Statistics and Application 统计学与应用, 2021, 10(5), 892-899
Published Online October 2021 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2021.105093

Hans汉斯

糖尿病患者慢性并发症及危险因素分析

李　超*，曾　莎，张晓良，瞿颖秋

重庆理工大学理学院，重庆

收稿日期：2021年9月21日；录用日期：2021年10月6日；发布日期：2021年10月21日

摘　要

本文对200例II型糖尿病患者慢性并发症及危险因素进行研究。首先，对患者年龄与性别构成比，以及并发疾病构成进行描述性统计分析。发现患者主要集中在40岁~69岁年龄段，且不同年龄段的性别构成差异较大；在糖尿病患者并发视网膜病变的情况下，再并发其他疾病，如并发高血压、肾病、动脉粥样硬化等比例非常大。然后，基于Logistic逐步回归、随机森林算法、相关性检验三种方法对糖尿病患者并发高血压的危险因素进行综合分析。结果表明：AGE (年龄)，BP_HIGH (收缩压)，SCR (血肌酐)，SUA (血清尿酸)，BMI (体重指数)，BU (血尿素)是关键的危险因素。

关键词

糖尿病及慢性并发症，Logistic回归，随机森林，相关性检验

Analysis of Diabetes with Chronic Complications and Its Risk Factors

Chao Li*, Sha Zeng, Xiaoliang Zhang, Yingqiu Qu

School of Science, Chongqing University of Technology, Chongqing

Received: Sep. 21st, 2021; accepted: Oct. 6th, 2021; published: Oct. 21st, 2021

Abstract

200 cases of type II diabetes with chronic complications and its risk factors were studied in this paper. First of all, descriptive statistical analysis was conducted on the age and sex ratio of patients, as well as the composition of concurrent diseases. It was found that the patients mainly concentrated in the age group of 40~69 years old, and the sex composition of different age groups was different; in the situation of diabetic patients with retinopathy, there are a large proportion of

*通讯作者。

文章引用: 李超, 曾莎, 张晓良, 瞿颖秋. 糖尿病患者慢性并发症及危险因素分析[J]. 统计学与应用, 2021, 10(5): 892-899. DOI: 10.12677/sa.2021.105093

other concurrent diseases, such as hypertension, nephropathy and atherosclerosis. Then, based on Logistic stepwise regression, random forest algorithm and test of association, the risk factors of hypertension in diabetic patients were comprehensively analyzed. The results showed that: AGE, BP_HIGH, SCR, SUA, BMI, and BU are the key risk factors.

Keywords

Diabetes and Chronic Complications, Logistic Regression, Random Forest, Test of Association

1. 引言

随着经济社会发展，人们的生活水平逐渐提高，一些慢性病的发病率也逐渐升高。例如糖尿病，糖尿病是一种代谢紊乱性疾病[1]，也是一种多因素疾病，受到遗传、社会、生活等多方面影响。因此，掌握糖尿病患病率和并发症情况，了解其危险因素的分布具有重要意义。

许多学者对 II 型糖尿病及其并发症进行了研究。谢利平等对 1436 例 II 型糖尿病患者慢性并发症进行研究，发现各种糖尿病并发症合并率约为 84.26%，主要危险因素包括年龄、糖尿病病程、糖化血红蛋白等[2]；邹飒枫等对糖尿病及并发症进行研究，发现高血压合并糖尿病患者多见于 60 岁及以上人群，并发视网膜病变、冠心病和脑卒中的危险有着随年龄增高而增加的趋势[3]；李纯净等通过变量筛选结果发现，AGE_ONSE、ACR、PLAS_CR 和 HB 这四个因素对心血管疾病的影响尤为显著[4]。此外，在与高血压有关的危险因素中，彭浩等[5]、吴云涛等[6]、张红叶等[7]的研究发现，血清尿酸增加是高血压前期人群进展至高血压的独立危险因素。在这些研究中，大多数学者都是采用 Logistic 逐步回归方法对糖尿病患者慢性并发症及危险因素进行统计分析，没有采用多种方法综合分析，因此本文考虑 Logistic 逐步回归、随机森林算法和相关性检验三种方法，对 200 例糖尿病并发视网膜病变再并发其他疾病及危险因素进行综合分析。

2. 数据来源及预处理

2.1. 数据来源

本文数据来源于“国家人口健康科学数据中心数据仓储 PHDA”，由中国人民解放军总医院提供，一共收集了 200 例糖尿病并发视网膜病变患者数据集[8]。其中，主要包含患病种类以及 32 种糖尿病并发症疾病患病数据，如高血压、高脂血、动脉粥样硬化、脑卒中、颈动脉狭窄等；同时，包含相关因素数据，如 AGE (年龄)，SEX (性别)，BP_HIGH (收缩压)，MARITAL_STATUS (婚姻状态)，HB (血红蛋白)等共 54 个因素。

2.2. 数据预处理

2.2.1. 缺失数据处理

首先，由于部分特征缺失严重，删除缺失率大于 0.2 的特征，剩下 36 个特征；其次，按性别分组分别对身高和体重用对应的均值对缺失值进行填充，然后根据 BMI 公式计算对应的体重指数 BMI；之后，其余特征的缺失数据，则是用对应均值进行填充。由于 BMI 综合反映了身高和体重的信息，危险因素分析中，只考虑 BMI，不考虑身高和体重，即共有 AGE (年龄)，SEX (性别)，BP_HIGH (收缩压)，MARITAL_STATUS (婚姻状态)，HB (血红蛋白)等 34 个特征纳入到危险因素分析中。

2.2.2. 数据标准化

为避免数据量纲或数量级差异造成的负面影响，需要对数据进行标准化处理。标准化处理范围主要是除二分类变量(如性别，婚姻状态，是否患病)外的特征，如 BMI (体重指数)，SCR (血肌酐)，SUA (血清尿酸)等。一般的标准化方法有 Min-max 标准化，z-score 标准化，中心化标准化。本文采用 z-score 标准化方法，计算公式为：$x^* = \frac{x-\mu}{\sigma}$，其中 x 表示观测值，μ 表示样本均值，σ 表示样本标准差。

3. 并发症以及危险因素分析

3.1. 描述性分析

3.1.1. 200 例 II 型糖尿病患者的年龄及性别构成比

本数据集包括 200 例糖尿病并发视网膜病变患者及其他并发症疾病相关数据，其中，患者最小年龄为 20 岁，最大年龄为 85 岁。对 200 例糖尿病并发视网膜病变患者年龄以及性别构成进行描述性统计分析，可以得到下表 1：

Table 1. Age and sex composition of 200 cases with type II diabetes
表 1. 200 例 II 型糖尿病患者的年龄及性别构成

年龄(岁)	男性(n)	占比(%)	女性(n)	占比(%)	合计(n)	合计占比(%)	性别比
20~29	2	1.6	3	4	5	2.5	0.67:1
30~39	6	4.8	1	1.33	7	3.5	6.00:1
40~49	35	28	11	14.67	46	23	3.18:1
50~59	40	32	18	24	58	29	2.22:1
60~69	37	29.6	29	38.67	66	33	1.28:1
70~79	5	4	10	13.33	15	7.5	0.50:1
80~89	0	0	3	4	3	1.5	0.00:1
合计	125	100	75	100	200	100	1.67:1

为了更直观地观察各年龄段中糖尿病患者男女构成情况，绘制各个年龄段患者构成情况的折线图如下：

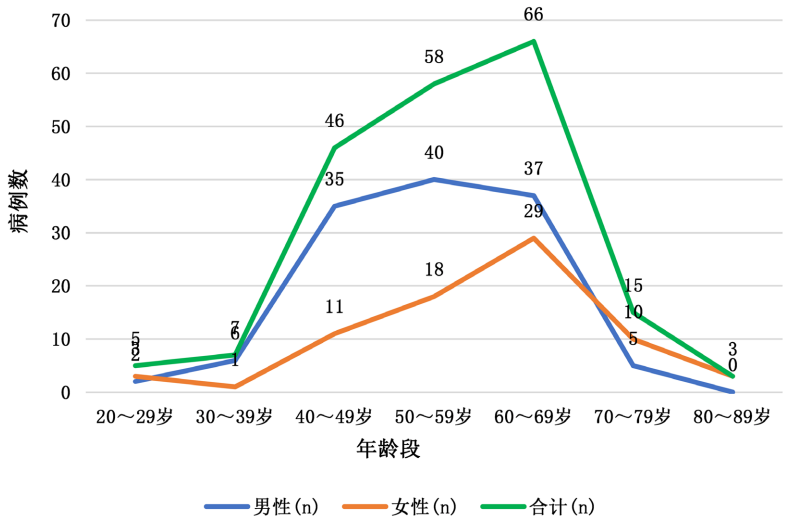

Figure 1. Age group composition of patients
图 1. 各个年龄段患者构成情况

由表 1 及图 1 可见，200 例糖尿病并发视网膜病变患者中，在年龄组方面，构成占比最高的是 60 岁~69 岁，其次 50 岁~59 岁，之后是 40 岁~49 岁。在性别构成方面，30 岁~70 岁年龄段的男性患者明显多于女性患者，而其他年龄组中女性患者多于男性患者。

总的来说患者主要集中在 40 岁~69 岁年龄段，占总人数的 85%。这提示，当年龄到达 40 岁之后，就要警惕糖尿病并发视网膜病变的发生。从性别比可见，在 30 岁~39 岁年龄组中性别差异最大，男性：女性性别比达到了 6:1，这也提示处于这个年龄组中的男性要警惕糖尿病并发视网膜病变的发生。

3.1.2. 200 例 II 型糖尿病患者慢性并发症的疾病构成

首先，为直观观察已并发视网膜病变糖尿病患者慢性并发症疾病构成，绘制柱状图如下：

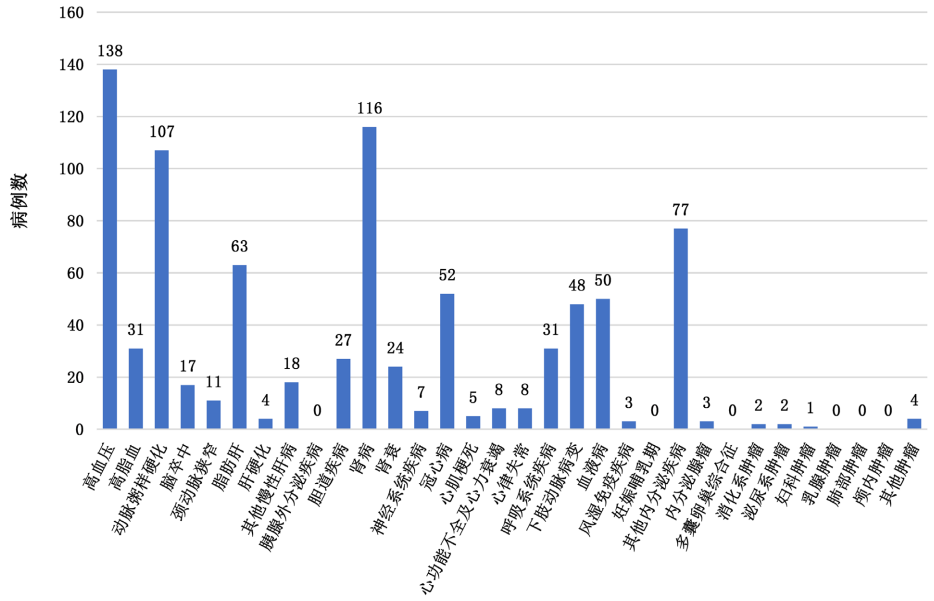

Figure 2. Constitution of diabetic complications
图 2. 糖尿病并发症疾病构成

从图 2 可发现，并发高血压，肾病，动脉粥样硬化的病例数较多。通过绘制柱状图对并发症情况有了初步了解，为了进一步探究糖尿病并发症疾病构成情况，计算每种疾病的占比后得到糖尿病并发症构成表如下：

Table 2. Composition of complications
表 2. 并发症疾病构成表

并发症	例数(n)	构成比(%)	并发症	例数(n)	构成比(%)
高血压	138	69	心功能不全及心力衰竭	8	4
肾病	116	58	心律失常	8	4
动脉粥样硬化	107	53.5	神经系统疾病	7	3.5
其他内分泌疾病	77	38.5	心肌梗死	5	2.5
脂肪肝	63	31.5	肝硬化	4	2
冠心病	52	26	其他肿瘤	4	2
血液病	50	25	风湿免疫疾病	3	1.5
下肢动脉病变	48	24	内分泌腺瘤	3	1.5

Continued

高脂血	31	15.5	消化系肿瘤	2	1
呼吸系统疾病	31	15.5	泌尿系肿瘤	2	1
胆道疾病	27	13.5	妇科肿瘤	1	0.5
肾衰	24	12	胰腺外分泌疾病	0	0
其他慢性肝病	18	9	妊娠哺乳期	0	0
脑卒中	17	8.5	多囊卵巢综合征	0	0
颈动脉狭窄	11	5.5	乳腺肿瘤	0	0

由表 2 可见，在糖尿病患者并发视网膜病变的情况下，再发生其他并发疾病的比例非常大。其中，并发高血压、肾病、动脉粥样硬化、其他内分泌疾病、脂肪肝的比例非常大，分别是 69%、58%、53.5%、38.5%、31.5%。因此，对糖尿病患者并发视网膜病变时，再并发其他慢性疾病危险因素的分析是十分有必要的。接下来，将对糖尿病患者并发视网膜病变合并并发高血压的危险因素进行分析。

3.2. 并发症危险因素的 Logistic 回归和随机森林分析

为了探究糖尿病患者并发高血压的危险因素，本文中，响应变量为分类变量 HYPERTENTION (高血压)，协变量包括 AGE (年龄)，SEX (性别)，BP_HIGH (收缩压)，MARITAL_STATUS (婚姻状态)，HB (血红蛋白)，SCR (血肌酐)，SUA (血清尿酸)，BMI (体重指数)，BU (血尿素)，PLT (血小板)等 34 个相关因素。

3.2.1. Logistic 回归分析

Logistic 回归是一种非线性回归模型，实际上是一种分类模型，并常用于二分类问题研究中。Logistic 回归可以表示为：

$$\ln\frac{p}{1-p}=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_p\beta_p+\cdots$$

使用 Logistic 回归逐步回归可以筛选出对响应变量有显著影响的协变量，因此，多数糖尿病与并发症及其危险因素的研究基于 Logistic 回归进行。

按照前述响应变量和协变量的设置，采用极大似然参数估计方法，使用 Logistic 回归，进行逐步回归分析，得到糖尿病并发高血压危险因素如下表 3 所示：

Table 3. Logistic stepwise regression: risk factors
表 3. Logistic 逐步回归：危险因素表

危险因素	Coefficient	P-value	OR	95%置信区间	
				下限	上限
AGE	0.483	0.015	1.621	1.097	2.395
BP_HIGH	1.201	0.000	3.322	1.951	5.658
BMI	0.415	0.047	1.514	1.006	2.281
SCR	2.209	0.009	9.107	1.743	47.598
SUA	0.448	0.081	1.565	0.946	2.590

由表 3 可见，根据 Logistic 逐步回归分析，有五个特征可以看做是糖尿病患者并发视网膜病变时，再合并并发高血压的危险因素，它们分别是 AGE (年龄)，BP_HIGH (收缩压)，BMI (体重指数)，SCR (血肌酐)，SUA (血清尿酸)。

3.2.2. 随机森林(Random Forest)

为了对糖尿病并发高血压危险因素进行对比研究，本文在 Logistic 回归的基础上，引入随机森林(Random Forest)算法，对糖尿病患者并发视网膜病变合并并发高血压的危险因素进行分析。

随机森林算法中，对于每个特征会给出对应的特征重要性。与系数不同，特征重要性是根据学习过程中使用特征信息的多少来决定，取值范围 0~1 之间，而且它们的和为 1。在本例中，按照前述响应变量和协变量设置，由特征重要性得出糖尿病患者并发视网膜病变合并并发高血压的危险因素如表 4。

Table 4. Random forest: Risk factors
表 4. 随机森林：危险因素表

特征	特征重要性
BP_HIGH	0.098161
BU	0.077872
SCR	0.051431
BMI	0.046114
HB	0.04197
LDH_L	0.041069
SUA	0.039595
AGE	0.038302
GGT	0.037713
ALB	0.032912

为便于和基于 Logistic 回归得到危险因素进行比较，选择随机森林得到的前五个危险因素，分别是：BP_HIGH (收缩压)、BU (血尿素)、SCR (血肌酐)、BMI (体重指数)、HB (血红蛋白)。这里可见，两种方法得到的危险因素，共同的部分有：BP_HIGH、BMI、SCR；差异部分为，Logistic 回归得到的危险因素包含有 AGE 和 SUA，而随机森林得到危险因素包含有 BU 和 HB。为进一步确定主要的危险因素，接下来考虑流行病学研究里基于渐进无条件方法的相关性检验，对上述 7 个危险因素进行研究，最后综合三种方法结果确定危险因素。

3.3. 相关性检验

当暴露和非暴露队列中的发病概率($\pi_i, i=1,2$)相等时，就说暴露和疾病没有关联。为了检验暴露和疾病是否有关，设原假设为：$H_0:\pi_1=\pi_2$。根据 Newman [9]，可以使用皮尔逊(Pearson)，Wald，Likelihood 相关性检验对原假设进行检验，检验统计量分别为：

$$X_w^2=\left(\log\widehat{OR}\right)^2\left(\frac{1}{\hat{e}_1}+\frac{1}{\hat{e}_2}+\frac{1}{\hat{f}_1}+\frac{1}{\hat{f}_2}\right)^{-1}=\frac{\left(\log\widehat{OR}\right)^2 r_1r_2m_1m_2}{r^3}$$

$$X_p^2=\frac{\left(a_1-\hat{e}_1\right)^2}{\hat{e}_1}+\frac{\left(a_2-\hat{e}_2\right)^2}{\hat{e}_2}+\frac{\left(b_1-\hat{f}_1\right)^2}{\hat{f}_1}+\frac{\left(b_2-\hat{f}_2\right)^2}{\hat{f}_2}$$

$$X_{lr}^2=2\left[a_1\log\left(\frac{a_1}{\hat{e}_1}\right)+a_1\log\left(\frac{a_2}{\hat{e}_2}\right)+b_1\log\left(\frac{b_1}{\hat{f}_1}\right)+b_2\log\left(\frac{b_2}{\hat{f}_2}\right)\right]$$

其中，$a_i, i=1,2$ 表示暴露和非暴露队列中观测患病人数，$\hat{e}_i, i=1,2$ 表示在原假设下，暴露和非暴露队列

中期望患病人数；$b_i, i=1,2$ 表示暴露和非暴露队列中观测未患病人数，$\hat{f}_i, i=1,2$ 表示在原假设下，暴露和非暴露队列中期望未患病人数；$\widehat{OR}$ 表示机会比的估计。在大样本时，上述统计量渐近服从于自由度为 1 的卡方分布。

本文在显著性水平 $\alpha=0.05$ 下，对 7 个危险因素进行相关性检验。首先根据暴露确定观测频数以及计算期望频数的 2×2 列联表，然后计算相应的检验统计量以及 p 值，对原假设进行检验。

对于 AGE (年龄)，将年龄大于 40 岁视为暴露，可以得到对应的观测频数和期望频数列联表：

Table 5. Age-hypertension observation (expectation) counts contingency table
表 5. AGE-高血压观测(期望)频数列联表

		暴露		
		是	否	合计
高血压	是	134 (129.72)	54 (58.28)	188 (188)
	否	4 (8.28)	8 (3.72)	12 (12)
	合计	138 (138)	62 (62)	200 (200)

根据表 5 的结果，可以计算得到相关性检验统计量以及 p 值的值如下：

$$X_p^2=\frac{(134-129.72)^2}{129.72}+\frac{(54-58.28)^2}{58.28}+\frac{(4-8.28)^2}{8.28}+\frac{(8-3.72)^2}{3.72}=7.49\ (p=0.0062)$$

$$X_w^2=(\log 4.96)^2\left(\frac{1}{129.72}+\frac{1}{58.28}+\frac{1}{8.28}+\frac{1}{3.72}\right)=6.12\ (p=0.0134)$$

$$X_{lr}^2=2\times\left[134\log\frac{134}{129.72}+54\log\frac{54}{58.28}+4\log\frac{4}{8.28}+8\log\frac{8}{3.72}\right]=6.92\ (p=0.0085)$$

类似地，对于其他相关因素分别根据暴露确定观测频数和计算期望频数列联表，再计算相应检验统计量与 p 值，将结果整理得到表 6。

Table 6. Pearson, Wald, LR Test Statistic and p value of 7 risk factors
表 6. 7 个危险因素的 Pearson，Wald，LR 检验统计量及其 p 值

危险因素	X_p^2	p 值	X_w^2	p 值	X_{lr}^2	p 值
AGE	7.49	0.0062	6.12	0.0134	6.92	0.0085
BP_HIGH	35.13	3.0841e−9	53.5	2.5875e−13	34.38	4.5336e−9
HB	4.13	0.0421	3.14	0.0764	3.76	0.0525
SCR	22.1	2.5881e-6	43.3	4.6958e−11	25.78	3.8263e−7
SUA	10.25	0.0014	11.2	8.1797e−4	10.7	0.0011
BMI	4.71	0.0300	4.42	0.0355	4.54	0.0331
BU	19.29	1.1229e−5	29.1	6.8737e−8	21.4	3.7277e−6

从表 6 见，在对 7 个相关因素进行的相关性检验中，HB 这个因素的相关性检验，三种检验的 p 值分别为 0.0421，0.0764，0.0525，认为在显著性水平 $\alpha=0.05$ 下，没有充分的证据支持拒绝原假设，即没有充分的证据支持高血压发病与 HB 是否暴露有关。显然，其余的 6 个相关因素的相关性检验中，检验的 p 值都远远小于显著性水平 $\alpha=0.05$，认为有充分的证据支持拒绝原假设，即该 6 个危险因素的暴露与否

对高血压发病有显著性影响。

综合三种方法对糖尿病患者并发视网膜病变再并发高血压的危险因素分析，认为 AGE (年龄)，BP_HIGH (收缩压)，SCR (血肌酐)，SUA (血清尿酸)，BMI (体重指数)，BU (血尿素)是关键的危险因素。

4. 总结与讨论

本文采用 200 例 II 型糖尿病患者及其并发症数据，首先对数据进行描述性统计分析，发现患者主要集中在 40 岁~69 岁年龄段，占总人数的 85%；在糖尿病患者并发视网膜病变的情况下，再发生其他并发疾病的比例非常大。其中，并发高血压、肾病、动脉粥样硬化、其他内分泌疾病、脂肪肝的比例排名靠前。然后，基于 Logistic 逐步回归、随机森林算法、相关性检验对糖尿病患者并发视网膜病变再并发高血压的危险因素进行分析，结果表明 AGE (年龄)，BP_HIGH (收缩压)，SCR (血肌酐)，SUA (血清尿酸)，BMI (体重指数)，BU (血尿素)是关键的危险因素。

本文仍可改进，特征筛选方法还包括 lasso 回归和 ridge 岭回归方法，但在本文中表现不好。因此，在分析其他并发症的危险因素或者提升样本量后，可以考虑这两种方法。

参考文献

[1] 楼雪勇. 2 型糖尿病患者慢性并发症患病率及危险因素分析[J]. 中国现代医生, 2011, 49(4): 120-121.

[2] 谢利平, 王国洪. 1436 例 2 型糖尿病患者慢性并发症合并率及相关因素的统计分析[J]. 中国病案, 2015, 16(8): 92-94.

[3] 邹飒枫, 乔晶. 4158 例高血压合并糖尿病病人现况及并发症分析[J]. 中华疾病控制杂志, 2010, 14(11): 1124-1125.

[4] 李纯净, 丁雪, 李芸, 等. 删失数据下部分线性模型对糖尿病并发症的统计分析[J]. 长春工业大学学报, 2020, 41(2): 105-111+209.

[5] 彭浩, 丁建松, 彭颖, 等. 女性人群血清尿酸水平与高血压及高血压前期的关系[J]. 中华高血压杂志, 2011, 19(3): 236-239.

[6] 吴云涛, 吴寿岭, 李云, 等. 血清尿酸对高血压前期人群血压转归的影响[J]. 中华高血压杂志, 2010, 18(6): 545-549.

[7] 张红叶, 李莹, 陶寿淇, 等. 血清尿酸与四年后血压变化及高血压发病的关系[J]. 高血压杂志, 2001(2): 72-75.

[8] 中国人民解放军总医院. 共享杯版_糖尿病并发症预警数据集[DB/OL] https://www.ncmi.cn/phda/dataDetails.do?id=CSTR:A0006.11.A0005.202006.001018, 2021-06-28.

[9] Newman, S.C. (2001) Biostatistical Methods in Epidemiology. 2nd Edition, John Wiley & Sons, Inc., New York, 94-98.