实验 14、统计评论

年级：2015 级　专业：生物信息学　评分：

学号：1530416025　姓名：刘伊娜　签名：

编号	一	二	三	四	总分	评阅人
得分						

一、目的要求：

1、加深对医学统计中常用统计量的理解;

2、熟悉并掌握医学统计中常用统计量的计算方法。

二、软硬件平台：

1. 硬件平台：（硬件配置）

(1). CPU: 1.6 GHz Intel Core i5

(2). 内存：4 GB 1600 MHz DDR3

(3). 硬盘：APPLE SSD 251 GB

2. 系统平台：（操作系统及其版本号）

Mac OS Sierra 10.12

3. 软件平台:（软件系统及其版本号,若是在线分析平台,还需要提供 URL 地址）

XAMPP 5.6.28-1

三、实验内容：

1、对某一基因突变相关的心血管疾病或肿瘤发病风险的独立研究报道进行统计评论。

所用文献：《TCF7L2 基因多态性与新疆维吾尔族人群 2 型糖尿病的相关性研究》

1.1、摘要评价：

(1)研究的问题是否重要?

答：很重要。糖尿病在全世界的发病率有逐年增高的趋势，已经成为世界上继肿瘤、心脑血管病之后第三位严重危害人类健康的慢性疾病。且糖尿病对人类健康危害最大的是在动脉硬化及微血管病变基础上产生的多种慢性并发症，主要是危害心、脑、肾、血管、神经、皮肤等。TCF7L2 基因是目前发现的与 2 型糖尿病相关性最强的基因。故可通过研究 2 型糖尿病发病风险与 TCF7L2 基因单核苷酸多态性位点之间的关系，针对携带不同基因或基因型的人群进行发病风险的有效预测，从而达到预防疾病发生的效果，提高人们的健康水平，故此研究关注的问题具有重要意义。

(2)统计结果描述是否足够清晰明确?

答：足够清晰。本研究中对统计结果的描述如下：

两组 rs12255372 多态性位点 3 种基因型，T、 G 等位基因频率比较，差异均有统计学意义（$P<0.05$）。非条件 Logistic 回归分析显示，T 等位基因携带者患 2 型糖尿病的风险为 G 等位基因携带者的 1.238 倍〔OR = 1.238，95% CI (1.049，1.461)，$P<0.05$〕；校正年龄、性别后 GT 基因型携带者患 2 型糖尿病的风险为 GG 基因型携带者的 1.221 倍〔OR = 1.221，95% CI (1.001，1.490)，$P<0.05$〕；GT + TT 基因型携带者患 2 型糖尿病的风险为 GG 基因型携带者的 1.252 倍〔OR = 1.252，95% CI (1.033，1.516)，$P<0.05$〕。两组 rs7085532 多态性位点 3 种基因型，A、G 等位基因频率比较，差异均无统计学意义（$P>0.05$）。

可看出在对统计结果进行描述时，研究人员对统计分析所得的理论数字进行了对其所含的统计学意义的挖掘。不仅对不同基因型进行了分析讨论，还对不同等位基因进行了分析讨论，说明了其所用的统计学方法并给出了 OR 值以及 95%CI；此外，在对 rs12255372 多态性位点进行分析时，发现了差异具有统计学意义后还对年龄和性别进行校正后再次进行了统计分析，使所得结果更加可靠。故此文献对统计结果描述还是较清晰明确的。

(3)统计结果能否支撑结论?

答：本研究结论如下：

TCF7L2 基因 rs12255372 多态性位点可能与新疆维吾尔族人群 2 型糖尿病相关，而 rs7085532 多态性位点可能与新疆维吾尔族人群 2 型糖尿病无相关性。

可看出本文中对统计结果是可以支撑结论的。

1.2、研究设计:

(1)用了何种研究类型?是否合适?

答：本研究使用的是 case-control 研究，是合适的研究方法。本研究中选定了患有 2 型糖尿病的维吾尔族病例组共 969 例，以及未患 2 型糖尿病的对照组 958 例，分别调查其暴露(如环境因素、遗传因素、内分泌作用以及保护因子的缺乏)于危险因子（GT+TT）的情况及程度，以判断暴露危险因子与 2 型糖尿病有无关联及关联程度大小的。

(2)采用了何种统计方法?是否合适?

答：通过阅读文献可知，本研究使用的统计学方法如下：

（1）采用 SPSS 16. 0 软件包对数据进行处理，计算 TCF7L2 基因型及等位基因频率，检验各群体是否符合 Hardy－Weinberg 遗传平衡原理，能否代表相应人群;

（2）计量资料以（ x ± s）表示，两组间比较采用 t 检验及协方差分析；
（3）等位基因及基因型频率比较采用 x 2 检验；
（4）采用非条件 Logistic 回归分析等位基因及基因型与疾病的相关性（以 P＜0. 05 为差异有统计学意义）。

本研究中分别运用了 t 检验，方差分析，卡方检验和非条件 Logistic 回归分析来研究其关心的不同方面，之后综合各方面的统计所得数据和研究结果对所研究的问题下结论，且在针对不同问题对统计方法的选择合理，故本研究所使用的统计方法是还较为得当的。

1.3、数据：

(1)样本是如何获得的?选样过程是否存在偏倚?样本量是否足够?

答：本研究所涉及的 2 型糖尿病组（T2MD 组）均来自 2012 年 3 月—2013 年 3 月在新医科大学第一附属医院就诊的维吾尔族 2 型糖尿病患者人群，共 969 例，其中男 606 例、女 363 例，平均年龄（51.2±9.7）岁。

选择同期在该院体检中心体检健康 者作为正常对照组（ NC 组），共 958 例，其中男 609 例、女 349 例，平均年龄（ 50. 3 ± 9. 8）岁。

此外，研究人员详细询问了受试者的病史和相关检查，排除了高血压、内分泌疾病、恶性肿瘤、冠心病、慢性肝肾疾病及其他慢性疾病。且受试者为无血缘关系并且在新疆地区居住时间≥20 年以上的常住人口，三代直系亲属均为维吾尔族，最大限度地排除了由于环境和地域差异而引起的基因变异。

从选样过程可看出样本量是较大的，同时关注两组样本的平均年龄，在很大程度上排除了年龄对 2 型糖尿病发病风险的影响。

(2)文章是否详细描述了数据收集、整理以及分析的过程?有没有什么问题?

答：文章对数据的来源，数据的收集，基因DNA提取的过程，TCF7L2 基因 SNP 分型检测均进行了较为详尽的描述，得到了两组样本中的基因型频率和等位基因频率数据，并进行了两组一般资料和实验室检查指标比较。此外，由于此研究涉及到了TCF7L2上多态性位点的研究，故其分别进行了位点 rs12255372 的 T2MD 组 和 NC组的针对一般资料和实验室检查指标在不同基因型的详细比较。

(3)这些数据对于研究问题来说是否合适?

答：是合适的。因为本研究的着眼点就是 TCF7L2 基因多态性与新疆维吾尔族人群 2 型糖尿病的相关性，因此选取两组样本的基因型频率和等位基因频率数据进行进一步统计分析，是解决问题的合理思路。此外，本研究还着重对环境和地域差异等因素引起的基因变异的排除，使研究所得结果更具可靠性。

1.4、结果和结论：

(1)文章清晰地呈现结果且与假设相关吗?

答：是的。文章通过表格的方式清晰呈现了其所收集的数据和统计分析所得的数据，并且分析了每个统计所得数据所反应的两组间的差异是否具有统计学意义，即是否可推翻原假设。

(2)结果是否支持文章得到的结论?

答：结果可以支持文章所得结论。

(3)统计图表提供的信息是否足够充分使你对该研究及结果有一个很好的把握?

答：可以。本研究共列出了六个表，且每个表下均对表中所涉及的符号及英文简写进行了解释，对我在阅读文章时的跟进感有很大的帮助作用。且六个表中涉及到的广泛的数据和其背后所反应出的研究考虑到的各种因素，也使我能对研究结果有更好的体会和认同。

(4)是否存在应该在文章讨论但没有讨论的偏差或缺陷?

答：本文并没有讨论其所存在的缺陷和不足之处。我认为除了遗传因素，影响 2 型糖尿病的发病的另一重要因素是人们的生活习惯。而本文虽努力排除了环境和地域因素对实验结果的影响，但并未对人们生活习惯的差异做过多考虑，这是有所欠缺的。

(5)文章的结果是否有实际意义?统计上是否显著?

答：文章的结果指出了 TCF7L2 基因 rs12255372 多态性位点可能与新疆维吾尔族人群 2 型糖尿病相关，而 rs7085532 多态性位点可能与新疆维吾尔族人群 2 型糖尿病无相关性，具有实际意义，且通过整合所得各种数据可得出研究结果具有显著的统计学意义。

(6)如果结果具有统计显著性,文章是否给出了效应值?

答：文章在所呈现的表格中详细罗列了统计检验所得的效应值，使作者在对统计数据进行分析时让读者是有据可查的。

(7)如果没有显著性,是否说明了结果的重要性?

答：本文在对TCF7L2 基因的另一多态性位点rs7085532 位点进行研究后，呈现了实验得到的具体数字，结合语言描述让读者知道了其所得的结果并没有显著性，故得出了rs7085532 多态性位点可能与新疆维吾尔族人群 2 型糖尿病无相关性的结论。

(8)根据所研究的问题,这一结果有意义吗?

答：有意义。研究所得的结果十分扣题，充分回答了研究开始时所提出的问题。

(9)文章是否清晰地说明了研究的局限性?

答：本文并没有讨论其所存在的缺陷和不足之处。

2、对某一基因位点多态性相关的心血管疾病或肿瘤发病风险系列研究报道的meta 分析(荟萃分析)文章进行统计评论。

所用文献：《The association of matrix metalloproteinase-9 promoter polymorphisms with gastric cancer risk: a meta-analysis 》

(1)是否明确提出了研究问题?

答：明确提出了问题：研究 MMP-9 基因与胃癌的发病风险。

(2)对该问题进行研究是否有必要?

答：很有必要。胃癌是全世界最常见的癌症之一，且其死亡率很高。很多因素如遗传因素和环境因素均会导致胃癌的发生,其中,遗传因素扮演着主要角色。当一系列相关基因发生突变后，不正常的细胞可能会导致肿瘤组织的形成。基质金属蛋白酶（MMP-9）是一种白明胶酶，与癌症的发生密切相关。一系列研究发现其启动子区域-1562 位点上 C->T 的突变可能会影响 MMP-9 基因的表达水平，但所得结果有所争议。故本 meta 研究通过整合前人所得的结果，旨在得到一个较为可靠的统一结论，是很有研究必要的。

(3)是否引用了发表时的最新研究成果?【可选内容】

(4)如何搜集相关研究报道的?方法是否合适?

答：本研究对 PubMed，CNKI，EMBASE，万方数据库进行了检索，之后对检索所得的文章进行了筛选，最终得到了可利用数据的 9 篇相关研究报道。
方法是合适的，且中英文的数据库均做了检索，故所得到的数据集还是比较全面的。

(5)怎么筛选文章的?是否有详细描述?

答：最初共收集到 33 篇文献，查重后还有 17 篇文献，之后对文章进行全文追查后找到了 16 篇文献的全文，进行数据审查后最终有 9 篇含有可利用数据的文献被纳入了研究中。

对于筛选文章的流程，文章中附有专门的流程图来描述，但我认为还不够详尽，还应说明判断数据是否可用的标准是什么。

(6)是否存在发表偏倚?

答：研究使用漏斗图和 Egger 检验来判断研究结果是否有潜在的发表偏移。漏斗图的结果显示研究结果并不存在潜在的发表偏移，Egger 检验所得的数据也进一步证明了这一点。

(7)文章清晰地呈现结果且与假设相关吗?

答：文章通过一系列的统计数字和对数字的解读，得出结论“MMP-9 基因启动子区域 1562 位点 C/T 会增加胃癌的发病风险”。此结果与假设相关，直接回答了所研究的问题。

(8)结果是否支持文章得到的结论?

答：支持。由统计所得的数字可看出，MMP-9的T等位基因与胃癌的发病风险密切相（OR = 1.150， 95% CI = 1.014－1.304），且TT基因型所导致的发病率比其他基因型高1.666倍(OR = 1.666, 95% CI = 1.127－ 2.461)。而结论是“MMP-9基因启动子区域1562位点C/T会增加胃癌的发病风险”。故研究所得结果是支持文章所得结论的。

(9)统计图表提供的信息是否足够充分使你对该研究及结果有一个很好的把握?

答：文章中放有一张关于文章筛选的流程图，一张含有所有所用文献的数据整合表，一张展示在 meta 分析过程中所得的主要的汇总 OR 值结果，四张可直观展示结果的森林图以及一张可直观展示结果的漏斗图。这些图表使研究过程和所得结果在我心中有了更加立体和直观的体现，也使我能更加清楚地知道文章结论是如何得出的。

(10)是否存在应该在文章讨论但没有讨论的偏差或缺陷?

答：文章涉及的样本绝大部分来自亚洲，一部分来自非洲印度，几乎没有纳入欧洲人，故其所涵盖的人种很有局限性，需要后续的研究数据进行进一步的补充。

(11)文章的结果是否有实际意义?统计上是否显著?

答：文章的结果有实际意义，且统计学意义显著。通过对统计结果的归纳整理得出了“MMP-9 基因启动子区域 1562 位点 C/T 会增加胃癌的发病风险”的结论。

(12)文章是否清晰地说明了研究的局限性?

答：本文在 Discussion 部分的最后讨论了其所存在的缺陷和不足之处。

其中指出了研究中所存在的一系列问题：

（1）研究所纳入文章数量较少；

（2）缺少其他影响因素的信息（如年龄，性别，吸烟程度，胃癌的等级，饮食和饮酒习惯等）对实验所得 OR 值进行校正；

（3）并未讨论基因之间的相互作用可能会造成的影响；

（4）没有对数据进行专门的种族划分导致不能得出针对不同人群的研究结论。

3、基因突变与心血管疾病或肿瘤发病风险或个性化用药等方面相关研究报道的 meta 分析【可选内容】

1.1、选择一个感兴趣的疾病类型，在万方数据、中文期刊网(CNKI)、维普数据、PubMed 等文献数据库中，联合“多态性”“基因”“遗传”等关键词，搜索该疾病;

1.2、快速浏览这些文献报道的标题，看看哪个基因出现频率最高;

1.3、以该疾病和基因作为关键此，限定在标题和摘要中出现，进行二次检索【限定在真正的研究中可能不适用】;

1.4、下载相关文献原文;

1.5、快速阅读摘要和/或全文，从中提取统计数据;

1.6、对这些研究报道中的统计结果进行深入分析，探讨一下这些针对同一主题的不同 研究报道中，是否存在统计学问题(如发表偏倚)。

四、讨论：

1、对某一基因突变相关的心血管疾病或肿瘤发病风险的独立研究报道进行统计评论。

所用文献:《TCF7L2 基因多态性与新疆维吾尔族人群 2 型糖尿病的相关性研究》

本研究以“2 型糖尿病和 TCF7L2 基因多态性的关系”为题进行研究，能为预测不同人群的 2 型糖尿的发病风险提供参考和借鉴作用，选题社会热点问题，体现出较强的时代特色性与和可操作性。文章逻辑结构严谨，层次分明，文笔流畅，表达清晰，且重点突出。文章格式符合学术规范，段落分明，条理易读。运用学术语言，但因有对专业内容的适当解释而使得文章通俗易懂。在研究过程中也有将专业知识原理与现实问题结合起来，且研究清晰呈现了一系列与假设相关的统计结果，且结果可明显支持文章得到的结论。但是文章并为指明研究中所存在的不足之处，如除了遗传因素，影响 2 型糖尿病的发病的另一重要因素是人们的生活习惯。而本文虽努力排除了环境和地域因素对实验结果的影响，但并未对人们生活习惯的差异做过多考虑，这是有所欠缺的。

2、对某一基因位点多态性相关的心血管疾病或肿瘤发病风险系列研究报道的 meta 分析(荟萃分析)文章进行统计评论。

所用文献：《The association of matrix metalloproteinase-9 promoter polymorphisms with gastric cancer risk: a meta-analysis 》

本文以“MMP-9 基因与胃癌的发病风险”为题，重点探讨了 MMP-9 基因启动子区域-1562 位点上 C->T 的突变可能会影响 MMP-9 基因的表达水平，从而增加胃癌的患病风险，选题具有很强的针对性和现实意义。文章结构安排合理， 层次清晰，写作时参考的相关文献资料与主题联系紧密，而且参考的资料较新。在写作过程中作者运用其专业基本知识来分析所得的一系列统计数字，在数字和结果之间建立桥梁，让读者可清楚知道研究结果的现实含义。文章中涉及了对五个大型中英文数据库的检索，故所得到的数据集是比较全面的；且文章中附有文献

筛选的相关流程图，但不足够详尽，若加入对判断数据是否可用的标准的说明会更加清晰有力;运用了两种方法漏斗图和 Egger 检验来判断研究结果是否有潜在的发表偏移，也体现出了 meta 分析的多重验证性；文章中的统计图表提供的信息使研究过程和所得结果在读者心中有了更加立体和直观的体现，也使读者能更加清楚地知道文章结论是如何得出的;且在文章的最后清楚讨论了其所存在的几点缺陷和不足之处，对后人的进一步研究有很实际的指导作用。