Smoking is the cause of lung cancer

Ilija Barukčić

1 Internist, Horandstrasse, DE-26441 Jever, Germany
Correspondence: Ilija Barukčić, Horandstrasse, DE-26441 Jever, Germany. Tel: 49-4466-333. E-mail: Barukcic@t-online.de

Received: January 25, 2019; Accepted: January 25, 2019; Published: January 25, 2019.

ABSTRACT

Objective: The aim of this study is to re-evaluate the relationship between smoking and lung cancer.

Methods: In order to clarify the relationship between cigarette smoking and lung cancer, a review and meta-analysis of appropriate studies with a total sample size of n = 48393 was conducted. The p-value was set to p < 0,05.

Results.

It was not possible to reject the null-hypothesis H₀: without smoking no lung cancer. Furthermore, the null-hypothesis H₀: No causal relationship between smoking and lung cancer was rejected.

Conclusions

Compared to the results from previous studies, the results of this study confirm previously published results. According the results of this study, without smoking no lung cancer. Smoking is the cause of lung cancer.

Keywords: Smoking, lung cancer, causal relationship
Introduction

Formerly, lung cancer (LC), was an obscure and uncommon disease. Hasse reported in the late 1840s about 22 ever-published cases of lung cancer\(^1\). Meanwhile, lung cancer (LC) is one of the deadliest and most prevalent human cancers. The incidence and mortality rates of lung cancer, the first among all cancer types\(^2\), are still high. About 2093876 new cases of lung cancer occurred globally in 2018\(^3\). Furthermore, in the year 2018 about 1761007 people died from lung cancer. To date, lung cancer is the leading cause of cancer related death worldwide.

Especially small-cell lung cancer (SCLC) is characterized by its rapid growth and high response rates to chemotherapy and radiotherapy. The prognosis of SCLC depends more or less on the tumor stage. By time, the five-year survival rates of lung cancer patients remain low at 10% and have only slightly improved during the past decade\(^4\). A series of investigations and risk analyses indicated that factors such as smoking, air pollution, and occupational exposure (e.g. asbestos) are somehow related to lung cancer but the etiology of lung cancer is not yet clear. Especially smoking has been closely linked to lung cancer. The tobacco smoke includes about 7000 kinds of chemical substance. Carcinogens such as N'-nitrosonornicotine (NNN), benzo[a]pyrene, and (methylnitrosamino)-1-(3-pyridyl)-1-butanone (NNK) are rich in the stream of tobacco smoke. The relationship between the use of tobacco smoke and lung cancer is discussed in literature since more than 80 years while the historical origins of the discovery that smoking is related to lung cancer are complex\(^5\). The first association was documented by a case–control study conducted in Germany in the 1930s by Müller\(^6,7\). Preliminary evidence has been provided that smoking cessation even after diagnosis of early stage lung cancer may improve the prognostic outcomes\(^8\).

Material and Methods

Search strategy

The articles that met inclusion criteria were identified by Google Scholar and by searching in PubMed. The reference lists of review-articles were manually scanned to identify additional relevant studies.
Study selection

To be eligible for inclusion, the papers published have the following inclusion criteria: (1) published in English language; (2) no data access barriers. The exclusion criteria were as follows: (1) sample size less than n=1000. The titles and abstracts of all the retrieved articles using the inclusion criteria were screened. Data extraction was performed on included articles.

### 1. Identification of records

<table>
<thead>
<tr>
<th>Records identified by searching in the databases</th>
<th>Size</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>PubMed</td>
<td>7798</td>
<td></td>
</tr>
<tr>
<td>Google Scholar</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Web of Science</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Additional records identified from other sources</td>
<td>2</td>
<td>7800</td>
</tr>
</tbody>
</table>

### 2. Clean up of search

| Records removed after verifying duplication | 1    |
| Records excluded by title                   | 6989 |
| Records excluded by the summary             | 523  |
| (Articles outside the inclusion criteria)  | 287  |

### 3. Eligibility

| Articles evaluated for eligibility | 287  |
| Articles excluded for various reasons   |      |
| - Sample size less than 1000           | 241  |
| - Data access barriers                 | 35   |

### 4. Included

| Articles included in the meta-analysis | 11   |

Figure 1.

Flow Diagram of the article selection process. Adopted from PRISMA\(^9\),\(^10\) 2009.
Data analysis

The following\(^6,11,12,13,14,15,16,17,18,19,20\) data were recorded for analysis.

Table 1. *Without* smoking *no* lung cancer

<table>
<thead>
<tr>
<th>Study ID</th>
<th>Year</th>
<th>N</th>
<th>Case_P</th>
<th>Case_T</th>
<th>Con_P</th>
<th>Con_T</th>
<th>Odds Ratio</th>
<th>OR lower</th>
<th>OR upper</th>
<th>p(SINE)</th>
<th>$X^2$ (SINE)</th>
<th>p Value (SINE)</th>
<th>k</th>
<th>R/U</th>
</tr>
</thead>
<tbody>
<tr>
<td>Müller</td>
<td>1939</td>
<td>172</td>
<td>83</td>
<td>86</td>
<td>30</td>
<td>86</td>
<td>15,03</td>
<td>177,45</td>
<td>0,92276</td>
<td>0,14042</td>
<td>0,1571</td>
<td>227,15</td>
<td>0,7463</td>
<td>0,4095</td>
</tr>
<tr>
<td>Doll &amp; Hill</td>
<td>1952</td>
<td>2716</td>
<td>1508</td>
<td>1357</td>
<td>1289</td>
<td>1187</td>
<td>4,66</td>
<td>22,23</td>
<td>0,95762</td>
<td>0,0361090641</td>
<td>0,4925</td>
<td>0,1371</td>
<td>0,1274</td>
<td></td>
</tr>
<tr>
<td>Lager et al.</td>
<td>1965</td>
<td>2000</td>
<td>1626</td>
<td>1600</td>
<td>928</td>
<td>980</td>
<td>5,04</td>
<td>15,53</td>
<td>0,95277</td>
<td>0,1806515385</td>
<td>0,6442</td>
<td>0,1975</td>
<td>0,4994</td>
<td></td>
</tr>
<tr>
<td>Wynder et al.</td>
<td>1970</td>
<td>10251</td>
<td>659</td>
<td>686</td>
<td>5136</td>
<td>5947</td>
<td>22,45</td>
<td>33,54</td>
<td>0,97534</td>
<td>0,957324801</td>
<td>0,3981</td>
<td>0,2115</td>
<td>22,45</td>
<td></td>
</tr>
<tr>
<td>Birkholz et al.</td>
<td>1985</td>
<td>1132</td>
<td>1108</td>
<td>1217</td>
<td>1192</td>
<td>1227</td>
<td>8,84</td>
<td>15,53</td>
<td>0,95277</td>
<td>0,1806515385</td>
<td>0,6442</td>
<td>0,1975</td>
<td>0,4994</td>
<td></td>
</tr>
<tr>
<td>Harris et al.</td>
<td>1990</td>
<td>5530</td>
<td>2828</td>
<td>2514</td>
<td>1987</td>
<td>1964</td>
<td>7,07</td>
<td>12,67</td>
<td>0,96427</td>
<td>2,951678123</td>
<td>0,1872</td>
<td>0,3058</td>
<td>0,480</td>
<td></td>
</tr>
<tr>
<td>Peto et al.</td>
<td>1994</td>
<td>7093</td>
<td>3087</td>
<td>2678</td>
<td>2414</td>
<td>2647</td>
<td>2,81</td>
<td>4,62</td>
<td>0,94906</td>
<td>1,1406037364</td>
<td>0,4555</td>
<td>0,0935</td>
<td>0,2159</td>
<td></td>
</tr>
<tr>
<td>Sobue et al.</td>
<td>1994</td>
<td>2197</td>
<td>1076</td>
<td>1035</td>
<td>1141</td>
<td>1141</td>
<td>3,90</td>
<td>5,46</td>
<td>0,94932</td>
<td>1,094090997</td>
<td>0,2954</td>
<td>0,1528</td>
<td>0,4069</td>
<td></td>
</tr>
<tr>
<td>Jöckel et al.</td>
<td>1998</td>
<td>2008</td>
<td>999</td>
<td>1008</td>
<td>768</td>
<td>1004</td>
<td>5,30</td>
<td>7,22</td>
<td>0,97201</td>
<td>3,012986272</td>
<td>0,8028</td>
<td>0,2561</td>
<td>0,1551</td>
<td></td>
</tr>
<tr>
<td>Kreuzer et al.</td>
<td>1998</td>
<td>1873</td>
<td>1657</td>
<td>1708</td>
<td>1351</td>
<td>1741</td>
<td>14,73</td>
<td>35,16</td>
<td>0,93909</td>
<td>0,283268535</td>
<td>0,5946</td>
<td>0,3294</td>
<td>0,1760</td>
<td></td>
</tr>
<tr>
<td>Bellika et al.</td>
<td>1999</td>
<td>12878</td>
<td>5506</td>
<td>5623</td>
<td>5855</td>
<td>5993</td>
<td>12,26</td>
<td>19,10</td>
<td>0,95901</td>
<td>2,437317951</td>
<td>0,1186</td>
<td>0,3304</td>
<td>0,2816</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>48393</td>
<td>17393</td>
<td>17826</td>
<td>22043</td>
<td>30567</td>
<td>0,9911</td>
<td>12,70</td>
<td>0,8911</td>
<td>0,8911</td>
<td>0,8911</td>
<td>12,70</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Alpha = 0,05

Degrees of freedom = 11

$X^2$ CRITICAL (SINE) = 19,68

$X^2$ Calculated (SINE) = 12,70

p value (SINE) = 0,31
Table 2. Smoking is the cause of lung cancer.

<table>
<thead>
<tr>
<th>Study ID</th>
<th>Year</th>
<th>N</th>
<th>Case_P</th>
<th>Case_T</th>
<th>Con_P</th>
<th>Con_T</th>
<th>X² (Sine)</th>
<th>k</th>
<th>k lower</th>
<th>k upper</th>
<th>p value (k)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Müller</td>
<td>1939</td>
<td>172</td>
<td>83</td>
<td>86</td>
<td>30</td>
<td>86</td>
<td>0,10</td>
<td>0,65</td>
<td>0,48</td>
<td>0,82</td>
<td>0,0000000000</td>
</tr>
<tr>
<td>Doll &amp; Hill</td>
<td>1952</td>
<td>2714</td>
<td>1350</td>
<td>1357</td>
<td>1289</td>
<td>1357</td>
<td>0,04</td>
<td>0,14</td>
<td>0,09</td>
<td>0,18</td>
<td>0,0000000000</td>
</tr>
<tr>
<td>Lombard et. al</td>
<td>1965</td>
<td>2080</td>
<td>1026</td>
<td>1040</td>
<td>928</td>
<td>1040</td>
<td>0,19</td>
<td>0,20</td>
<td>0,15</td>
<td>0,25</td>
<td>0,0000000000</td>
</tr>
<tr>
<td>Wynder et. al</td>
<td>1979</td>
<td>10231</td>
<td>659</td>
<td>684</td>
<td>5156</td>
<td>9547</td>
<td>0,91</td>
<td>0,21</td>
<td>0,19</td>
<td>0,24</td>
<td>0,0000000000</td>
</tr>
<tr>
<td>Benhamou et al.</td>
<td>1985</td>
<td>3132</td>
<td>1184</td>
<td>1217</td>
<td>1392</td>
<td>1915</td>
<td>0,89</td>
<td>0,31</td>
<td>0,27</td>
<td>0,35</td>
<td>0,0000000000</td>
</tr>
<tr>
<td>Harris et. al</td>
<td>1993</td>
<td>5530</td>
<td>2829</td>
<td>2916</td>
<td>1997</td>
<td>2614</td>
<td>2,60</td>
<td>0,31</td>
<td>0,28</td>
<td>0,34</td>
<td>0,0000000000</td>
</tr>
<tr>
<td>Pershagen et. al</td>
<td>1994</td>
<td>3983</td>
<td>1100</td>
<td>1136</td>
<td>2607</td>
<td>2847</td>
<td>1,14</td>
<td>0,09</td>
<td>0,06</td>
<td>0,13</td>
<td>0,0000000000</td>
</tr>
<tr>
<td>Sobue et. al</td>
<td>1994</td>
<td>2197</td>
<td>1022</td>
<td>1056</td>
<td>1013</td>
<td>1141</td>
<td>1,09</td>
<td>0,15</td>
<td>0,11</td>
<td>0,20</td>
<td>0,0000000000</td>
</tr>
<tr>
<td>Jöckel et. al</td>
<td>1998</td>
<td>2008</td>
<td>949</td>
<td>1004</td>
<td>768</td>
<td>1084</td>
<td>3,01</td>
<td>0,26</td>
<td>0,21</td>
<td>0,31</td>
<td>0,0000000000</td>
</tr>
<tr>
<td>Kreuzer et. al</td>
<td>1998</td>
<td>3470</td>
<td>1687</td>
<td>1709</td>
<td>1358</td>
<td>1761</td>
<td>0,28</td>
<td>0,33</td>
<td>0,29</td>
<td>0,37</td>
<td>0,0000000000</td>
</tr>
<tr>
<td>Boffetta et al.</td>
<td>1999</td>
<td>12876</td>
<td>5504</td>
<td>5621</td>
<td>5505</td>
<td>7255</td>
<td>2,44</td>
<td>0,31</td>
<td>0,29</td>
<td>0,33</td>
<td>0,0000000000</td>
</tr>
</tbody>
</table>

Total 48393 17393 17826 22043 30567 12,70 0,32 0,31 0,32635 0,0000000000

Alpha = 0,05

Degrees of freedom = 11

X² CRITICAL (k) = 19,68

X² Calculated (k) = 4837,90

p value (SINE) = 0,00
Statistical Analysis

All the statistics analyses were conducted by Microsoft® Excel® for Mac® version 16.2 (181208) software (© 2018, Microsoft GmbH, Munich, Germany). A p < 0.05 was considered significant on statistical analyses.

Definitions

Definition. The 2x2 Table

Consider the case of Bernoulli trials (period of time) with probability \( p(a_t) \) for success. Let \( a_t = 1 \) if the \( t-th \) outcome is a success and 0 if it is a failure. Then \( a = (a_1 + a_2 + \ldots + a_n) \) is the number of successes in \( n \) trials (period of time) \( t \). It is \( p(a_t) = p(A_t \cap B_t) \) the joint probability of \( A_t \) and \( B_t \) and

\[
a = \left( a_1 + a_2 + \cdots + a_n \right) \equiv \sum_{t=1}^{t=n} a_t
\]  

(1)

Let \( b_t = 1 \) if the \( t-th \) outcome is a success and 0 if it is a failure. Then \( b = (b_1 + b_2 + \ldots + b_n) \) is the number of successes in \( n \) Bernoulli trials (period of time) \( t \). It is \( p(b_t) = p(A_t \cap B_t) \) the joint probability of \( A_t \) and \( B_t \) and

\[
b = \left( b_1 + b_2 + \cdots + b_n \right) \equiv \sum_{t=1}^{t=n} b_t
\]  

(2)

Let \( c_t = 1 \) if the \( t-th \) outcome is a success and 0 if it is a failure. Then \( c = (c_1 + c_2 + \ldots + c_n) \) is the number of successes in \( n \) Bernoulli trials (period of time) \( t \). It is \( p(c_t) = p(A_t \cap B_t) \) the joint probability of \( A_t \) and \( B_t \) and

\[
c = \left( c_1 + c_2 + \cdots + c_n \right) \equiv \sum_{t=1}^{t=n} c_t
\]  

(3)

Let \( d_t = 1 \) if the \( t-th \) outcome is a success and 0 if it is a failure. Then \( d = (d_1 + d_2 + \ldots + d_n) \) is the number of successes in \( n \) Bernoulli trials (period of time) \( t \). It is \( p(d_t) = p(A_t \cap B_t) \) the joint probability of \( A_t \) and \( B_t \) and
\[ d \equiv (d_1 + d_2 + \cdots + d_n) \equiv \sum_{t=1}^{t=n} d_t \quad (4) \]

Let \(A\) denote another binomial random variable with the probability \(p(A_t)\). It is \(A_t = (a_t + b_t)\) at the same Bernoulli trial (period of time) \(t\) and

\[ A \equiv \left( (a_1 + b_1) + (a_2 + b_2) + \cdots + (a_n + b_n) \right) \equiv \sum_{t=1}^{t=n} A_t \quad (5) \]

Let \(\bar{A}\) denote the complementary random variable of the binomial random variable \(A\) with the probability \(p(\bar{A}_t)\). It is \(\bar{A}_t = (c_t + d_t)\) at the same Bernoulli trial (period of time) \(t\) and

\[ \bar{A} \equiv \left( (c_1 + d_1) + (c_2 + d_2) + \cdots + (c_n + d_n) \right) \equiv \sum_{t=1}^{t=n} \bar{A}_t \quad (6) \]

Let \(B\) denote another binomial random variable with the probability \(p(B_t)\). It is \(B_t = (a_t + c_t)\) at the same Bernoulli trial (period of time) \(t\) and

\[ B \equiv \left( (a_1 + c_1) + (a_2 + c_2) + \cdots + (a_n + c_n) \right) \equiv \sum_{t=1}^{t=n} B_t \quad (7) \]

Let \(\bar{B}\) denote the complementary random variable of the binomial random variable \(B\) with the probability \(p(\bar{B}_t)\). It is \(\bar{B}_t = (c_t + d_t)\) at the same Bernoulli trial (period of time) \(t\) and

\[ \bar{B} \equiv \left( (b_1 + d_1) + (b_2 + d_2) + \cdots + (b_n + d_n) \right) \equiv \sum_{t=1}^{t=n} \bar{B}_t \quad (8) \]

At each Bernoulli trial it is
\[ n_t \equiv (a_t + b_t + c_t + d_t) \equiv A_t + A_t \equiv B_t + B_t \] (9)

and the sample size \( n \) itself equal to

\[ n \equiv \sum_{t=1}^{n} (a_t + b_t + c_t + d_t) \equiv \sum_{t=1}^{n} A_t + A_t \equiv \sum_{t=1}^{n} B_t + B_t \] (10)

The meaning of the abbreviations \( a, b, c, d, n \) et cetera are explained by following 2 by 2-table (Table 3).

Table 3. The sample space of a contingency table

<table>
<thead>
<tr>
<th>Condition A (risk factor)</th>
<th>Condition B (Outcome)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes = +1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>b</td>
<td>A</td>
</tr>
<tr>
<td>c</td>
<td>d</td>
<td>A</td>
</tr>
<tr>
<td>Total</td>
<td>B</td>
<td>B</td>
</tr>
</tbody>
</table>

In this context, it is \( p(A_t) = p(a_t) + p(b_t) \) or \( p(A_t) = p(A_t \cap B_t) + p(b_t) \) or \( p(A_t) = p(A_t \cap B_t) + p(A_t \cap B_t) \) while \( p(A_t) \) is not identical with \( p(a_t) \). Thus far, it is \( p(B_t) = p(a_t) + p(c_t) \) or \( p(B_t) = p(A_t \cap B_t) + p(c_t) \) and equally \( p(B_t) = 1 - p(B_t) \) or \( p(B_t) = p(b_t) + p(d_t) \). Since the joint probability of \( A_t \) and \( B_t \) is denoted in general by \( p(A_t \cap B_t) \), it is \( p(A_t \cap B_t) = p(A_t) - p(b_t) \) or \( p(A_t \cap B_t) = p(B_t) - p(c_t) \) or in other words \( p(B_t) + p(b_t) - p(c_t) = p(A_t) \). In general, it is \( p(a_t) + p(c_t) + p(b_t) + p(d_t) \). The following table may show the relationship in more details.
Table 4. The probabilities of a contingency table

<table>
<thead>
<tr>
<th>Conditioned B</th>
<th>Yes = +1</th>
<th>No = +0</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes =+1</td>
<td>p(a₁)</td>
<td>p(0)</td>
<td>p(A₁)</td>
</tr>
<tr>
<td>No = +0</td>
<td>p(c₁)</td>
<td>p(d₁)</td>
<td>p(A₂)</td>
</tr>
<tr>
<td>Total</td>
<td>p(B₁)</td>
<td>p(B₂)</td>
<td>1</td>
</tr>
</tbody>
</table>

**Definition. Index of unfairness**

The index of unfairness (IOU) is defined as

\[
IOU = \left( \frac{A + B}{n} \right) - 1
\]  

(11)

**Definition. Independence**

Let \( A_t \) denote random variable at a Bernoulli trial (period of time) \( t \). Let \( B_t \) denote another random variable at the same Bernoulli trial (period of time) \( t \). Let \( p(A_t) \) denote the probability of \( A_t \). Let \( p(B_t) \) denote the probability of \( B_t \). Let \( p(A_t \cap B_t) \) denote the joint probability of \( A_t \) and \( B_t \). In the case of independence\(^{21,22} \) of \( A_t \) and \( B_t \) it is generally valid that

\[
p(A_t \cap B_t) = p(A_t) \times p(B_t)
\]  

(12)
Definition. Sufficient Condition (Conditio per Quam)

The mathematical formula of the sufficient condition relationship \((\text{conditio per quam})\) of a population is defined as

\[
p(A_t \rightarrow B_t) \equiv \frac{(a_t) + (c_t) + (d_t)}{N_t} = 1
\]

\[
\equiv p(a_t) + p(c_t) + p(d_t)
\]

\[
\equiv p(B_t) + p(d_t)
\]

\[
\equiv p(a_t) + p(A_t)
\]

\[
\equiv +1.
\]

and is used to prove the hypothesis: if \(A_t\) then \(B_t\) or is taken to express that the occurrence of an event \(A_t\) is a sufficient condition\(^{32,33}\) for existence or occurrence of an event \(B_t\). The occurrence of an event \(A_t\) is a sufficient condition for occurrence of the event \(B_t\) or \(B_t\) is a necessary condition for \(A_t\). In other words, sufficient and necessary conditions are converse relations.

Definition. The \(X^2\) Test of Goodness of Fit of a Sufficient Condition

A random sample of observations can come from a particular distribution (sufficient condition distribution) but must not. The \(X^2\) test of goodness-of-fit is an appropriate method for testing the null hypothesis that a random sample of observations comes from a specific distribution (i.e. the distribution of a sufficient condition) against the alternative hypothesis that the data have some other distribution. The additive property of \(X^2\) distribution may sometimes be used as an additional test of significance. In this case, the continuity correction should be omitted from each \(X^2\) value. Under conditions where the chi-square goodness of fit test cannot be used it is possible to use an approximate and conservative (one sided) confidence interval known as
**the rule of three.** The $X^2$ distribution is a particular type of a gamma distribution and widely applied in the field of mathematical statistics. The applicability of using the Pearson chi-squared statistic in cases where the cell frequencies of a $2 \times 2$ contingency table are not greater than five is widely discussed\(^{34}\) in literature and the use of Yate’s\(^ {35}\) continuity correction is proposed. However, studies provided evidence that incorporating Yate’s continuity correction\(^ {36}\) is not essential\(^ {37}\). Still, using the *continuity correction*, the chi-square value of a conditio per quam relationship is derived\(^ {23, 24, 25, 26, 27, 28, 29, 30, 31}\) as

\[
X^2 \left( (A \rightarrow B) \mid A \right) \equiv \frac{\left( \frac{b}{A} - \frac{1}{2} \right)^2}{A} + 0 = 0
\]

or alternatively as

\[
X^2 \left( (A \rightarrow B) \mid B \right) \equiv \frac{\left( \frac{b}{B} - \frac{1}{2} \right)^2}{B} + 0 = 0
\]

**Definition. Necessary Condition (Conditio Sine Qua Non)**

Among the many generally valid natural laws and principles under which nature or matter itself assures its own self-organization, a relationship between events denoted as a necessary\(^ {23, 24, 25, 26, 27, 28, 29, 30, 31}\) condition (a conditio sine qua non) is one among the most important. A necessary (or an essential) event or condition $A_t$ for some event $B_t$ is a condition that must be satisfied in order to obtain $B_t$. In this respect, to say that an event $A_t$ with its own probability $p(A_t)$ is at the same (period of) time $t$ a necessary condition for another event $B_t$ with its own probability $p(B_t)$ is equivalent to say that it is impossible to have $B_t$ without $A_t$. In other words, *without* $A_t$ *no* $B_t$ or the absence of $A_t$ guarantees the absence of $B_t$. The mathematical formula of the *necessary* condition relationship (conditio sine qua non) of a population is defined as
\[
p(A_t \leftarrow B_t) \equiv \frac{(a_t) + (b_t) + (d_t)}{N_t} = 1
\]
\[
\equiv p(a_t) + p(b_t) + p(d_t)
\]
\[
\equiv p(A_t) + p(d_t)
\]
\[
\equiv p(a_t) + p(B_t) = p(a_t) + (1 - p(B_t))
\]
\[
\equiv +1.
\]

**Definition. The \( X^2 \) Test of Goodness of Fit of a Necessary Condition**

Under conditions where the chi-square goodness of fit test cannot be used it is possible to use an approximate and conservative (one sided) confidence interval known as the *rule of three*. Using the *continuity correction*, the chi-square value of a *conditio sine qua non* distribution before changes to

\[
X^2 \left( (A \leftarrow B) \mid B \right) \equiv \frac{(c) - (1/2)^2}{B} + 0 = 0
\]

Depending upon the study design, another method to calculate the chi-square value of a *conditio sine qua non* distribution (while using the *continuity correction*) is defined as

\[
X^2 \left( (A \leftarrow B) \mid A \right) \equiv \frac{(c) - (1/2)^2}{A} + 0 = 0
\]

**Definition. Exclusion (\( A_t \) Excludes \( B_t \) and Vice Versa Relationship)**

The mathematical formula of the *exclusion* relationship (\( A_t \) excludes \( B_t \) and vice versa) of a population was defined before changes to

\[
\]
Ilija Barukčić - Smoking is the cause of lung cancer

\[ p(A_t \mid B_t) \equiv \frac{(b_t) + (c_t) + (d_t)}{N_t} = 1 \]

\[ \equiv p(b_t) + p(c_t) + p(d_t) \]

\[ \equiv p(b_t) + p(A_t) = p(b_t) + \left(1 - p(A_t)\right) \]  \hspace{1cm} (18)

\[ \equiv p(c_t) + p(B_t) = p(c_t) + \left(1 - p(B_t)\right) \]  \hspace{1cm} (19)

\[ \equiv +1. \]

and used to prove the hypothesis: \( A_t \) excludes \( B_t \) and vice versa. Why should \( A_t \) exclude \( B_t \) and vice versa? Under which conditions can such a relationship be given?

**Definition. The \( X^2 \) Test of Goodness of Fit of the Exclusion Relationship**

The chi square value with degree of freedom 2-1=1 of the exclusion relationship \(^{23, 24, 25, 26, 27, 28, 29, 30, 31}\) with a *continuity correction* can be calculated as

\[ X^2 \left( (A \mid B) \mid A \right) = \frac{((a) - \left(\frac{1}{2}\right))^2}{A} + 0 = 0 \]  \hspace{1cm} (19)

Depending upon the study design, another method to calculate the chi-square value of a *conditio sine qua non* distribution is defined as

\[ X^2 \left( (A \mid B) \mid B \right) = \frac{((a) - \left(\frac{1}{2}\right))^2}{B} + 0 = 0 \]  \hspace{1cm} (20)

The chi square Goodness of Fit Test of the exclusion relationship examines how well observed data compare with the expected theoretical distribution of an exclusion relationship.
Definition. The Mathematical Formula of the Causal Relationship $k$

The mathematical formula of the causal relationship $k$ is defined at every single event, at every single Bernoulli trial $t$, as

$$k(A_t, B_t) = \left( \frac{p(A_t \cap B_t) - p(A_t) \times p(B_t)}{\sqrt{p(A_t) \times (1 - p(A_t)) \times p(B_t) \times (1 - p(B_t))}} \right)^2$$

where $A_t$ denotes the cause and $B_t$ denotes the effect. Under some certain circumstances, the chi-square distribution can be applied to determine the significance of causal relationship $k$. Pearson’s concept of correlation is not identical with causation. Causation as such is not identical with correlation. This has been proved many times and is widely discussed in many publications.

Definition. The 95% Confidence Interval of the Causal Relationship $k$

A confidence interval (CI) of the causal relationship $k$ calculated from the statistics of the observed data can help to estimate the true value of an unknown population parameter with a certain probability. Under some conditions, the 95% interval for the causal relationship $k$ is derived as

$$\left\{ k(A_t, B_t) - \sqrt{\frac{5}{N}} ; k(A_t, B_t) + \sqrt{\frac{5}{N}} \right\}$$

(22)
Definition. The rule of three

Under some specified conditions (i.e. the dataset analyzed is large enough or \( n \), the sample size, is \( n \sim 30 \) and more), a Chi-square goodness of fit test is able to provide evidence whether a sample distribution observed is identical with a theoretical distribution expected. Formally, the Chi-square goodness of fit test is defined as \( X^2 = ((\text{sample distribution}) - (\text{theoretical distribution}))^2/(\text{theoretical distribution}) \) or something like \( X^2 = ((\text{observed}) - (\text{expected}))^2/(\text{expected}) \). An approximate and conservative (one sided) confidence interval as discussed by and known as the rule of three can be of practical value if the Chi-square goodness of fit test cannot be applied. Under some circumstances, the rule of three derived as

\[
p_{\text{critical}} = 1 - \left(\frac{3}{n}\right)
\]

while \( n \) is the sample size is one way to calculate the probability of events which occur with a probability near 1. Another and a very simple path to calculate the probability of an event can be performed by the following method.

Definition. The unknown population proportion \( \pi_{\text{upper}} \)

Tests of hypotheses concerning the sampling distribution of the sample proportion \( p \) (i.e. conditio sine qua non \( p(\text{SINE}) \), conditio per quam \( p(\text{IMP}) \) et cetera) can be performed using the normal approximation. The calculation of the rejection region based on the sample proportion to construct a confidence interval for an unknown population proportion \( \pi_{\text{upper}} \) can be performed under conditions of sampling without replacement (Sachs, 1992) by the formula

\[
p_{\text{critical upper}} = \left(p - \frac{1}{2 \times n}\right) - \left(Z \times \sqrt{\frac{p \times (1 - p)}{n}} \times \frac{N - n}{N - 1}\right)
\]

while the term ((\( N-n \))/(\( N-1 \))) denotes the finite population correction.
Definition. Odds Ratio

The odds\(^{44,45,46,47}\) ratio, abbreviated as OR\((A,B)\), is a very commonly used measure of association for 2×2 contingency tables and given by

\[
OR(A, B) \equiv \frac{a}{c} \div \frac{b}{d} \equiv \frac{a \times d}{c \times b}
\]  

(25)

Although severely and justifiably criticized especially by Karl Pearson (1857–1925), the longtime and rarely challenged leader of statistical science and Heron\(^{48}\), Odds ratio is still regularly referred to. The standard error and 95% confidence interval of the Odds ratio (OR) can be calculated according to Altman\(^{49}\). Given the severely limited character of odds ratio, the standard error of the log Odds ratio is calculated as

\[
SE\left(\ln\left(OR(A, B)\right)\right) \equiv \sqrt{\left(\frac{1}{a}\right) + \left(\frac{1}{b}\right) + \left(\frac{1}{c}\right) + \left(\frac{1}{d}\right)}
\]  

(26)

where \(\ln\) denotes the logarithmus naturalis. The 95% confidence interval of the odds ratio is given by

\[
95\% CI \equiv \exp\left(ln\left(OR(A, B)\right) - 1.96 \times SE\left(ln\left(OR(A, B)\right)\right)\right)
\]  

to

\[
\exp\left(ln\left(OR(A, B)\right) + 1.96 \times SE\left(ln\left(OR(A, B)\right)\right)\right)
\]  

(27)

Definition. The Chi-square goodness-of-fit test

A Chi-Square goodness-of-fit test is one of commonly used methods of statistical inference an originally proposed by Karl Pearson. Given some conditions (simple random sampling, categorical random variable, expected value of the number of sample observations is at least 5 et cetera), the chi-square goodness of fit test can be applied to determine whether (sample
distribution) data observed are consistent with (theoretical distribution) hypothesized data. The degrees of freedom (d.f.) of a chi-square goodness of fit test is equal to the number of levels (k) of the categorical variable minus 1. In general, the chi-square goodness of fit test is given by

\[ X^2 = \sum_{t=1}^{k} \left( \frac{(x_t - (n \times p(x_t)))^2}{(n \times p(x_t))} \right) \]  

(28)

Example.

Suppose, a coin is tossed 100 times with the results given in Table 3.

Table 5. A fair coin.

<table>
<thead>
<tr>
<th>Event</th>
<th>Observed (x_t)</th>
<th>Expected (n\times p(x_t))</th>
<th>((x_t)- (n\times p(x_t)))</th>
<th>(((x_t)-(n\times p(x_t)))^2/ (n\times p(x_t))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heads</td>
<td>40</td>
<td>50</td>
<td>-10</td>
<td>(-10)^2/50 = 2</td>
</tr>
<tr>
<td>Tails</td>
<td>60</td>
<td>50</td>
<td>+10</td>
<td>(+10)^2/50 = 2</td>
</tr>
<tr>
<td>n</td>
<td>100</td>
<td>100</td>
<td></td>
<td>(X^2 = 4)</td>
</tr>
</tbody>
</table>

In this context, the chi-square goodness of fit test requires to state a null hypothesis (H_0) and an alternative hypothesis (H_A). In point of fact, it is p=p(Heads) and q=p(Tails) and (p +q) = 1 or (p(Heads) + p(Tails)) = 1 or p(Tails) = 1 – p(Heads). In our present case (\(\alpha = 0.05\)), for a chi-square goodness of fit test of this example, the hypotheses take the following form.

Null hypothesis: The data are consistent with a specified distribution or p(Heads)=0.5

The null hypothesis claims equally that p(Heads) = 1 – p(Tails) = 0.5

Alternative hypothesis: The data are not consistent with a specified distribution. The Null hypothesis is not true.

The value of the test statistics as calculated before is

\[ X^2 = \sum_{t=1}^{k} \left( \frac{(x_t - (n \times p(x_t)))^2}{(n \times p(x_t))} \right) = \frac{(40 - 50)^2}{50} + \frac{(60 - 50)^2}{50} = \frac{100}{50} + \frac{100}{50} = 2 + 2 = 4 \]  

(29)

with d. f. = k-1=2-1 = 1. Unfortunately, the p-value of \(X^2=4\) is less than the significance level (0.05). We accept the alternative hypothesis and reject the null-hypothesis. The sample data do
not provide support for the hypothesis that the coin tossed is fair. **In general, it is not necessary** that \( p = q \), to be able use the chi square goodness-of-fit test which is the mathematical the foundation of the chi square goodness of fit test of the necessary condition, of a sufficient condition et cetera with \( d. f. = k-1=2-1 = 1 \).

**Definition. The Chi Square Distribution**

The following critical values of the chi square distribution as visualized by Table 4 are used in this publication.

<table>
<thead>
<tr>
<th>p-Value</th>
<th>One sided ( X^2 )</th>
<th>Two sided ( X^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1000000000</td>
<td>1.642374415</td>
<td>2.705543454</td>
</tr>
<tr>
<td><strong>0.0500000000</strong></td>
<td><strong>2.705543454</strong></td>
<td><strong>3.841458821</strong></td>
</tr>
<tr>
<td>0.0400000000</td>
<td>3.06490172</td>
<td>4.217884588</td>
</tr>
<tr>
<td>0.0300000000</td>
<td>3.537384596</td>
<td>4.709292247</td>
</tr>
<tr>
<td>0.0200000000</td>
<td>4.217884588</td>
<td>5.411894431</td>
</tr>
<tr>
<td>0.0100000000</td>
<td>5.411894431</td>
<td>6.634896601</td>
</tr>
<tr>
<td>0.0010000000</td>
<td>9.549535706</td>
<td>10.82756617</td>
</tr>
<tr>
<td>0.0001000000</td>
<td>13.83108362</td>
<td>15.13670523</td>
</tr>
<tr>
<td>0.0000100000</td>
<td>18.18929348</td>
<td>19.51142096</td>
</tr>
<tr>
<td>0.0000010000</td>
<td>22.59504266</td>
<td>23.92812698</td>
</tr>
<tr>
<td>0.0000001000</td>
<td>27.03311129</td>
<td>28.37398736</td>
</tr>
<tr>
<td>0.0000000100</td>
<td>31.49455797</td>
<td>32.84125335</td>
</tr>
<tr>
<td>0.0000000010</td>
<td>35.97368894</td>
<td>37.32489311</td>
</tr>
<tr>
<td>0.0000000001</td>
<td>40.46665791</td>
<td>41.82145620</td>
</tr>
</tbody>
</table>
Results

**Theorem. Without smoking no lung cancer**

**Claims.**

**Null hypothesis:**
Smoking is a necessary condition (a conditio sine qua non) of lung cancer. In other words, the sample distribution of the study analyzed agrees with the hypothetical (theoretical) distribution of a necessary condition.

**Alternative Hypothesis:**
Smoking is not a necessary condition (a conditio sine qua non) of lung cancer. In other words, the sample distribution of the study analyzed does not agree with the hypothetical (theoretical) distribution of a necessary condition.

The significance level (Alpha) below which the null hypothesis will be rejected is $\alpha = 0.05$.

**Proof.**

The results of the data reviewed and re-analyzed by this article which investigated the relationship between smoking and lung cancer are viewed by the table (**Table 1**). Altogether, 11 studies with a sample size of $n = 48393$ were meta-analyzed while the level of significance was $\alpha = 0.05$. In toto, all studies re-analyzed provide significant evidence of a conditio sine qua non relationship ($X^2$ Calculated (SINE) = 12.70 and is less than $X^2$ Critical (SINE) = 19.68) between a smoking and lung cancer. All studies analyzed were able to provide evidence of a significant, positive cause effect relationship. In other words, the null-hypothesis cannot be rejected, the data analyzed support the null-hypothesis: without smoking no lung cancer.

**Quod erat demonstrandum.**
Theorem. Smoking is the cause of lung cancer

Claims.

Null Hypothesis:
Smoking is not the cause of lung cancer. In other words. k = 0.

Alternative Hypothesis:
Smoking is the cause of lung cancer. In other words. k ≠ 0.

The significance level (Alpha) below which the null hypothesis will be rejected is alpha=0.05.

Proof.
The results of the re-analyses of the data reviewed by this article (Table 2) which investigated the causal relationship between smoking and lung cancer are viewed by the table (Table 2). Altogether, 11 studies were meta-analyzed while the level of significance was alpha = 0.05. In toto, 11 from 11 studies provided significant evidence of a causal relationship between a smoking and human lung cancer. In the same respect, smoking is a necessary (Table 1) condition of lung cancer. In other words, without smoking no lung cancer. Thus far, the conclusion is inescapable: smoking of tobacco is the cause of human lung cancer (k ~ 0.32, X² Calculated (k) = 4837.90 and is greater than X² Critical (k) =19.68).

Quod erat demonstrandum.

4. Discussion
Based on the results of this study, smoking is a necessary condition (a conditio sine qua non) of lung cancer. In other words, without smoking of tobacco no lung cancer (Table 1). In the same respect the cause-effect relationship k (Table 2) is highly significant. Firstly, without smoking of tobacco no lung cancer will not develop. Secondly. There is a highly significant cause effect relationship between smoking and lung cancer. Thus far, we are authorized to deduce that smoking of tobacco is not only one cause of lung cancer but smoking of tobacco is the cause of human lung cancer. Still, the lung cancer risk as associated with secondhand
smoking has not been addressed by this review in an appropriate way. The association between passive smoking and lung cancer has been investigated by several other studies. With regard to this problem, the results of several other studies clearly indicate that non-smokers exposed to Environmental Tobacco Smoke (ETS) are at increased risk\textsuperscript{51, 52, 53} of lung cancer. In particular, about 433 from 48393 patients (Table 7) have been treated as \textit{never smoker} while in reality it cannot be excluded that these patients were exposed to passive smoking.

<table>
<thead>
<tr>
<th>Table 7.</th>
<th>Statistical analysis.</th>
<th>Index of unfairness = 0.183</th>
</tr>
</thead>
<tbody>
<tr>
<td>The studies re-analysed</td>
<td>causal relationship $k = 0.316181947$. 95% CI ($k$): (0.30–0.326)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$p$-value ($k$</td>
<td>HGD) = 0</td>
</tr>
<tr>
<td></td>
<td>Odds ratio (OR) = 15.55</td>
<td>14.07</td>
</tr>
<tr>
<td>Smoking</td>
<td>YES</td>
<td>17393</td>
</tr>
<tr>
<td></td>
<td>NO</td>
<td>433</td>
</tr>
<tr>
<td></td>
<td>17826</td>
<td>30567</td>
</tr>
</tbody>
</table>

The suspicion appears to be justified that through the years of smoking an active or passive smoker transfers his own lungs into kind of a “hazardous waste landfill” for a wide variety of cancerogenic toxins with all the consequences which might follow by time. In order to further clarify the association between smoking and lung cancer any exposure to environmental tobacco smoke (ETS) should be considered by the studies performed.

5. Conclusion

The list of studies which provided striking evidence on the relationship between smoking and lung cancer is long enough and justifies to take a short way around. A total ban on smoking is necessary. In any case, smoking is the cause of lung cancer.
Financial support and sponsorship

Nil.

Conflict of interest statement

The authors declare that no conflict of interest exists according to the guidelines of the International Committee of Medical Journal Editors.

Acknowledgements

The public domain software program Gnuplot was used to draw the figures in this publication. The open source, independent and non-profit Zotero Citation Manager was used to create and manage references and bibliographies.
References


22. Kolmogoroff A. Grundbegriffe Der Wahrscheinlichkeitsrechnung, Berlin, Heidelberg: Springer Berlin Heidelberg; 1933. doi: https://doi.org/10.1007/978-3-642-49888-6


26. Barukčić I. Helicobacter Pylori is the Cause of Gastric Cancer, Mod Health Sci, 2018; 1(1):43-50. doi: https://doi.org/10.30560/mhs.v1n1p43


38. Pearson K. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, Lond Edinb Dublin Philos Mag J Sci, 1900; 50(302):157-175.


