Smoking is the cause of lung cancer

Ilija Barukčić[1]

[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.

ABASTRACT

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 H_0: *without* smoking *no* lung cancer. Furthermore, the null-hypothesis H_0: 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	Size	Total
Records identified by searching in the databases		
PubMed	7798	
Google Scholar	0	
Web of Science	0	
Additional records identified from other sources	2	7800
2. Clean up of search		
Records removed after verifying duplication	1	
Records excluded by title	6989	
Records excluded by the summary	523	287
(Articles outside the inclusion criteria)		
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

Study ID	Year	N	Case_P	Case_T	Con_P	Con_T	Odds Ratio	OR lower	OR upper	p(SINE)	X2 (SINE)	p Value (SINE)	k	IOU
Müller	1939	172	83	86	30	86	51,64	15,03	177,45	0,98256	0,1046511628	0,7463	0,6491	0,1570
Doll & Hill	1952	2714	1350	1357	1289	1357	10,17	4,66	22,23	0,99742	0,0361090641	0,8493	0,1371	0,4724
Lombard et. al	1965	2080	1026	1040	928	1040	8,84	5,04	15,53	0,99327	0,1884615385	0,6642	0,1975	0,4394
Wynder et. al	1979	10231	659	684	5156	9547	22,45	15,03	33,54	0,99756	0,9137426901	0,3391	0,2135	-0,3648
Benhamou et al.	1985	3132	1184	1217	1392	1915	13,48	9,40	19,33	0,98946	0,8948233361	0,3442	0,3138	0,2110
Harris et. al	1993	5530	2829	2916	1997	2614	10,05	7,97	12,67	0,98427	2,5956790123	0,1072	0,3089	0,4000
Pershagen et. al	1994	3983	1100	1136	2607	2847	2,81	1,97	4,02	0,99096	1,1408450704	0,2855	0,0935	0,2159
Sobue et. al	1994	2197	1022	1056	1013	1141	3,80	2,58	5,60	0,98452	1,0946969697	0,2954	0,1529	0,4069
Jöckel et. al	1998	2008	949	1004	768	1004	5,30	3,89	7,22	0,97261	3,0129482072	0,0826	0,2561	0,3551
Kreuzer et. al	1998	3470	1687	1709	1358	1761	22,76	14,73	35,16	0,99366	0,2832065535	0,5946	0,3294	0,3700
Boffetta et al.	1999	12876	5504	5621	5505	7255	14,95	12,36	18,10	0,99091	2,4353317915	0,1186	0,3104	0,2916
	Total	48393	17393	17826	22043	30567				0,9911	12,70			

Alpha =	0,05
Degrees of freedom =	11
X² CRITICAL (SINE) =	19,68
X² Calculated (SINE) =	12,70
p value (SINE) =	0,31

Table 2. Smoking[6, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20] is the cause of lung cancer.

Study ID	Year	N	Case_P	Case_T	Con_P	Con_T	X^2 (Sine)	k	k lower	k upper	p value (k)
Müller	1939	172	83	86	30	86	0,10	0,65	0,48	0,82	0,0000000000
Doll & Hill	1952	2714	1350	1357	1289	1357	0,04	0,14	0,09	0,18	0,0000000000
Lombard et. al	1965	2080	1026	1040	928	1040	0,19	0,20	0,15	0,25	0,0000000000
Wynder et. al	1979	10231	659	684	5156	9547	0,91	0,21	0,19	0,24	0,0000000000
Benhamou et al.	1985	3132	1184	1217	1392	1915	0,89	0,31	0,27	0,35	0,0000000000
Harris et. al	1993	5530	2829	2916	1997	2614	2,60	0,31	0,28	0,34	0,0000000000
Pershagen et. al	1994	3983	1100	1136	2607	2847	1,14	0,09	0,06	0,13	0,0000000001
Sobue et. al	1994	2197	1022	1056	1013	1141	1,09	0,15	0,11	0,20	0,0000000000
Jöckel et. al	1998	2008	949	1004	768	1004	3,01	0,26	0,21	0,31	0,0000000000
Kreuzer et. al	1998	3470	1687	1709	1358	1761	0,28	0,33	0,29	0,37	0,0000000000
Boffetta et al.	1999	12876	5504	5621	5505	7255	2,44	0,31	0,29	0,33	0,0000000000
	Total	**48393**	17393	17826	22043	30567	12,70	0,32	0,31	0,32635	0,0000000000

Alpha =	0,05
Degrees of freedom =	11
X^2 CRITICAL (k) =	19,68
X^2 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 $a = (a_1 + a_2 + ... + a_n)$ 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

$$a \equiv (a_1 + a_2 + \cdots + a_n) \equiv \sum_{t=1}^{t=n} a_t \tag{1}$$

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

$$b \equiv (b_1 + b_2 + \cdots + b_n) \equiv \sum_{t=1}^{t=n} b_t \tag{2}$$

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

$$c \equiv (c_1 + c_2 + \cdots + c_n) \equiv \sum_{t=1}^{t=n} c_t \tag{3}$$

Let d_t = 1 if the t-*th* outcome is a success and 0 if it is a failure. Then $d = (d_1 + d_2 + ... + d_n)$ is the number of successes in *n* Bernoulli trials (period of time) t. It is $p(d_t)= p(\underline{A}_t \cap \underline{B}_t)$ the joint probability of ($\underline{A}_t$ and $\underline{B}_t$) 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

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

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

$$\underline{A} \equiv ((c_1 + d_1) + (c_2 + d_2) + \cdots + (c_n + d_n)) \equiv \sum_{t=1}^{t=n} \underline{A}_t \quad (6)$$

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

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

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

$$\underline{B} \equiv ((b_1 + d_1) + (b_2 + d_2) + \cdots + (b_n + d_n)) \equiv \sum_{t=1}^{t=n} \underline{B}_t \quad (8)$$

At each Bernoulli trial it is

$$n_t \equiv (a_t + b_t + c_t + d_t) \equiv A_t + \underline{A}_t \equiv B_t + \underline{B}_t \tag{9}$$

and the sample size n itself equal to

$$n \equiv \sum_{t=1}^{n} (a_t + b_t + c_t + d_t) \equiv \sum_{t=1}^{n} A_t + \underline{A}_t \equiv \sum_{t=1}^{n} B_t + \underline{B}_t \tag{10}$$

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

		Conditioned B (Outcome)		
		Yes = +1	**No = +0**	**Total**
Condition A	Yes =+1	a	b	A
(risk factor)	No = +0	c	d	$\underline{A}$
	Total	B	$\underline{B}$	n

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 $p(A_t) = p(A_t \cap B_t)+p(A_t \cap \underline{B}_t)$ 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 $p(\underline{B}_t) = 1- p(B_t)$ 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 probabitlities of a contingency table

		Conditioned B		
		Yes = +1	**No = +0**	**Total**
Condition A	Yes =+1	$p(a_t)$	$p(b_t)$	$p(A_t)$
	No = +0	$p(c_t)$	$p(d_t)$	$p(\underline{A}_t)$
	Total	$p(B_t)$	$p(\underline{B}_t)$	**1**

Definition. Index of unfairness

T*he index of unfairness* (IOU) is defined as

$$IOU \equiv \left(\left(\frac{A \quad + B}{n}\right) - 1\right) \tag{11}$$

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

$$p(A_t \cap B_t) \equiv p(A_t) \times p(B_t) \tag{12}$$

Definition. Sufficient Condition (Conditio per Quam)

The mathematical formula of the *sufficient* condition relationship [23, 24, 25, 26, 27, 28, 29, 30, 31] (*conditio per quam*) of a population is defined as

$$
\begin{aligned}
p(A_t \rightarrow B_t) &\equiv \frac{(a_t) + (c_t) + (d_t)}{N_t} = 1 \\
&\equiv p(a_t) + p(c_t) + p(d_t) \\
&\equiv p(B_t) + p(d_t) \\
&\equiv p(a_t) + p(\underline{A}_t) \\
&\equiv +1.
\end{aligned}
\tag{13}
$$

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× 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

$$X^2\left(\left(A \quad \rightarrow B \quad\right)|A \quad\right) \equiv \frac{\left(\left(b \quad\right)-\left({}^{1}/_{2}\right)\right)^2}{A}+0=0 \qquad ()$$

or alternatively as

$$X^2\left(\left(A \quad \rightarrow B \quad\right)|\underline{B} \quad\right) \equiv \frac{\left(\left(b \quad\right)-\left({}^{1}/_{2}\right)\right)^2}{\underline{B}}+0=0 \qquad (14)$$

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

$$p(A_t \leftarrow B_t) \equiv \frac{(a_t) + (b_t) + (d_t)}{N_t} = 1$$
$$\equiv p(a_t) + p(b_t) + p(d_t)$$
$$\equiv p(A_t) + p(d_t) \tag{15}$$
$$\equiv p(a_t) + p(\underline{B}_t) = p(a_t) + (1 - p(B_t))$$
$$\equiv +1.$$

Definition. The X² 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 [23, 24, 25, 26, 27, 28, 29, 30, 31] before changes to

$$X^2\left((A \leftarrow B)|B\right) \equiv \frac{\left((c) - (1/2)\right)^2}{B} + 0 = 0 \tag{16}$$

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

$$X^2\left((A \leftarrow B)|\underline{A}\right) \equiv \frac{\left((c) - (1/2)\right)^2}{\underline{A}} + 0 = 0 \tag{17}$$

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 [23, 24, 25, 26, 27, 28, 29, 30, 31] as

$$
\begin{aligned}
p(A_t \mid B_t) &\equiv \frac{(b_t) + (c_t) + (d_t)}{N_t} = 1 \\
&\equiv p(b_t) + p(c_t) + p(d_t) \\
&\equiv p(b_t) + p(\underline{A}_t) = p(b_t) + \left(1 - p(A_t)\right) \\
&\equiv p(c_t) + p(\underline{B}_t) = p(c_t) + \left(1 - p(B_t)\right) \\
&\equiv +1.
\end{aligned}
\tag{18}
$$

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=1of the exclusion relationship [23, 24, 25, 26, 27, 28, 29, 30, 31] with a *continuity correction* can be calculated as

$$
X^2\left((A \mid B) \mid A\right) \equiv \frac{\left((a) - (1/2)\right)^2}{A} + 0 = 0 \tag{19}
$$

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

$$
X^2\left((A \mid B) \mid B\right) \equiv \frac{\left((a) - (1/2)\right)^2}{B} + 0 = 0 \tag{20}
$$

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 [23, 24, 25, 26, 27, 28, 29, 30, 31] is defined *at every single event, at every single Bernoulli trial t,* as

$$k(A_t, B_t) \equiv \frac{p(A_t \cap B_t) - \left(p(A_t) \times p(B_t)\right)}{\sqrt[2]{p(A_t) \times \left(1 - p(A_t)\right) \times p(B_t) \times \left(1 - p(B_t)\right)}} \tag{21}$$

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

$$\left\{k(A_t, B_t) - \sqrt[2]{\frac{5}{N}} \quad ; \quad k(A_t, B_t) + \sqrt[2]{\frac{5}{N}}\right\} \tag{22}$$

Definition. The rule of three

Under some specified conditions (i. e. the dataset analyzed is large enough or *n*, the sample size, is *n* ~ 30 and more), a Chi-square[38] goodness of fit test is able to provide evidence whether a *sample* distribution observed is identical with a *theoretica*l distribution expected. Formally, the Chi-square goodness of fit test is defined as X^2 = ((sample distribution) - (*theoretica*l distribution))^2/(*theoretica*l distribution) or something like X^2=((observed)-(*expected*))^2/(*expected*). An approximate and conservative (one sided) confidence interval as discussed[39,40,41,42] 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

$$p_{Critical} \quad = \quad 1 - \left(\frac{3}{n}\right) \tag{23}$$

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 π_{upper}

Tests of hypotheses concerning the sampling distribution of the sample proportion **p** (i. e. conditio sine qua non p(SINE), conditio per quam p(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 π_{upper} can be performed under conditions of sampling without replacement (Sachs, 1992) by the formula

$$p_{critical\,upper} \quad = \quad \left(p - \frac{1}{2 \times n}\right) - \left(Z \times \sqrt[2]{\left(\frac{p \times (1-p)}{n}\right) \times \left(\frac{N-n}{N-1}\right)}\right) \tag{24}$$

while the term ((N-n)/(N-1)) denotes *the finite*[43] *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

$$OR(A\ \ ,B\ \ ) \equiv \frac{a\ \ /b}{c\ \ /d} \equiv \frac{a\ \ \times d}{c\ \ \times b} \tag{25}$$

Although severely and justifiably criticized especially by Karl Pearson (1857–1925), the long-time 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

$$SE\left(ln\left(OR(A\ \ ,B\ \ )\right)\right) \equiv \sqrt[2]{\left(\frac{1}{a}\right)+\left(\frac{1}{b}\right)+\left(\frac{1}{c}\right)+\left(\frac{1}{d}\right)} \tag{26}$$

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

$$\begin{aligned} 95\,\%\,CI \equiv\ & exp\left(ln\left(OR(A\ \ ,B\ \ )\right)-\left(1.96\times SE\left(ln\left(OR(A\ \ ,B\ \ )\right)\right)\right)\right) \\ & to \\ & exp\left(ln\left(OR(A\ \ ,B\ \ )\right)+\left(1.96\times SE\left(ln\left(OR(A\ \ ,B\ \ )\right)\right)\right)\right) \end{aligned} \tag{27}$$

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

$$X^2 \equiv \sum_{t=1}^{k} \frac{\left(\left(x_t\right) - \left(n \times p\left(x_t\right)\right)\right)^2}{\left(n \times p\left(x_t\right)\right)} \tag{28}$$

Example.

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

Table 5. A fair coin.

Event	Observed (x_t)	Expected ($n \times p(x_t)$)	((x_t)- ($n \times p(x_t)$))	(((x_t)- ($n \times p(x_t)$))^2)/ ($n \times p(x_t)$)
Heads	40	50	-10	$(-10)^2/50 = 2$
Tails	60	50	+10	$(+10)^2/50 = 2$
n	100	100		$X^2 = 4$

In this context, the chi-square goodness of fit test[50] 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

$$X^2 \equiv \sum_{t=1}^{k} \frac{\left(\left(x_t\right) - \left(n \times p\left(x_t\right)\right)\right)^2}{\left(n \times p\left(x_t\right)\right)} = \frac{(40-50)^2}{50} + \frac{(60-50)^2}{50} = = \frac{100}{50} + \frac{100}{50} = 2 + 2 = 4 \tag{29}$$

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 6. The critical values of the chi square distribution (degrees of freedom: 1)

	p-Value	One sided X²	Two sided X²
The chi square distribution	0.1000000000	1.642374415	2.705543454
	0.0500000000	**2.705543454**	**3.841458821**
	0.0400000000	3.06490172	4.217884588
	0.0300000000	3.537384596	4.709292247
	0.0200000000	4.217884588	5.411894431
	0.0100000000	5.411894431	6.634896601
	0.0010000000	9.549535706	10.82756617
	0.0001000000	13.83108362	15.13670523
	0.0000100000	18.18929348	19.51142096
	0.0000010000	22.59504266	23.92812698
	0.0000001000	27.03311129	28.37398736
	0.0000000100	31.49455797	32.84125335
	0.0000000010	35.97368894	37.32489311
	0.0000000001	40.46665791	41.82145620

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 <u>is not</u> the cause of lung cancer. In other words. $k = 0$.

Alternative Hypothesis:

Smoking <u>is</u> the cause of lung cancer. In other words. $k \neq 0$.

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 ($k \sim 0.32$, X^2 Calculated (k) = 4837.90 and is greater than X^2 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[51, 52, 53] of lung cancer. In particular, about 433 from 48393 patients (Table 7) have been treated as *never smoker* while in reality it cannot be excluded that these patients were exposed to passive smoking.

Table 7.					Statistical analysis.		Index of unfairness =	0,183
The studies re-analysed					causal relationship k =	0,316181947.	95 % CI (k): (0,30- 0,326)	
		Lung cancer			p-value (k | HGD) =	0	X^2 (k) =	4838
		YES	NO		Odds ratio (OR) =	15,53	14,07	17,14
Smoking	YES	17393	22043	39436	p (SINE) =	0,991052425	X^2 (SINE) =	12,7
	NO	433	8524	8957				
		17826	30567	**48393**				

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

1. Hasse KE. An Anatomical Description of the Diseases of the Organs of Circulation and Respiration, London, Sydenham Society; 1846. http://archive.org/details/ananatomicaldes02hassgoog. Accessed January 25, 2019.

2. Ferlay J, Colombet M, Soerjomataram I, et al. Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods, Int J Cancer, December 2018. doi: https://doi.org/10.1002/ijc.31937

3. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA Cancer J Clin, 2018; 68(6):394-424. doi: https://doi.org/10.3322/caac.21492

4. Sant M, Allemani C, Santaquilani M, Knijn A, Marchesi F, Capocaccia R. EUROCARE-4. Survival of cancer patients diagnosed in 1995–1999. Results and commentary, Eur J Cancer, 2009; 45(6):931-991. doi: https://doi.org/10.1016/j.ejca.2008.11.018

5. Thun MJ. Early landmark studies of smoking and lung cancer, Lancet Oncol, 2010; 11(12):1200. doi: https://doi.org/10.1016/S1470-2045(09)70401-2

6. Müller FH. Tabakmißbrauch und Lungencarcinom, Z Für Krebsforsch, 1940; 49(1):57-85. doi: https://doi.org/10.1007/BF01633114

7. Morabia A. Quality, originality, and significance of the 1939 “Tobacco consumption and lung carcinoma” article by Mueller, including translation of a section of the paper, Prev Med, 2012; 55(3):171-177. doi: https://doi.org/10.1016/j.ypmed.2012.05.008

8. Parsons A, Daley A, Begh R, Aveyard P. Influence of smoking cessation after diagnosis of early stage lung cancer on prognosis: systematic review of observational studies with meta-analysis, BMJ, 2010; 340(jan21 1):b5569-b5569. doi: https://doi.org/10.1136/bmj.b5569

9. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA Statement, Open Med Peer-Rev Indep Open-Access J, 2009; 3(3):123-130.

10. Liberati A, Altman DG, Tetzlaff J, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration, PLoS Med, 2009; 6(7):1000100. doi: https://doi.org/10.1371/journal.pmed.1000100

11. Doll R, Hill AB. Study of the Aetiology of Carcinoma of the Lung, Br Med J, 1952; 2(4797):1271-1286. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2022425/. Accessed January 22, 2019.

12. Lombard HL. An epidemiological study in lung cancer, Cancer, 1965; 18(10):1301-1309.

13. Wynder EL, Stellman SD. Impact of long-term filter cigarette usage on lung and larynx cancer risk: a case-control study, J Natl Cancer Inst, 1979; 62(3):471-477.

14. Benhamou S, Benhamou E, Tirmarche M, Flamant R. Lung cancer and use of cigarettes: a French case-control study, J Natl Cancer Inst, 1985; 74(6):1169-1175.

15. Harris RE, Zang EA, Anderson JI, Wynder EL. Race and Sex Differences in Lung Cancer Risk Associated with Cigarette Smoking, Int J Epidemiol, 1993; 22(4):592-599. doi: https://doi.org/10.1093/ije/22.4.592

16. Pershagen G, Akerblom G, Axelson O, et al. Residential radon exposure and lung cancer in Sweden, N Engl J Med, 1994; 330(3):159-164. doi: https://doi.org/10.1056/NEJM199401203300302

17. Sobue T, Suzuki T, Fujimoto I, et al. Case-control study for lung cancer and cigarette smoking in Osaka, Japan: comparison with the results from Western Europe, Jpn J Cancer Res Gann, 1994; 85(5):464-473.

18. Jöckel KH, Ahrens W, Jahn I, Pohlabeln H, Bolm-Audorff U. Occupational risk factors for lung cancer: a case-control study in West Germany, Int J Epidemiol, 1998; 27(4):549-560.

19. Kreuzer M, Kreienbrock L, Gerken M, et al. Risk factors for lung cancer in young adults, Am J Epidemiol, 1998; 147(11):1028-1037.

20. Boffetta P, Pershagen G, Jockel K-H, et al. Cigar and Pipe Smoking and Lung Cancer Risk: a Multicenter Study From Europe, JNCI J Natl Cancer Inst, 1999; 91(8):697-701. doi: https://doi.org/10.1093/jnci/91.8.697

21. Moivre A de [1667-1754]. The Doctrine of Chances or a Method of Calculating the Probability of Events in Play, London: printed by W. Pearson for the author; 1718. doi: https://doi.org/10.3931/e-rara-10420

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

23. Barukčić I. Die Kausalität, 1. Aufl. Hamburg: Wiss.-Verl.; 1989.

24. Barukčić I. The Mathematical Formula of the Causal Relationship k, Int J Appl Phys Math, 2016; 6(2):45-65. doi: https://doi.org/10.17706/ijapm.2016.6.2.45-65

25. Barukčić I. Theoriae Causalitatis Principia Mathematica, Norderstedt: Books on Demand; 2017.

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

27. Barukčić I. Human Papillomavirus—The Cause of Human Cervical Cancer, J Biosci Med, 2018d; 06(04):106-125. doi: https://doi.org/10.4236/jbm.2018.64009

28. Barukčić I. Human Cytomegalovirus is the Cause of Glioblastoma Multiforme, Mod Health Sci, 2018; 1(2):19. doi: https://doi.org/10.30560/mhs.v1n2p19

29. Barukčić I. The Equivalence of Time and Gravitational Field, Phys Procedia, 2011; 22:56-62. doi: https://doi.org/10.1016/j.phpro.2011.11.008

30. Barukčić I. Epstein-barr virus is the cause of multiple sclerosis, Int J Curr Med Pharm Res, 2018; 4(9 (A)):3674-3682. doi: https://doi.org/10.24327/23956429.ijcmpr20180538

31. Barukčić K, Barukčić JP, Barukčić I. Epstein-Barr virus is the cause of rheumatoid arthritis, Romanian J Rheumatol, 2018; 27(4):148-163. https://view.publitas.com/amph/rjr_2018_4_art-02/page/1. Accessed December 29, 2018.

32. Wertheimer R. Conditions, J Philos, 1968; 65:355–364. http://www.jstor.org/stable/2023797.

33. Gomes G. Are Necessary and Sufficient Conditions Converse Relations, Australas J Philos, 2009; 87:375–387. doi: https://doi.org/10.1080/00048400802587325

34. Fisher RA. On the Interpretation of χ 2 from Contingency Tables, and the Calculation of P, J R Stat Soc, 1922; 85(1):87. doi: https://doi.org/10.2307/2340521

35. Yates F. Contingency Tables Involving Small Numbers and the χ 2 Test, J R Stat Soc Suppl, 1934; 1(2):217. doi: https://doi.org/10.2307/2983604

36. Grizzle JE. Continuity Correction in the χ 2 -Test for 2 × 2 Tables, Am Stat, 1967; 21(4):28. doi: https://doi.org/10.2307/2682103

37. Conover WJ. Some Reasons for Not Using the Yates Continuity Correction on 2×2 Contingency Tables, J Am Stat Assoc, 1974; 69(346):374-376. doi: https://doi.org/10.1080/01621459.1974.10482957

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.

39. Rumke CL. Implications of the Statement: No Side Effects Were Observed, N Engl J Med, 1975; 292(7):372-373.

40. Louis TA. Confidence Intervals for a Binomial Parameter after Observing No Successes, Am Stat, 1981; 35(3):154-154.

41. Hanley JA. If Nothing Goes Wrong, Is Everything All Right?, JAMA, 1983; 249(13):1743.

42. Jovanovic BD, Levy PS. A Look at the Rule of Three, Am Stat, 1997; 51(2):137-139.

43. Isserlis L. On the Value of a Mean as Calculated from a Sample, J R Stat Soc, 1918; 81(1):75. doi: https://doi.org/10.2307/2340569

44. Fisher RA. The Logic of Inductive Inference, J R Stat Soc, 1935; 98(1):39. doi: https://doi.org/10.2307/2342435

45. Cornfield J. A method of estimating comparative rates from clinical data; applications to cancer of the lung, breast, and cervix, J Natl Cancer Inst, 1951; 11(6):1269-1275.

46. Edwards AWF. The Measure of Association in a 2 × 2 Table, J R Stat Soc Ser Gen, 1963; 126(1):109. doi: https://doi.org/10.2307/2982448

47. Mosteller F. Association and Estimation in Contingency Tables, J Am Stat Assoc, 1968; 63(321):1. doi: https://doi.org/10.2307/2283825

48. Pearson K, Heron D. On Theories of Association, Biometrika, 1913; 9(1-2):159-315. doi: https://doi.org/10.1093/biomet/9.1-2.159

49. Altman DG. Practical Statistics for Medical Research, Boca Raton, Fla: Chapman & Hall/CRC; 1999.

50. Sachs L. Angewandte Statistik, Berlin, Heidelberg: Springer Berlin Heidelberg; 1992.

51. Taylor R, Cumming R, Woodward A, Black M. Passive smoking and lung cancer: a cumulative meta-analysis, Aust N Z J Public Health, 2001; 25(3):203-211.

52. Sheng L, Tu J-W, Tian J-H, Chen H-J, Pan C-L, Zhou R-Z. A meta-analysis of the relationship between environmental tobacco smoke and lung cancer risk of nonsmoker in China, Medicine (Baltimore), 2018; 97(28):e11389. doi: https://doi.org/10.1097/MD.0000000000011389

53. Kim A-S, Ko H-J, Kwon J-H, Lee J-M. Exposure to Secondhand Smoke and Risk of Cancer in Never Smokers: A Meta-Analysis of Epidemiologic Studies, Int J Environ Res Public Health, 2018; 15(9):1981. doi: https://doi.org/10.3390/ijerph15091981

