# INTERNATIONAL EVIDENCE

# ON

# SMOKING AND LUNG CANCER

# (PROJECT IESLC)

## A FIRST REPORT

## PART I: THE DATABASES; METHODS USED TO COLLECT AND ANALYSE THE DATA AND SCOPE OF THE INFORMATION OBTAINED

Peter N Lee
Barbara A Forey and
Katharine J Young

P N Lee Statistics and Computing Ltd
17 Cedar Road
Sutton
Surrey, SM2 5DA
UK

March 2003

# EXECUTIVE SUMMARY

Based on papers published up to the end of 1999, 296 studies have been identified which provide information from epidemiological case-control or prospective studies involving 100 or more lung cancer cases. Two linked databases have been set up. One contains details of the characteristics of each study, while the other contains relative risk data relating to certain aspects of smoking status (current/ex/ever vs never/non current), product (cigarette; pipes, cigars and combinations), cigarette type (manufactured/ handrolled, filter/plain, menthol) and amount smoked. For each study, the study database contains details of the study itself, the potential confounding variables considered and the smoking exposure indices for which results are available (including such aspects as tar level, black/blond tobacco, inhalation, age of starting to smoke, duration for which relative risks have not been entered at this stage). For each of the 9551 relative risks included, the relative risk database contains not only the relative risks and 95% confidence intervals, but precise details of their definition and information on how they were derived.

This report starts by describing the methods used to identify relevant papers, which involved examination of almost 6000 papers, and to try to classify them into separate studies. It then describes in detail the structure of the databases and the methods used for entry and checking of data. The methods by which relative risks were derived from data presented in various ways are also described. Although the intention was to have non overlapping studies, this could not always be achieved without marked loss of useful data. There were 277 independent principal studies, with 19 subsidiary studies where data will only be used in meta-analyses where equivalent results are not available from the principal studies.

The 277 principal studies were conducted in 37 countries, with 8 starting before 1940. 79% were of case-control design, with 58% providing data for both males and females. The largest study involved 22161 lung cancer cases with a further 3 studies involving 10000 or more and a further 33 1000 or more. 193 of the studies relate to the general population, others relating to subjects with specific characteristics (e.g. in defined

occupations). Full histological confirmation was only carried out in 25% of studies. 31% provide data by histological type. Data on ever smoking are available for 92% of the studies, while data on current smoking are available for 54% and on ex-smoking for 47%. Data on amount smoked are available for 60% of the studies. Other aspects of smoking for which results are stored on the relative risk database are less frequently available (e.g. pipe/cigar 36%, filter/plain 14%, handrolled 6%). Data are also available on risk by pack-years, duration, age of starting to smoke, years since stopped smoking and inhalation for at least 40 studies, but have not yet been included on the relative risk database. The potential non smoking confounding variables most commonly taken into account are age (143 studies), risky occupational exposures (27), education (21) and race (19). Fuller details of the studies are given in this report.

Of the 9551 relative risks, 8934 relate to the principal studies and 617 to the subsidiary studies. The number of relative risks per principal study varies widely, from only 1 in 22, to over 100 in 20, the largest being one with 428 relative risks entered. Of the relative risks, 67% are for males, 27% for females and 7% for sexes combined. 80% relate to results for the full age range of the study, while 20% are age-specific. 78% relate to all races within the country studied, and 22% are race-specific. 69% relate to all lung cancer types with others relating to specific types. 50% of the risks relate to cigarette smoking (regardless of other product), with 22% relating to smoking of any product and 17% to smoking of cigarettes only. 544 relative risks relate specifically to manufactured or handrolled cigarettes, 676 to filter or plain cigarettes and 23 to menthol cigarettes. 41% of relative risks are for specific amount smoked. 39% are adjusted for at least one variable. 325 have a relative risk value with no confidence interval available. Only 35% of the relative risks and confidence intervals are as given originally or calculated directly from the numbers in the relevant 2 × 2 table. The rest involve more complex calculations. Fuller details of the relative risks are given in the report.

The report ends by describing techniques for conducting meta-analyses and the format of the tables presenting the results. The process of selecting which relative risks to include in an an analysis is described in detail. It has to be quite complex to ensure that all the relevant data are included, while at the same time avoiding double-counting.

Results from a variety of meta-analyses will be described in Part II of this report, and plans for further work in Part III.

# INDEX

Tables Page

Appendices

## 1. Introduction

The objective of the IESLC project is to collect and summarize published epidemiological evidence relating smoking to lung cancer, with a view to assessing how the strength of the association varies by the index of exposure to smoking considered and by the characteristics of the study reporting the findings.

The work, which started in 1997, has involved a number of stages. These included:

i) **Identification of the studies** Attention has been restricted to epidemiological case-control or prospective studies involving 100 or more lung cancer cases, and so far to papers published up to the end of 1999.

ii) **Development of software** Available inhouse software (ROELEE) was extended to allow entry of data in a suitable format and to carry out selected data summaries and meta-analyses conveniently.

iii) **Setting up databases to allow entry of relevant data** The structure involves two linked databases, one containing study details, with a record for each study, the other containing relative risk details, with a record for each relative risk (RR). The study database contains details of the study itself (e.g. location, timing, design, type of controls used, proxy use, response rate), the potential confounding variables considered, the smoking exposure indices for which results are available and the demographic variables by which results are broken down. The relative risk database contains all RRs reported relevant to the exposure indices of "major interest" (*vide infra*), for the whole population and broken down by the more important demographic variables, with sufficient detail stored to define the RR precisely.

iv) **Entry and checking of data** As envisaged at the start of the project, RRs have so far only been entered for four smoking exposure indices of "major interest".

| | |
|---|---|
| smoking status | (Current/ex/ever vs never/non-current) |
| product | (Cigarette, pipes, cigars and combinations) |
| cigarette type | (Manufactured/handrolled, filter/plain, menthol) |
| amount smoked | |

For these indices, data were entered, where available, for total lung cancer and by histological type, for the whole population, and broken down by age, sex and race, and, for prospective studies only, for different lengths of follow up period.

For all other smoking exposure indices, including tar level, time of plain/filter switch, tobacco type (black/blond etc), inhalation, age of starting to smoke, duration, pack-years, years since stopped smoking, information was recorded on the study database to indicate whether RRs were available, but at this stage no data have been entered.

v) **Carrying out analyses** Although a certain amount of analysis using the study database has been carried out to summarise the characteristics of the studies considered and the quantity and type of data available, the main work has involved carrying out numerous meta-analyses to meet the main objectives of the project.

This report describes the work carried out in fuller detail and presents the results of the analyses so far conducted. It also considers how the databases might be further used and most usefully extended in the future. Part I of the report describes the method of identifying the studies, the databases and the methods used to carry out meta-analyses, Part II presents and discusses results, while Part III considers further work.

## 2. Identifying the studies

The objective was to identify epidemiological studies of prospective or case-control design (uncontrolled case studies not being included, as RRs cannot be calculated) which involved a total of 100 lung cancers or more and which either reported RRs relating any of the four aspects of smoking to lung cancer or which provided data from which such RRs could be calculated.

To obtain papers describing such studies, the extensive files on smoking and health accumulated by P N Lee Statistics and Computing Ltd (PNLSC) were examined. Papers in those files which were at all likely to contain material of interest for the project were examined to see if they either provided relevant information and/or cited papers not already on the PNLSC reference system that were of possible relevance. Such cited papers were then obtained, added to the PNLSC reference system and then examined as above. Ultimately, a position was reached whereby no paper examined cited a paper of possible relevance that had not already been examined. MEDLINE searches were also carried out to detect whether any possibly relevant papers had been missed, any found being obtained and examined as above.

Attention was restricted to papers published by the end of 1999,[a] but no restriction was made on language. Where necessary (principally Chinese and Japanese papers), English translations were obtained, although as far as possible dictionaries were used to identify key information from non-English papers.

Overall, 5993 papers were identified, of which 5749 could be obtained and examined. Of these, 687 contained data relevant to the project, 175 described studies which were not relevant because the number of lung cancers considered was less than 100, and the remaining 4888 did not provide relevant data at all.

[a] Exceptionally, a 2000 conference paper of which the abstract had been published in 1997, and a 2001 reprint of a 1943 paper were included.

The next step was to take the papers that contained relevant data and classify them into the separate studies they described, taking account of the fact that some papers described results from more than one study, and that results from the same study were often described in multiple publications. Ultimately, for each study identified, a file was built up of papers relevant to that study, the files being sorted by continent, by country within continent, and by state within the USA. This sorting made it easier to ensure that the studies identified as separate really were so, though on occasion (as described in the next section), there were some problems in deciding whether or not papers described results from the same or different studies. Ultimately, files of papers relating to 296 studies were obtained.

Appendix A gives certain details of the 296 studies, the 6-character reference used to identify the study, a longer study title (which includes information on the location and timing of the study), the reference key to the principal publication used to extract data and the reference keys to other relevant publications. Reference keys are those used in the PNLSC reference system. Appendix B gives all the reference keys used, in alphabetical order, together with the associated full references.

## 3. The databases

### 3.1 Structure of the two databases

There are two linked databases. The first, the study database, contains one record for each study. This record is identified by a unique six-character reference (REF), and holds information relevant to the study as a whole, described more fully in §3.3. The second, the relative risk (RR) database, holds the detailed results, and can contain multiple records for each study. Each record refers to a specific comparison, and contains the information describing that comparison (e.g. current cigarette smokers vs. never smoked at all, for a particular sex, age, race and lung cancer type) and the actual results. Each record also contains the study REF, which links it to the relevant record in the study database. The RR database is described more fully in §3.4.

[Note that it is possible for a study to be entered in the study database but to have no corresponding records in the RR database. This is the case for two studies which were initially thought to provide relevant data but in fact did not. These studies are not included in any of the tables in this report.]

### 3.2 Data entry and checking

Before data entry on computer, master copies of the papers in the study file were read through closely to identify the information that would need to be entered, highlighting this by a marker pen (and making notes on the paper where necessary) to facilitate later checking. Where multiple papers were available for the same study, a principal publication was selected to provide most of the information, though details of interest not described in the principal publication but available elsewhere were also entered. The principal publication was usually that which provided information on the largest number of lung cancer cases, for example based on longer follow-up for a prospective study or avoiding interim results from a case-control study. On occasion, descriptions of some aspects of the study conflicted between different papers – where necessary, the most likely version was determined by consultation between the authors of this report, with notes of the problem being recorded on the database.

The study data and the RR data given directly in the paper were entered on the computer initially, usually by KJY, and then checked, usually by BAF. Further RR data were then derived and entered, usually by BAF, and checked by KJY. These checks were carried out partly by reference back to the original papers, and partly by running an automatic checking program which investigated the completeness and consistency of the data entered. See Appendix C for details of the automated checks.

## 3.3 The study database

### 3.3.1 Structure of the database

As described in more detail in Appendix D, the study database contains one record for each study, with each record consisting of “fields” within “cards”. The “cards” separate the different main classes of information recorded, while the “fields” contain the individual data items within each class. Each field may contain data of various types, including:

presence : the item may be present or absent,

graded : the item may have one or more discrete levels defined in its associated grading system (graded >0 is used for items which have to have a positive grade)

measured : the item may take any integer value within the specified range (measured +v is used for items which must be positive)

character : the item is text with up to the defined number of characters

real : the item may take any decimal value within the defined range (only the RR database in fact contains decimal data)

For all field types, data items may be entered as missing or not applicable.

The six cards used for data entry, together with a brief description of the fields included in each, are as follows:

**Study description** This includes the study short and full title, details of possible overlaps or links with other studies on the database, whether the study is restricted to men or to women or is unrestricted, the age range and the race of the population considered, the location of the study, the period of the study, the year and reference key of the principal publication and the number of lung cancer cases and controls (or at risk for prospective studies). A free text comment also contains additional detail where required, including the reference keys of other related publications.

**Study design** This includes the study type (case/control, prospective, nested case/control or case/cohort), the type of controls used (e.g. healthy, diseased/hospital), the disease categories included in the controls, the type of population studied, details of proxy use, and differences between cases and controls in respect of hospitals they come from, interview setting and vital status. It also includes details on the extent of histological confirmation, response rates and variables used to match cases and controls on.

**Results presented** This includes a series of presence fields, indicating whether results are presented by histological type and for the various smoking exposure variables of interest (ex-smokers, current smokers, ever smokers, cigarette type, handrolled cigarette smoking, pipe smoking, cigar smoking, pipe and cigar smoking combined but not separately, and by amount smoked) for which RRs are to be recorded. It also indicates whether some more detailed results on the primary smoking exposure variables are available, but have not been entered. Definitions of handrolled cigarettes and pipes are also entered (which vary regionally, e.g. Western pipe vs water pipe), as are definitions of ex-smokers (by time of giving up).

**Other smoking/tobacco-related results** This includes a series of presence fields indicating whether data are available in the papers for various other smoking exposure indices, for which data have not been entered on the RR database at this stage. There are 32 of these, some relating to quite common indices such as age of starting to smoke or pack-years, others to quite rarely used indices such as smoking before breakfast. All smoking exposure indices found in the studies were included.

**Other non-smoking related results** The first field indicates whether the study provides data on diseases other than lung cancer. The remaining 35 fields indicate whether or not results are subdivided by specific variables such as education, social class, etc. Note that these fields do not indicate whether they paper provides information on the relationship of the variables to lung cancer, only whether it provides information on how the association of smoking with lung cancer varies by level of the variable considered. Note that smoking/lung cancer RRs by level of these variables are not entered in the database; the only such stratifying variables by which RRs are entered are sex, age and race.

**Confounders considered** The first field gives the total number of potential confounding variables for all the RRs entered in the RR database. The remaining fields indicate whether adjustment has occurred for 45 separate potential confounders. On most occasions, data entry is 0 for confounder not adjusted for or 1 for confounder adjusted for. Exceptionally, a higher number than 1 indicates that the confounder was adjusted for by use of more than 1 variable (e.g. diet by several specific foods).

Further **Derived fields** cards are used to preserve on the database certain fields created from other fields, but not entered directly.

The record itself is uniquely identified by a six character study reference, usually based on the principal author's name.

For some studies, certain subsets of the subjects were omitted from the relevant results; for instance some studies included both sexes but only analysed males, while other studies omitted subjects with incomplete data from all analyses. In these cases, the description of the study as analysed was entered.

### 3.3.2 The study data

The data recorded on the study database for each of the 296 studies is presented in Appendix E. This is in the form of a computer-generated report. Note that this report is based only on fields which provide positive information. Thus, for example, in the card “results presented”, the report only shows those smoking exposure indices for which results were presented. Results for other indices for this card, for which no output is shown, are taken not to be available.

### 3.3.3 Problems with overlapping studies

In theory, RRs being meta-analysed should come from independent studies involving distinct lung cancer cases; if some lung cancer cases feature in more than one study, they will be "double-counted" in any meta-analysis which includes results from both studies. In practice, avoidance of such double-counting is difficult and may not always be the most desirable solution. For example, suppose study A describes a case-control study conducted in 1970-80 involving all hospital cases in town X admitted with lung cancer, while study B describes a similar case-control study in the same town conducted in 1978-88. Including results from both studies would involve some double-counting, of deaths in 1978-80, but avoiding this would require totally ignoring results from one study (or both), with a substantial loss of power, which would seem to be less desirable than allowing some double-counting. Even omitting study B if it had been conducted in 1975-77 (totally within the period for study A) may not necessarily be appropriate, if the paper describing study B reports data for some exposure indices not considered in the paper describing study A. One would not want to include results from both studies in analysis of the same exposure index (and would omit study B if both RR estimates were available), but one might want to use data from either study if only one provides the required RR. There are other possibilities too that need to be borne in mind; for example, studies of overlapping regions or studies which do not completely describe where or when they were conducted and may overlap other studies.

In entering data from individual studies, care was taken to avoid double-counting by, for example, not entering results for the same exposure index for all cases and for a study subset. Nevertheless, there were some sets of studies which were noted on the database as having overlaps or links. For the purposes of analysis, these sets of studies were grouped into two categories.

The first category are studies with a modest degree of overlap, which cannot be disentangled and which it was decided to ignore. These sets are described below briefly:

1. CHAN contains data from a case-control study in 5 hospitals in Hong Kong in 1976-77, involving 397 cases, while the LAMWK2 case-control study contains data involving 480 cases from one of these hospitals for 1976-80.
2. KOO contains data from a case-control study in 8 hospitals in Hong Kong in 1981-83 involving 120 lung cancer cases, LAMWK contains data from a case-control study in one of these hospitals for 1981-84 involving 163 cases, while LAMTH contains data from another case-control study in 8 hospitals in Hong Kong in 1983-86 involving 445 cases, apparently not from the hospital used by LAMWK. [Note that KOO and LAMTH clearly overlap little (1983) and that, unlike LAMTH and LAMWK which appeared to involve all the available cases, KOO only involved a proportion.]

3. SOBUE2 contains data from a case-control study conducted in Osaka in 1965-83 involving 2083 cases diagnosed at one Center, while MATSUD contains data from a case-control study conducted in certain regions of, and companies in, Osaka in 1965.
4. GOODMA contains data from a case-control study conducted in Oahu, Hawaii in 1983-85 involving 326 cases with no restriction on race or sex, while CHYOU contains data from a prospective study also conducted in Oahu where 7961 Japanese-American men were interviewed in 1965-68 and followed until 1990.
5. CPSI contains 12 year follow up data from all the US states participating in the million person study, while ENSTRO contains 28 year follow up data only from Californian participants in this study.

The second category contains sets of studies which clearly do overlap, where one or more members of the set (“principal studies”) contain the most appropriate data (and do not themselves overlap) and where, for other members (“subsidiary studies”), RRs should only be included in meta-analyses if equivalent results are not available from the principal studies. These sets are also described below:

1. XIANGZ (principal study) is a cohort study of Yunnan tin miners followed from 1976-87, including 983 cases while QIAO and LUBIN (subsidiary studies) are case-control studies of Yunnan tin miners involving, respectively, 107 cases occurring in 1976-84 and 427 cases occurring in 1984-88.
2. MRFIT (principal study) involves 9 year follow up of all 361662 men screened in this study, while MRFITR (subsidiary study) involves 12 year follow up of only those 12866 men taking part in the intervention trial proper.
3. LUBIN2 (principal study) contains combined results from a large multicentre case-control study in 7 centres in 5 West European countries

whilst VUTUC - Austria, BENHAM - France, BERRIN and PISANI - Italy and GILLIS - Scotland (subsidiary studies) present results from individual countries not including Germany.

4. BOFFET (principal study) contains combined results from a multicentre case-control study in 7 centres in Germany, Sweden and Italy, whilst JAHN (subsidiary study) present results from one of the German centres.
5. AKIBA (principal study) contains data from the Atomic bomb survivors cohort study, whilst ISHIMA (subsidiary study) presents data from a case-control study from within this cohort.
6. CASCOR (principal study) is a case-control study in Berlin involving 389 cases, 220 of which were already considered in ROOTS (subsidiary study), another Berlin case-control study involving 270 cases in all.
7. KAISER (principal study) is a cohort study of Kaiser health check attendees in California involving men and women interviewed in 1964-73 and followed up to 1980, by which time 714 had lung cancer, whilst OSANN2 (subsidiary study) is a nested case-control study within the Kaiser population involving 217 female cases enrolled in 1964-77 and diagnosed in 1969-77.
8. TVERDA (principal study) contains results from a Norwegian cohort study involving 44,290 men and 24,535 women aged 35-49 examined in 5 areas of Norway between 1972 and 1978 and followed up until 1988. VEIERO (subsidiary study) contains results of follow up until 1983 of 25,956 men and 25,496 women in 3 of these areas examined a second time starting in 1977.
9. HEIN and LANGE (principal studies) contain results from, respectively, the Copenhagen Male Cohort study population, interviewed in 1970-71 and followed to 1988 and the Copenhagen City Heart Cohort study population, interviewed in 1976-78 and followed to 1989. PRESCO (subsidiary study) contains results from a combined analysis, with follow up to 1993, of data from 3 cohort studies in Copenhagen including the two referred to above. [Note that, because of the overlap, only data by age and

by amount smoked, not available for the principal studies, were entered on the database for PRESCO.]

10. WALD, BENSHL and HOLE (principal studies) contain results from follow up of 3 UK cohort studies, the BUPA study, the Whitehall study and the Paisley-Renfrew study. TANG2 (subsidiary study) contains results from a combined analysis, specifically with respect to the filter/plain comparison, of 4 UK cohort studies including the three referred to above.

11. GRAHAM and BROSS (principal studies) contains results from case-control studies conducted at the Roswell Park Memorial Institute in, respectively, 1956-60 and 1960-66. BYERS1 (subsidiary study) contains results from cases at the same institute interviewed in 1957-65. [Note that, because of the overlap, only data on histological type of lung cancer, not considered in the two principal studies, were entered on the database for BYERS1.]

12. WYNDER6 (principal study) contains results from a case-control study continuously ongoing in various US hospitals for 1969-96, whilst WYNDER5, WYNDER7 and WYNDER8 (subsidiary studies) contains results from the same study for interviews conducted in, respectively, 1969-76, 1977-84 and 1985-90.

    [Note that there were an extremely large number of papers covering results from various periods in various sets of hospitals and that results were selected to choose what seemed the most appropriate analyses, whilst attempting to avoid overlap. Note also that, when considering results for filter/plain, one can include results from both WYNDER5 and WYNDER6, without overlap.]

Note that for sets 9 and 10, the individual studies are chosen to be the principal studies as they were originally intended to be separate and the combined analysis occurred later. For sets 3 and 4, the combined studies are chosen to be the principal studies as they were planned as multicentre studies.

### 3.3.4 Study characteristics

Table 1 gives the distribution of various selected study characteristics by study type and overall. Note that the distributions are based, not on all the original 296 studies, but on the 277 principal studies, excluding the 19 subsidiary ones. For some variables, footnotes indicate where differing characteristics of subsidiary studies from their associated principal studies would have an effect on the distribution.

**Design** Of the 277 principal studies, 218 (78.7%) are of case-control design and 53 (19.1%) are of prospective design, with a further 5 (1.8%) being of nested case-control design and 1 (0.4%) being of case-cohort design. In both the nested case-control and the case-cohort design, cases are drawn from within a prospective study, the difference being that in the case-cohort design controls are selected at baseline while in the nested case-control design controls are matched to the cases after the disease is diagnosed.

**Sexes considered** In the majority of the studies, 58.1%, subjects included both sexes,[a] while 34.7% considered males only and 7.2% considered females only.

**Age of subjects** In the case of 187 studies, mostly of case-control design, there was no lower age limit on the study population or the lower age limit was not stated. Specific lower age limits were set in 90 studies (including 42 of the 53 prospective studies), the highest being 55. Thus, none of the studies were restricted specifically to the elderly.

In 198 studies, there was no upper age limit on the study population or the upper age limit was not stated. Two studies had an upper age limit of 49 so were restricted to relatively young subjects,[b] but no other studies had an upper age limit

[a] Including one study that had only male controls.

[b] One other study reported two non-contiguous age groups, the lower of which had an upper limit of 45.

less than 55. The upper age limit in prospective studies is based on the age at baseline, so clearly the subjects may have been older than this when they got lung cancer.

**Location** Studies were most commonly conducted in North America (33.6%), West Europe or Scandinavia (29.2%) and Asia (26.0%), and less commonly conducted in East Europe or the Balkans (5.1%), South or Central America (3.6%), Africa (1.8%) and Australasia (0.7%). It was notable that prospective studies were particularly likely to be conducted in North America (47.2%) and West Europe or Scandinavia (35.8%), with none at all being conducted in South or Central America or in Africa.

Of the 93 studies conducted in North America, 81 were conducted in the USA and 10 in Canada, with 2 involving both these countries.

Of the 81 studies conducted in West Europe or Scandinavia, 22 were conducted in the UK, 13 in Germany and 10 in Sweden with studies also conducted less commonly in a further 10 countries. There were 3 studies conducted in multiple countries (not considered in the counts for the individual countries).

Of the 72 studies conducted in Asia, 36 were conducted in China (excluding Hong Kong), 19 in Japan and 5 in Hong Kong with studies also conducted less commonly in a further 5 countries.

Of the 31 studies conducted in other areas, 5 were conducted in Poland, with no more than 3 in any other country.

Overall, studies were conducted in 37 countries.

**Race of subjects** In 232 studies, there was no selection on race though clearly variation in the location of the study would cause major variation in the racial distribution. In 28 studies (21 in North America, 6 in Europe and 1 in Africa), subjects were specifically restricted to whites, in one of these studies (in the US) hispanics being excluded from the definition of whites. In 11 studies, subjects were specifically restricted to one race (blacks in 4 African studies, Chinese in 2 Hong Kong and 2 Singapore studies, Japanese in 1 Japanese and 1 US study, and Scandinavian in 1 Swedish study. In 6 further studies (all in the US), certain races were included and others excluded.

**Timing** The earliest period considered by any study was JARUP which was a case-control study based on Swedish smelter workers dying in 1928-1981. The earliest non-occupational study[a] was DAVEYS, which contains results from two German case-control studies conducted in 1930-41. Six more case-control studies started between 1936-1940, and the number of studies starting gradually accelerated, with 15 studies starting in 1941-50, 32 in 1951-60, 41 in 1961-70, 65 in 1971-80 and 82 in 1981-90. Prospective studies did not start until the early 1950s.

**Study size** The distribution of the number of lung cancer cases was very skew with the median being 322. 239 studies involved less than 1000 cases, 33 between 1000 and 9999 cases and 4 studies 10000 or more cases. The three largest studies were all conducted in the USA, STOCKW involving 22161 cases in Florida, KELLER involving 15038 in Illinois and BROWN2 involving 14596 in Missouri. The largest study in Asia was LIU4, involving 10000 cases in China, whilst the largest in Europe were the multicentre studies LUBIN2 (7804 cases) and BOFFET (5621 cases). The prospective studies involving most cases were CPSI (5138 cases), DORN (5097 cases) and CPSII (3229 cases), CPSI and CPSII being the largest in terms of study population, both involving over a million persons.

[a] The study by Müller,[5] often cited as the first study to report the link between smoking and lung cancer, is not included as it had less than 100 cases.

**Population studied** Of the 270 studies where the type of persons studied was known, 193, mainly case-control studies, were of the general population with no restriction stated and 26, commonly prospective studies, were of people employed in specific industries. The remaining 52 studies involved a wide variety of populations. Some were very specific, e.g. Atomic bomb survivors, war veteran pensioners or Lutheran insurance holders, while others were general population subject to restrictions, e.g. English speaking, holding driving licence, long-term residents or volunteers for screening programs.

**Nature of controls** Of the 226 studies that were not prospective, 95 used only diseased (hospital) controls, 74 used only healthy (population) controls, 28 used decedent controls and 20 used a mixture of types of control. Additionally, 1 study (DORANT - the case-cohort study) used a subcohort as controls whilst 6 did not define their controls.

The 142 studies that described diseased and/or decedent controls were classified according to whether these controls included specific smoking-related and non smoking-related diseases. 44 of these (31.2%) included smoking-related cancers, while 52 (36.6%) included respiratory disease and 86 (60.6%) included some smoking-related diseases, often heart disease, not excluded from the control group in 77 (54.6%) studies.

**Proxy use** 77 studies (mainly case-control) obtained information at least partly from proxies, such as next-of-kin. Of these 77 studies, there were 33 where it was possible to establish that there was a substantial difference in proxy use between the cases and controls, with proxy use always being higher for cases than for controls.

**Case-control differences** Of case-control studies where the comparison was possible, in 15.5% the cases and controls came from different hospitals, in 21.8%

the cases and controls were interviewed in a different setting and in 15.4% the cases and controls were of different vital status.

**Histological confirmation** Full histological confirmation was carried out in 69 (24.9%) of studies. These are predominantly case-control studies, only 2 of 53 (3.8%) prospective studies insisting on full confirmation. None of the studies were based on autopsy diagnosis.

**Response rate** For the 39 prospective, or nested case-control and case-cohort studies for which this was known, the response rate varied between 23% and 100% with a median of 80%. For the 130 case-control studies for which the response rate for the cases was known, this varied between 27 % and 100% with a median of 90%. For the 110 case-control studies for which this was known, the response rate for the controls varied between 23% and 100% with a median of 89.0%. For the 107 case-control studies for which the response rate for both cases and controls were known, the rate did not differ significantly.

**Matching factors** As shown in Table 2, the commonest matching factors used in case-control studies were sex (72.1% of those studies which involved both sexes) and age (66.7% of studies). 39.7% of studies matched for factors such as interviewer, hospital, timing of interview etc, while 11.9% of studies, mainly in the USA, matched for race. Other factors were rarely matched for.

**Available results** Table 3 provides details on the extent to which studies provide information on some of the aspects of smoking that have been recorded on the relative risk database. Of the 277 principal studies, 31.0%, mainly case-control studies, provide data by histological type. Data on ever smoking[a] are available for 91.7% of the studies, with data on current smoking available for

[a] Includes studies where the definition of smoking excluded long-term ex smokers.

54.2% and ex-smoking for 46.6%. Ex-smoking is variously defined, as shown in Table 3. Data on amount smoked are available for 59.9% of the studies. The other aspects of cigarette smoking for which data are recorded on the relative risk database are available from a smaller percentage of studies, with data on cigarette type (filter/plain) available for 14.1%, on pipe and/or cigar smoking for 35.7% and on handrolled smoking[a] for 5.8%. Note that the definition of handrolled smoking and of pipe smoking varies according to local custom. Note also that for some principal studies where data on one of the listed aspects is not available, data may be available on occasion from the linked subsidiary studies.

**Further aspects of smoking** Table 4 provides details on the extent to which studies provide information on other aspects of smoking, for which data have not so far been recorded on the relative risk database. It is clear that substantial numbers of studies (>40) provide information on risk in relation to age of starting to smoke, inhalation, years since stopped smoking, duration of smoking and pack-years (the product of duration and amount smoked). There are also a moderate number of studies that provide information on risk in relation to the proportion of the cigarette smoked, other aspects of cigarette type than the simple filter/plain comparison for which data have so far been entered (e.g. tar level or time of product switch) and other types of tobacco (chewing tobacco and snuff). Some aspects of smoking have only been studied in very few studies.

**Other stratifying variables** So far only sex, age and race have been considered as stratifying variables in the relative risk database. However, some studies give details on how the association of smoking with lung cancer varies by level of other stratifying variables. Table 5 presents details of which stratifying variables have been considered in at least 3 studies. By far the commonest are risky occupational exposures, with 29 studies reporting results stratified on this. Results by occupation and by risky non-occupational exposure are also reported in 8 and 6 studies respectively. Other reasonably common stratifying variables

[a] Non-conventional manufactured cigarette smoking (bidi, pilli) is included with handrolled.

are region/residence/place of birth, diet, current/previous medical conditions and genetics/family history of lung cancer.

**Other diseases** There are 45 studies for which the papers in the file present results relating smoking to diseases other than lung cancer.

**Confounders** Table 6 provides information on the extent to which potential confounding variables have specifically been taken into account in analysis. Of the 277 studies, 110 (39.7%) did not adjust for any variable at all in analysis (though many of these will have matched for age and/or other factors at the design stage). Only 66 studies (23.8%) adjusted for 3 or more potential confounders, 17 being the largest number of factors taken account of in analysis.

Table 6 also shows all those variables taken account of in at least 3 studies. Age is by far the commonest, with 143 studies adjusting for it. The next most common are risky occupational exposures, considered in 27 studies, aspects of smoking considered in 26 studies (mainly those comparing risk in filter and plain cigarette smokers), education (21 studies) and race (19 studies).

## 3.4 The relative risk database

### 3.4.1 Structure of the database

As described in more detail in Appendix F, the relative risk database contains one record for each relative risk. Again, each record consists of “fields” within “cards.” The three cards used for data entry, together with a brief description of the fields included in each, are as follows:

**RR description** This includes an RR identification number which is unique within the study, together with details defining the RR. These include the sex, age range, race, lung cancer type and (for prospective studies) the follow-up period. The smoking exposure is defined by the smoking status (ever, current and ex), the smoking product (e.g. any, cigarettes only, cigarettes +/- others, pipes only), the cigarette type (e.g. any, manufactured cigarettes, manufactured cigarettes +/- handrolled, filter only, plain only, menthol), the lowest and highest number of cigarettes/day (e.g. 21 to 39 - both entered as 0 if the RR applies regardless of amount smoked) and the denominator (e.g. never smoker, nonsmoker, never cigarettes and, for filter/plain comparisons, ever plain, always plain, etc.). See Appendix F for fuller details of the possible levels of the grading systems used for smoking product, cigarette type and denominator.

**RR adjustment** This includes whether or not the RR is adjusted for sex, age, race or other confounders, and in the case of other confounders, the number of variables adjusted for. The actual other confounders adjusted for are given in a text comment if they are less than the full set already defined in the study database.

**RR data** For unadjusted results only, this includes the 2 × 2 table, i.e. the number of exposed and unexposed cases, and the number of exposed and unexposed controls, at-risk population or man-years at risk. For all results, it includes the RR estimate itself and its upper and lower 95% confidence limits. For unadjusted data the RR and 95% confidence limits are calculated from the

2 × 2 table (if available). For adjusted data, they may be as given in the source papers or as derived by other means, a further variable indicating the method of derivation. The possible methods of derivation are described in §3.4.5.

The record includes the six character study reference linking it to the corresponding record on the study database.

### 3.4.2 Identifying which relative risks to enter

In identifying what RRs to enter, four aspects – smoking index, lung cancer type, confounders adjusted for, and strata – were considered and these are discussed in the following sections. RRs relating to all combinations of these aspects were entered.

As discussed above, it is important in meta-analyses to avoid "double counting", and this applies equally within studies. Although in some circumstances it is quite legitimate for more than one RR from a study to be included in a meta-analysis (for instance by strata such as sex and contiguous age groups, and for independent disease groups compared to separate control groups), in other circumstances it is not (for instance when disease groups are compared to a shared control group, that control group would be double counted; and if current and ex smokers were each compared to never smokers, including both in a meta-analysis of ever smokers would double count the never smokers). For a simple stratifying variable, it is readily apparent at the analysis stage whether or not inclusion of multiple RRs is valid. However for the other aspects it is not. It was therefore decided that, with the exception of the straightforward strata of sex and race, all valid combinations would be constructed at the outset. This resulted in a considerably larger numbers of RRs being entered for some studies than had been presented in the original papers.

### 3.4.2.1 Smoking indices

For each RR it was necessary to define the smoking exposure of the numerator and of the denominator separately, exposure being defined according to four indices of major interest - smoking status, product, cigarette type and amount smoked.

When identifying the numerator, smoking status was defined as current smoker, ex smoker or ever smoker.

Product was defined as one of nine levels:

| | |
|---|---|
| 1 | all products |
| 2 | cigarettes (with or without other products) |
| 3 | other products but not cigarettes |
| 4 | cigarettes only |
| 5 | both cigarettes and others |
| 6 | other products (with or without cigarettes) |
| 7 | pipe only |
| 8 | cigar only |
| 9 | both pipe and cigar (but not cigarettes) |

Where the product related to cigarette smokers (i.e. product levels 2,4,5), one of the following levels was additionally selected to indicate cigarette type:

| | |
|---|---|
| 1 | all types |
| 2 | manufactured (with or without hand-rolled) |
| 3 | hand-rolled (with or without manufactured) |
| 4 | manufactured only |
| 5 | hand-rolled only |
| 6 | both hand-rolled and manufactured |
| 7 | mainly manufactured (but some hand-rolled) |
| 8 | mainly hand-rolled (but some manufactured) |
| 9 | menthol |
| 10 | filter only |
| 11 | plain only |
| 12 | mainly filter |
| 13 | mainly plain |
| 14 | always filter |
| 15 | always plain |

| | |
|---|---|
| 16 | ever filter |
| 17 | ever plain |
| 18 | both plain and filter |
| 19 | plain and filter equally |

The RR was further described as relating to the whole group of smokers so far defined (e.g. current smokers of manufactured cigarettes only) or to a category within that group by number of cigarettes smoked (e.g. 1-10, 11-20, etc. per day). The categories used vary considerably from study to study, and have been entered as given in the paper, except that in exceptional circumstances they may have been combined together (for instance if the study originally presented a very detailed breakdown for males but a less detailed breakdown for females, then the less detailed form may have been entered for both sexes). Where appropriate, the categories were noted to refer to numbers of "cigarette equivalents", not cigarettes, per day.

When identifying the denominator, attention was usually restricted to just four groups: never smokers, non smokers, never smoked cigarettes (of any type) and non smokers of cigarettes. [We use nonsmoker to refer to those not currently smoking, i.e. to never and ex smokers combined.] However when the numerator related to the smoking of filter, hand-rolled or menthol cigarettes then the denominator could also be defined as relating to plain, manufactured or non-menthol smokers respectively. Other denominators such as "never smoked or smoked <5 cigarettes per day" or "never smoked or gave up more than 10 years ago" were used only if none of the four main denominators were available.

All valid combinations of the above definitions of numerators and denominators were used. Thus RRs were entered for:

| | | | |
|---|---|---|---|
| current cigarette smoker | of each and every (cigarette) product group / cigarette type / amount | versus | never smoker |
| current cigarette smoker | " | " | non smoker |
| ex cigarette smoker | " | " | never smoker |
| ever cigarette smoker | " | " | never smoker |
| current other smoker | of each and every (other) product group / amount | versus | never smoker |
| current other smoker | " | " | non smoker |
| ex other smoker | " | " | never smoker |
| ever other smoker | " | " | never smoker |
| current cigarette smoker | of each and every (cigarette) product group / cigarette type / amount | versus | never cigarette smoker |
| current cigarette smoker | " | " | non cigarette smoker |
| ex cigarette smoker | " | " | never cigarette smoker |
| ever cigarette smoker | " | " | non cigarette smoker |

Also for

| | | | | |
|---|---|---|---|---|
| filter only | of each and every smoking status / (cigarette) product group / amount | versus | plain only/always | of the same smoking status / (cigarette) product group / amount |
| filter only | " | " | ever plain | " |
| filter only | " | " | mainly plain | " |
| mainly filter | " | " | plain only | " |
| always filter | " | " | plain only | " |
| ever filter | " | " | plain only | " |
| both plain and filter | " | " | plain only | " |
| plain and filter equally | " | " | plain only | " |

and similarly for the hand-rolled/manufactured comparison.

It may be useful to note some examples of combinations that have not been entered:

mixed products versus cigarettes only
pipe smoker versus non smoker of pipes
high amount versus low amount (except if no comparison with never/non smokers is available)

Note also that ever smoker versus non smoker would not be valid, as exsmokers would be counted in both the numerator and denominator.

For some studies the smoking product was poorly defined. This often arose for studies conducted in countries where cigarettes were the predominant product, so that papers (and also the original questionnaires used in the studies) refer only to “smoking” and, without specific local and historical knowledge, it is difficult to know whether this should be interpreted as meaning cigarettes only, or all products, or that these are in any case the same thing.

In addition, for studies where it was clearly stated that there were no smokers of other products, three of the product levels (all products, cigarettes and cigarettes only) are identical, as are certain smoking status levels for the denominator (never smoked and never smoked cigarettes; and non smoker and non smoker of cigarettes). Thus each basic result should in principle be entered under six separate definitions. However, this was considered to represent excessive duplication and instead the policy adopted was as follows:

a) If only “smoking” is referred to (even if tables are subdivided by numbers of cigarettes or pack-years) then define RRs as relating to any product (numerator) vs never/non smoker (denominator).

b) If the terms “cigarette smoking” and “smoking” are used as if synonymous with no mention at all of other products, then define RRs as relating to cigarettes vs never/non cigarettes,

c) But if there is any hint that other products were asked about in the questionnaire, then define the RRs as relating to cigarettes, vs either never/non smoked if the smokers of other products have been excluded from the analysis, or never/non smoked cigarettes if they have been included with the unexposed group (e.g. if pipe/cigar only smokers have been combined with the never cigarette smokers to form the denominator).

d) Only enter RRs using the “cigarettes only” level if mixed smokers have specifically been excluded (i.e. do not use it when there were no mixed smokers).

e) When there are no smokers of other products, enter as all products vs never/non smokers – with the exception that when there are male but no female smokers of other products, at least the main results for females should be entered twice, both for all products and for cigarettes.

3.4.2.2 Lung cancer type

Results were entered for all lung cancers, for Kreyberg I (as originally presented, or by combining squamous, small and large) and Kreyberg II (as originally presented, or by combining adenocarcinoma and other – i.e. confirmed but not squamous, small or large) and for squamous, small, large and adenocarcinoma separately. In addition, the following groups were constructed if not originally presented:

Squamous or nearest equivalent

Adenocarcinoma or nearest equivalent

All lung cancers or nearest equivalent – but at least squamous and adenocarcinoma.

[At the start of the project, other individual lung cancer types and groupings were entered, but this was subsequently discontinued.]

3.4.2.3 Confounders adjusted for

Results were entered unadjusted, and adjusted for the most available confounders. If available, then results adjusted for less confounders were also usually entered. Exceptionally, for prospective studies, unadjusted results were sometimes omitted provided that results adjusted for age only were available.

3.4.2.4 Strata

Three strata were considered – sex, age and race. Results were entered for males and females separately when available. Combined sex results were only entered when the equivalent results (i.e. for the same smoking indices, confounders, age and race) were not available. Results were entered both for all ages combined, and for individual age groups. The age groups used vary considerably from study to study, and have been entered as found, except that adjacent groups may have been combined together in exceptional circumstances, for instance to avoid very small number of cases. Results were entered for all races (if originally presented combined), and for individual racial groups. On occasion, results for combined races were also derived and entered.

### 3.4.3 Derivation of the relative risks

Adjusted RRs and their 95% CIs are entered as given when available. Unadjusted RRs are calculated from their 2 × 2 table, if available, otherwise entered as given. The 2 × 2 table may be constructed by summing groups (e.g. adding current and ex smokers to obtain ever smokers, or adding over lung cancer types, or adding over other stratifying factors), from a percentage distribution or from a matched pairs table. If the numbers of cases are denoted by $a_i$ and the numbers of controls (or the at risk population in a prospective study) by $b_i$, where the subscript $i$ = 0 refers to the unexposed (non smoking) group and $i$ = 1 refers to the exposed (smoking) group, then the RR and it confidence limits are calculated by:

$$\text{RR} = (a_1\, b_0) / (a_0\, b_1)$$
$$\text{LCL} = \text{RR} / \phi$$
$$\text{UCL} = \text{RR}\ \phi$$

where $\phi$, a factor based on the variance of the RR, is given by

$$\ln(\phi) = 1.96\sqrt{((1/a_0) + (1/a_1) + (1/b_0) + (1/b_1))} \quad \text{for a CC study,}$$

or $$\ln(\phi) = 1.96\sqrt{((1/a_0) + (1/a_1) - (1/b_0) - (1/b_1))} \quad \text{for a prospective study.}$$

If both a 2 × 2 table and an unadjusted RR/CI were presented originally, then the RR/CI calculated as above is used, and any discrepancy from that originally given is noted in the database.

A variety of other methods are used to provide estimates of the RR and CI in other circumstances. The main methods are described briefly here, and fuller details are given in Appendix G. Calculations were mainly carried out using Quattro Pro spreadsheets.

**Correction for zero cell** When a 2 × 2 table has one cell with value zero (which usually occurs when there are no never smoker cases) then the RR and CI cannot be calculated by the usual formula. The method used is to add a correction of 0.5 to each of the four cells, and then to apply the formula. [No calculation is made if a table has two zero cells. Usually such RRs have not been entered at all, for instance if they occur in a breakdown to very narrow age groups, then adjacent age groups would simply be combined. However they have been entered on a very few occasions, for instance if a product code such as pipe smoking has been entered for males but has no exposure for females.]

**Combining independent RRs** Combining RRs over strata uses the method of Fleiss and Gross,[1] the same method as for meta-analysis. The resulting estimate is adjusted for the stratifying variable. When this combined RR is subsequently used in a meta-analysis, the end result will be exactly the same as if all the original RRs had been included. This method is also used for combining RRs for individual diseases groups, provided they are independent estimates (i.e. each disease group has a separate control group)

**Combining non-independent RRs** When non-independent RRs are to be combined, for instance if adjusted RRs are available for current and ex smokers, each versus never smokers, then the method of Fry and Lee[2] is used to provide a combined estimate for ever smokers. This method starts from a source table giving adjusted RRs and CIs for $n$ smoking groups relative to a single nonsmoking base group. The hypothetical underlying 2 × ($n$ + 1) table of numbers of "adjusted cases and controls" is estimated, these then being summed to give the required groups for the numerator and denominator, and the resulting 2 × 2 table used with the usual formula to estimate the adjusted RR and CI. A variation of the method allows non-independent disease groups to be combined (i.e. when

RRs for several disease groups are given, each relative to a single shared control group).

**Ratio of rates** Prospective studies often present mortality rates for both exposed and unexposed groups. The RR is estimated simply by the ratio of the two rates.

**CI estimated from crude numbers** When an adjusted RR was presented originally without a CI, but the corresponding 2 × 2 table is available, then the original RR is used and its confidence interval is estimated by assuming its width is the same as the width of the interval for the equivalent unadjusted RR. In fact, the estimated interval will be narrower than the true one (since adjustment widens the interval[3]), and thus this method will increase the weight that the estimate is given when entered into a meta-analysis. However this will usually be a small effect and the only alternative is to omit the RR altogether from all meta-analyses.

### 3.4.4 Characteristics of the relative risks

A total of 9551 relative risks are entered on the database, of which 8934 relate to the principal studies and 617 to the subsidiary studies. Among the 277 principal studies, 22 (7.9%) have only one RR, and a further 103 (37.2%) have between 2 and 10 RRs, while 20 (7.2%) have over 100 RRs, the highest number being 428. The median number of RRs per principal study is 12.0 (Table 7).

Table 8 gives the distribution of various selected RR characteristics by study type and overall, based on all the 296 studies.

**Sex** 6365 (66.6%) RRs are for males, 2547 (26.7%) for females and 639 (6.7%) for sexes combined. Of the 161 principal studies that included both sexes, 45 (28.0%) gave no sex-specific RRs.

**Age** 7631 (79.9%) RRs refer to the full age range of the study. Among the remaining age-specific results, the lowest age group studied is <40, and the highest 80+. Only 46 (16.6%) of the principal studies give any age-specific results (with a further 3 studies having results in the associated subsidiary studies).

**Race** 7470 (78.2%) RRs refer to all races (within the country studied). Of those 2081 where a restriction applied, either in the conduct of the study or in the analysis, the majority (1576 – 75.5%) refer to whites. Only 13 (4.7%) of the principal studies (or their subsidiaries) gave separate results for different racial groupings.

**Lung Cancer type** 6612 RRs (69.2%) refer to lung cancers of all types, including most (2811 – 94.1%) of RRs from prospective studies. A further 1589 RRs refer specifically to one of the four main types (478 – squamous, 198 – small, 321 – large and 592 – adenocarcinoma), while there are 503 and 293 RRs for Kreyberg I and II groups respectively. Nearly all of the principal studies present results for all lung cancer types combined, with 8 (2.9%) others having a near equivalent (at least squamous and adeno), while 5 (1.8%) do not. Results for squamous (or combinations not including adeno) are available for 90 (32.5%) of the studies (either principal studies or their subsidiaries), while results for adeno (or combinations not including squamous) are available for 93 (33.6%) studies.

**Smoking status** For prospective studies, the majority of results refer to current smoking (i.e. current at baseline), with 2136 (71.5%) of RRs, and 45 (84.9%) of studies (or their subsidiaries) having such results. For the other study designs, there are somewhat more results for ever smoking, with 3492 (53.2%) RRs, than for current smoking, with 2480 (37.8%), and they are available for substantially more studies, 206 (92.0%) for ever smoking, and 104 (46.4%) for current smoking. There are fewer results for ex smoking among the other study designs (593 RRs), but they are still available for 89 (39.7%) of studies.

**Smoking product** Half the RRs (4801 – 50.3%) refer to smoking cigarettes (irrespective of whether other products were also smoked). Most of the others refer to smoking of any product (22.4%) or to smoking of cigarettes only (17.1%). Virtually all studies (270) have results for either cigarettes or all products, with the remaining 7 studies having results for cigarettes only.

**Cigarette type** Among the 6695 RRs relating to cigarettes, most (81.3%) refer to all types of cigarette, while 544 (8.1%) refer to either manufactured or hand-rolled cigarettes, 679 (10.3%) to filter or plain cigarettes and 23 (0.3%) to menthol cigarettes. These come from 19 (6.9%) principal studies (or their subsidiaries) for manufactured or hand-rolled cigarettes, 34 (12.3%) studies for filter or plain cigarettes and 4 (1.4%) studies for menthol cigarettes.

**Number of cigarettes smoked** 3896 (40.8%) RRs are for specific amounts smoked. These include 51 for occasional smokers (entered on the database as smoking 0.5 cigarettes per day), 68 for smoking a specific number of cigarettes (56 for smoking 20 per day and 12 for smoking 40 per day), 2571 for smoking in a specified range (the interval width having median 9.0 and maximum 44), and 1206 for smoking in an open-ended range (the lower boundary varying from 6 to 75 with median 25.0). 81.1% of prospective studies (or their subsidiaries) have amount-specific results, compared with 56.4% of case control studies.

**Denominator** 8806 RRs (92.2%) use the four main denominators, and 517 (5.4%) RRs from 42 (15.9%) studies have the denominators referring to specific cigarette types. Only 228 (2.4%) RRs have the non-standard denominators (which were only entered for comparisons where the main denominators were not available). 21 (7.6%) studies use these denominators, with 12 studies having no other denominators.

**Follow up period** 1458 (48.8%) RRs for prospective studies relate to the whole study period, with the remainder relating to some shorter follow up period. 25 (47.2%) of prospective studies present results for specific follow-up periods.

**Adjustment** 3739 RRs have some adjustment, including 61.7% from prospective studies and 28.9% from case-control or other studies. Of sexes combined RRs, 29.7% are adjusted for sex. Among the adjusted RRs, a large majority (95.1%) are adjusted for age, but relatively few for race (5.5%) or other factors (40.3%). The adjusted RRs come from 166 studies (or their subsidiaries) (83.0% of prospective studies and 55.4% of other studies) and 27 studies have only adjusted RRs.

**2 × 2 table** The full 2 × 2 table is available for 5562 (95.7%) of the unadjusted RRs. Of these, 183 have one zero cell and 7 have two zero cells.

**RR and CI** 3 RRs have no values for the RR or CI but only a statement of non-significance (as well as the 7 RRs mentioned above which have no values for the RR or CI because they have two zero cells in the 2 × 2 table). 325 RRs have a RR value but no CI, including 22 with a statement of significance or non-significance. There are 7 studies which have no complete RR/CIs. The RR values range from 0 to 316.

**Derivation method** 3320 (34.8%) RRs are either as given originally or are calculated directly from the numbers in the 2 × 2 table. For a further 160 (1.7%) RRs where both the 2 × 2 table and the RR and CI were originally available, the RR and CI are recalculated because of a discrepancy, 3172 (33.2%) are calculated after summing categories to obtain a 2 × 2 table, and 183 (1.9%) are calculated using a zero cell correction. 220 (2.3%) RRs are calculated by other straightforward methods (inverting, converting from 90% CI, symmetry, ratio of rates, SMRs, expected values, combining from independent estimates). The method of Fry and Lee[2] for combining non-independent estimates is used for 696 (7.3%) RRs. Other methods, or combinations of methods (but not estimation of adjusted CIs from crude numbers) are used for 655 (6.9%) RRs. The remaining 1134 (11.9%) RRs involve estimation of the CI from crude numbers, and two studies have only RRs with this type of estimation.

## 4. Carrying out meta-analyses

### 4.1 Selecting the relative risks for the meta-analyses

The process of selecting which RRs to include in an analysis can be quite complex as it has to address two main objectives – to include all the relevant data but at the same time to avoid double counting. The rules used when entering data will ensure that double counting is avoided if (1) within each study, values of the stratifying fields (sex, age, race) are non-overlapping; (2) within each strata only one set of values of the smoking indices, the lung cancer type, the follow-up period and the number of confounders adjusted for is chosen; and (3) either a principal study or its subsidiary but not both are included.

When defining the relevant data for a particular analysis, it may be possible to choose a single specific value of a smoking index (e.g. for an analysis of ‘pipe only’ smokers). Only RRs with that value will be included, and studies without any such RRs will be excluded altogether. However more commonly, a number of values may be acceptable in the analysis (e.g. in an analysis of filter versus plain cigarette smoking, ‘only filter’, ‘mainly filter’ and ‘ever filter’ may all be acceptable). An order of priority is defined, so that one value only will be chosen from those studies which had RRs entered for more than one acceptable value. In a similar way, preferred values of lung cancer type can be chosen, and the number of adjusting variables can be chosen to be the minimum or maximum available.

The choice between principal and subsidiary studies can be specified in a similar way, except that the preference is now implemented over the group of linked studies. RRs from the subsidiary study will only be allowed if there are no eligible RRs from the principal study.

For the stratifying variables of age and race, RRs may have been entered on the database for the whole study, or for individual strata, or both. For many analyses, results for the whole study will be preferred if available. However where only strata-specific RRs are available then the widest available strata will be preferred. For example, if a study included ages 25+, but reported filter/plain results only for ages 35-74, and moreover additionally presented these results split into age groups 35-44, 45-54, 55-64 and 65-74, then an analysis of filter/plain irrespective of age would choose the RR for age 35-74, whereas a filter/plain analysis restricted to subjects aged up to 55 would include the two RRs for ages 35-44 and 45-54.

When specifying ‘preferences’ on a number of fields, the order in which they are implemented may affect the outcome. For instance, suppose an analysis of cigarette smoking for squamous LC is required. The smoking exposures ‘cigarettes (regardless of other products)’ and ‘cigarettes only’ are defined as 1st and 2nd preferences respectively, as are LC types ‘squamous’ and ‘KI’. Further supposing that a study has two RRs, (1) for ‘cigarettes (regardless of other products)’ and ‘KI’, and (2) for ‘cigarettes only’ and ‘squamous’. If the preference on smoking is implemented first, then RR 1 will be chosen, whereas if the preference on LC type is implemented first, then RR 2 will be chosen. Therefore, attention is first restricted to those RRs which have acceptable values for all the preferencing fields. Preferences for the most important aspects of the analysis, usually the smoking exposures, are implemented next, while the less important aspects, usually the demographic strata and the principal/subsidiary study status, are implemented later.

It was decided at the outset that single-sex results would be preferred to combined-sex results, and the latter have only been entered on the database when the former are not available. For single-sex results, the smoking results that are available are sometimes different for the two sexes (e.g. a study may present male results for many different product definitions, but restrict female results to cigarette only smokers; or a principal study may present only male results while a subsidiary has results for both sexes). For these reasons, all setting of preferences is done within sex, and then the choice between sex-specific or sexes-combined is implemented afterwards. A further complication is that some studies present unadjusted results for the separate sexes but adjusted results only for sexes combined, or other combinations. To handle this situation, the final stage is to choose in the following order of preference:

for an analysis of 'most adjusted' – both MA and FA; CA; both MA and FU; both MU and FA; MA; FA; both MU and FU; CU; MU; FU.

for an analysis of 'least adjusted' – both MU and FU; CU; both MA and FU; both MU and FA; MU; FU; both MA and FA; CA; MA; FA.

(where U and A refer to least and most adjusted results respectively, and M, F and C refer to males, females and sexes combined).

## 4.2 Combining the relative risks

The method used to carry out the meta-analysis of the selected relative risks is as described by Fleiss and Gross[1]. Both fixed-effects and random-effects meta-analysis have been carried out to form combined estimates of the individual independent risks. Fixed-effects meta-analysis assumes a common underlying relative risk estimate and only takes into account within-study variability in calculating the combined relative risk estimate and its 95% confidence limit. Random-effects meta-analysis also takes into account between-study variability. Where there is no evidence of heterogeneity between the sets of estimates, the two analyses give the same results.

The notation used in some of the output is the same as that used by Fleiss and Gross[1]. Thus we have:

N the number of relative risks being combined

NS the number of studies from which the relative risks are taken (except when the analysis is subdivided into factor levels (see "Section 3" in §4.3) NS in the Total column is the sum of the values in the individual columns, i.e. the number of study × factor levels from which the relative risks are taken) s the individual relative risk estimate being combined (s = 1, …N)

$Y_s$ the logarithm of relative risk estimate s

$W_s$ the associated weight, calculated as the inverse of the variance of the logarithm of the relative risk

Fixed RR the fixed effects relative risk estimate, calculated by
exp ((3 $W_sY_s$)/( 3 $W_s$) = exp ($\overline{Y}$)
summation being over s = 1, … N

Fixed RRl the lower 95% confidence limit of the fixed effects relative risk estimate, calculated by exp($\overline{Y}$ -1.96/$\sqrt{\sum W_s}$ )

Fixed RRu the upper 95% confidence limit of the fixed effects relative risk estimate, calculated by exp($\overline{Y}$ +1.96/$\sqrt{\sum W_s}$ )

$Q_s$ the study's contribution to the heterogeneity estimate, calculated by $W_s(Y_s - \overline{Y})^2$. Where N is large, this can be regarded as a chisquared on 1 d.f..

$P_s$ the associated probability value, used to indicate outliers.

Het Chi (or Q in Fleiss and Gross notation) the heterogeneity chisquared on N-1 d.f., calculated by 3$Q_s$. If Q $\leq$ N-1, the random effects and fixed effects estimates are the same, but if Q > N-1 they differ.

Het df the degrees of freedom corresponding to Het Chi (= N-1)

Het P the probability value associated with Het Chi

Random RR,
Random RRl,
Random RRu The random effects relative risk estimate and its lower and upper 95% confidence limits. The method for deriving this, originally described by DerSimonian and Laird[4], is most conveniently given by Fleiss and Gross[1].

Note that the method of testing for heterogeneity between individual relative risk estimates can also be used for testing for heterogeneity between sets of combined estimates, e.g. between different locations. This is shown on the output as Between Chi, Between df and Between P. When more than two sets are compared, each pair-wise comparison is shown, except that where the sets represent increasing levels of a factor (e.g. start year), each set is compared with the first, and the chisquared value, degrees of freedom and associated probability value are also shown for a test of linear trend. The trend statistic is defined as $T = 3w_jd_j(m_j\text{-}\overline{m})$ where $w_j$ is the combined weight for level j of the factor, $m_j$ is the logarithm of the fixed-effects relative risk estimate for that level, $\overline{m}$ is the logarithm of the combined estimate over levels and $d_j$ is a "dose" variable, taken to be 1,2,3 … for successive levels of j. Note that all the between p values, for pairwise comparisons, trend or overall variation between factor levels are based on fixed-effects assumptions.

## 4.3 Detailed output

For each meta-analysis, the full detailed output comes in eight sections preceded by a cover page. All the pages for the meta-analysis are given the same main table number and main heading (describing the analysis), with the section number blank for the cover page and 1 to 8 for the specific section (e.g. Table 3-5 is section 5 within Table 3). The content of each section is as follows:

Cover page : This shows

(i) restrictions on the data included,

(ii) the order of preference for selecting relative risks to be included, and

(iii) a short description of the contents of the table

Note that Sections 1 to 3 concern adjusted data, with relative risks adjusted for the most potential confounders chosen from a study, while Sections 4 to 6 concern unadjusted data, with relative risks adjusted for the least potential confounders chosen from a study.

Section 1 : For each adjusted relative risk selected, a listing of the relevant characteristics of those relative risks. This includes the values of all the variables used to select the relative risk and used as “factors” in Section 3, as well as the two key identifiers of the relative risk: the study 6-character reference (REF) and the number of the relative risk within that study (NRR).

Section 2 : For each adjusted relative risk selected, the output shows in the first part of the section the sex, the number of potential confounding variables adjusted for, the 2 × 2 table of results (where available), the relative risk with its 95% CI, and in the second part of the section $Y_s$, $W_s$, $Q_s$ and $P_s$ (as defined in §4.2). Where multiple independent estimates are available for a study (typically different sexes or age groups), combined results are also shown for the study. Note that the 2 × 2 table is headed “exposed/non-exposed” × “case/control”. Exposed and non-exposed are as defined in the cover page and include any comparison (e.g. filter vs plain). Control will be numbers at risk or man-years for prospective studies, indicated by an asterisk (*) in the left-hand margin. Relative risks calculated by adding 0.5 to each cell (where a zero is present) are indicated by a tilde (~).

Section 3 : This gives the results of fixed effects and random effects meta-analyses of the adjusted data. For the overall data and for data subdivided by sex, and for data subdivided by various other factors within sex, the output shows, for each factor level, the number of estimates combined (N), the number of studies from which these estimates come (NS), the combined weight for the studies combined (Wt) as well as the relative risks and confidence limits themselves (RR, RRl, RRu) and coded P values testing for heterogeneity and for variation between factor level: P values are coded as +++, --- or *** p<0.001; ++, -- or ** p<0.01; +, - or * p<0.05; (+), (-) or (*) p<0.1 and N.S. p>0.1, with plus signs indicating significant positive differences or relative risks greater than 1, minus signs indicating significant negative differences or relative risks less than 1, and asterisks indicating significant non-directional heterogeneity.

Sections 4 to 6 : As Sections 1 to 3 but for unadjusted data.

Section 7 : This lists the studies excluded from consideration, together with information on the stage at which they were excluded. The stage refers back to the various restriction and selection stages described in the cover page. A study is excluded when no relative risk can be found to satisfy the criteria required.

Section 8 : This lists potentially overlapping studies for which data have been included, and also any results which would have been included in preference except that they had incomplete data (typically a relative risk without confidence interval).

Note that the main results are given in Sections 3 and 6 while Sections 1, 2, 4, 5, 7 and 8 mainly provide detailed information only required when one wants to see the individual estimates or to check the program is correctly selecting the data. Accordingly, when results are presented, the full output is shown in Appendices with only selected parts of the Section 3 and 6 results given in the main tables.

An example full output is shown in Appendix H.

Results of the analyses are described separately in Part II of this report.

## TABLE 1 Characteristics of the 277 principal studies

| Characteristic | Level | Study type<br>Case-control | Prospective | Nested Case-control[a] | Total |
|---|---|---|---|---|---|
| Total | | 218 | 53[b] | 6 | 277 |
| Sexes considered : | Males only | 64[c] | 29 | 3 | 96 |
| | Females only | 18 | 2 | 0 | 20 |
| | Both | 136[d] | 22 | 3 | 161 |
| Lowest age in study : | Not known | 8 | 1 | 1 | 10 |
| | Unrestricted | 159 | 3 | 1 | 163 |
| | Unrestricted, but has been employed | 5 | 7 | 0 | 14 |
| | 16-24 | 3 | 3 | 0 | 6 |
| | 25-34 | 25 | 12 | 0 | 37 |
| | 35-44 | 17 | 18 | 0 | 35 |
| | 45-55 | 1 | 9 | 2 | 12 |
| Highest age in study : | Not known | 8 | 1 | 1 | 10 |
| (at baseline for | Unrestricted | 162 | 24 | 2 | 188 |
| prospective studies) | 49-64 | 2 | 13 | 0 | 15 |
| | 65-69 | 7 | 6 | 1 | 14 |
| | 70-74 | 7 | 3 | 2 | 12 |
| | 75-79 | 16 | 2 | 0 | 18 |
| | 80+ | 16 | 4 | 0 | 20 |
| Continent : | North America | 66 | 25 | 2 | 93 |
| | West Europe/Scandinavia | 60 | 19 | 2 | 81 |
| | Asia | 65 | 6 | 1 | 72 |
| | East Europe/Balkans | 11 | 2 | 1 | 14 |
| | South or Central America | 10 | 0 | 0 | 10 |
| | Africa | 5 | 0 | 0 | 5 |
| | Australasia | 1 | 1 | 0 | 2 |
| Country within : | USA | 57 | 22 | 2 | 81 |
| North America | Canada | 8 | 2 | 0 | 10 |
| | USA and Canada | 1 | 1 | 0 | 2 |
| Country within : | Multiple countries | 3[e] | 0 | 0 | 3 |
| West Europe/ | UK | 13 | 9 | 0 | 22 |
| Scandinavia | Germany | 13 | 0 | 0 | 13 |
| | Sweden | 9 | 1 | 0 | 10 |
| | Finland | 4 | 2 | 1 | 7 |
| | Italy | 6 | 0 | 0 | 6 |
| | Norway | 2 | 3 | 0 | 5 |
| | Denmark | 0 | 3 | 0 | 3 |
| | France | 3 | 0 | 0 | 3 |
| | Netherlands | 2 | 0 | 1 | 3 |
| | Spain | 2 | 0 | 0 | 2 |
| | Switzerland | 2 | 0 | 0 | 2 |
| | Iceland | 0 | 1 | 0 | 1 |
| | Belgium | 1 | 0 | 0 | 1 |

TABLE 1 Characteristics of the 277 principal studies (Continued)

| Characteristic | Level | Study type | | | |
|---|---|---|---|---|---|
| | | Case-control | Prospective | Nested Case-control[a] | Total |
| Country within : Asia | China (not Hong Kong) | 33 | 3 | 0 | 36 |
| | Japan | 16 | 2 | 1 | 19 |
| | Hong Kong | 5 | 0 | 0 | 5 |
| | India | 4 | 0 | 0 | 4 |
| | Taiwan | 4 | 0 | 0 | 4 |
| | Singapore | 2 | 0 | 0 | 2 |
| | South Korea | 1 | 0 | 0 | 1 |
| | Thailand | 1 | 0 | 0 | 1 |
| Country within : East Europe/Balkans | Poland | 5 | 0 | 0 | 5 |
| | Greece | 3 | 0 | 0 | 3 |
| | Hungary | 1 | 1 | 0 | 2 |
| | Turkey | 2 | 0 | 0 | 2 |
| | Czechoslovakia | 0 | 1 | 0 | 1 |
| | USSR (Russia) | 0 | 0 | 1 | 1 |
| Country within : South or Central America | Argentina | 3 | 0 | 0 | 3 |
| | Brazil | 2 | 0 | 0 | 2 |
| | Cuba | 2 | 0 | 0 | 2 |
| | Uruguay | 2 | 0 | 0 | 2 |
| | Colombia | 1 | 0 | 0 | 1 |
| Country within : Africa | South Africa | 3 | 0 | 0 | 3 |
| | Zimbabwe/Rhodesia | 2 | 0 | 0 | 2 |
| Country within : Australasia | Australia | 1 | 1 | 0 | 2 |
| Races considered : | Not known | 2 | 0 | 0 | 2 |
| | Unrestricted | 178 | 47 | 5 | 230 |
| | Whites (including hispanics) | 21 | 5 | 1 | 27 |
| | Whites (excluding hispanics) | 1 | 0 | 0 | 1 |
| | Blacks | 4 | 0 | 0 | 4 |
| | Chinese | 4 | 0 | 0 | 4 |
| | Japanese | 1 | 1 | 0 | 2 |
| | Scandinavian | 1 | 0 | 0 | 1 |
| | Various combinations (not all) | 6 | 0 | 0 | 6 |
| Start year of study[f] : | Not known | 16 | 0 | 0 | 16 |
| | 1928-1930 | 2 | 0 | 0 | 2 |
| | 1931-1940 | 6 | 0 | 0 | 6 |
| | 1941-1950 | 14 | 1 | 0 | 15 |
| | 1951-1955 | 7 | 9 | 0 | 16 |
| | 1956-1960 | 10 | 5 | 1 | 16 |
| | 1961-1965 | 11 | 10 | 0 | 21 |
| | 1966-1970 | 9 | 10 | 1 | 20 |
| | 1971-1975 | 13 | 8 | 1 | 22 |
| | 1976-1980 | 39 | 4 | 0 | 43 |
| | 1981-1985 | 33 | 3 | 1 | 37 |
| | 1986-1990 | 41 | 2 | 2 | 45 |
| | 1991-1997 | 17 | 1 | 0 | 18 |

TABLE 1 Characteristics of the 277 principal studies (continued/2)

| Characteristic | Level | Study type | | | |
|---|---|---|---|---|---|
| | | Case-control | Prospective | Nested Case-control[a] | Total |
| Numbers of lung cancers : | Not known | 1 | 0 | 0 | 1 |
| | 100-249 | 85 | 25 | 3 | 113 |
| | 250-499 | 61 | 14 | 2 | 77 |
| | 500-999 | 40 | 8 | 1 | 49 |
| | 1000-2499 | 21 | 2 | 0 | 23 |
| | 2500-4999 | 3 | 2 | 0 | 5 |
| | 5000-9999 | 3 | 2 | 0 | 5 |
| | 10000- | 4 | 0 | 0 | 4 |
| Nature of controls : | Not applicable | 0 | 53 | 0 | 53 |
| | Not known | 6 | - | 0 | 6 |
| | Diseased (hospital) | 95 | - | 0 | 95 |
| | Healthy | 70 | - | 4 | 74 |
| | Decedents | 27 | - | 1 | 28 |
| | Healthy + Diseased/Decedents | 16 | - | 0 | 16 |
| | Diseased + Decedents | 4 | - | 0 | 4 |
| | Subcohort | 0 | - | 1 | 1 |
| Diseased/decedent : controls | Total | 142 | - | 1 | 143 |
| | Inclusions not known | 1 | - | 0 | 1 |
| | Include smoking related cancer | 44 | - | 0 | 44 |
| | Include respiratory disease | 52 | - | 1 | 53 |
| | Include heart disease | 77 | - | 1 | 78 |
| | Include other smoking related disease | 78 | - | 1 | 79 |
| | Include non smoking related cancer | 91 | - | 0 | 91 |
| | Include orthopaedic/trauma patients | 116 | - | 1 | 117 |
| | Include other non smoking related disease | 110 | - | 1 | 111 |
| | Include any smoking related disease | 86 | - | 1 | 87 |
| Type of population : | Not known | 7 | 0 | 0 | 7 |
| | General population (no restriction stated) | 177 | 14 | 2 | 193 |
| | General population but with minimum residence time restriction | 15 | 0 | 0 | 15 |
| | Employed in specific industries | 10 | 14 | 2 | 26 |
| | Other restrictions[g] | 9 | 25 | 2 | 36 |
| Proxy use : | Not known | 12 | 0 | 0 | 12 |
| | No[h] | 131 | 51 | 6 | 188 |
| | Yes | 75 | 2 | 0 | 77 |
| Cases and controls : from different hospitals | Not applicable | 104 | 53 | 6 | 163 |
| | Not known | 11 | - | - | 11 |
| | No | 87 | - | - | 87 |
| | Yes | 16 | - | - | 16 |

TABLE 1 Characteristics of the 277 principal studies (continued/3)

| Characteristic | Level | Study type<br>Case-control | Prospective | Nested<br>Case-control[a] | Total |
|---|---|---|---|---|---|
| Cases and controls :<br>different interview setting | Not applicable | 0 | 53 | 0 | 53 |
| | Not known | 25 | - | 0 | 25 |
| | No | 151 | - | 6 | 157 |
| | Yes | 42 | - | 0 | 42 |
| Cases and controls :<br>different vital status[i] | Not applicable | 0 | 53 | 0 | 53 |
| | Not known | 10 | - | 0 | 10 |
| | No | 176 | - | 6 | 182 |
| | Yes | 32 | - | 0 | 32 |
| Full histological confirmation : | No | 153 | 51 | 4 | 208 |
| | Yes | 65 | 2 | 2 | 69 |
| Response rate at baseline : | Not applicable | 218 | 0 | 0 | 218 |
| | Not known | - | 17 | 3 | 20 |
| | 23-50 | - | 3 | 2 | 5 |
| | 51-70 | - | 7 | 0 | 7 |
| | 71-80 | - | 8 | 0 | 8 |
| | 81-90 | - | 7 | 0 | 7 |
| | 91-99 | - | 10 | 0 | 10 |
| | 100 | - | 1 | 1 | 2 |
| Response rate for cases : | Not applicable | 0 | 53 | 1 | 54 |
| | Not known | 88 | - | 3 | 91 |
| | 27-50 | 6 | - | 0 | 6 |
| | 51-70 | 16 | - | 0 | 16 |
| | 71-80 | 21 | - | 0 | 21 |
| | 81-90 | 24 | - | 1 | 25 |
| | 91-99 | 39 | - | 0 | 39 |
| | 100 | 24 | - | 1 | 25 |
| Response rate for controls : | Not applicable | 0 | 53 | 1 | 54 |
| | Not known | 108 | - | 3 | 111 |
| | 23-50 | 6 | - | 0 | 6 |
| | 51-70 | 17 | - | 0 | 17 |
| | 71-80 | 16 | - | 0 | 16 |
| | 81-90 | 17 | - | 0 | 17 |
| | 91-99 | 34 | - | 1 | 35 |
| | 100 | 20 | - | 1 | 21 |

## TABLE 1 Characteristics of the 277 principal studies (continued/4)

[a] Includes one case-cohort study.

[b] In three of the prospective studies, additional data are available from subsidiary studies of case-control or nested case-control design (see §3.4).

[c] In one of the case-control studies, the principal study reports data only for males but the related subsidiary study reports data for both sexes. Includes two studies for which sex was not stated but has been assumed males only.

[d] Includes one study with cases of both sexes and male only controls.

[e] The three studies and the countries included are:

| | |
|---|---|
| LUBIN2 | - Austria, France, Germany, Italy, UK/Scotland |
| BOFFET | - Germany, Italy, Sweden |
| POFFIJ | - Belgium, France, Germany, Luxemburg. |

(These countries are not included in the counts shown in the Table.)

[f] For retrospective case control studies, refers to the earliest deaths included.

[g] Includes volunteers for screening programs.

[h] Includes routinely collected data such as hospital or employment records.

[i] For retrospective studies where information may have been collected from several sources, refers to the main source of smoking data.

TABLE 2 Matching factors used in principal case-control studies (including nested case-control studies)

| Factor | Relevant studies with data | Matched | % |
|---|---|---|---|
| Sex | 136[a] | 98 | 72.1 |
| Age | 219 | 146 | 66.7 |
| Race | 219 | 26 | 11.9 |
| Socioeconomic status | 219 | 3 | 1.4 |
| Urban/rural residence | 219 | 3 | 1.4 |
| Education | 219 | 1 | 0.5 |
| Occupation | 219 | 3 | 1.4 |
| Marital status | 219 | 1 | 0.5 |
| Other factors[b] | 219 | 87 | 39.7 |

[a] Restricted to studies which involved both sexes.

[b] These are factors such as interviewer, hospital, timing of interview, etc.

TABLE 3 Results on smoking presented on relative risk data base (277 principal studies)

| Results by/for | Case-control studies | Prospective studies | Nested case-control studies[a] | Total |
|---|---|---|---|---|
| Total studies | 218 | 53 | 6 | 277 |
| Histological type | 75 | 9 | 2 | 86 |
| Ex-smokers | 85 | 40 | 4 | 129 |
| Current smokers | 98 | 45 | 6 | 149 |
| Ever smokers | 202 | 48 | 4 | 254 |
| Cigarette type | 33 | 6 | 0 | 39 |
| Handrolled cigarettes | 14 | 2 | 0 | 16 |
| Pipe smokers | 21 | 14 | 0 | 35 |
| Cigar smokers | 14 | 11 | 0 | 25 |
| Pipe and cigar smokers (not separately) | 27 | 11 | 1 | 39 |
| Amount smoked | 123 | 39 | 4 | 166 |
| Handrolled defined as: | | | | |
| Western handrolled | 8 | 2 | 0 | 10 |
| Bidi | 2 | 0 | 0 | 2 |
| Black tobacco handrolled | 1 | 0 | 0 | 1 |
| Chinese tobacco handrolled | 2 | 0 | 0 | 2 |
| Pilli (with cardboard tube) | 1 | 0 | 0 | 1 |
| Type of pipe: | | | | |
| Western pipe | 44 | 21 | 1 | 66 |
| Water pipe | 1 | 1 | 0 | 2 |
| Bamboo, water or long stem | 0 | 1 | 0 | 1 |
| Kizami | 2 | 0 | 0 | 0 |
| Definition of ex-smoker: | | | | |
| Unspecified | 22 | 12 | 4 | 38 |
| Given up for any period | 14 | 21 | 0 | 35 |
| Gave up at least 1 month | 1 | 3 | 0 | 4 |
| Gave up at least 1 year | 27 | 2 | 0 | 29 |
| Gave up at least 2 years | 11 | 1 | 0 | 12 |
| Gave up at least 5 years | 5 | 1 | 0 | 6 |
| Gave up at least 10 years | 5 | 0 | 0 | 5 |

[a] Includes one case-cohort study.

TABLE 4 Other aspects of smoking for which results are available but which have not been entered on the relative risk database (principal studies[a])

| Aspect of smoking | Studies |
|---|---|
| Pack-years[b] | 76 |
| Duration of smoking | 76 |
| Age of starting to smoke | 58 |
| Years since stopped smoking | 57 |
| Inhalation (depth or frequency) | 40 |
| Tar level | 10 |
| Fraction of length smoked, or butt length | 10 |
| Black/blond tobacco | 8 |
| Product switch | 7 |
| Filter/plain switch (time or proportion) | 6 |
| Chewing tobacco | 4 |
| Snuff | 3 |
| First cigarette (time after waking, or before breakfast) | 3 |
| Change in amount smoked | 2 |
| Cigarette size | 2 |
| CO delivery | 2 |
| Snuff and/or chewing (combined) | 2 |
| Reasons for giving up | 2 |
| Maximum number smoked | 2 |
| Cigarette holder use | 1 |
| Smoking index[c] | 1 |
| Nonsmoking interval | 1 |
| Removal from mouth between puffs | 1 |
| Imported/national cigarettes | 1 |

[a] In addition, recent change in smoking habit, and nicotine delivery were each available for 1 subsidiary study.

[b] Or other indices representing the product of duration of smoking and amount smoked.

[c] Amount × Duration/Age of starting.

TABLE 5 Stratifying variables (other than sex, age and race) considered in more than two studies

| Stratifying variable | Studies |
|---|---|
| Risky occupational exposures | 29 |
| Regions within study area | 13 |
| Dietary factors | 13 |
| Urban/rural residence | 12 |
| Current medical condition | 12 |
| Occupation | 8 |
| Genetic factors | 8 |
| Cooking/heating practices | 6 |
| Risky exposure (non-occupational) | 6 |
| Previous medical history | 6 |
| Religion | 5 |
| Socio-economic status | 5 |
| Place of birth | 4 |
| Family history of lung cancer | 4 |
| Alcohol consumption | 4 |
| Air pollution | 4 |
| Site of lung cancer | 3 |

TABLE 6 Confounders taken into account (277 principal studies)

| | Number | Studies |
|---|---|---|
| Total number of confounding variables considered per study: | 0 | 111 |
| | 1 | 57 |
| | 2 | 45 |
| | 3 | 22 |
| | 4-5 | 20 |
| | 6-7 | 18 |
| | 8-10 | 3 |
| | 11-17 | 3 |

Specific factors adjusted for in 3 or more studies:

| Factor | Studies | Factor | Studies |
|---|---|---|---|
| Age | 143 | Current medical conditions | 6 |
| Risky occupational exposure | 27 | Non-alcoholic drinks (tea, coffee, etc.) | 5 |
| Aspects of smoking | 26 | ETS exposure | 5 |
| Education | 21 | Body mass index | 5 |
| Race | 19 | Source of cases | 4 |
| Region | 15 | Religion | 4 |
| Medical history | 13 | Hospital admission time | 4 |
| Socio-economic status | 12 | Family history of lung cancer | 3 |
| Diet | 12 | Marital status | 3 |
| Sex | 11 [a] | Family history of cancer | 3 |
| Urban/rural residence | 10 | Hospital | 3 |
| Alcohol consumption | 9 | Time period in study | 3 |
| Occupation (general) | 8 | Air pollution | 3 |
| Residence | 7 | Date of death | 3 |
| Cooking/heating | 7 | | |

[a] Restricted to 45 studies which presented only combined sex results.

TABLE 7 Numbers of relative risks per study

| Number | Principal studies | Subsidiary studies |
|---|---|---|
| 1 | 22 | 1 |
| 2-5 | 59 | 2 |
| 6-10 | 44 | 5 |
| 11-15 | 28 | 1 |
| 16-20 | 14 | 1 |
| 21-50 | 64 | 5 |
| 51-100 | 26 | 3 |
| 101-200 | 14 | 1 |
| >200 | 6 | 0 |
| Median | 12.0 | 16.0 |

## TABLE 8 Characteristics of the 9551 relative risks

| | | Study type | | | |
|---|---|---|---|---|---|
| Characteristic | Level | Case-control | Prospective | Nested case-control[a] | Total |
| Total | | 6407 | 2986 | 158 | 9551 |
| Sex | Combined | 578 | 34 | 27 | 639 |
| | Male | 3953 | 2354 | 58 | 6365 |
| | Female | 1876 | 598 | 73 | 2547 |
| Age | All in study | 5775 | 1710 | 146 | 7631 |
| | Age-specific | 632 | 1276 | 12 | 1920 |
| Race | All in country | 5155 | 2161 | 154 | 7470 |
| | White (including hispanics) | 769 | 803 | 4 | 1576 |
| | Black | 240 | 4 | 0 | 244 |
| | Chinese | 138 | 0 | 0 | 138 |
| | Japanese | 40 | 10 | 0 | 50 |
| | Other and combinations | 65 | 8 | 0 | 73 |
| Lung cancer type | All | 3710 | 2811 | 91 | 6612 |
| | Squamous | 436 | 32 | 10 | 478 |
| | Large | 184 | 7 | 7 | 198 |
| | Small/oat | 282 | 31 | 8 | 321 |
| | Adeno | 527 | 55 | 10 | 592 |
| | Kreyberg I | 478 | 5 | 20 | 503 |
| | Kreyberg II | 277 | 4 | 12 | 293 |
| | Other including squamous and adeno | 148 | 2 | 0 | 150 |
| | Other including squamous but not adeno | 180 | 39 | 0 | 219 |
| | Other including adeno but not squamous | 61 | 0 | 0 | 61 |
| | Other | 124 | 0 | 0 | 124 |
| Smoking status | Ever | 3443 | 579 | 49 | 4071 |
| | Current | 2397 | 2136 | 83 | 4616 |
| | Ex | 567 | 271 | 26 | 864 |
| Smoking product | All | 1765 | 351 | 21 | 2137 |
| | Cigarettes | 3375 | 1290 | 136 | 4801 |
| | Cigarettes only | 749 | 886 | 0 | 1635 |
| | Other[b] | 79 | 54 | 0 | 133 |
| | Other[b] only | 132 | 91 | 1 | 224 |
| | Cigarettes and other[b] | 110 | 149 | 0 | 259 |
| | Pipe only | 106 | 77 | 0 | 183 |
| | Cigar only | 70 | 65 | 0 | 135 |
| | Pipe and cigar (not cigarettes) | 21 | 23 | 0 | 44 |
| Cigarette type | All | 3141 | 2163 | 136 | 5440 |
| | Manufactured[c] | 266 | 14 | 0 | 280 |
| | Hand-rolled[c] | 171 | 37 | 0 | 208 |
| | Manufactured and hand-rolled | 45 | 11 | 0 | 56 |
| | Filter[c] | 413 | 60 | 0 | 473 |
| | Plain[c] | 136 | 23 | 0 | 159 |
| | Filter and plain | 51 | 5 | 0 | 56 |
| | Menthol | 11 | 12 | 0 | 23 |
| Amount | All | 3973 | 1564 | 118 | 5655 |
| | Amount-specific | 2434 | 1422 | 40 | 3896 |

TABLE 8 Characteristics of the 9551 relative risks (continued)

| | | Study type | | | |
|---|---|---|---|---|---|
| | | Case- | | Nested | Total |
| Characteristic | Level | control | Prospective | case-control[a] | |
| Denominator | Never smoker | 3262 | 1522 | 35 | 4819 |
| | Nonsmoker | 501 | 461 | 3 | 965 |
| | Never smoked cigarettes | 1537 | 583 | 83 | 2203 |
| | Non smoker of cigarettes | 450 | 336 | 33 | 819 |
| | Manufactured cigarette smoker[c] | 100 | 30 | 0 | 130 |
| | Plain cigarette smoker[c] | 323 | 42 | 0 | 365 |
| | Non-menthol smoker | 10 | 12 | 0 | 22 |
| | Other[d] | 224 | 0 | 4 | 228 |
| Follow-up period | All of study period | - | 1458 | 10 [e] | 1468 |
| | Part of study period | - | 1528 | 0 [e] | 1538 |
| Adjustment | None | 4535 | 1144 | 133 | 5812 |
| | Any | 1872 | 1842 | 25 | 3739 |
| | Adjusted for sex | 173 | 17 | 0 | 190 |
| | Adjusted for age | 1703 | 1830 | 22 | 3555 |
| | Adjusted for race | 165 | 41 | 0 | 206 |
| | Adjusted for 1 other | 569 | 217 | 0 | 786 |
| | Adjusted for 2-3 others | 390 | 73 | 3 | 466 |
| | Adjusted for 4+ others | 125 [f] | 130 | 0 | 255 [f] |
| 2 × 2 table | Not applicable | 1875 | 1842 | 25 | 3742 |
| | Any missing cell | 146 | 101 | 0 | 247 |
| | All 4 cells present | 4386 | 1043 | 133 | 5562 |
| | With zero cell | 104 | 77 | 9 | 190 |
| RR and CI | All missing | 9 | 1 | 0 | 10 |
| | RR present but CI missing | 184 | 141 | 0 | 325 |
| | All present | 6214 | 2844 | 158 | 9216 |

TABLE 8 Characteristics of the 9551 relative risks (continued/2)

| | | Study type | | | |
|---|---|---|---|---|---|
| Characteristic | Level | Case-control | Prospective | Nested case-control[a] | Total |
| Derivation method | Original given | 811 | 325 | 19 | 1155 |
| | Original read from graph or chart | 21 | 11 | 0 | 32 |
| | Original RR without CI, but significance or non significance stated | 23 | 2 | 0 | 25 |
| | Direct from 2 × 2 table[g] | 1744 | 305 | 59 | 2108 |
| | From 2 × 2 table - discrepancy | 157 | 0 | 3 | 160 |
| | Summed over smoking groups | 1528 | 434 | 65 | 2027 |
| | Summed over disease groups | 154 | 0 | 0 | 154 |
| | Other sum | 715 | 276 | 0 | 991 |
| | With correction for zero cell | 98 | 76 | 9 | 183 |
| | Inverted, CI converted from 90%, symmetry | 32 | 13 | 0 | 45 |
| | Ratio of rates | 72 | 80 | 0 | 152 |
| | SMRs, expecteds | 10 | 3 | 0 | 13 |
| | Adjustment (by Fleiss and Gross) | 8 | 2 | 0 | 10 |
| | Fry & Lee over smoking groups | 423 | 206 | 3 | 632 |
| | Fry & Lee over disease groups | 20 | 0 | 0 | 20 |
| | Fry & Lee, other | 44 | 1 | 0 | 44 |
| | Other and combinations[h] | 155 | 510 | 0 | 655 |
| | RR original, CI from crude numbers | 149 | 97 | 0 | 246 |
| | Ratio of rates or inverted, with CI from crude numbers | 24 | 158 | 0 | 182 |
| | Fry & Lee with CIs from crude numbers | 95 | 18 | 0 | 113 |
| | Other with CI from crude numbers | 124 | 469 | 0 | 593 |

[a] Includes one case-cohort study.

[b] Pipe and/or cigars.

[c] Only, mainly, ever, etc.

[d] Variously include light smokers, long-term ex smokers/smokers who started recently or smokers with amount smoked unknown.

[e] Case-cohort study only.

[f] Includes 6 adjusted RRs for which the adjusting factors are unknown.

[g] Includes 2 × 2 table derived from percentage distribution or from matched pairs table, and adjusted RR from a 2 × 2 × n table. Includes RR only given originally, or RR/CI given to less than 2 decimal places but agrees so far as given with calculated values.

[h] But not zero cell or CI estimated from crude numbers.

## References

1. Fleiss JL, Gross AJ. Meta-analysis in epidemiology, with special reference to studies of the association between exposure to environmental tobacco smoke and lung cancer: a critique. *J Clin Epidemiol* 1991;**44**:127-39.

2. Fry JS, Lee PN. Revisiting the association between environmental tobacco smoke exposure and lung cancer risk. I. The dose-response relationship with amount and duration of smoking by the husband. *Indoor Built Environ* 2000;**9**:303-16.

3. Lee PN. Simple methods for checking for possible errors in reported odds ratios, relative risks and confidence intervals. *Stat Med* 1999;**18**:1973-81.

4. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Control Clin Trials* 1986;**7**:177-88.

5. Müller FH. Tabakmißbrauch und Lungencarcinom. *Z Krebsforsch* 1939;**49**:57-85.