娄 培,刘 莉,陈先来,等. 基于问卷调查的医疗数据分类分级研究[J]. 中华医学图书情报杂志,2018,27(6):22-27,80.

DOI:10.3969/j.issn.1671-3982.2018.06.004

·研究与探讨·

基于问卷调查的医疗数据分类分级研究

娄 培,刘 莉,陈先来,安 莹,李忠民

[摘要]目的:针对医疗卫生机构在运营过程中产生信息众多而医疗行业尚没有明确的数据保护方案的问题,对底层数据进行分类分级,进而保护患者隐私等重要信息。方法:分类整理电子病历中涉及的数据项,对每类中包含的数据项进行抽象,设计医疗数据分类分级的调查问卷对相关人员进行调研。结果:根据问卷结果,结合数据分析方法,确定医疗数据的分类分级,对人口学、标识、卫生机构人员、卫生费用、药物、检查等 11 个大类赋予 0 级可公开、1 级数据关系个人信息安全、2 级数据涉及机构信息安全 3 种不同级别。结论:为后续的访问控制或跨库联合识别奠定了基础。

[关键词]大数据;分类分级;数据安全;医疗数据;问卷调查

[中图分类号]TP309.2;R197.323 [文献标志码]A [文章编号]1671-3982(2018)06-0022-07

Questionnaire investigation-based categorization and classification of medical data

LOU Pei, LIU Li, CHEN Xian-lai, AN Ying, LI Zhong-min

(Central South University Information Security and Big Data Institute, Changsha 410000, Hunan Province, China)

Corresponding author: CHEN Xian-lai

[**Abstract**] **Objective** To categorize and classify the basic medical data and protect the important information of patients such as their privacy as massive information was generated in medical and health institutions and no definite data protection plan has been worked out by the medical industry. **Methods** The data were extracted from EMR and the related persons were investigated by designing questionnaire for investigating the categorization and classification of medical data. **Results** The data extracted from EMR were categorized and classified according to the questionnaire investigation results combined with data analysis methods. The 11 categories of the data extracted from EMR, such as demography, markers, health institution personals, health expenses, drugs and inspection, were given 0 level for publicity, 1 level for data-related individual information security and 2 level for data involving institutional information security. **Conclusion** Questionnaire investigation-based categorization and classification of data extracted from EMR lays a solid foundation for the successive access-controlled identification or cross database joint identification.

[**Key words**] Big data; Categorization and classification; Data security; Healthcare data; Questionnaire survey

大数据分析和挖掘会带来巨大的商业价值,也不可避免地会泄露人们的隐私[1]。如何进行数据保护,在不泄露用户隐私的前提下,提高大数据的利用率,挖掘大数据的价值,是目前大数据研究的关键问题之一。目前,我国医疗信息化建设飞速发展,多地、多个行业都建设了大数据中心,并致力于区域内的医疗数据共享[2]。大数据中心包含海量的数据,在推动数据资源共享开放的同时,加强数据资源的

[基金项目]国家重点研发计划"精准医学研究"专项"精准医学大数据处理和利用的标准化技术体系建设"项目子课题《精准医学大数据体系的规范化应用与评价》(2016YFC0901705)研究成果之一

[作者单位]中南大学信息安全与大数据研究院,湖南 长沙 410000

[作者简介]娄 培(1994-),女,河北保定人,在读硕士研究生,研究方向为信息分析。

[通讯作者]陈先来(1970-),男,湖南新田人,博士,教授,主、参编著作 10 余部,发表论文 40 余篇。E-mail:chenxianlai@csu.edu.cn

安全性,实行数据资源分级、分类管理就变得非常重要。数据分类分级是从隐私安全与保护成本的角度出发,对数据进行分类和等级划分,进而根据不同需要对关键数据进行重点防护。

1　医疗数据分类分级的国内现状

在大数据建设中,国内外对医疗数据的分类分级标准尚未明确。中国把医学大数据研究与应用作为重要的发展战略,一些大数据规范文件正在积极地制定中。全国信息安全标准化技术委员会发布的《大数据安全标准化白皮书》将医疗数据的安全风险分为静态数据的安全风险和动态数据的安全风险。对静态数据的安全风险,要设置访问权限控制和安全风险的分级分类管理策略;对动态数据的安全风险,要设置加密和动态审计,要对重要敏感数据(如涉及个人隐私的电子病历、电子健康档案、人口健康等数据)进行分级、标识等[3]。

国务院印发的《"十三五"国家信息化规划》中提出建设统一开放的大数据体系,强化数据资源管理,要推动数据资源的分类分级管理[4]。大数据安全标准特别工作组已启动的大数据安全国家标准制定项目《信息安全技术大数据安全管理指南》中提出,从大数据安全需求、数据分类分级等方面开展数据保护的管理工作[5]。各行各业都在积极响应政府号召,如贵州省发布《贵州省政府数据数据分类分级指南》把政府数据按主题、行业和服务进行分类,又将数据的安全等级划分为 6 级[6],但没有给各类赋予级别,在分类和分级的合并上还存在不足;《中国移动 IDC 维护管理规定-数据安全管理分册》中对移动公司网络系统中的数据分为用户身份、服务内容、衍生数据、运营管理 4 个大类,并按客户的重要程度定义了 4 种安全级别[7];赵鹏等人提出了银行数据资产安全管理体系框架,并在相关步骤中对数据项分类、数据资产保密分级标准和数据资产备份分级标准给出了示例[8]。

由于医疗行业数据的复杂性和行业的特殊性,我国对健康医疗领域相关数据安全和隐私保护的立法相对比较滞后,分类分级标准还在研制中。《"健康中国 2030"规划纲要》关于推进健康医疗大数据应用提出:"加强健康医疗大数据相关法规和标准体系建设,强化国家、区域人口健康信息工程技术能力,制定分级分类分域的数据应用政策规范[9]"。美国在立法方面相对完善,对数据保护的相关法律要求分散在各法律法规的条款中。例如《健康保险携带和责任法案》中明确规定了个人隐私数据保护的具体范围和披露原则[10],《隐私盾协议》提出用于商业目的的个人数据从境外传输到美国后必须明确告知数据采集、传输和使用的流程及目的[11],《联邦隐私法案》中对政府机构应当如何收集个人信息及什么内容的个人信息能够收集、储存、或向公众开放的权利等都做出了比较详细的规定[12]。

2　数据来源及方法

确定大数据环境下底层数据的安全,可以更好地保护用户隐私,促进医学研究和数据共享。对数据进行分类分级是数据保护的第一步,也是关键的一步。本文在对医学数据进行分类的基础上设计调查问卷,调查医学大数据中心主要用户对数据类的等级划分意见,并在此基础上结合数据分析方法、国内外现有法规等对级别进行调整,确定最终分类分级标准。

2.1　医疗数据分类分级设计

2.1.1　分类设计

结合卫生部电子病历基本数据集与中南大学医学大数据平台中的数据项,采用面分类法和线分类法将所有医疗数据分为 11 个大类。这 11 个大类涵盖了住院病历记录、转诊记录、医疗机构信息等数十张表单的 700 多个数据项。

医疗卫生机构在运营过程中获取、管理和利用的首要信息为患者个人医疗信息,包括诊疗过程中收集的人口学、健康史、手术、药物、检查、诊断及住院信息。除此之外,医疗资源信息与服务价格信息等也是人们关注的重要信息。其中,医疗资源信息是反映医疗机构的人力、物力资源的信息,把这部分数据归类到卫生机构人员中;服务价格信息,如门诊就诊费、检查检验费、医药费等,则归类到卫生费用中[13]。为方便存储和查阅,医院信息系统会为医疗过程中产生的每张表单赋予单号,为每项检查赋予编号。在本次研究中,把这类数据归类到医疗信息标识中。

2.1.2　分级设计

参考《保密法》《信息安全技术信息安全事件分

类分级指南》等分级准则，将上述数据划分为 5 个安全级别：0 级表示被调查者认为可以公开的数据；1 级为危害个人，表示会对患者和医务人员的工作和生活造成影响；2 级为危害机构，表示对医疗机构的权益造成损害；3 级为危害社会，指对社会秩序和公共利益造成损害；4 级为危害国家，指对国家安全造成损害。

2.2 问卷调查

2.2.1 调查对象的选取

医学大数据中心的用户主要来自医疗卫生机构、卫生行政管理部门、医学科研机构等，涉及 9 个群体，包括行政管理人员、医务人员、信息管理人员、科研人员、医学教育工作者等。在正式调查前进行了小范围的预调研，最终通过问卷星发放网络问卷。

本次调查共发放 402 份问卷，回收问卷 326 份，其中有效问卷 323 份。对问卷结果进行信度和效度分析，Cronbach a 系数为 0.97，大于 0.7；进行 KMO 和 Bartlett 检验，效度系数为 0.949，显著性 sig 值为 0.000，表明差异是显著的。因此，问卷整体的信度和效度理想，问卷数据可靠有效。

被调查者的基本信息分析如图 1 所示。被调查人员来自医疗相关机构，其中以医疗卫生机构人员最多（占 35.28%），其次是医学教育机构（占 16.87%）。调查群体中明确填写所属人群的人员占 70%，其中以医务人员和信息管理人员为主（分别占比 19.94% 和 20.25%），其他人员包括行政管理人员、医疗保险人员、患者、药品器械公司人员、医学科研人员、医学教育工作者；明确选择使用数据目的的占 70%；使用临床数据进行医药卫生研究、临床医疗和医学教育的人数最多（占一半以上），其次是使用数据进行行政管理和保险方面的研究；绝大部分人认为保护医学数据很重要，占 88.85%。

图 1 被调查者基本信息分析

2.2.2 调查问卷设计框架

问卷分为两大部分，第一部分主要用于收集被调查人员的基本情况及其对数据公开的看法等，问题包括被调查者所属群体、使用医疗数据的主要目的、性别、年龄、所属的机构、对医学数据保护的重要性的看法和对医疗数据公开的看法，问题分为多选和单选；第二大部分为数据分级部分，包含 42 个问题，设 5 个安全级别（表 1），相关群体根据背景知识结合自身理解对所设问题的安全级别进行选择，在分级设计的 0–4 级中选择一级，其中家族史及之后的数据中不包含能够识别患者信息的数据（即为脱敏后的数据）。

表 1　医疗数据分类分级

序号	题目\选项	0 可以公开	1 危害个人	2 危害机构	3 危害社会	4 危害国家
1	身份证号	13(4.02%)	243(75.23%)	13(4.02%)	29(8.98%)	25(7.74%)
2	姓名	48(14.86%)	241(74.61%)	11(3.41%)	12(3.72%)	11(3.41%)
3	性别	219(67.8%)	89(27.55%)	4(1.24%)	7(2.17%)	4(1.24%)
4	出生日期	111(34.37%)	186(57.59%)	8(2.48%)	9(2.79%)	9(2.79%)
5	民族	235(72.76%)	51(15.79%)	3(0.93%)	21(6.5%)	13(4.02%)
6	联系电话	13(4.02%)	259(80.19%)	12(3.72%)	29(8.98%)	10(3.1%)
7	籍贯	174(53.87%)	114(35.29%)	9(2.79%)	15(4.64%)	11(3.41%)
8	家庭地址	11(3.41%)	250(77.4%)	12(3.72%)	33(10.22%)	17(5.26%)
9	职业	160(49.54%)	112(34.67%)	23(7.12%)	14(4.33%)	14(4.33%)
10	工作单位	47(14.55%)	158(48.92%)	71(21.98%)	31(9.6%)	16(4.95%)
11	血型	190(58.82%)	107(33.13%)	6(1.86%)	9(2.79%)	11(3.41%)
12	婚姻状况	148(45.82%)	157(48.61%)	4(1.24%)	9(2.79%)	5(1.55%)
13	家庭成员	29(8.98%)	240(74.3%)	8(2.48%)	37(11.46%)	9(2.79%)
14	患者联系人信息	10(3.1%)	240(74.3%)	17(5.26%)	37(11.46%)	19(5.88%)
15	患者类型代码	165(51.08%)	115(35.6%)	18(5.57%)	15(4.64%)	10(3.1%)
16	医疗信息标识	65(20.12%)	187(57.89%)	27(8.36%)	29(8.98%)	15(4.64%)
17	医疗卫生机构信息	188(58.2%)	42(13%)	67(20.74%)	17(5.26%)	9(2.79%)
18	医疗卫生人员信息	116(35.91%)	121(37.46%)	61(18.89%)	17(5.26%)	8(2.48%)
19	患者的医疗保险类型和付费方式	176(54.49%)	82(25.39%)	22(6.81%)	27(8.36%)	16(4.95%)
20	收费项目名称和总金额	185(57.28%)	63(19.5%)	39(12.07%)	22(6.81%)	14(4.33%)
21	签名信息	80(24.77%)	148(45.82%)	51(15.79%)	27(8.36%)	17(5.26%)
22	家族史	127(39.32%)	164(50.77%)	7(2.17%)	16(4.95%)	9(2.79%)
23	疾病史	129(39.94%)	164(50.77%)	11(3.41%)	13(4.02%)	6(1.86%)
24	传染病史	119(36.84%)	155(47.99%)	14(4.33%)	25(7.74%)	10(3.1%)
25	过敏史	174(53.87%)	126(39.01%)	8(2.48%)	11(3.41%)	4(1.24%)
26	手术名称	186(57.59%)	101(31.27%)	19(5.88%)	12(3.72%)	5(1.55%)
27	其他手术信息	180(55.73%)	102(31.58%)	28(8.67%)	9(2.79%)	4(1.24%)
28	助产记录	145(44.89%)	134(41.49%)	25(7.74%)	15(4.64%)	4(1.24%)
29	药物信息和使用方法	184(56.97%)	88(27.24%)	32(9.91%)	14(4.33%)	5(1.55%)
30	药物过敏信息	196(60.68%)	97(30.03%)	19(5.88%)	5(1.55%)	6(1.86%)
31	体格检查项目信息	184(56.97%)	117(36.22%)	12(3.72%)	6(1.86%)	4(1.24%)
32	检查检验报告结果	152(47.06%)	139(43.03%)	19(5.88%)	8(2.48%)	5(1.55%)
33	其他检查检验相关信息	152(47.06%)	136(42.11%)	21(6.5%)	11(3.41%)	3(0.93%)
34	患者主诉	156(48.3%)	138(42.72%)	17(5.26%)	9(2.79%)	3(0.93%)
35	患者入院科室	160(49.54%)	109(33.75%)	43(13.31%)	8(2.48%)	3(0.93%)
36	患者出院情况	167(51.7%)	110(34.06%)	34(10.53%)	10(3.1%)	2(0.62%)
37	转诊记录	159(49.23%)	109(33.75%)	42(13%)	9(2.79%)	4(1.24%)
38	病程记录	144(44.58%)	113(34.98%)	51(15.79%)	12(3.72%)	3(0.93%)
39	医嘱信息	171(52.94%)	94(29.1%)	45(13.93%)	10(3.1%)	3(0.93%)
40	诊断信息	156(48.3%)	120(37.15%)	32(9.91%)	12(3.72%)	3(0.93%)
41	护理记录	173(53.56%)	102(31.58%)	35(10.84%)	10(3.1%)	3(0.93%)
42	日期时间	185(57.28%)	97(30.03%)	25(7.74%)	12(3.72%)	4(1.24%)

3 结果

3.1 分布分析

对各选项的选择人数进行统计。从表 1 可以看出,超过半数的人认为以下信息可以公开:性别、民族、籍贯、血型、患者类型、卫生机构信息、患者的医疗保险类型和付费方式、收费项目名称和总金额、过敏史、手术名称、其他手术信息、药物信息和使用方法、药物过敏信息、体格检查项目信息、出院情况、医嘱信息、护理记录、日期时间。它们中包括间接描述患者人口学的信息、费用信息和住院过程中产生的诊断信息,这些数据单独识别个人的风险很小,大部分人倾向于公开此类信息。

半数以上人认为公开后可能危害个人的信息类别有:身份证号、姓名、出生日期、联系电话、家庭地址、家庭成员、患者联系人信息、医疗信息标识、家族史、疾病史。这部分信息以人口学信息为主,单个数据项直接识别患者个人的风险很大。

被调查者的选择主要集中在 0 级可以公开和 1 级危害个人。随着级别的增加,选择人数呈减少的趋势。

产生这种现象的原因有两点:一是更多人可能只关注到个人隐私的层面,对于机构、社会信息安全的关注度较低;二是问卷中提到关于诊断、检查等信息是脱敏的,对于此类不包含个人信息的数据大部分人选择公开。

3.2 聚类分析

统计每题中各选项的人数,使用 k-means 方法对数据进行分类。实验发现,k 取 4 和 5 时对人口学大类中的数据划分过于细致,不利于级别的划分,各类中的案例数差异较大,数据不平衡;k 取 3 时聚类效果最好。所以把级别从之前的 5 级更改为 3 级,即 0 级表示可以公开的数据;1 级表示数据关系个人信息安全,数据泄露会对患者或医务人员的工作和生活造成影响;2 级表示数据涉及机构信息安全,数据泄露会对机构的权益造成损害,可能影响社会秩序。

对数据结果进行整理,划分为 1 级的有身份证号、姓名、联系电话、家庭地址、家庭成员、患者联系人信息,划分为 2 级的有出生日期、工作单位、婚姻状况、医疗信息标识、医疗卫生人员信息、签名信息、家族史、疾病史、传染病史、助产记录。

3.3 结果调整

在对问卷结果进行分析的基础上,参照相关法规、标准,对得到的初步分类分级结果进行调整。国家标准 GB/T 35273-2017《信息安全技术个人信息安全规范》明确将生育信息、以往病史、过敏信息、传染病史等个人的健康信息纳入个人敏感信息,民族、出生日期、家庭关系等纳入个人信息[14]。

国标 GB/Z28828-2012《信息安全技术公共及商用服务信息系统个人信息保护指南》对个人敏感信息的定义为“一旦遭到泄露或修改,会对标识的个人信息主体造成不良影响的个人信息”。个人敏感信息包括身份证号码、手机号码、种族、政治观点、宗教信仰、基因、指纹等[15]。

参照上述标准,本次研究将民族的安全等级由 0 调整为 1,出生日期、婚姻状况、健康史(家族史、疾病史、传染病史、过敏史)、助产记录信息的安全等级由 2 调整为 1。调整部分见表 2。

表 2 数据分类分级结果调整

分类	二级类	分级
人口学	工作单位	2
	身份证号、姓名、出生日期、民族、联系电话、家庭地址、婚姻状况、家庭成员、患者联系人信息	1
	性别、籍贯、职业、血型、患者类型代码	0
卫生机构人员	医疗卫生人员信息	2
	医疗卫生机构信息	0
健康史	家族史、疾病史、传染病史、	1
	过敏史	0
手术	助产记录	1
	手术名称、其他手术信息	0

4 结论

根据表 2 可以看出,调整后人口学、卫生机构人员、健康史、手术 4 个大类中包含的二级类数据安全级别不同,如人口学大类中包含 15 个二级类,15 个二级类又划分了 3 种不同的级别。

对于同一大类中二级类分级不同的情况,其保密等级以不低于最高保密等级的原则进行调整。

得出的最终分类分级结果为：人口学、标识、卫生机构人员、日期时间及签名类为 2 级，健康史、手术信息为 1 级，卫生费用、药物、检查、诊断及住院信息类为 0 级（表 3）。

表 3　最终建议的数据分类分级结果

分类	二级类	分级
人口学	工作单位、身份证号、姓名、出生日期、民族、联系电话、家庭地址、婚姻状况、家庭成员、患者联系人信息、性别、籍贯、职业、血型、患者类型代码	2
标识	医疗信息标识	2
卫生机构人员	医疗卫生人员信息、医疗卫生机构信息	2
卫生费用	患者的医疗保险类型和付费方式、收费项目名称和总金额	0
签名信息	签名信息	2
日期时间	日期时间	0
健康史	家族史、疾病史、传染病史、过敏史	1
手术	助产记录、手术名称、其他手术信息	1
药物	药物信息和使用方法、药物过敏信息、	0
检查	体格检查项目信息、检查检验报告结果、其他检查检验相关信息	0
诊断及住院信息	患者主诉、患者入院科室、患者出院情况、转诊记录、病程记录、医嘱信息、诊断信息、护理记录	0

5　结语

本文通过对相关专业人员的问卷调查，得到初步分类分级结果，并结合已出台的标准及各种规章制度中对个人隐私信息的界定对结果进行了修改，经过综合分析得出每一个分类的安全等级。在实际应用中，用户所需的数据项组成复杂，不同数据项的安全等级不同，因此需要设计一种规则模型，在应用的过程中按照设计好的模型对数据集进行分级。如由多个数据项组成的数据集，其等级为组成它的所有数据项的最高等级；由多条数据项组成的数据集，通过统计、分析这些记录可以获得某项指标，则该数据集的等级等于该指标的等级等。

利用上述规则模型，可以在使用过程中对数据集进行级别划分，进而将其应用于访问控制或者跨库联合识别中，防止有人故意申请使用一些低等级的数据，通过联合大量的有关联关系的低等级数据，用数据挖掘或者其他方法得到等级较高的数据，从而造成重要信息或隐私信息的泄露。

对医疗数据的分类分级不是一成不变的，随着科技进步和大数据技术的发展，将会有更多的数据项加入进来，也会有更细致的分级策略。希望本文的初步探索可以对同行有所启发。

【参考文献】

[1] 方滨兴，贾　焰，李爱平，等. 大数据隐私保护技术综述[J]. 大数据，2016，2(1)：1-18.

[2] 陈　青，熊晓峰. 基于大数据的医院数据中心建设思考[J]. 科技资讯，2016，14(20)：8-10.

[3] 大数据安全标准化白皮书(2018 版)[EB/OL]. (2018-04-16)[2018-04-25]. http://www.cesi.ac.cn/201804/3789.html.

[4] “十三五”国家信息化规划. 国发〔2016〕73 号 000014349/2016-00257[R]. 北京：国务院，2016：12.

[5] 国家标准草案发布：《信息安全技术 大数据安全管理指南》[EB/OL]. (2017-05-26)[2017-12-25]. http://toutiao.manqian.cn/wz_16nn02AwfHb.html.

[6] 贵州省质量技术监督局. 政府数据数据分类分级指南(试行) DB52/T 1123-2016[S]. 贵州，2016：9.

[7] 中国移动通信有限公司网络部. 中国移动 IDC 维护管理规定云计算资源管理分册(2016 版)[EB/OL]. [2018-04-25]. http://www.doc88.com/p-7992592652581.html.

[8] 赵　鹏. 银行数据资产安全分级标准与安全管理体系建设方法[C]//中国软科学研究会. 第七届软科学国际研讨会论文集中国卷(上). 中国软科学研究会：中国软科学研究会，2012：9.

[9] 中共中央、国务院. “健康中国 2030”规划纲要[EB/OL]. (2016-12-30)[2017-12-25]. http://www.mohrss.gov.cn/SYrlzyhshbzb/zwgk/ghcw/ghjh/201612/t20161230_263500.htm.

(下转第 80 页)