Predicting Health Care Risk with Big Data Drawn from Clinical Physiological Parameters

Honghao Wei[1,*], Yang Yang[1], Huan Chen[2],
Bin Xu[1], Jian Li[3], Miao Jiang[4,*], and Aiping Lu[4]

[1] Tsinghua University, Beijing, China
[2] Beijing University of Posts and Telecommunications, Beijing, China
[3] School of Basic Medical Sciences, Beijing University of Chinese Medicine, Beijing, China
[4] Institute of Basic Research of Clinical Medicine,
China Academy of Chinese Medical Sciences, Beijing 100700, China
O_HENRYWILL@126.com, miao_jm@vip.126.com

Abstract. Fatty liver often afflicts patients seriously and jeopardizes the health of human race with high possibility of deteriorating into cirrhosis and liver cancer, which motivates researchers to detect causes and potential influential factors. In this paper, we study the problem of detecting the potential influential factors in workplaces and their contributions to the morbidity. To this end, gender and age, retirement status and department information are chosen as three potential influential factors in workplaces. By analyzing those factors with demographics, Propensity Score Matching and classic classifier models, we mine the relationship between the workplace factors and morbidity. This finding indicates a new domain of discussing the causes of fatty liver which originally focuses on daily diets and lifestyles.

Keywords: Fatty liver, gender and age, retirement status, department information.

1 Introduction

Fatty liver disease (FLD), or simply fatty liver, is a very common chronic disease, whose prevalence continues to increase, especially in the "Western" world. According to the research by DeNoon (2013) [7], fatty liver disease occurs in 33% of European-Americans, 45% of Hispanic-Americans, and 24% of African-Americans. While large vacuoles of fat accumulate in liver cells via the process of steatosis, patients with fatty liver first suffer from an enlarged liver or vague right upper abdominal pains. Subsequently, in a subpopulation of patients, the disease progresses to more severe liver diseases, such as cirrhosis or total liver failure.[1] It eventually deteriorates into liver cancer, which damages tissues over time, significantly reduces the quality of life and ultimately leads to premature death.

Fortunately, fatty liver is reversible. If patients are diagnosed and treated promptly, the risk of the disease would decline significantly. In recent studies, alcoholic abuse,

* Corresponding authors.

H.-Y. Huang et al. (Eds.): SMP 2014, CCIS 489, pp. 88–98, 2014.

together with obesity, are found to be the most influential factors. Though breakthroughs have been made in the pathology of the disease, more undetected causes are there. They are hard to ignore since they can increase the incidence magnificently.

Currently, vast majority of passionate researches focus on the daily diets and lifestyles. Previous researchers have found direct impacts of alcohol and obesity. Teli (1995) [15] reported the case of alcohol-induced fatty liver and suggested the hazard of excessive intake of alcohol. Ueno (1997) [17] noticed the increasing incidence of obese patients with fatty liver and proposed restricted diet and exercise therapy as useful methods to combat the disease. Similar work has been done by Tominaga (1995) [16], who elicited a direct relationship between the degree of obesity and fatty liver in children.

Influence analysis has attracted considerable research interests and is becoming a popular research topic (Wanless 1990[18], Angulo 2002[3], Marchesini 2003[12], Fabbrini 2010[8],). However, most exciting works focused on daily diets and lifestyles. On the contrary, the influential factors in workplaces have been largely ignored. Since the working hours occupy approximately half of an adult's time, it is reasonable to involve workplace factors in the exploration of causes of fatty liver.

In this paper, we aim to quantitatively study the influential factors in workplaces and how they affect the incidence of fatty liver. Our objective is to effectively and efficiently discover the underlying influence pattern related to workplace.

The rest of the paper is organized as follows: Sect.2 formally formulates the problem; Sect.3 proposes relevant methods; Sect.4 introduces evaluation aspects and classic classifier models, and presents experimental results validating our assumptions; Sect.5 displays new discoveries and offers the potential explanations; Sect.6 concludes.

2 Problem Definition

In this section, we define several relevant concepts in order to formulate the fatty liver prediction problem.

The following concepts are introduced:

2.1 Definition 1. Lab Test Records

We define the lab test record of patient n to be $\mathrm{X}_n = \{X_I\}$ where X_I denotes a lab test performed on the patient n. We also use set $\{\mathrm{a}_1, \mathrm{a}_2, \dots, \mathrm{a}_{46}\}$ to represent the lab test items evaluating the health condition of the patient, and for each patient we finally choose 46 major items to denote their health status among a range of alternative attributes. Therefore, we have $X_I = \{\mathrm{a}_1, \mathrm{a}_2, \dots, \mathrm{a}_{46}\}$. We denote $\mathrm{Y_n}$ as a label node that indicates whether the patient n has fatty liver.

In our experiment, we use the classical models roughly to evaluate the potentially relevant lab test items and select 46 items of them to be the attributes to forecast the results. The selected attributes are listed as follows in table 1:

We screen fatty liver cases from people in several stable organizations of Chaoyang District, Beijing, China. The data are obtained from their physical examinations. Meanwhile, a patient could have several lab tests in different organizations and the results could be different. While we could deduce the differences of health conditions of the patient, the lab tests from different organizations with different features and instructions could make a difference. In order to guarantee the integrity of the data and avoid extra turbulence of different tests, we acquire the data of four years from one single medical organization. In addition, all the people are instructed to receive all the lab test items so that our 46 attributes are ensured to exist in the test report.

Table 1. Attribute in test

Attribute	Gender	Age	Systolic Pressure	Weight	Height
	BMI	Pro	Diastolic Pressure	Thyroid	LYM
	MONO	NEUT	LYM#	MONO	NEUT
	RBC	MCV	Hematocrit	MCH	MCHC
	PLT	MPV	PCT	PDW	RDW
	HGB	DB	GLU(UR)	KET	SG
	OB	pH	Waistline	URO	NIT
	WBC	ALT	Cr	URIC	GLU
	CHO	TG	Gallbladder	AST	HDL
	Kidney				

Consequentially, we select the lab test results from 2010 to 2013 and utilize our proposed models to forecast the prevalence and uncover its potential relationship with influential factors in workplaces.

2.2 Definition 2. Fatty Liver

In this paper, we define fatty liver as the condition in which large vacuoles of triglyceride fat accumulate in liver cells. This is usually accompanied by defect of the ability to respond to metabolic stress. With the impairment of this ability, a number of diseases and disorders can be incurred.

All the patients diagnosed with fatty liver have undertaken physical exams, blood tests and ultrasound examinations. Liver enzymes higher than normal level can be found and fat in the liver has been detected in a computed tomography scan or magnetic resonance imaging test. In clinical practice, patients with these clinical traits are also considered as having the disease and we adopt the same criterion in our study[11].

3 Approach

In this section, we briefly introduce the Logistic Regression and Propensity Score Matching.

3.1 Logistic Regression

Logistic Regression is a type of probabilistic statistical classification model. Considering p independent variables $x = \{x_1, x_2, \ldots, x_p\}$, we define conditional probability $P(Y = 1|x) = p$ as the possibility of occurrence based on measurement. The explanation of Logistic Regression begins with the possibility of occurrence. We define it as follows:

$$P(Y = 1|x) = \frac{1}{1 + e^{-g(x)}}$$

Where $g(x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$

The possibility of non-occurrence refers to be:

$$P(Y = 0|x) = 1 - P(Y = 1|x) = 1 - \frac{e^{g(x)}}{1 + e^{g(x)}} = \frac{1}{1 + e^{g(x)}}$$

In this way, the odds of experiencing an event (the odds of the possibility of occurrence and non-occurrence) is

$$\frac{P(Y = 1|x)}{P(Y = 0|x)} = \frac{p}{1-p} = e^{g(x)}$$

The logarithmic term of the above is the linear function that we demand.

In case we have n observation samples, the observed values refer to $\{y_1, y_2, \ldots, y_n\}$. The probability of occurrence could be written as

$$y_i = p_i^{y_i}(1 - p_i)^{1 - y_i}$$

Since all observations are independent, we get the likelihood function

$$l(\beta) = \Pi_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1 - y_i}$$

Using Maximum likelihood estimation, such as Newton-Raphson's method [21], we determine the coefficient before each variable.

3.2 Propensity Score Matching

In order to estimate the influence of a single attribute, we create a dynamic matched sample of treated and untreated nodes. The treated and untreated groups feature differently in the targeted attribute. Bias might arise because of the apparently different outcomes of these two groups of units. In experiments, the randomization enables unbiased estimation of treatment effects by the law of large numbers. However, unfortunately, in observational studies, the assignment of treatments to research subjects is not randomized. For the purpose of eliminating the bias between treated and untreated units and narrowing down the differences, we adopt Propensity Score Matching. Full details regarding Propensity Score Matching methods are provided in Caliendo's (2008) [5] guidance for the implementation.

In the implementation of Propensity Score Matching, we match treated nodes with untreated ones that are likely to have the same attributes, such as similar health condition etc. For every examinee (we label him or her as the t th examinee in the treated group), we estimate pit, the propensity to have been treated, using a logistic

regression of the likelihood of health status, which is conditional on a vector of observable characteristics and clinical traits (X). It is listed as follows:

$$p_{it} = P(x_{it} = 1 \mid X_{it}) = \frac{e^{g(x)}}{1 + e^{g(x)}}$$

Where $g(x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$

We drop matched pairs of which the distance of propensity scores exceeds two standard deviations. For all treated examinee i, we choose an untreated match j such that $\left|\left|p_{it} - p_{jt}\right|\right|$ is minimized to conform to $\min\left|\left|p_{it} - p_{jt}\right|\right| \leq 2\,\sigma_d$ where σ_d is the standard deviation of the distance $\left|\left|p_{it} - p_{jt}\right|\right|$. We then compare fractions of treated (n+) and untreated (n-) adopters. Once the odd exceeds integer 1, the targeted attribute is believed to have impacts on the final result [4], [14].

4 Experiment

We use a collection of real medical records from a famous hospital in Beijing. The data set spans four years, containing 5535 medical records corresponding to 1985 individual patients and 91 kinds of lab tests in total.

We view each examinee's record as an instance, and aim to infer whether the corresponding indicators have potential effects on and contributions to fatty liver. Three influential factors are taken into account in our experiments: gender and age, retirement status, and department information.

We use demographic methods to demonstrate the influence of age and gender. In the meantime, for the purpose of good matching, Propensity Score Matching is proposed to evaluate the influence of retirement status. Last but not least, we adopts several classic classifier models to test the department attribute.

4.1 Evaluation Aspects

We evaluate our methods from the following two aspects:

Precision and Recall. In pattern recognition and information retrieval with binary classification, precision is the fraction of retrieved instances that are relevant, while recall is the fraction of relevant instances that are retrieved.

F-measure. The F-measure can be interpreted as a weighted average of the precision and recall, where an F-measure reaches its best value at 1 and worst value at 0. The f-measure is defined as:

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall},$$

We compare the results of fatty liver forecast generated by following methods:

Logistic Regression. Lab test results are treated as features and LR (Alison 1999) [2] is employed as the classification model for fatty liver disease forecasting.

Naive Bayes. A traditional Naive Bayes model (Frank, Trigg, Holmes, Witten 2000) [10] is used for prediction.

J48 Decision Tree. C4.5 Decision Tree is considered as well. For convenience, we employ the J48 Decision Tree provided by Weka 3.6 (Frank 2004) [9].

All algorithms are implemented in Java and C++, and all experiments are performed on a Lenovo laptop running Win 7 with AMD 1.4 GHz and 4 GB of memory.

4.2 Gender and Age

We imply the potential differences among distinct age and gender groups on the possibility of getting fatty liver. In our demographic test, we select 5535 instances to demonstrate the relationship. Among them, male examinees and female examinees are approximately half-and-half, which guarantees the representativeness and universality of the data.

By way of eliminating extreme samples, we set a series of criterion to screen the data (see Table 2). All selected female and male participants should meet the following standards in Table 2.

The demographic results are showed in Fig.1 and Fig.2. With these figures, we infer that age factor shares linear relationship with the incidence of fatty liver. Therefore, we analyze the data with linear fitting. For female participants, the correlation coefficient is 0.9221. In contrast, the correlation coefficient of the male participants is only 0.7081. The large gap between the correlation coefficients of two gender groups needs further discussion.

Table 2. Criterion for examinee enrollment

Clinical item	Female	Male
Systolic Pressure	>= 110 and <= 140	>= 110 and <= 140
Diastolic Pressure	>= 70 and <= 90	>= 70 and <= 90
Weight	>= 45 and <= 75	>= 60 and <= 80
Height	>= 155 and <= 175	>= 165 and <= 185
BMI	>= 15 and <= 25	>= 19 and <= 25
Blood Sugar	<= 6	<= 6
Cholesterol	<= 6	<= 6
Glycerol acid	<= 2	<= 2

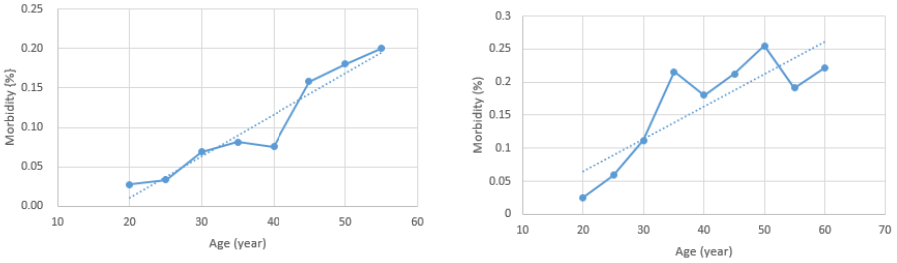

Fig. 1. Morbidity and Age (female participants (left) and male participants (right))

4.3 Retirement Status

In the interests of large number laws, the number of unretired and retired employees involved in the matching exceeds 800. Additionally, we utilize Propensity Score Matching to create reasonable pairs of the retired and unretired with similar health conditions and limited age differences.

Fig. 3 demonstrates the result of Propensity Score Matching. Obviously, all the final odds each year exceed integer 1. Though the odds fluctuates due to the different data each year, it is still convincing that retirement status contributes to prevalence of fatty liver significantly.

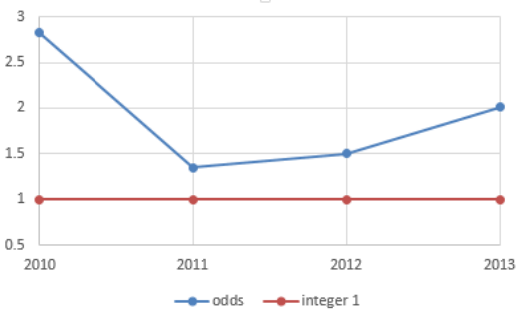

Fig. 2. The odds of Propensity Score Matching in different years

4.4 Department Information

The demographics of the rate of fatty liver suggest the potential relationship between career and the prevalence of illness. To illustrate the situation, 7 typical departments have been chosen and the corresponding morbidity has been calculated.

From table 3, the differentiation of morbidity among departments has been evidently displayed. The incidence of Human Resources Department is five times as much as that of Research and Development center.

Table 3. Incidence of fatty liver in different departments

Department	Incidence
Technique Department	31.25%
HR Department	61.36%
Machining Department	31.19%
Administrative Department	27.86%
Finance Department	26.32%
Express Mail Services	23.08%
Research and Development Center	13.79%

To further clarify the influence of careers and working departments, we randomly take HR Department and Technique Department as an instance. Three classic classifier models are used to test the examples. They are Logistic Regression, Naive Bayes and J48 Decision Tree.

Table 4 confirms the theory and all these three models exhibit that applying career and department information as attributes would increase the F1-measure. The most apparent improvement of the performance rests in Logistic Regression. The F1-measure increases up to 14.16%.

Table 4. Prediction results about department information

Class	Method	Precision	Recall	F1
With	LR	0.650	0.648	0.629
Department	Naive Bayes	0.636	0.639	0.636
Information	J48	0.584	0.593	0.561
Without	LR	0.561	0.571	0.551
Department	Naive Bayes	0.616	0.620	0.615
Information	J48	0.560	0.574	0.547

5 Discussion

In this section, we mainly deal with the results in the experiment part. Through further analysis, we draw the conclusion on the relationship between the workplace factors and morbidity.

5.1 Gender and Age

The linear relationship displayed in Fig. 1 and Fig. 2 clarifies the potential contributions of gender and age to the morbidity of fatty liver. We notice that the morbidity increases with age. It conforms to the regular senescence patterns of human beings.

However, there are three more interesting questions worth noting: 1) Why the correlation coefficient of male examinees are much lower than that of the female examinees; 2) Why the apparent skewing of female and male participants occurs at different ages; 3) Why at those ages, the incidence of fatty liver within the female units inclines to decrease while the incidence within male units increases sharply?

To better answer these questions, we need to understand the different career choice and social roles of different genders. In the sample selected, male participants tend to take career more physically and psychologically demanding, such as manual work or

scientific research. Female participants, on contrary, usually work in civil services, such as accountants. The pressure of the career differs and the external interferences are more likely to disrupt the male workers' original internal senescence patterns, thus resulting in the decline of correlation coefficient. The obesity status recognizes this assumption. Table 5 illustrates the BMI categories for male and female. It is worth noting that much higher percentage of male participants (46.41%) than female (21.79%) are overweight or even obesity. It is ultimately consistent with fatty liver studies since obesity is one of the major causes of the illness (Centis 2013) [6].

Table 5. BMI categories for male and female

Class	Male	Female
Underweight (BMI $\leq$18.5)	2.65%	7.69%
Normal weight (BMI = 18.5 $\sim$ 24.9)	50.95%	70.51%
Overweight (BMI = 25 $\sim$ 29.9)	40.23%	18.91%
Obesity (BMI $\geq$30)	6.18%	2.88%

To be more convincing, we analyze the important timing point and age bracket of these two units. For women, age 40 is a watershed. The incidence of fatty liver in the early 40s is significantly under prediction by linear regression while during late 40s the prevalence soars high. We deduce that, by the age of 40, women have fulfilled the responsibility about marriage and conception. The era of promoting in career has passed as well (O'Brien 1993) [13]. Therefore, the 40s women tends to be more relaxed, tolerant and resilient which aid to confront the metabolic diseases such as fatty liver. Additionally, the physical condition of women deteriorates swifter than men. It usually takes place at the age of 40s when obesity begins to increase. Clinical study has also proved that the fatty liver peak for women is usually 40s with endocrine disorders (Yan 2013) [20]. Therefore, the rate of fatty liver is raised as expected in the late 40s.

For male participants, the sudden rush of morbidity takes place at age 35 and 50. They are the important preparation periods for man: one for the golden development era in 40s, another for the last chance of promotion in 50s. The relevant performances would be irregular and greasy diet, insufficient rest time and work in overload. It is found in clinical practices that the intake of edible oil and alcohol increases while the sleeping hour decreases significantly during the period (welsh 2012) [19]. This suits the career pattern of a man and strengthens the relationship between the morbidity and workplace factors.

5.2 Retirement Status

By comparing the senior citizens with similar age and health condition, yet with different retirement status in Propensity Score Matching, the contribution of retirement status becomes obvious. The odds of data in four years are larger than the integrity 1.

It is noticed that those examinees who retire at their 60's usually occupy the important position in workplaces in the last years of their careers. They are either the leader or the person in charge. The pressure and demands, which probably account for the temporary decline of morbidity, varies greatly before and after retirement.

5.3 Department Information

To paraphrase the effect of department information, we need to take these elements into consideration: gender rate, average age and the specialty of career. Firstly, the specialty of career distinguishes the morbidity in different professions. For instance, the department with highest incidence of fatty liver are HR department and machining department. Both of these jobs are physically demanding and exhausting. What's more, the sex rate and average age directly make accounts in tackling the problem. The average age in Research and Development Center is 32 while the average age in Machining is 39. The percentage of female employees in Machining is below 14% while the percentage in Administrative Department is up to 90%. From the last two parts, we conclude that women have comparatively lower morbidity than men and the rate increases with age. Therefore, the situation in different departments differs.

6 Conclusion

In the framework of data analysis, this paper aims to investigate the potential causes of fatty liver in workplaces regarding gender and age, retirement status and department information. Data from 5535 instances suggests the potential relationship between the influential factors in workshop and the prevalence of fatty liver. The skewing from linear relation of female in early 40s and of male in 35 and 50s suits the career pattern of two units. Additionally, the subtle effect of retirement status on decreasing the morbidity temporarily has been detected. Furthermore, we explore the different morbidity in distinct departments and the situation is in line with our assumption.

Based on these findings, in order to improve the prediction of the disease and take precautions, female employees in late 40s and male in 35 and 50s should be suggested to take regular physical examination and regular life pattern is needed. Moreover, seniors before the retirement age should be instructed to avoid work overload and learn to release stress. Last but not least, the health condition of particular departments should be paid special attention and the employers should be provided with more opportunities for body check and rest.

Regarding some limitations of this study, further research is necessary.

First, this study aims to find the causes in workplaces. To confirm the conclusion of this study, further studies are needed to investigate more factors especially those unique ones in workplaces.

Next, the relationship among the influential factors has not yet been discussed. It is unclear what the most influential factors is in workplaces and whether some of the factors would affect others (e.g. the influence of gender differences on retirement status regarding the morbidity).

At last, the targeted solutions and their impacts should be explored and examined. Further guidelines for early warning and diagnosis are needed as well, to develop a better understanding of fatty liver and share this with others.

References

1. Adler, M., Schaffner, F.: Fatty liver hepatitis and cirrhosis in obese patients. The American Journal of Medicine 67(5), 811–816 (1979)
2. Allison, P.D.: Logistic regression using the SAS system. SAS Institute. Inc., Cary (1999)
3. Angulo, P.: Nonalcoholic fatty liver disease. New England Journal of Medicine 346(16), 1221–1231 (2002)
4. Aral, S., Muchnik, L., Sundararajan, A.: Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. Proceedings of the National Academy of Sciences 106(51), 21544–21549 (2009)
5. Caliendo, M., Kopeinig, S.: Some practical guidance for the implementation of propensity score matching. Journal of Economic Surveys 22(1), 31–72 (2008)
6. Centis, E., Moscatiello, S., Bugianesi, E., et al.: Stage of change and motivation to healthier lifestyle in non-alcoholic fatty liver disease. Journal of Hepatology 58(4), 771–777 (2013)
7. De Noon, D.J.: Fatty liver disease: Genes affect risk (2013)
8. Fabbrini, E., Sullivan, S., Klein, S.: Obesity and nonalcoholic fatty liver disease: biochemical, metabolic, and clinical implications. Hepatology 51(2), 679–689 (2010)
9. Frank, E., Hall, M., Trigg, L., et al.: Data mining in bioinformatics using Weka. Bioinformatics 20(15), 2479–2481 (2004)
10. Frank, E., Trigg, L., Holmes, G., et al.: Technical note: Naive Bayes for regression. Machine Learning 41(1), 5–25 (2000)
11. Liver, F.: Guidelines for diagnosis and treatment of nonalcoholic fatty liver diseases. Chinese Journal of Hepatology 3 (2006)
12. Marchesini, G., Bugianesi, E., Forlani, G., et al.: Nonalcoholic fatty liver, steatohepatitis, and the metabolic syndrome. Hepatology 37(4), 917–923 (2003)
13. O'Brien, K.M., Fassinger, R.E.: A causal model of the career orientation and career choice of adolescent women. Journal of Counseling Psychology 40(4), 456 (1993)
14. Peikes, D.N., Moreno, L., Orzol, S.M.: Propensity score matching. The American Statistician 62(3) (2008)
15. Teli, M.R., James, O.F.W., Burt, A.D., et al.: The natural history of nonalcoholic fatty liver: A follow-up study. Hepatology 22(6), 1714–1719 (1995)
16. Tominaga, K., Kurata, P.J.H., Chen, Y.K., et al.: Prevalence of fatty liver in Japanese children and relationship to obesity. Digestive Diseases and Sciences 40(9), 2002–2009 (1995)
17. Ueno, T., Sugawara, H., Sujaku, K., et al.: Therapeutic effects of restricted diet and exercise in obese patients with fatty liver. Journal of Hepatology 27(1), 103–107 (1997)
18. Wanless, I.R., Lentz, J.S.: Fatty liver hepatitis (steatohepatitis) and obesity: An autopsy study with analysis of risk factors. Hepatology 12(5), 1106–1110 (1990)
19. Welsh, J.A., Karpen, S., Vos, M.B.: Increasing prevalence of nonalcoholic fatty liver disease among United States adolescents, 1988-1994 to 2007-2010. The Journal of Pediatrics 162(3), 496–500, e1 (2013)
20. Yan, J., Xie, W., Ou, W., et al.: Epidemiological survey and risk factor analysis of fatty liver disease of adult residents, Beijing, China. Journal of Gastroenterology and Hepatology 28(10), 1654–1659 (2013)
21. Jennrich, R.I., Sampson, P.F.: Newton-Raphson and related algorithms for maximum likelihood variance component estimation. Technometrics 18(1), 11–17 (1976)