光电阅读机在大人群现场调查中数据处理及质量控制的应用*

李新艳 吴玉璘 姜志欣 陈 伟 邹文霓 石 慧 潘 丽 李 瑛△

(江苏省生殖健康检验中心 江苏省计划生育科学技术研究所 江苏 南京 210036)

摘要 目的:探讨光电阅读机在大样本现场调查项目质量控制和数据录入中的作用。方法:使用光电阅读机读入《子宫颈癌危险因素调查表》信息并进行表格质量评估,核查调查表中调查对象编码缺失、重复及进行读入信息一致性比对。结果 在录入的70160份数据中,调查对象编码缺失率为0.87%,调查对象编码重复率为0.79%,信息一致率为99.47%。3个项目点连续三个月质量控制,表格不合格率略有下降。结论 在大人群现场调查项目中通过光电阅读机进行质量控制及数据录入,能快速准确的将纸质信息转化为电子信息,并对电子信息进行核查,能迅速将调查表评估结果和建议反馈给调查表填写人员,及时指导调查员改进调查项目的编写方法和调查表内容的填写方法,从而提高了调查表信息采集质量,促进项目科研数据真实可信。

关键词 光电阅读机;现场调查;质量控制 数据录入

中图分类号:R179.324 **文献标识码**:A **文章编号**:1673-6273(2012)22-4360-05

Application of Optical Mark Reader in Data Entry and Quality Control in Field Work Based on Large Population*

LI Xin-yan, WU Yu-lin, JIANG Zhi-xin, CHEN Wei, ZOU Wen-ni, SHI Hui, PAN Li, LI Ying△

(Jiangsu Institute of Planned Parenthood Research, Jiangsu Clinical Laboratory of reproductive health, Nanjing Jiangsu 210036, China)

ABSTRACT Objective: To approach the role of the Optical Mark Reader (OMR) in quality control of site interviewing and data entry in a large sample project of field work. Methods: The OMR was used to data entry of the questionnaires of risk factors for cervical cancer, several indexes were checked such as the code missing, the code repetition and the consistency of data information. Results: In the total number of 70160 data entry, the rate of code missing was 0.87%, the code repetition rate was 0.79%, and the information consistency was 99.47%. The percent of disqualified questionnaires was declined after quality control for three months. Conclusion: It was fast and accurate to transform paper-based to electronic information by using OMR for data entry in large sample project. After verifying electronic information according to the paper-based information, the results and feedback suggestions were given to the site interviewers, to guide the interviewers to improve the compilation of survey items and the questionnaire content. These improved the quality of survey information and promoted the research data authentic.

Key words: Optical Mark Reader; Field work; Quality control; Data entry

Chinese Library Classification(CLC): R179.324 Document code: A

Article ID:1673-6273(2012)22-4360-05

光电阅读机(Optical Mark Reader,OMR)是一种专用计算机输入设备,它可以直接读入信息卡上的涂写内容,并传送到计算机中去,这为计算机处理数据提供了方便。光电阅读机能够在许多计算机统计管理系统中使用,充分显示出数据快速输入的优越性[1]。它不仅在高考阅卷中使用,在其他领域中也得到推广,如卫生保健中用于体检卡输入,病疫检测中用于病情调查卡输入,教育系统用于学籍管理、教学质量评估,还在知识竞赛等方面都产生了很好的效果[2,3]。但是光电阅读机应用于大型现场调查还比较少见,本文旨在利用光电阅读机的信息读入快速便捷、准确性高等特点,而将其引入到大人群宫颈癌筛查项目中进行质量控制、数据录入等,探讨光电阅读机在人群大样本现场调查中应用的可行性。

1 资料和方法

1.1 研究对象

2010年参加江苏省已婚妇女宫颈癌早期筛查项目的9个项目县65岁以下已婚妇女,共70160人次。

1.2 研究方法

利用Visual FoxPro桌面级数据库自主开发录入软件,采用南昊S43FBSA型光电阅读机读入2010年9个项目县《子宫颈癌危险因素调查表》70160份,共4279760个信息点。对调查表中调查对象编码缺失、调查对象编码重复、信息一致性等指标进行核查。每次质控后根据出现的问题提出整改意见,并及时将整改意见反馈给信息采集点。如连续质控合格率低于95%,

* 基金项目:江苏省科技支撑计划(社会发展)项目(BS2007080)

作者简介:李新艳(1984),女,学士,助理工程师,主要研究方向:计算机技术与应用,

电话:13851771837,E-mail:xinyanli1985@163.com

△通讯作者:李瑛,E-mail: liying2008@jsmail.com.cn

(收稿日期:2011-11-21 接受日期:2011-12-18)

则增加培训并加大信息采集现场质控力度[4]。选择其中 3 个项目县连续跟踪三个月，观察质控的效果。

1.2.1 数据读入系统 对《子宫颈癌危险因素调查表》中的变量进行定义和赋值，利用 Visual FoxPro 桌面级数据库系统编写《子宫颈癌危险因素调查表》数据录入系统并进行调试，软件中包含与需手工录入的姓名以及细胞学编码的对接功能、DBF 文件和 EXCEL 文件自动转换功能、数据信息查询修改等功能。软件开发完毕后试读入 100 份信息表格，计算读入速度，并核实信息表格中 61 个信息点录入的准确性。因信息点繁多细小，开发的过程会出现部分信息点缺失以及填涂不规范等情况考虑不周的问题，还会出现部分信息点二进制或十进制设置不当的问题，这些情况通过多次的测试和矫正得以解决。

1.2.2 数据采集人员培训及现场调查 编制培训教材，采用集中培训和现场指导的方式对项目点信息调查员进行培训，确保每一位数据采集人员了解设计原则，熟悉调查表中每个变量的定义及表格填涂原则与要求，调查员经考核合格后方可参加信息收集工作。2010 年 3 月—11 月，对 9 个项目县《子宫颈癌危险因素调查表》共 70160 例妇女进行年龄、职业、婚育史、生育史、避孕史、肿瘤史等进行调查[5-11]，同时对调查对象的健康检查及实验室检查情况共 4279760 个信息进行收集。

1.2.3 数据读入过程中质控 在读入过程中对调查表污损、调查对象编码缺失、调查对象编码重复等情况进行质量评估。

1.2.4 读入信息一致性比对 比较《子宫颈癌危险因素调查表》中纸质信息与读入电子信息是否完全一致。信息读入完成后，将纸质调查表上个人信息、一般情况、月经、婚育史、性行为史及卫生习惯、吸烟、饮酒和饮茶情况、饮食情况、避孕史、生殖道感染既往患病史、家庭及本人恶性肿瘤史、体格检查结果、妇科问诊等 15 项核心指标信息与读入的电子信息进行一致性比对。

2 结果

2.1 录入表格时间

录入 70160 份调查表数据资料，共耗时 14 小时，包括对污损表格、填涂错误表格等问题表格的处理。

2.2 数据采集人员培训及现场调查

采用集中培训和现场指导的方式对项目点信息调查员进行培训，9 个项目县共进行一次集中培训、9 次现场指导培训，共培训调查员 126 名，确保每一位工作人员了解设计原则，熟悉调查内容和工作流程。2010 年 3 月 -11 月，共完成 9 个项目县 70160 例《子宫颈癌危险因素调查表》的现场采集工作。

2.3 数据读入过程中质控结果

70160 份数据中，调查表污染折损未被读入表格共 337 例（0.48%），调查对象编码缺失无法读入表格共 607 例（0.87%），调查对象编码重复无法读入表格共 557 例（0.79%）。

表格数据读入过程中出现以下报错情况：

（l）读入过程中表格污损、卡纸现象。如读入过程中弹出小窗口，显示 A 传感器格式标记错、A 传感器卡纸同步错、B 传感器同步计数超界，则表明调查表被污染或者折损，此份调查表无法读入，见图 1、图 2。

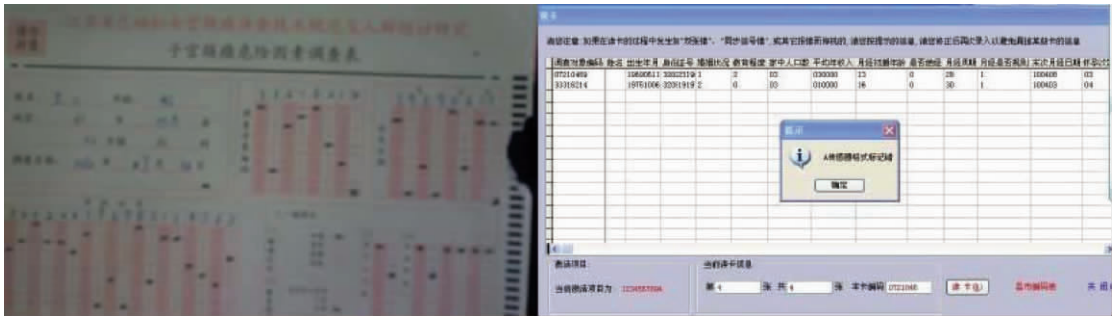

图 1 调查表 A 面识别线部分缺失，导致读入系统显示 A 传感器格式标记错，无法读入(1:5)

Fig.1 Data entry system displays A sensor format errors of the card that can not read due to the excalation of side A logo line(1:5)

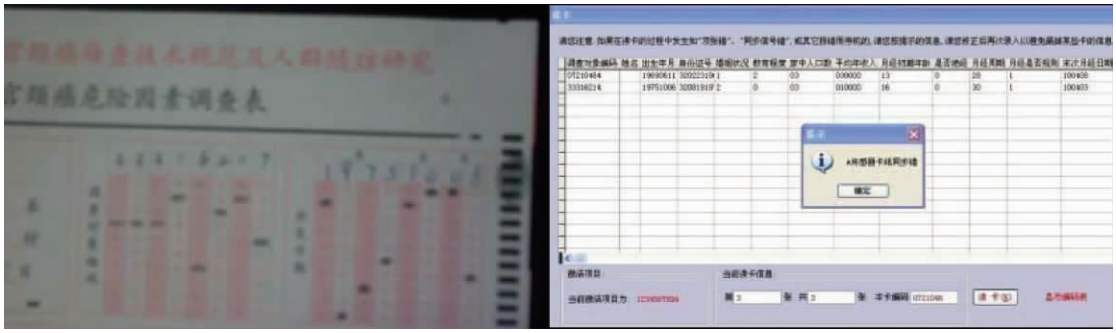

图 2 调查表 A 面标识线部位被污染，导致读入系统显示 A 传感器卡纸同步错，无法读入(1:5)

Fig.2 Data entry system displays A sensor paper jam and synchronization errors of the card that can not be read due to contamination of side A logo line (1:5)

⑵ 读入过程中调查对象编码缺失。如读入过程中弹出小窗口，显示调查对象编码读出有误，则表明调查对象编码缺失，见图 3。

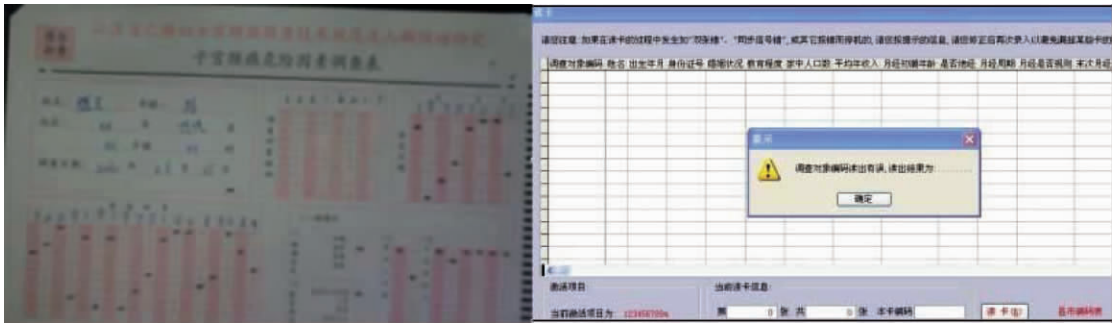

图 3 调查表中调查对象编码缺失，导致读入系统显示调查对象编码读出有误，无法读入(1:5)

Fig.3 Data entry system displays an error code of the survey that can not read due to the card coding missing(1:5)

⑶ 读入过程中调查对象编码重复。如读入过程中弹出小窗口，显示此卡信息已存在是否替换，则表明调查对象编码重复。

2.4 读入信息一致性比对结果

70160 份数据中，填涂不规范造成误读表格共 167 例(0.24%)，填涂较浅造成漏读表格共 50 例(0.07%)，填涂过深造成反面错读表格共 158 例(0.23%)。

读入错误的原因包括书写或填涂部分不符合规范造成误读，需及时改正，若无法改正部分，则和信息采集地取得联系，共同核实出正确信息。填涂过程中有因填涂超过指定范围或填涂错误无正确处理造成误读、因填涂过浅无法被机器识别造成漏读、因填涂过深造成反面错读等现象，见图 4。

图 4 因填涂不规范，导致读入系统误读的表格(1:5)

Fig.4 Misreading of the card by data entry system is caused by the non-standard filling in(1:5)

表 1 数据质控后不合格表情况

Table 1 The failed quality of survey after the quality control

Category	First month n(%)	Second mounh n(%)	Third month n(%)
Number of processing card	13983	12892	3389
Rate of defaced card	73(0.52)	58(0.45)	14(0.41)
Rate of card coding missing	130(0.93)	76(0.59)	27(0.80)
Rate of card coding repeating	125(0.89)	86(0.67)	11(0.32)
Rate of filling in non-standard	28(0.20)	16(0.12)	7(0.21)
Rate of filling the deep or shallow	59(0.42)	38(0.29)	6(0.18)

2.5 质控结果反馈前后对比

3 个项目县连续三个月质控观察，读入过程中质控后表格质量情况（见表 1）和读入信息一致性比对质控后表格质量情况（见表 2）均显示，3 个项目点经过质控以后不合格表格数下降，表格质量上升的趋势。

表 2 数据质控后表格质量情况

Table 2 Quality of the survey after the quality control

	First month n(%)	Second month n(%)	Third month n(%)
Failure rate	415 (2.97)	274 (2.13)	65 (1.92)
Qualified rate	13568 (97.03)	12618 (97.87)	3324(98.08)

3 讨论

步入信息社会，越来越注重信息资源的开发与利用。计算机可以快速和准确地处理各种信息，对各种信息进行统计、分析和打印，用以指导我们的工作，从而大大提高了工作效率。在计算机的应用过程中，遇到的最大难题就是数据处理的“瓶颈效应”，即一方面是计算机的快速和高效，另一方面是人工信息录入的慢速和低效。而光电阅读机的出现，解决了大量数据录入计算机的难题，同时也大大减轻了人的负担[12]。本项目筛查和随访的样本数将达到 100 万人次，如果所有调查表采用人工双份录入并进行录入数据的比对、核查、校对，这将是一个巨大的工作量，需要的人力和时间都是十分惊人的。而且由于人员长期重复同样的工作，也常常会出现人为的错误，降低了数据的可信度，增加了数据比对核查校对的工作量。光电阅读机的使用解决了大样本现场调查项目数据录入的难题，节约了人力物力，同时也提高了数据录入的质量。

在医学科学研究中，研究者对实验和数据收集过程十分重视，但对数据记录过程不够重视，由此造成的数据缺失和错误极为常见，因此数据管理和质量控制是医学科学研究中十分重要的环节[13-15]。本项目共有 100 万对象的数据资料，若加入质量控制，需要耗费了大量的人力和物力[16]。由于项目数据量大，项目点分散，如果采用人工随机抽样质控的办法，需要较长时间和大量的人力投入才能完成，不能及时将质量评估结果反馈给现场数据采集人员。而在光电阅读机读入数据的同时加入数据质量控制的环节，能够及时获得数据质量的信息，可以帮助现场组织者和调查员寻找数据采集中质量不高的原因，及时改进数据采集方法，提高数据质量。本项目利用自主研发的数据读入系统与光电阅读机配套使用，能够在完成数据读入的同时完成调查表是否污损、调查对象编码是否缺失、调查对象编码是否重复、调查表纸质信息与电子信息是否一致等问题的核查，这样就可以在最短的时间内将不合格表格的主要原因和存在的主要问题反馈给现场数据采集人员，并提出改进的建议和要求，数据采集人员可根据反馈信息和建议及时改进数据采集方法。

南昊 S43FBSA 型光电阅读机读卡速度是 3-5 张 / 秒，因录入过程会有部分信息表出现卡纸、双张报错、调查对象编码缺失或者重复的现象，录入 70160 份调查表数据资料，共耗时 14 小时，读卡速度 1.4 张 / 秒。从结果可以看出，通过光电阅读机读入数据，大大缩短了数据录入时间[17]，一人一机即可完成原本需要两人双份录入和比对、校对工作，而且在高速的同时达到了高效。3 个项目县连续 3 个月数据质控的结果说明，污损表格、对象编码缺失表格、对象编码重复表格、填涂不规范表格、填涂较浅和较深表格的百分率均有所下降，不合格表格数量也有下降趋势，数据的质量有所提高。另外，由于本项目筛查人群样本大且现场调查时间相对集中，信息采集现场秩序有时比较混乱，造成了表格折叠、污染或丢弃，这些情况在光电阅读机读入数据时均可发现，数据处理人员能及时与信息采集地取得联系并提出整改建议，通过连续三个月录入过程中质控，表格的污损率明显下降。调查表中部分内容因位数过多极易造成填涂错误，如身份证编码、末次月经日期、全身一般情况等项目，通过连续三个月一致性质控，这些易出错的项目的错误率也明显下降。

综上所述，将光电阅读机应用到大样本现场调查项目现场质控和数据录入，不仅可以完成数据的准确快速录入，还能及时反馈质控信息、及时改正数据采集中存在的问题，提高了科研数据的质量。但同时要认识这种读卡设备的局限性，它只能对信息卡中固定部分的涂写状态进行识别，使用前先要对所输入的数据进行规范化，根据数据的要求设计出专用的信息卡，对信息卡中所有填涂部位进行定义，标注出填涂位置，同时要用套色板统一印刷。因此，只有在充分考虑光电阅读机的利弊的条件下，进行大规模的数据输入时才能显示和发挥它的优势。

参考文献（References）

[1] 徐萌. 光电阅读机的发展与展望 [J]. 山东大学学报, 1993, 28(3): 371-372
Xu Meng. Optical mark reader of developments and prospects[J]. Journal of Shandong University, 1993, 28(3):371-372(In Chinese)

[2] 翁功平.光标阅读机 OM R 原理的设计与实现[J]. 工业控制计算机, 2010, 23(4):61-62
Weng Gong-ping. Real ization and Design of Optical Mark Reader[J]. Industrial Control Computer, 2010, 23(4):61-62(In Chinese)

[3] 裴文俊. 浅谈光学标记阅读机（OMR）在 PETS 考试管理中的应用 [J]. 教育信息化, 2002, (8):46-47
Pei Wen-jun. Application of optical mark reader in PETS test management[J]. China Education Info, 2002 (8):46-47(In Chinese)

[4] 林宁,潘丽,蔡瑞芬,等.宫颈癌筛查中标本质量的评估方法[J].中国妇幼保健,2009, 24(33):4691-4694
Lin Ning, Pan Li, Cai Rui-fen, et a1. Evaluation method of specimen quality in cervical cancer screening[J]. Maternal and Child Health Care

of China, 2009, 24(33): 4691-4694(In Chinese)

[5] Smith JS, Green J, Berrington de Gonzalez A, et al. Cervical cancer and use of hormonal contraceptives: a systematic review [J]. Lancet, 2003, 361(9364): 1159-1167

[6] Liu J, Rose B, Huang X, et al. Comparative analysis of characteristics of women with cervical cancer in high-versus low-incidence regions [J]. Gynecol Oncol, 2004, 94(3): 803-810

[7] Edelstein ZR, Madeleine MM, Hughes JP, et al. Age of Diagnosis of Squamous Cell Cervical Carcinoma and Early Sexual Experience[J]. Cancer Epidemiol Biomarkers Prev, 2009, 18(4): 1070-1076

[8] Domingo EJ, Dy Echo AV. Epidemiology, prevention and treatment of cervical cancer in the Philippines[J]. J Gynecol Oncol, 2009, 20(1): 11-16

[9] Negri E, La Vecchia C, Bosetti C, et al. Risk of cervical cancer in women with a family history of breast and female genital tract neoplasms [J]. Int J Cancer, 2005, 117(5): 880-881

[10] Louie KS, de Sanjose S, Diaz M, et al. Early age at first sexual intercourse and early pregnancy are risk factors for cervical cancer in developing countries[J]. Br J Cancer, 2009, 100 (7): 1191-1197

[11] Garland SM, Brotherton JM, Skinner SR, et al. Human Papillomavirus and Cervical Cancer in Australasia and Oceania: Risk-factors, Epidemiology and Prevention[J]. Vaccine, 2008, 26suppl(12): M80-M88

[12] **徐振林,郑波.用好** OMR **推动我国信息产业发展**[J].**电子计算机与外部设备**, 1998, 22(1):42-46
Xu Zhen-lin, Zheng Bo. Use OMR to promote the development of china's information industy[J]. CHIP, 1998, 22(1):42-46(In Chinese)

[13] **张坚,李红,满青青,等.中国居民营养与健康状况调查中血脂检测的质量控制**[J].**中国慢性病预防与控制**, 2008, 16(5):445-447
Zhang Jian, Li Hong, Man Qing-qing, et al. Quality Control of Plasma Lipids Measurement in CNHS 2002[J]. Chin J Prev Contr Chron Non-commun Dis, 2008, 16(5):445-447(In Chinese)

[14] **金丕焕. 医学科学研究中的数据管理和分析集** [J]. **中华预防医学杂志**, 2002, 36(1):68-70
Jin Bu-huan.Medical research in data management and analysis [J]. Chin J Prev Med, 2002, 36(1):68-70(In Chinese)

[15] Horri I. Data Management for Toxicological Studies [J]. Envi ron Health Perspect, 1994, 102(1): 71-75

[16] **葛爱华, 魏永跃, 李瑛, 等. 大型流行病学调查资料的数据管理和质量控制**[J]. **南京医科大学学报(自然科学版)**, 2011, 31(4):544-548
Ge Ai-hua, Wei Yong-yue, Li Ying, et al. Data management and quality control of the large-scale epidemiological survey study [J]. Acta Universitatis Medicinalis Nanjing, 2011, 31(4):544-548(In Chinese)

[17] **彭向东,吕筠,吕青. 流行病学调查中常用数据录入方法的选择**[J]. **中华流行病学杂志**, 2007, 28(11):1147
Peng Xiang-dong, Lv Jun, Lv Qing. Data entry methods of epidemiological survey study [J]. Chin J Epidemiol, 2007, 28(11): 1147(In Chinese)

(上接第 4337 页)

[16] **甄延城, 耿红, 周成超, 等. 流动肺结核病人结核病防治机构利用情况分析**[J]. **中国公共卫生**, 2010, 26(2): 157-159
Zhen Yan-cheng, Geng Hong, Zhou Cheng-chao, et al. The analysis of utilizing TB prevention and control of the floating tuberculosis patients[J]. Chinese Public Health, 2010, 26 (2): 157-159

[17] **邓肖英, 李齐芳, 龚明成. 灵山县** 2005-2008 **年肺结核病流行病学分析**[J]. **右江民族医学院学报**, 2010, 32(2): 167-169
Deng Xiao-ying, Li Qi-fang, Gong Ming-cheng. The epidemiology of tuberculosis in Lingshan, 2005-2008 [J]. Youjiang Medical College, 2010, 32 (2): 167-169

[18] **陈潇潇, 黄明豪, 李小宁等. 苏北地区农民结核病防治知识调查**[J]. **中国公共卫生**, 2007, 23 (6): 693-695
Chen Xiao-xiao, Huang Ming-hao, Li Xiao-ning, et al. The survey on the knowledge of TB in Northern farmers [J]. Chinese Public Health, 2007, 23 (6): 693-695

[19] **刘鸽, 冯学山, 詹绍康. 我国流动人口结核病流行现状与防制策略**[J]. **中国公共卫生**, 2007, 23(6): 701-703
Liu Ge, Feng Xue-shan, Zhan Shao-kang. The status of the floating population and the prevalence of TB control strategies in China [J]. Chinese Public Health, 2007, 23 (6): 701-703

[20] **付强. 胶南市结核病疫情分析及防治措施**[J]. **中国现代医生**, 2011, 49 (18): 175-176
Fu Qiang. Analysis and Prevention of Tuberculosis Analysis in Jiaonan City[J]. China Modern Doctor, 2011, 49 (18): 175-176