《疾病监测》审稿意见与作者答复

题目：2014－2016 年云南省宣威市死因监测数据库清洗探讨

作者： 万霞,刘利群,杨功焕

——————————审稿专家意见与作者答复——————————

初审专家意见及作者修改说明：

1、没有大的差别” “次年或第三年上报”此语表达不准确。为什么单说肺癌？考虑是否存在”霍桑效应”。

答：专家您好，您提的很好，当时计算的时候，忽略了死亡时间在 12 月份的病例，因此，重新在数据库里计算了，扣除了死亡时间在 12 月份且上报时间在 1 个月内的病例 237 例。其他问题已修改。

2、第二步：描述音同字不同的如何处理?

答：专家您好，我们是先根据这些条件，筛选出记录，通过人工判断是否为同一记录，如果是，则删除其中一条记录。根据您提的疑问，我们觉得可能我们遗漏了删除的原则的描述，故作了相应的补充。

3、四分位间距（即 Q3-Q1）为 234 天。

答：由于这 131 例重复上报的病例两次上报的死亡时间间隔不满足正态分布，因此，在做趋势描述时，不能采用均数和标准差来描述，只能采用了中位数和四分位间距来描述

4、“第二次上报死亡时间主要是在次年 1—3 月”为什么?

答：在文章的 4.2.2 部分中提到“通过与地方 CDC 工作人员沟通了解到，为了减少漏报，国家 CDC 允许各基层 CDC 在次年 1-3 月还可以补充报告前一年的死亡记录……对于重复报告的病例，多是由村医调查回来后，或根据死亡医学证明书填报，或根据死因推断的”因此，有些死亡病例的重得就是有可能是在次年 1-3 月时由于村医没有回顾之前医院上报的记录，而是通过调查得来的，有些调查对象不是亲密的家属，所以可能对死亡时间、死亡诊断、死亡地点等信息回忆的并不十分准确。

5、3.3 描述不妥。

答：在背景部分补充了对宣威肺癌高发乡镇的描述；想说明的是正是由于死在家中的病例容易漏报。

6、“78.4%的肺癌病例来自来宾镇、宛水街道、龙潭镇、西宁街道、双龙镇、龙场镇、海岱镇、东山镇及虹桥街道，而这些乡镇/街道经证实是于肺癌高发区[4]。” 偏离文章主题。

答：由于近 80%的补充的病例，都是集中在肺癌高发区，因此，从数据的角度，我们考虑是否存在有“霍桑效应”。

复审专家意见及作者修改说明：

1、存在文字口语化，语句不通顺问题；

答：专家您好，谢谢您的建议！我们对部分文字作了修改。

2、存在某些主观判断的结论性描述，如“1.2 数据清洗及其重要性介绍”中最后一段的描述

答：专家您好！这部分描述是笔者平时与硕/博士的讨论的体会，发现他们对数据清洗存在误区，因此，觉得特别有必要把数据清洗的过程描述出来，让研究生们有一个感性的认识。这也是写这篇文章的初衷之一。

3、3.1 数据清洗结果描述建议列表，包括删除记录原因构成，异常值情况等

答:专家您好,谢谢您的建议!已根据您的建议在文中增加了“表 1 数据清洗情况列表（n=24979）”。

4、如果要对 “霍桑效应”问题进行特定描述与分析，建议在结果中另起一段，并在讨论中加以描述并分析原因给予改进建议

答：专家您好，谢谢您的建议！已根据您的建议在文中增加了“4.2.2 对于次年补充上报的病例，需要有严格审核机制”中增加了如下描述“同时，对于新补充的病例，一定需要避免“霍桑效应”[9]。本研究现场为宣威市，宣威在过去的 20 年一直是肺癌高发区，本死因监测数据库中发现次年新补充的 78.4%的肺癌病例来自于肺癌高发区。因此，这些地区的死亡病例是否真的为肺癌，是否存在“霍桑效应”，有待于现场的进一步核实。这也提示了应加强基层工作人员的培训，尽量及时上报死亡病例，同时，对于那些报告死在“家中”的病例，在死因上报

平台中先进行初步核查，确认死者是否真的死在家中，是否之前被医疗机构上报过，死因是否正确，最大限度地保证死因填写的准确性。”

5、用死因构成是否稳定来说明数据质量比较牵强

答：专家您好！用死因构成来说明数据质量有两个方面的原因：1）未明原因的比例越低，则死因监测数据质量越好；2）从人口学的角度，如果没有特别的原因，连续年代的大类疾病相对来说是比较稳定；如果有部分年代某些疾病的构成发生了大的变化，就需要考虑是否有什么原因导致了疾病的构成发生了变化（例如：某年发生了地震，则“伤害”的比例则会升高）；或者是数据收集过程中发生了什么问题，导致数据出现了异常。因此，我们在数据预处理中，通常都需要先分析一下数据中各死因的构成的情况。从本监测数据来看，各年代的各大类疾病构成没有大的变化，且当地也没有出现什么特殊的危险因素，因此，笔者认为这个指标是可以间接反映数据质量的。

6、讨论中对发现的问题的原因分析和建议提出比较泛泛，如为什么会出现上报不及时？原因是什么？要如何去改进？

答：专家您好！根据您的建议，在文中相应的部分作了如下的修改：“按照死因监测报告流程要求，县及县级以上医疗机构在 7 天内完成对卡片的审核和网络直报，县级以下医疗机构则需要在 30 天内完成审核及网络报告，且县级疾控部门、妇幼保健机构的死亡报告管理人员应于 7 天内通过网络进行审核确认。并且将死因及时报告率及时审核率作为上级疾控部门对下级疾控部门考评内容之一[8]。但是，从数据可以看出，68.8%的病例未能实现及时上报。出现未能及时上报的原因，主要可能是近 95%的病例是在医院外的死亡。在死因监测系统中，要求基层防保医生对在医院外的死亡的病例，开展入户调查后才能填写《死亡医学证明书》，完成网络上报工作[9]。而对于在医院外死亡的病例，基层防保医生不可能实时地获取死亡病例的死亡信息，这使得报告的及时性受到制约。因此，需要加强部门间的协作，例如公安与计生部门的协作，信息共享，及时互通死者的信息，使基层防保医生尽早获取死者死亡信息，尽早开展调查，促进死因监测数据的及时上报。”

7、为什么建议用出生日期作为年龄分析变量，而不用身份证号码+出生日期相结合的方式？

答：专家您好！从本监测系统来看，有 8.12%的身份证号是缺失的，而且对于未出现缺失的身份证号，有 27 例出现身份证号中出生日期是不正确的。但是，本系统中出生日期这个变量无一缺失，且从格式来看，没有出现不正确的日期。

通过有正确身份证号码的 22649 条记录与出生日期进行比较，发现只有 9 条记录出生日期不符，其中有 2 条记录计算的年龄超过 1 岁；通过出生日期计算的年龄与上报年龄相比，也仅有 131 例（0.6%）不一致，其中 121 例相差不超过 1 岁。

因此，从本研究中，可以看出出生日期与有正确身份证号基本上是一致的，但是，出生日期更完整，因此，推荐使用出生日期。当然，如果系统平台升级后，身份证号码更准确时，可以考虑采用身份证号码+出生日期相结合的方式。这也是我们对平台升级建议中所提出的对身份证号码作进一步的升级，提高身份证号码的准确性。这样，未来也许只需要收集身份证号就可以了。

8、平台升级建议是否合理？特别是对身份证号码中月份和日期变量的限定，是否能够实现？请斟酌

答：专家您好！根据笔者对编程的经验，从程序的角度对身份证号码中月份和日期变量的限定，应该是很容易实现的。因此，如果能在平台升级中增加这些功能的限定，将会大大提高数据的准确，且减少很多基层人工核对的工作。所以，我们认为是合理的，我们也很欢迎及感谢专家帮我们提出更多关于平台升级的宝贵建议。

——————————————定稿会意见与作者答复——————————————

定稿会意见：

本文经这次修改后，基本达到要求，可以发表，谢谢！