Table 2 Summary results on leukaemia mortality in both cohorts and in selected subgroups

Cohort	Subgroup	No of Men	Person-years*	Obs	Exp	SMR (95% CI)
NSF	All men	3812	66742	3	7·1	0·42 (0·09, 1·23)
	Byproducts	662	11167	1	1·3	0·76 (0·02, 4·29)
	Coke ovens	1661	28476	1	3·0	0·34 (0·00, 1·86)
	Maintenance	1061	18510	1	1·7	0·58 (0·01, 3·28)
BSC	All men	2708	46969	2	4·9	0·41 (0·05, 1·47)
	Byproducts	631	10384	1	1·0	0·98 (0·02, 5·57)
	Coke ovens	1688	27840	1	2·9	0·35 (0·01, 1·92)

*The total amount of observation time accumulated during the follow up period.

posure to coke oven fumes. Of the two BSC men who died from leukaemia, one had some byproducts experience, and the other had worked in coke oven jobs.

We have also reviewed unpublished tabulations of leukaemia mortality over 17 years (1967 to the end of 1983) in an associated study of 287 male industrial workers employed on 1 January 1967 at four tar distillation plants and one benzene refinery. Results showed no deaths from leukaemia. Expected deaths were not calculated, but are likely to be about 0·5. Details of the cohort and mortality from other causes 1967–83, have been reported.[7]

Discussion

The experience of the combined cohort over 20 years amounted to a reasonably large scale study of men occupationally exposed to low concentrations of benzene. No evidence was found for excess mortality from leukaemia in either the NSF or BSC cohort overall (standardised mortality ratio (SMR) 42 and 41 respectively), or among men known to have worked in the coke oven battery jobs in particular (SMR 34 and 35 respectively). The SMRs for byproduct workers, though somewhat higher, did not suggest excess mortality from leukaemia (SMR 76 and 98 respectively), but were based on only 2·3 expected deaths.

The results are therefore consistent with the other evidence reviewed recently[1]: whereas excess leukaemia mortality has been shown in relation to high exposure to benzene, several large scale studies of men exposed at concentrations below 10 ppm have shown negative results. Furthermore, these preliminary analyses by exposure group, together with the available, albeit recent, hygiene measurements, go some way towards overcoming the two limitations of the available low exposure studies highlighted in that review,[1] that is, the lack of quantitative exposure data, and the possibility that a proportion of the 'exposed cohort' has not been exposed to benzene at all.

J FINTAN HURLEY
JOHN W CHERRIE
WILLIAM MACLAREN
Institute of Occupational Medicine Ltd, 8 Roxburgh Place, Edinburgh EH8 9SU

1 Swaen GMH, Meijers JMM. Risk assessment of leukaemia and occupational exposure to benzene. *Br J Ind Med* 1989;46:826–30.
2 Brunekreef B, Swaen GMH, Meijers JMM. Correspondence. *Br J Ind Med* 1990;47:717–8.
3 Redmond CK, Strobino BR, Cypess RH. Cancer experience among coke by-product workers. *Ann NY Acad Sci* 1976;217:102–15.
4 Swaen GMH, Slangen JJM, Volovics A, Hayes RB, Scheffers T, Sturmans F. Mortality of coke plant workers in the Netherlands. *Br J Ind Med* 1991;48:130–5.
5 Hurley JF, Archibald RMcL, Collings PL, Fanning DM, Jacobsen M, Steele RC. The mortality of coke workers in Britain. *Am J Ind Med* 1983;4:691–704.
6 National Smokeless Fuels Ltd. *Pollution at coke works.* National Smokeless Fuels Ltd, 1986. (Final report on ECSC project No 7257-14/340/08.)
7 Drummond L, Luck R, Afacan AS, Wilson HK. Biological monitoring of workers exposed to benzene in the coke oven industry. *Br J Ind Med* 1988;45:256–61.
8 Maclaren WM, Hurley JF. Mortality of tar distillation workers. *Scand J Work Environ Health* 1988;13:404–11.

Bootstrap estimate of the variance and confidence interval of kappa

Sir,—Methodological research in clinical medicine and epidemiology is often concerned with measuring the extent of agreement between two methods or observers for rating an outcome, or the reproducibility of a method or observer for rating an outcome on two different occasions. The kappa coefficient, as originally described by Cohen,[1] quantifies agreement or reproducibility when the outcome is dichotomous. The coefficient was subsequently extended to nominal and ordinal outcome variables with three or more categories.[2 3]

The computation of kappa, including a test of null hypothesis and an estimation of confidence interval based on normal distribution theory, is summarised by Fleiss.[4] Although the normal theory procedure is reliable for testing of null hypotheses, the procedure is often not reliable for constructing a confidence interval. As the kappa coefficient is bounded by 1 (perfect agreement), its sampling distribution is highly skewed when strong agreement or reproducibility exists. (Note that kappa can also be negative (observed agreement less than chance expected agreement) in which case it is bounded by -1). A better alternative to determine the confidence interval of kappa is therefore based on the empirical sampling distribution generated by the computer intensive bootstrap resampling method.[5]

We have written a computer program in the BASIC language (compiled by Microsoft's QBASIC version 4.5) to carry out the bootstrap estimate of the variance and confidence interval of kappa. The program works on the IBM compatible PC with or without a math coprocessor and it supports CGA, EGA, and VGA. Because the program is entirely menu driven, it is easy to run. The user types "KAPPA" as a DOS command and then simply responds to a series of question prompts by the program. One question asks the user to state C, the number of categories in the outcome variable. If C is greater than 2 the user is then asked to state whether the outcome variable is nominal or ordinal.

The ordinary kappa coefficient is computed when C is 2. If C is greater

than 2 and the outcome is nominal, the program computes one kappa for each category v the remaining categories, and also the overall kappa, which is the weighted average of the individual kappas. If C is greater than 2 and the outcome is ordinal, the program computes the weighted-kappa.[3] All of these procedures are discussed in Fleiss.[4]

The program outputs the observed and chance expected proportion of agreement, kappa and its variance, statistical test of Ho:KAPPA = 0, and confidence interval of KAPPA based on the bootstrap resampling approach. The program optionally gives a graphical depiction of the empirical sampling distribution of kappa generated by bootstrap. Although the program sets a minimum number of bootstrap samples to ensure a reasonably smoothed sampling distribution, the user can increase this number. Thus when kappa is greater than 0·7 at least 1500 bootstrap samples are required to obtain a smoothed sampling distribution especially at the tail ends.

It should be pointed out that "statistical significance" of kappa does not indicate the strength of agreement. It would indeed be surprising if kappa, which measures the agreement of two methods for rating the same outcome, were not statistically different from zero. What is more pertinent is the quantitative significance of kappa. Landis and Koch[6] suggest the following guideline: A kappa coefficient greater than 0·75 indicates strong beyond chance agreement or reproducibility; a value below 0·40 indicates poor agreement; and a value between 0·40 and 0·75 represents fair agreement. Further commentaries on the use of kappa are expounded elsewhere.[7 8]

The computer program is available upon request. Please send either a 5·25 or 3·5 inch diskette with sufficient money to cover airmail postage to Dr James Lee.

K P FUNG
Physician in private practice, Hong Kong
JAMES LEE
Division of Biostatistics and Health Informatics, Department of Community, Occupational and Family Medicine, National University of Singapore, NUH, Lower Kent Ridge Road, Singapore 0511

1 Cohen J. A coefficient of agreement for nominal scale. *Educ Psychol Meas* 1960;**20**:37–46.
2 Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull* 1968;**70**:213–20.
3 Cicchetti DV, Allison T. A new procedure for assessing reliability of scoring EEG sleep recordings. *Amer J EEG Technol* 1971;**11**:101–9.
4 Fleiss JL. *Stastistical methods for rates and proportions, 2nd ed.* New York: Wiley, 1981 (chapter 13).
5 Efron B, Tibshirani R. Bootstrap methods for standard errors, confidence intervals and other measures of statistical accuracy. *Statist Sci* 1986;**1**:54–77.
6 Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; **33**:159–74.
7 Maclure M, Willett WC. Misinterpretation and misuse of the kappa statistic. *Am J Epidemiol* 1987;**126**:161–9.
8 Yoshizawa CN, LeMarchand LL. Re: Misinterpretation and misuse of the kappa statistic. *Am J Epidemiol* 1988; **128**:1179–80.