**审稿意见与作者修改说明（稿号：2020-0213）**

|  |
| --- |
| **初审专家意见与作者修改说明** |
| 专家1意见：问题1：文章应按照第三人称来写，不要说“我们”，改为“项目组”或者课题组。回复1：已经通篇检查并修改。问题2：文章语句的表达拷贝官方指南的痕迹过重，请项目组工作人员认真阅读修改。回复2：本课题组对文章的文字描述进行了精炼和修订。问题3：2.2.3最后一句话是不是想说明“以上统计分析采用由北京协和医学院统计学人员使用SAS 9.4软件完成。”？回复3：不是。这句话是想说明本研究数据分析使用的SAS软件为正版。为了避免误解，现已修改为“以上统计分析使用SAS 9.4软件（北京协和医学院软件使用许可）完成。”问题4：统计表中文字不使用斜体。回复4：已经通篇检查并修改。问题5：附表2a和2b改为三线表格式。回复5：已经修改，附表2现调整为附表3。问题6：请精简文字。回复6：本课题组调整了文章内容，精炼了文字描述，修改了不够通顺的语句。目前，文章摘要字数为599，英文摘要字数为404；正文（第1-6部分，不包括参考文献）字数为7922，比上一稿减少了19%。问题7：本论文实际工作量较大，但是创新性、科学性及文字表达均一般。回复7：在精炼和修订文字表达的基础上，在这一稿中，本课题组结合数据库（质量评价与改进后）的现实应用价值，修改了文章题目，增加了关键词，重述了研究目的，对于本研究的方法学和创新点，尝试进行了更清晰的概括和提炼。医保数据是基础的、覆盖全体居民的、蕴含发病和患病信息的重要资料；基于本课题组的认知，对此类数据进行质量评价和改进的工作尚很少见。本研究对县级新农合数据进行了质量评价和改进，且这一工作是参考WHO《健康机构数据质量报告卡》中的框架，规范进行的。通过本研究过程，课题组也探索和初步形成了数据质量评价和改进的标准化操作程序。此外，还有一项在一般数据清洗等过程中不常见的创新性过程，即数据字典的编制。本课题组严格依照ICD-10对于疾病的编码规范，编制了“原始诊断-标准诊断-ICD10编码”数据字典，利用这一流程将诊断从非结构化的文本形式转换为结构化的编码形式。这个过程虽然耗费大量人工和时间，但可以统一形成标准诊断文本和编码时的注意事项，从原则上提升了疾病判断的一致性，大大改进了新农合数据中诊断的质量。并且，形成的成果可以在其他数据库中重复利用。所以编制数据字典是必要的且很有意义的**审稿意见二：**问题1：“研究数据集中的诊断为文本型变量，经过观察有大量填写为“无”的记录和少量空白记录；由于缺少诊断信息则无法进一步分析，因此，对这些记录进行了删除。”。被删除的记录，“无”的记录和“空白”记录各多少？能否知道删掉了多少个体？ 回复1：在198662条记录中，198650条的诊断均为“无”，仅12条的诊断为空白。本文中现已修改了对这一点的描述，请详见2.2.1小节第1行，及第3.1小节第1段第2行。本课题组将分析记录数据库和删除记录数据库合并在一起后，进行了精确查重和模糊查重，以确定其中包含多少个体。过程为：第一步，提取出有身份证号的记录，查找“姓名+身份证号+就诊证号”，三者都相同的判定为同一个体，给予唯一的精确查重后编号。第二步，对所有记录查找“姓名+性别+年龄（正负2岁或以内）+就诊证号+家庭地址”，五者都相同的判定为同一个体，之后与第一步精确查重的结果相比对。完全一致的跳过；部分一致的表示同一个体的所有报销记录中，有些填写了身份证号而有些未填写，确认无误后使用精确查重后的编号；完全不一致的，给予唯一的模糊查重后编号（且不能与精确查重后编号重复）。在这一过程后，每名个体都被给予了一个唯一编号。如表i所示，有18542名个体仅出现于删除记录数据库；有1422名个体在两个库中均出现，说明他们的报销记录部分被删除，部分进入了分析；有171796名个体仅出现于分析记录数据库。问题2：“我们首先对诊断进行了格式整理和编码的工作，形成了“诊断-标准诊断-ICD-10编码”字典。”。流行病学专业科研工作人员和两名公共卫生专业研究生是否经过ICD编码培训？该工作人员抽查两名研究生的结果，10%为抽查了多少（量可能很大）？复核后，报告复核质量指标。回复2：这三名人员在本项工作中，主要通过阅读ICD-10上的相关说明文字，尤其是第二卷《指导手册》和第一卷《类目表》中每个三位数类目和四位数亚目起始处的注解，来掌握了ICD10编码的基本原则。除此以外，对于没有把握的编码，还通过查询互联网及相互讨论，最终确定。10%的抽查比例对应1012条记录，其中40条记录，出现该名工作人员的编码结果与两名研究生的编码结果不一致的情况，占比4%。本句话内容已经补充进正文，请见当前2.2.2小节最后一句话。3.常规数据清洗三步后，是否有抽样核查以检验清洗的准确性？ 回复3：是。常规数据清洗由本课题组工作人员与一家数据技术服务公司合作完成。每一步均是先由本课题组提出清洗需求，再由公司通过程序实现并返回结果。每一步的清洗过程均有往复多轮的讨论，以确保公司清楚明确地理解并实现了课题组需要进行的清洗工作内容。每步常规清洗完成后，公司将返回当时的数据库，课题组工作人员抽查大约5%的记录内容。选择抽查记录的方式为从上到下平均选取，以使得行号在前、中、后的记录都被抽到。问题4：表1，性别缺失记录数2，性别(修订后)28，表达不清楚。建议修改表1，增加可读性。 回复4：已经对表1进行修改和精简。问题5：表4重点疾病或疾病大类总花费的描述性统计的单位是什么？是住院的总费用还是门诊的总费用？是次均费用？ 回复5：是住院或门诊就诊的次均总花费。已经对表4做了相应修订，在3.3小节的第2段第3-6行中也有提及。问题6：本文重点是“2013-2015年河南省某县新型农村合作医疗数据库质量评价”，表4与数据质量评价关系不大。 回复6：本课题组对修订后的数据库中的两个关键变量——诊断（疾病大类）和费用（次均就诊总花费）进行描述性统计，观察结果是否与对当地的一般认知或其他研究结果保持一致，来评价外部一致性。即，就费用来说，不同疾病/疾病大类门诊或住院的次均总费用，其范围、均值和中位数，以及在疾病间的比较情况，如果符合大众的一般认知和当地的一般消费水平，则可以认为本数据集中的数据资料较为可靠。因此，目前还是保留了表4，及方法、结果、讨论中与之相关的内容，但对内容进行了精简，请分别详见2.2.3小节第2段，3.3小节第2段，4.2小节。问题7：建议斟酌和修改表格，以增加其可读性；字数太多，文字精炼描述。 回复7：在本轮修改过程中，对表1、表3、表4的内容/表述方式进行了修订和适当精简。由于正文内容的调整，附表顺序也进行了相应调整，对调整后的附表2、附表3、附表4的内容和形式进行了修订。本课题组调整了文章内容，精炼了文字描述，修改了不够通顺的语句。目前，文章摘要字数为599，英文摘要字数为404；正文（第1-6部分，不包括参考文献）字数为7922，比上一稿减少了19%。问题8：表3的统计按住院和门诊人次，能否知道是多少人？其一年的花费是多少。回复8：本课题组对每名个体每年因为表3中的疾病去就诊，门诊或住院的花费进行了统计。结果如下方表ii所示。由于本篇文章的主要内容是对数据库进行质量评价，所以暂未将这些结果列进正文或附表。本课题组后续将对疾病给个人带来的经济 |
| **复审专家意见与作者修改说明** |
| 复审意见：表4 重点疾病或疾病大类就诊次均总花费的描述性统计，在表的下方，注释：次均总花费为门诊和住院费用。由于把门诊和住院费用放在一起，造成均数和标准差很大。建议表4删掉均数和标准差，提供P25和P75。回复5：感谢专家对表4的修改建议。“次均总花费”并非指门诊和住院费用之和，而是分别统计不同疾病病因下，门诊每次或住院每次的花费。为了避免歧义，现在表4中将“次均总花费”改为“每次就诊花费”。并添加了注释“每次就诊花费为分别统计不同疾病病因下，每次门诊或每次住院的花费”。由于次均花费在讨论部分4.2小节中，与其他研究结果进行了比较，所以没有从表4中删除均数和标准差。补充提供了P25和P75。 |
| **定稿会意见与作者修改说明** |
| 专家1意见：同意发表 |
| 注: 请针对专家意见（或批注）逐条回复，先表明是否同意专家意见，然后说明原因或者是否修改，修改位置（如见1.2）或者内容。多个审稿意见请自行在该模板按照序号依次列出，并将表附在修改稿首页，便于编辑核查。 |