·述评·

胃结直肠癌术后腹腔感染性并发症多中心、前瞻性数据库的建立及数据质量控制

王琦 吴舟桥 刘子宁 李子禹 季加孚
北京大学肿瘤医院暨北京市肿瘤防治研究所胃肠肿瘤中心一病区 恶性肿瘤发病机制及转化研究教育部重点实验室，北京 100142
通信作者：李子禹，Email：ziyu_li@hsc.pku.edu.cn

【摘要】 腹腔感染性并发症作为导致二次手术和术后死亡的主要原因，其发生率在国内不同医疗中心差异显著。由于缺乏来自全国范围的数据，尚无法合理评估并制定相应的诊断与治疗策略。为实现并发症规范化登记并为临床研究提供良好的数据平台，建立胃结直肠癌术后腹腔感染性并发症多中心、前瞻性数据库（PACAGE 数据库）。基于全国 20 家医疗中心的信息管理系统，利用电子病历报告表（e-CRF）采集网站收集胃结直肠癌手术患者的医疗信息，通过现场审计和 R 软件对数据进行核查及清洗。在数据清洗完成后，由主要研究者、数据管理员共同对数据库内数据进行核对和评价。当所有数据质询与疑问均已进行校正和解答后，对数据库进行锁定，形成最终的 PACAGE 数据库。PACAGE 数据库信息资源丰富，结合质量控制结果，其数据质量高，有望成为良好的并发症登记和临床研究的数据平台。

【关键词】 胃肿瘤； 结直肠肿瘤； 术后并发症登记； 数据质量； 数据清洗
基金项目：北京市医院管理中心青年人才培养“青苗”计划（QML20191103）；北京大学临床医学+X 青年项目（PKU2022LCXQ038）

Establishment and data quality control of a multicenter prospective database for prevalence of abdominal complications after gastroenterological surgery

Wang Qi, Wu Zhouqiao, Liu Zining, Li Ziyu, Ji Jiafu
Department of Gastrointestinal Surgery, Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education), Peking University Cancer Hospital & Institute, Beijing 100142, China
Corresponding author: Li Ziyu, Email: ziyu_li@hsc.pku.edu.cn

【Abstract】 As the main cause of secondary operation and postoperative death, the incidence of intraperitoneal infectious complications varies significantly in different medical centers in China. Due to the lack of national data, it is not possible to assess and develop appropriate diagnosis and treatment strategies properly. To provide a high-quality data platform for complication registration and clinical research, a multicenter prospective database for the Prevalence of Abdominal Complications After GastroEnterological surgery was established. Based on the Hospital Information System (HIS)of 20 medical centers in China, the electronic case reporting form (e-CRF) listed on the website was used to collect medical information of patients undergoing gastric or colorectal cancer surgery. The data were verified by on-site auditing, and data cleaning was performed by R software. After the data cleaning, the data in the database was checked and

DOI：10.3760/cma.j.cn441530-20221214-00524
收稿日期 2022-12-14 本文编辑 万晓梅
引用本文：王琦, 吴舟桥, 刘子宁, 等. 胃结直肠癌术后腹腔感染性并发症多中心、前瞻性数据库的建立及数据质量控制[J]. 中华胃肠外科杂志, 2023, 26(2): 154-159. DOI: 10.3760/cma.j.cn441530-20221214-00524.

evaluated by the principle investigators and data administrators. When all data queries and questions were corrected and answered, the database was locked to establish a multicenter prospective database for postoperative abdominal infectious complications (the PACAGE database). The PACAGE database has rich information resources and high data quality and is a good data platform for complication registration and clinical research.

【**Key words**】 Stomach neoplasms; Colorectal neoplasms; Postoperative complication registration; Data quality; Data cleaning

Fund programs: Beijing Hospital Management Center Young Talents Cultivation "Young Talents" Program (QML20191103); Peking University Clinical Medicine + X Youth Program (PKU2022LCXQ038)

一直以来，术后并发症作为判断手术质量的重要观察指标之一，备受国内外临床医生的关注。根据中国胃肠肿瘤外科联盟的统计数据，腹腔感染性并发症（吻合口漏）是导致二次手术和术后死亡的主要原因，且国内不同医疗中心所报道的并发症发生率差异明显[1]。由于缺乏来自全国范围的数据，我们无法合理评估腹腔感染性并发症的发生率，并制订相应的循证诊断与治疗策略。考虑到我国在术后并发症诊断登记方面仍缺乏经验，也缺乏能够反映我国胃肠道术后并发症发生、干预和转归情况的全国多中心数据库，本中心联合全国20家医疗中心开展了“胃结直肠癌术后腹腔感染性并发症现状研究”（Prevalence of Abdominal Complications After GastroEnterological surgery, PACAGE）[2]。2018年12月，PACAGE研究启动会在北京举行，会议讨论了病例报告表（case reporting form，CRF）的内容调整、降低入组偏倚的策略及入组标准的细节调整等方面的内容，为研究的顺利开展奠定了夯实的基础。启动会后，各参与中心在获得了所在单位伦理委员会的批准后，相继开始纳入研究患者，至2020年9月，提前完成入组计划。2020年12月，完成数据的核查及清洗，至2021年1月，再次数据审核后，形成最终的胃结直肠癌术后腹腔感染性并发症多中心、前瞻性数据库（以下简称PACAGE数据库）。现将PACAGE数据库的构建及数据质量控制情况报告如下。

一、数据来源

1. 医院来源：来自全国各区域的20家医疗中心参与了本数据库的建设，具体包括（按单位名称汉语拼音顺序排列）：北京大学肿瘤医院、北京协和医院、大连医科大学附属第一医院、福建医科大学附属协和医院、复旦大学附属华山医院、复旦大学附属中山医院、复旦大学附属肿瘤医院、广东省人民医院、吉林大学第二医院、南昌大学第一附属医院、南方医科大学南方医院、青岛大学附属医院、青海大学附属医院、厦门大学附属第一医院、山东省立医院、上海交通大学医学院附属瑞金医院、西安交通大学第一附属医院、浙江大学医学院附属杭州市第一人民医院、浙江大学医学院附属邵逸夫医院、中山大学附属第一医院。

2. 数据采集及录入：PACAGE研究已通过北京大学肿瘤医院及医学部伦理审查委员会审查批准（伦理号：2018YJZ56）；注册号为NCT03828266。各参与中心在获得所在单位伦理委员会的批准后，根据研究方案的纳入排除标准相继开始数据采集及录入工作[2]。各参与单位使用唯一用户名及密码，主要研究人员经培训后即可登录网站（http://219.239.107.20: 2000/cra/base/list.do?prPk=46）填写电子病历报告表（electronic case reporting form，e-CRF），研究人员需每个入组患者平均输入约100个变量。患者的核心信息、手术信息、术后信息均为前瞻性登记，在患者出院前完成录入，出院后只能录入部分出院时尚未报告病理情况的病理信息。与此同时，为了简化快速登记过程，及时、规范地上报临床数据，大多数数据可以通过下拉框选择，而只有年龄、身高、体质量、C-反应蛋白（C-reaction protein，CRP）、降钙素原（procalcitonin，PCT）和住院费用等指标需输入数值。每个单位的主要研究人员只能访问他们自己的数据集，而为了便于数据督查，研究管理员可以访问所有单位的数据集。

二、数据质量控制

1. 数据核查：数据核查主要从数据逻辑性、数据完整性、数据真实性及方案依从性等方面进行[3]。由两位数据核查员对所有登记数据进行现场双重核查，同时借鉴日本国家临床数据库（National Clinical Database，NCD）的数据审计经验，采用现场审计的方式，至少在3个参与中心，分别随机选择20例患者进行核查[4-5]。数据核查过程中，核查员对每一个录入数据值都结合原始病历资

料进行核对，如有异议，则与分中心研究员进行沟通，核对后进行相应修改、剔除、补充。尽管所有录入数据均进行核对，但核查过程中对以下6个方面的内容进行更为细致的核对[6]：(1)入组资质核实：对于病种、手术方式、手术范围等不符合研究方案纳入标准的重要指标数据，进行进一步疑问质询与核实。(2)数据取值范围：如身高，普通人群成年男性身高144~188 cm，成年女性140~180 cm[7]。超出此范围的数据需进一步进行数据疑问质询与核实。(3)异常值核实：通过对关键变量进行简单的描述性分析，例如术后CRP明显升高或长时间使用抗生素却未登记并发症的患者，将结合病历资料进行核查。(4)变量间的矛盾核实：如某患者手术方式与标本信息或病种不符，一般情况下需进一步核实。(5)缺失数据：年龄、性别、住院费用及其他关键变量缺失，需通过疑问质询进行补充。(6)关键日期与时间的核实：日期是否存在数量级错误，如手术日期是否在入组时间之前或出院时间之后等。

2. 数据清洗：数据清洗也是保证数据质量的重要环节。数据清洗也叫数据清理，是指从数据库或数据表中更正和删除不准确数据记录的过程。广义地说，数据清洗包括识别和替换不完整、不准确和不相关或有问题的数据和记录。而这些问题数据将直接影响结论的可靠性。为保证数据库质量，需要对重复数据、错误数据、异常值与缺失值进行清洗，具体分为以下5种情况[8]：(1)剔除不符合入组标准的患者数据。(2)对重复数据进行简化、合并或删除。(3)对数据录入错误等原因造成的错误数据进行舍弃或修正。(4)应用R软件(R-3.6.2)，采用基于正态分布的格拉布检验或切比雪夫定理来检测异常值，由具有临床背景的研究者进行目测检查，对于超出正常范围或与同列其他数据在逻辑上不一致的异常值，进行进一步数据疑问质询与核实。(5)对缺失数据，各中心研究人员分别根据原始病历进行补充修正，修正后仍然不可信或缺失，或不可追踪，则被标记为缺失值。以上数据清洗过程运用R软件(R-3.6.2)进行处理。

3. 数据审核及锁定：在数据清理完成后，由主要研究者和数据管理员共同对数据库内数据进行核对和评价。当所有数据质询和疑问均已进行校正和解答后，对数据库进行锁定。锁定后的数据文件不允许再作变动，以防止产生误操作及未经授权的修改。

三、PACAGE数据库的基本概况

1. 数据录入情况：2018年10月至2020年12月，20个参与中心共录入了4 131例胃结直肠癌患者数据。各月份入组登记情况见图1和图2。

2. 数据核查结果：由于新型冠状病毒感染(Corona Virus Disease 2019，COVID-19)的影响，研究者只抽取了入组数量前两名的中心(北京大学肿瘤医院、青海大学附属医院)进行了现场数据审计。从两个中心的数据中分别随机抽取20例患者，共计40例进行审计。通过检索病历资料，发现38个(0.81%，38/4 695)登记错误，并进行了修订。

3. 数据清洗结果：205例患者因各种原因被剔除。0.16%(774/471 609)的数据经过数据清理后

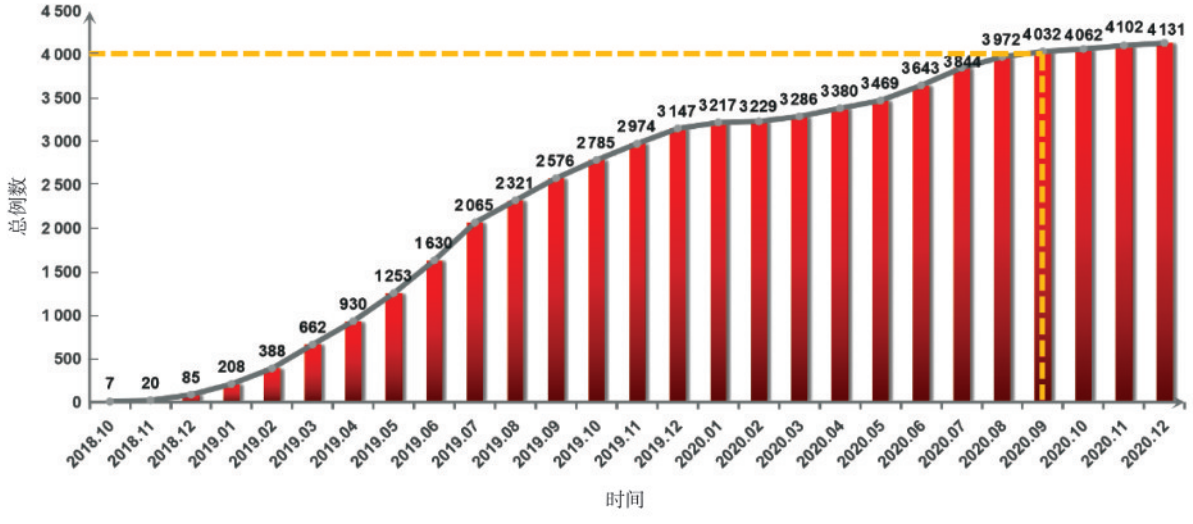

图1 胃结肠癌术后感染性并发症现状研究数据库单月入组情况

被纠正。其中，缺失数据为0.038%（181/471 609），见图3。

4.PACAGE数据库：该数据库包括全国5个区域（数据地区分布见图4），20个参与中心，共3 926例符合质量标准的有效医疗数据，471 609个有效数据条目。其中，胃癌2 271例（57.8%），结直肠癌1 655例（42.2%）。

四、PACAGE数据库的意义

回顾中国胃肠肿瘤外科并发症规范化诊断登记的历程，2015年中国胃肠肿瘤外科联盟数据库建立，2018年中国胃肠肿瘤外科联盟制订《中国胃肠肿瘤外科术后并发症诊断登记规范专家共识》，2018—2020年胃肠联盟内20家中心联合开展《胃结直肠癌术后腹腔感染性并发症的现状研究（PACAGE研究）》，建立PACAGE数据库，我们正在一步步实现胃癌术后并发症诊断、登记在更大范围内的标准统一[2,9-11]。PACAGE研究首次汇报了我国胃肠术后腹腔感染性并发症的现状，为后续相关研究提供了基本数据和标准。

五、PACAGE数据库及国际大型数据库数据质量控制经验

作为国内首个以并发症登记为主要目的的多

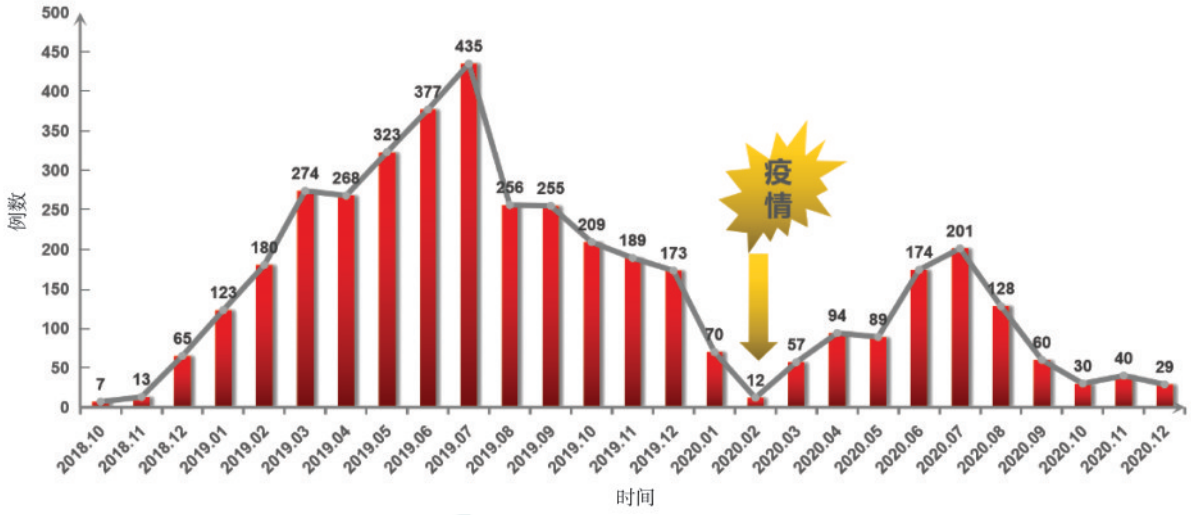

图2　胃结肠癌术后感染性并发症现状研究数据库单月新增入组情况

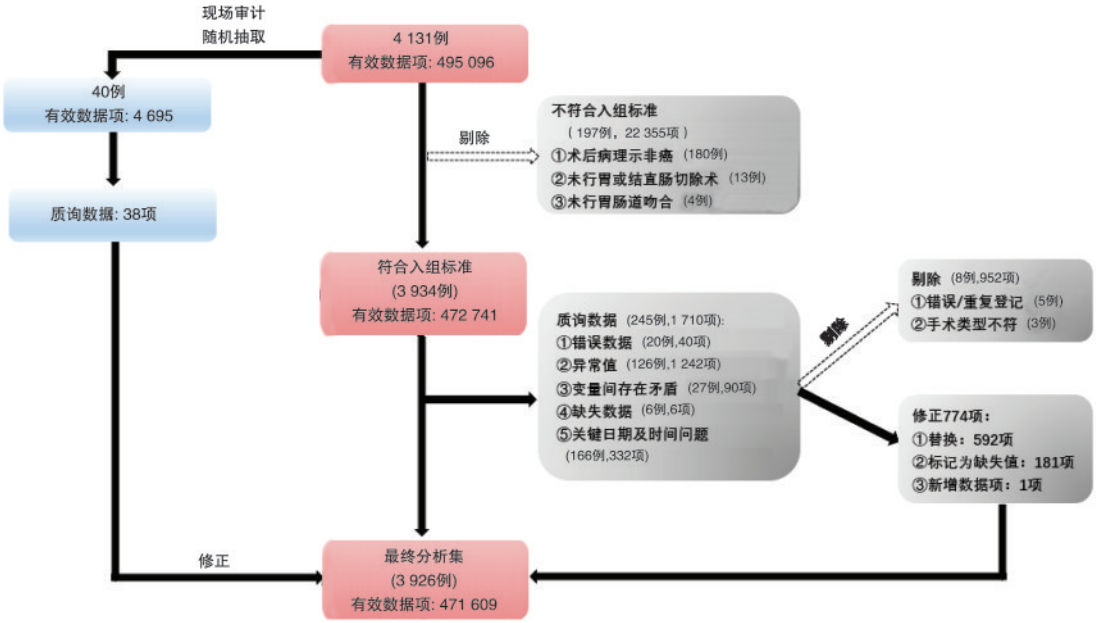

图3　胃结肠癌术后感染性并发症现状研究数据库现场审计及数据清洗的流程图

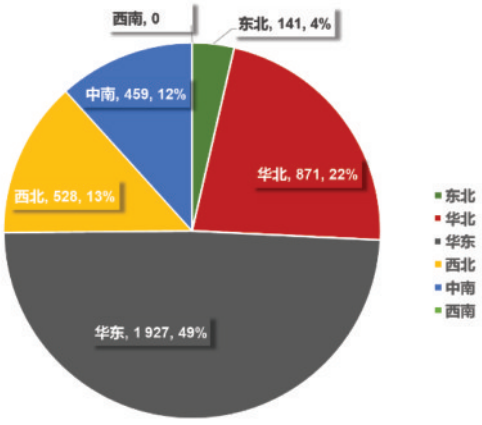

图 4 胃结肠癌术后感染性并发症现状研究数据库的数据地区分布

中心数据库，PACAGE 数据库的数据质量从数据核查和清洗结果而言，较为满意。现场数据审计结果显示，数据一致性为 99.2%（4 657/4 695），缺失数据为 0.038%（181/471 609）。而据文献报道，美国外科学院（American College of Surgeons，ACS）的国家手术质量提高计划（National Surgical Quality Improvement Program，NSQIP）数据库的数据一致性为 96.8%~98.4%（2005—2008 年）[12]；日本 NCD 的数据一致性为 98.1%[5]；瑞典国家食管癌和胃癌登记（National Register for Oesophageal and Gastric Cancer，NREV）数据库的数据一致性为 91.1%，缺失数据为 2.4%[13]；荷兰临床审计研究所（the Dutch Institute for Clinical Auditing，DICA）审计结果示数据一致性为 88.2%~100%，缺失数据为 0.6%~2.8%（2013—2015）[14]。PACAGE 研究数据一致性较高的原因主要在于相应录入变量简单，相较于 NSQIP 数据库或 NCD 而言，每例患者仅需要录入约 100 个变量，并且均尽可能通过点选，而避免了填报数字或开放型作答。从我们的结果中也能够看到，大多数修改的数据主要为包括 CRP 或 PCT 等录入数值。当然，本研究由于录入数据较少，与更为成熟的国际大型数据库相比，仍存在明显差距。

对比我国与西方发达国家的胃癌、结直肠癌诊疗现状可以发现，西方发达国家的肿瘤专科化建设和规范化诊疗较为成熟，而我国开展的临床数据库建设和临床研究的数量以及质量，与发达国家相比还有很大差距，故长久以来只能遵循西方或日韩的指南开展临床实践。近年来，随着我国外科医生对于开展临床研究理念和思路的进步，越来越多的医疗中心致力于开展高质量的临床研究，以获得更高级别的循证医学证据，书写适合我国人群及胃肠肿瘤现状的临床指南。而数据质量和临床研究水平是相辅相成的，想要开展高质量的临床研究，数据质量尤为重要。

临床数据审计是医疗保健质量评估和改善的主要手段，有助于改善患者的预后，对数据库数据的准确性和完整度有较为直观的判断[15-17]。美国胸外科学会（the Society for Thoracic Surgeons，STS）国家数据库主要通过培训专业的站点数据管理人员、自动化数据完整性验证系统、外部审核人员以及随机选择的现场审核来验证收集的数据质量[18]；NCD 主要是通过每年选定 45~46 家医院，各随机选取 20 例患者，通过现场审计（现场审核病例记录）或远程审计（将病例记录邮寄至日本胃肠外科学会办公室）进行数据质量审核[5]；而荷兰上消化道癌症审计（the Dutch Upper Gastrointestinal Cancer Audit，DUCA）数据库本身便是基于临床审计和监管的目的而建立的，主要通过以下 4 种方式验证和提升数据质量：（1）基于网络调查的数据核实系统（数据录入员在输入数据时即可收到关于错误、缺失或不可信的数据项的直接反馈）；（2）每家医院的电子报告表（总结缺失的变量和那些可能存在错误的变量）；（3）每周发送给临床专家最新审计报告；（4）第三方外部审计[14,17]。借鉴这些国际大型数据库的管理经验，PACAGE 研究在有限的人力、物力和财力的支持下进行了内部数据审核及清洗。也期待未来在更多支持下，我们可以通过培训专人进行病例资料填报及第三方审计，定期审核，进行严格的质量控制，提升临床数据的准确性及可靠性，进一步推动治疗的标准化及规范化。

在对 PACAGE 数据库进行数据清洗时，我们应用 R 软件，进行缺失值、逻辑错误及异常值的检测。数据缺失对于数据库的影响是尤为严重的，虽然依靠统计方法能够在一定程度上弥补数据缺失对分析造成的影响，但仍会影响到结论的可靠性，甚至造成无法分析的情况。对于检测出的缺失值，我们首先会反馈至各参与中心进行补充，确实无法查询或补充的，最终标记为缺失值；设置预定规则对逻辑错误进行检测，例如手术日期先于入组日期、手术日期晚于出院日期等，对检出信息进行进一步核查及修正；利用正态分布的格拉布检验或切比雪夫定理来检测异常值的数据信息，对于检测出的异常

值信息进行数据质询，由相应参与中心进行核查及修正。尽管例如日期逻辑错误等看似为简单错误，但实际上在包括荷兰等国家数据库中均无法彻底避免。在PACAGE筛检过程中，我们通过软件大大地节省了人力的时间成本，降低了人力筛检的错判和遗漏。然而，临床数据有时并无绝对界限，非常依赖临床医生及数据分析员的沟通及判断。在多中心数据研究中，当参考文献缺乏或不一致时，具备临床背景的审计员和数据分析员是数据清洗和设定参考范围的前提条件。虽然R语言对比其他方法，其可编程性和自由度更高，代码可重复使用．对需实现特有算法和批量数据处理的工作有较大优势，但是这种基于临床背景的判断和处理是R语言无法取代的[19]。

六、PACAGE数据库的局限性

PACAGE数据库作为国内并发症登记的首次大规模尝试，尚存一定局限性：(1)本数据库主要采用各中心自我报告的形式进行并发症登记和录入，将不可避免地影响并发症的报告率，特别是那些分级较低的并发症病例。(2)由于各参与单位的仪器、试剂及检验环境不同，可能对检测结果产生影响。(3)本数据库虽纳入了各区域多家大型诊疗中心的高质量医疗数据，但仍存在相应汇报偏倚。(4)本研究数据质控由研究人员实施内部核查，但实际上与第三方核查相比，内部核查仍可能存在相应偏倚，我们期待能够在未来进一步优化。

我们期待将来可以有全国更多的医疗中心参与，在严格把控数据库质量的同时，不断扩充数据量。同时，我们也鼓励各参与中心积极进行多中心合作，开展高质量研究，为中国胃肠肿瘤外科的临床实践提供更多、更高级别的循证医学证据。

利益冲突　所有作者均声明不存在利益冲突

参 考 文 献

[1] 苗儒林,李子禹,武爱文．中国胃肠肿瘤外科联盟数据报告(2014-2016)[J]．中国实用外科杂志,2018,38(1):90-93. DOI: 10.19538/j.cjps.issn1005-2208.2018.01.20.

[2] 王琦, 吴舟桥, 侯士阳, 等．胃结直肠癌术后腹腔感染性并发症的现状研究(PACAGE研究)[J]．中华消化外科杂志,2019, 18(3):229-234. DOI: 10.3760/cma.j.issn.1673-9752.2019.03.007.

[3] 刘智, 商洪才, 翟静, 等．基于网络的中医药临床试验在线数据核查[J]．天津中医药，2014, 31(11): 670-673. DOI: 10.11656/j.issn.1672-1519.2014.11.09.

[4] Kanaji S, Takahashi A, Miyata H, et al. Initial verification of data from a clinical database of gastroenterological surgery in Japan[J]. Surg Today, 2019,49(4):328-333. DOI: 10.1007/s00595-018-1733-9.

[5] Hasegawa H, Takahashi A, Kanaji S, et al. Validation of data quality in a nationwide gastroenterological surgical database: The National Clinical Database site-visit and remote audits, 2016-2018[J]. Ann Gastroenterol Surg, 2021,5(3):296-303. DOI: 10.1002/ags3.12419.

[6] 潘岳松．临床研究的数据管理与质量控制[J]．协和医学杂志，2018, 9(5): 458-462. DOI: 10.3969/j. issn. 1674-9081. 2018.05.016.

[7] 李思汉．我国北方地区成人各类体型不同身高的体重正常值的探讨[J]．营养学报,1986,8(2):98-109.DOI:10.13325/j.cnki.acta.nutr.sin.1986.02.002.

[8] 蔡晓谕, 林洁, 王超, 等．多中心临床术前贫血数据库构建[J]．临床输血与检验,2019,21(4):337-340,358. DOI:10.3969/j.issn.1671-2587.2019.04.001.

[9] «中国实用外科杂志»编辑部."中国胃肠肿瘤外科联盟"成立并发布初步数据[J]．中国实用外科杂志,2016,36(10):1077.

[10] 李子禹,吴舟桥,季加孚．中国胃肠肿瘤外科术后并发症诊断登记规范专家共识(2018版)[J]．中国实用外科杂志, 2018, 38(6):589-595.DOI:10.19538/j.cjps.issn1005-2208.2018.06.01.

[11] 吴舟桥, 李子禹, 季加孚．对胃癌术后并发症的再认识[J]．中华胃肠外科杂志,2017,20(2):121-124.DOI: 10.3760/cma.j.issn.1671-0274.2017.02.001.

[12] Shiloach M, Frencher SK, Steeger JE, et al. Toward robust information: data quality and inter-rater reliability in the American College of Surgeons National Surgical Quality Improvement Program[J]. J Am Coll Surg, 2010, 210(1): 6-16. DOI: 10.1016/j.jamcollsurg.2009.09.031.

[13] Linder G, Lindblad M, Djerf P, et al. Validation of data quality in the Swedish National Register for Oesophageal and Gastric Cancer[J]. Br J Surg, 2016, 103(10): 1326-1335. DOI: 10.1002/bjs.10234.

[14] van der Werf LR, Voeten SC, van Loe C, et al. Data verification of nationwide clinical quality registries[J]. BJS Open, 2019,3(6):857-864. DOI: 10.1002/bjs5.50209.

[15] Patel NK, Sarraf KM, Joseph S, et al. Implementing the national hip fracture database: an audit of care[J]. Injury, 2013, 44(12): 1934-1939. DOI: 10.1016/j. injury. 2013. 04.012.

[16] van Leersum NJ, Snijders HS, Wouters MW, et al. Evaluating national practice of preoperative radiotherapy for rectal cancer based on clinical auditing[J]. Eur J Surg Oncol, 2013,39(9):1000-1006. DOI: 10.1016/j.ejso.2013.06.010.

[17] Busweiler LA, Wijnhoven BP, van Berge Henegouwen MI, et al. Early outcomes from the Dutch Upper Gastrointestinal Cancer Audit[J]. Br J Surg, 2016,103(13):1855-1863. DOI: 10.1002/bjs.10303.

[18] Welke KF, Ferguson TB, Coombs LP, et al. Validity of the society of thoracic surgeons National Adult Cardiac Surgery Database[J]. Ann Thorac Surg, 2004,77(4):1137-1139. DOI: 10.1016/j.athoracsur.2003.07.030.

[19] 何贵成, 张华, 万毅．R语言在全国取水许可台账数据清洗中的应用[J]．电脑编程技巧与维护,2016,(12):71-73. DOI: 10.16184/j.cnki.comprg.2016.12.030.