of the LSS cohort. Moreover, unlike the permutation test, it is insensitive to unequal censoring. In addition, it is simple and inexpensive to compute.

**シミュレーション検定**

漸近的対数階数検定及び修正 Wilcoxon 検定の一般に使用される発見的弁明に刺激を受け，同じ漸近性を有するが，少数の対象例でも統計的有意性の過大評価を起こしにくい代わりとなる簡素な検定を提案する．この検定は permutation 検定とは異なり，等しい censoring という仮定の下では正確でないが，寿命調査対象集団の場合のように censoring が重大なときに，大変正確であると考えられる．更に permutation 検定とは異なり，この検定は等しくない censoring の影響を受けない．また，この検定は簡単で計算に大きな労力を必要としない．

As in the previous section, it is sufficient to consider the log-rank trend test. Let $n_{ij}$ be the number of subjects at risk in the j-th exposure group at the time of the i-th tumor. As before,

前の項目の場合と同様に，対数階数傾向検定を考慮するだけで十分である．$n_{ij}$を第i番目の腫瘍時の第j番目の被曝群における観察対象者数とする．

let $d_1, \ldots, d_g$ be doses associated with the exposure groups, and let $s_i = d_j$, where $j = J_i$ is the treatment group in which the i-th tumor occurs. If there is no dose effect, it is natural to assume that given the numbers of subjects $n_{i1}, \ldots, n_{ig}$ at risk in the g dose groups at the time of the i-th tumor, that $J_i$ has a multinomial distribution, $M(1; p_{i1}, \ldots, p_{ig})$, where $p_{ij} = n_{ij}/n_{i+}$. This motivates calling $e_i = \Sigma_{j=1}^{g} d_j p_{ij}$ and $v_i^2 = \Sigma_{j=1}^{g} (d_j - e_i)^2 p_{ij}$ the conditional expectation and variance of $s_i$. The log-rank trend test statistic is then

前回のように，$d_1, \ldots, d_g$ を被曝群に関連する線量とし，$s_i = d_j$ とする．ただし $j = J_i$ は第 i 番目の腫瘍が起こった治療群である．線量影響がないとするならば，第 i 番目の腫瘍時の g 線量群における観察対象者数を $n_{i1}, \ldots, n_{ig}$ とするならば，$J_i$ は多項分布 $M(1; p_{i1}, \ldots, p_{ig})$ を有する．ただし $p_{ij} = n_{ij}/n_{i+}$ である．したがって，$e_i = \Sigma_{j=1}^{g} d_j p_{ij}$ 及び $v_i^2 = \Sigma_{j=1}^{g}(d_j - e_i)^2 p_{ij}$ を $s_i$ の条件つき平均値及び平方偏差と呼ぶ．その場合，対数階数傾向検定統計量は下記のとおりになる．

$$\chi_T^2 = (s - e)^2 / v^2 \quad , \qquad (8)$$

where $s = \Sigma_{i=1}^{m} s_i$, $e = \Sigma_{i=1}^{m} e_i$, $v^2 = \Sigma_{i=1}^{m} v_i^2$, and m is the total number of tumors. Under the assumption of no dose effect, the statistic $\chi_T^2$ is assumed to have a $\chi^2$ distribution with one degree of freedom.

ただし $s = \Sigma_{i=1}^{m} s_i$，$e = \Sigma_{i=1}^{m} e_i$，$v^2 = \Sigma_{i=1}^{m} v_i^2$ 及び m は腫瘍の合計数である．線量影響がないと仮定すると，統計量 $\chi_T^2$ は自由度 1 の $\chi^2$ 分布をもつと考えられる．

The approach in this study to approximating the small-sample distribution of $\chi_T^2$, is to assume that conditioned on all of the $n_{ij}$ values, the $J_i$ above are independent and have the multinomial distributions identified there. While this is not strictly true, because the distribution of $J_i$, given the number of subjects at risk at the time of the i-th and (i+1)-th tumors, in general will not be the same as the distribution of $J_i$, given only the number of subjects at risk at the time of the i-th tumor. In the present context, where with one exception, 100 or more deaths occur between multiple myelomas, the assumption is very nearly correct. One may obtain the exact distribution of $\chi_T^2$ under the conditional multinomial assumption, by simply convolving the multinomial distributions above. It is this distribution that will be used as an approximation to the true distribution of $\chi_T^2$.

$\chi_T^2$ の少数対象例分布を近似するため本研究で使用する方法は，上記 $J_i$ が $n_{ij}$ 値すべてに条件付けられるが，独立したものであり，上記で識別した多項分布をもつと仮定することである．これは厳密には真ではない．なぜならば，第 i 番目及び第 i + 1 番目の腫瘍時における観察対象者数の場合の $J_i$ の分布は，第 i 番目の腫瘍時における観察対象者数のみの場合の $J_i$ の分布と通常は同じでないからである．本報告においては，1 例を除き多発性骨髄腫発生の間には100以上の死亡があり，この仮定は真実に非常に近い．上記の多項分布を単に合成することにより，条件付き多項仮定の下で正確な $\chi_T^2$ 分布を得ることができるかもしれない．真の $\chi_T^2$ 分布の近似として使用するのがこの分布である．

To simplify the computations in the general case of unequally spaced doses, one further approximation will be made. Rather than compute the exact distribution of $\chi_T^2$ under the conditional multinomial assumption, instead it will be simulated. Actually, as in the previous section, one-sided tests will be dealt with and the log-rank score, $U = s - e$, found in the numerator of the definition (8) of $\chi_T^2$ will be simulated. This is identical to the score used in the previous section. To be specific, the value U will be computed from the actual data, then 999 additional values of s will be simulated, by

間隔が不均一な線量を扱う一般的な場合の計算を簡単にするために，更に別の近似法を検討する．条件付き多項仮定の下で正確な $\chi_T^2$ 分布を計算する代わりに，それをシミュレートする．実際には，前項のように，片側検定を行い，$\chi_T^2$ の定義(8)の分子に見られる対数階数スコア $U = s - e$ をシミュレートする．これは前項で使用したスコアと同一である．具体的には，U 値を実際のデータから計算し，各々の U 値を計算し上記のとおり識別した多項分布から

sampling the required $J_i$ from the multinomial distributions identified above computing a value of U for each. The number of U values which are as large as or larger than the U value for the real data divided by 1,000 is the P value for what will be called the simulated log-rank trend test. This simulated P value is denoted by PS.

必要な $J_i$ を抽出することによって，更に 999 の s 値をシミュレートする．実際のデータの U 値と同じ，又はそれ以上の大きさの U 値数を 1,000で割ったものが，いわゆるシミュレートした対数階数傾向検定の P 値である．このシミュレートした P 値を PS によって示す．

Using this procedure on the data in Table 2 gives a PS value of 0.004, which is very close to the PP value of 0.003 previously obtained using the permutation test. These values are about what would be expected from an examination of the data. Note that the computing burden required to obtain the PS value is substantially less than that required to obtain the permutation value. To obtain one randomly simulated value of U for Table 2, one needs to compute and deal with two random numbers, whereas obtaining a random value of U by the more efficient method of the previous section requires computing and dealing with 143 random numbers.

この手順を表2のデータに適用すると，PS 値は 0.004 となり，先に permutation 検定により得た PP 値 0.003 に非常に近い．データの検討により得られる値は，これらの値に近いと考えられる．PS 値を求めるのに必要な計算労力は， permutation 値を求めるのに必要な労力より実質的に少ないことは注目に値する．表2の無作為にシミュレートした1個の U 値を求めるために計算し取り扱わなければならない無作為数は2個であるが，前項の効率の高い方の方法により無作為の U 値を求めるためには143の無作為数を計算し取り扱わなければならない．

As in the previous section, the pooled score statistic $U_+$ for the data in Table 3 is the sum of the score statistics $U_1, \ldots, U_g$ for the nine city/sex/age blocks which contain multiple myelomas. Since already simulation is "within tumor", nothing additional is required to simulate within blocks. The data in Table 3 are simply treated as a larger data set containing all 20 tumors and a PS value is computed in exactly the same manner as was done with the data in Table 2. Only 20 random values are utilized for each simulation, which again makes the computing of the PS value much less expensive (here, by a factor of 16) than the permutation value. The simulation test P value for Table 3 was PS = 0.008, which again very close to the value PP = 0.009 obtained from the permutation test.

前項の場合と同様に，表3のデータの pool したスコア統計量 $U_+$ は，多発性骨髄腫を含む9個の都市/性/年齢ブロックのスコア統計量 $U_1, \ldots, U_g$ の合計である．シミュレーションは既に"腫瘍内"のものであるので，ブロック内でシミュレートするために付加的なことは何も必要でない．表3のデータは腫瘍20例すべてを含む更に大きなデータ・セットとしてのみ扱い，PS 値を，表2のデータ計算と全く同じ方法で計算する．各シミュレーションに使用する無作為値は20個だけなので，この場合も permutation 値の計算と比較し PS 値の計算に要する労力はかなり少ない（ここでは 1/16である.）．表3のシミュレーション検定 P 値は PS = 0.008であり，これも permutation 検定で得た値 PP = 0.009に大変近似するものであった．

To verify the simulation test, a small Monte Carlo study in a simplified setting that reflected some of the important characteristics of the LSS cohort was conducted. Two markedly unequally sized groups of subjects were used, the first with 190 subjects and the second with 10; both with very high censoring rates. The survival and censoring distributions were the same for both groups and assumed to be exponential, with the censoring distribution having 49 times the intensity of the survival distribution. The probability is then 1 in 50 that a given individual will die from a specific tumor before being

シミュレーション検定の妥当性を証明するために，寿命調査対象集団の重要な特性の幾つかを反映する単純化した状況における小 Monte Carlo 調査を実施した．対象者数190人と10人の著しく大きさの異なる2群を使用した．両群とも非常に高い censoring 率を有する．両群の生存分布及び censoring 分布は同一であり，指数関数的と考えられ，censoring 分布の強度は生存分布の49倍であった．その場合，特定の個人が censoring を受ける前に特定の腫瘍で死亡する確率は 1/50である．対象例数1,000人の集合を作った．

censored. A set of 1,000 samples were generated. As expected, the samples gave rise to some inordinately small asymptotic P values; 2 were less than $10^{-6}$, and 10 were less than $10^{-4}$. The asymptotic P values would have led to 99 rejections at the 5% level, while at this level, the simulated distributions, which were computed exactly here, led to 43 rejections. Since 50 rejections were expected, the simulation test performed well. On the other hand, the standard asymptotic log-rank test yielded unacceptable results. In this 2-group setting, a half-unit continuity correction would improve the results of the asymptotic test; however it is not clear how such a correction should be made in a more general setting.

A technical problem with both the permutation and simulation tests is that different random number generator seeds will lead to different values of PP and PS. For the data in Table 2, three different seeds led to the PS values of 0.004, 0.005, and 0.009. Such differences can be reduced by increasing the number of simulations or permutations used. In the case of the simulation test, the problem can often be eliminated entirely by computing the exact distribution of the U scores under the approximate conditional multinomial assumption. This is easily done, for example, when the doses are equally spaced. On the other hand, it is seldom possible to compute exactly the permutation distribution at a reasonable cost.

## Conclusion

The epidemiologic example of multiple myeloma clearly illustrates the statistical mistakes which are possible through the application of large-sample results to small-sample problems. Although the conclusion regarding a positive association of multiple myeloma with radiation exposure was not changed, the P value was increased by an order of magnitude. The small Monte Carlo study further supported the use of small-sample procedures. Even though the study was extremely limited, it indicated that the asymptotic test yielded an actual alpha level twice the nominal level. This is quite unacceptable by any standard. These considerations provide justification for the recommendation of the present report, to routinely include a simulation test in analyses whenver one is dealing with studies involving only small numbers of cases.

予想どおり，これらの対象例には，漸近的P値が極端に小さいものがあり，2例においては$10^{-6}$以下，10例においては$10^{-4}$以下であった．シミュレートした分布は，ここで計算したとおり，5%のレベルで43の棄却を導き出したが，漸近的P値であれば同レベルで99の棄却を導き出したであろう．期待した棄却数は50であったので，シミュレーション検定の結果は妥当であったと言える．一方，標準的漸近的対数階数検定は受け入れられない結果を生み出した．この2群の場合には，半単位連続性修正をすれば漸近検定の結果が改善されるであろう．しかし，より一般的な状況においてそのような修正を行う方法は明らかでない．

Permutation 検定，及びシミュレーション検定に伴う技術的問題としては，異なる無作為数生成素が異なるPP値及びPS値を導くことである．表2のデータでは，異なる3個の生成素がPS値0.004，0.005及び0.009を導いた．このような差異を減少させるためには，使用するシミュレーション数又はpermutation数を増加する．シミュレーション検定の場合には条件付き近似多項仮定の下でUスコアの正確な分布を計算することにより，この問題は完全に解決される．例えば線量の間隔が等しいときなど，これは容易に行うことができる．一方，大きな労力をかけずに，permutation分布を正確に計算することはほとんど不可能である．

## 結　語

多発性骨髄腫の疫学例は，多数対象例の結果を少数対象例の問題に適用することから起こり得る統計的誤りを明確に示す．多発性骨髄腫と放射線被曝との明確な関連に関する結論に変化はないが，P値は1桁倍増加した．更に，小 Monte Carlo 調査により，少数対象例手順の有用性が明らかになった．この調査は非常に限定されたものであったが，漸近検定が生み出す実際のアルファ・レベルは名目上のレベルの2倍になることを示した．これはいかなる基準からも受け入れることはできない．以上の点を考慮すると，本報告に述べたように，少数症例のみを扱う研究を実施する際，解析の一環としてシミュレーション検定を行うことが妥当であると考える．

## REFERENCES

## 参考文献

1. BEEBE GW, KATO H, LAND CE: Studies of the mortality of A-bomb survivors. 6. Mortality and radiation dose, 1950-74. Radiat Res 75:138-201, 1978 (RERF TR 1-77)

2. KATO H, BROWN CC, HOEL DG, SCHULL WJ : Studies of the mortality of A-bomb survivors. Report 7. Mortality, 1950-78: Part 2. Mortality from causes other than cancer and mortality in early entrants. Radiat Res 91:243-64, 1982 (RERF TR 5-81)

3. MANTEL N: Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemother Rep 50:163-70, 1966

4. COX DR: Regression models and life tables (with discussion). J R Statist Soc B 26:187-220, 1972

5. GEHAN E: A generalized Wilcoxon test for comparing arbitrarily single censored samples. Biometrika 52:203-23, 1965

6. BRESLOW N: A generalized Krustal-Wallis test for comparing K samples subject to unequal patterns of censorship. Biometrika 57:579-94, 1970