**ICS** 11.020

|  |
| --- |
| **C** 07 |
|  |

团体标准

**T/BRA-CDCHE/ XXXX-XXXX**

|  |
| --- |
|  |

医学队列平台数据技术

队列数据管理 数据融合质控要求

Medical Cohort Platform Data Technology——

Cohort Data Managements——Quality Control Requirements of Data Integration

|  |
| --- |
| (草案) |
|  |

20XX- XX - XX发布

20XX - XX - XX实施

中国医院协会 发布

目  次

[前  言 III](#_Toc168887844)

[1． 范围 1](#_Toc168887845)

[2． 规范性引用文件 1](#_Toc168887846)

[3． 术语和定义 1](#_Toc168887847)

[4． 融合前 2](#_Toc168887854)

[4.1 融合的目的 2](#_Toc168887855)

[4.2 队列数据集的基本信息 2](#_Toc168887856)

[4.2.1 来源及研究起止日期 2](#_Toc168887857)

[4.2.2 研究目的 2](#_Toc168887858)

[4.2.3 研究设计 2](#_Toc168887859)

[4.2.4 研究对象 2](#_Toc168887860)

[4.2.5 变量 2](#_Toc168887861)

[4.2.6 样本量 2](#_Toc168887862)

[4.3 队列数据集的数据特征 2](#_Toc168887863)

[4.3.1 数据清单 2](#_Toc168887864)

[4.3.2 缺失数据 2](#_Toc168887865)

[4.3.3 核心变量和其他变量 2](#_Toc168887866)

[4.4 获得队列数据的使用许可 3](#_Toc168887867)

[4.4.1 联系数据负责人 3](#_Toc168887868)

[4.4.2 提供跨队列研究信息 3](#_Toc168887869)

[4.4.3 请求正式同意 3](#_Toc168887870)

[5． 融合中 3](#_Toc168887871)

[5.1 清理拟融合的原始队列数据 3](#_Toc168887872)

[5.1.1 识别并确认原始数据问题 3](#_Toc168887873)

[5.1.2 处理缺失数据 3](#_Toc168887874)

[5.1.3 处理异常数据 3](#_Toc168887875)

[5.1.4 处理重复数据 3](#_Toc168887876)

[5.1.5 与研究目标的权衡 4](#_Toc168887877)

[5.1.6 记录数据清理过程 4](#_Toc168887878)

[5.1.7 数据质量验证 4](#_Toc168887879)

[5.2 根据研究需求设定数据融合的规则 4](#_Toc168887880)

[5.2.1 选择需要融合的关键变量和次要变量 4](#_Toc168887881)

[5.2.2 设定数据质量评估标准 4](#_Toc168887882)

[5.2.3 设定预期满足标准的数据变量的百分比 4](#_Toc168887883)

[5.3 根据设定的数据融合标准评估所需的变量并进行定量评分 4](#_Toc168887884)

[5.4 验证融合结果 4](#_Toc168887885)

[6． 融合后 4](#_Toc168887886)

[6.1 使用文档记录过程 4](#_Toc168887887)

[6.2 确保过程的可追溯性和透明度 4](#_Toc168887888)

[6.3 定期评估和更新融合后的数据集 4](#_Toc168887889)

[参考文献 5](#_Toc168887890)

前  言

本文件按照GB/T 1.1—2020《标准化工作导则 第1部分：标准化文件的结构和起草规则》的规定起草。

本文件由北京大学第六医院提出。

本文件由北京慢性病防治与健康教育研究会归口。

本文件起草单位：北京大学第六医院、北京大学、中国医学科学院肿瘤医院、中国疾病预防控制研究中心、山东大学齐鲁医院、北京大学第一医院、天津市安定医院、中国电子技术标准化研究院。

本文件主要起草人：刘肇瑞、黄雨、魏文强、孙可欣、丁若溪

本文件其它起草人：张婷婷、陈园生、吕明、张媛、李航、徐广明、李瑞琪、王悦、张同超、白倩倩、潘鹏、颜国利。

队列数据管理 数据融合质控要求

1. 范围

本文件规定了跨队列研究实施过程中数据融合的质控要求等内容。

本文件适用于指导拟开展跨队列研究的数据融合工作，包括但不限于社区人群队列、区域性人群队列、针对某一疾病种类或基于特殊机构建立的人群队列。

1. 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

WS/T 306 卫生信息数据集元数据规范

WS 363卫生信息数据元目录（系列标准）

WS 364 卫生信息数据元值域代码（系列标准）

WS 365 城乡居民健康档案基本数据集

WS/T 370 卫生信息基本数据集编制规范

WS 372-2012 疾病管理基本数据集

WS 375 疾病控制基本数据集（系列标准）

GB/T 41814 信息技术 生物特征识别校准、增强和融合数据（系列标准）

1. 术语和定义

下列术语和定义适用于本文件。



队列 cohort

根据某个或某些共同特征而组建的一组特定人群。

注：特征包括但不限于：暴露因素、疾病或健康状态、出生时间或年代、地域、干预措施。



跨队列 cross-cohort

队列（3.1）间进行特征数据比较、融合和分析。

注：跨队列形式包括：a）横向跨队列：在不同元数据的队列间进行比较、融合和分析；b）纵向跨队列：在相同元数据的队列间进行比较、融合和分析。



数据融合 data fusion

将来自不同来源和不同格式的数据进行整合和合并。



变量 variable

数据集中最基本的元素，用于存储数据。

1. 融合前需明确的内容
   1. 融合的目的

明确研究问题和目的，说明开展数据融合的原因和目标。

* 1. 队列数据集的基本信息

基于数据的可靠性、适用性和完整性，按照融合的目的确定拟进行数据融合的队列数据集。收集或数据集中的相关信息，包括并不限于以下内容：

* + 1. 来源及研究起止日期

包括开展队列数据收集的单位、负责人、经费资助，以及数据采集的起始日期和完成日期。

* + 1. 研究目的

明确队列数据的研究目的，了解研究内容及预期结果。

* + 1. 研究设计

明确队列数据的研究设计类型，包括但不限于：现况调查、诊断试验、病例对照研究、队列研究、纵向研究、实验性研究等。根据不同的研究设计类型，记录相应的设计要素，包括但不限于抽样方法、随机方法、暴露因素、观测指标、干预措施、结局指标等内容。

* + 1. 研究对象

明确研究对象的特征、纳入标准、排除标准以及观察的期限。

* + 1. 变量

明确队列数据中的变量定义、数据类型、表示格式、数据单位以及取值范围。

* + 1. 样本量

根据研究目的选择样本，计算样本量是否可满足统计学检验的要求。

* 1. 队列数据可用性

根据队列数据集的数据特征，评估拟进行数据融合的队列数据的可用性，评估内容包括但不限于以下内容：

* + 1. 数据清单

创建一个清单，列出所有拟进行融合数据的队列数据源，包括其来源、格式、结构和可用性，从而对数据的整体情况进行全面了解。

* + 1. 缺失数据

对每个数据源进行缺失数据的识别和分析。确定每个数据源中缺失数据的变量数量、比例以及缺失原因。原因包括但不限于记录不完整、人为错误、技术问题或其他因素等。

* + 1. 核心变量和其他变量

评估每个变量的重要性和对研究问题的贡献，分析变量是否为主要观测指标或直接影响研究的主要结果。核心变量是对研究目的至关重要的变量，通常直接与研究假设或主要研究问题相关。确定哪些变量对于研究目的至关重要的核心变量，哪些是次要的其他变量。明确核心变量和其他变量的缺失比例，在决定如何处理缺失数据时根据变量的重要程度进行权衡。获得队列数据的使用许可

* + 1. 联系数据负责人

与数据负责人建立联系，申请队列数据使用许可。可以通过电子邮件、正式书面函件或面对面会议等适当的方式进行。

* + 1. 提供跨队列研究信息

向数据负责人提供跨队列研究的详细信息，包括研究问题、研究目的、数据提取需求、数据融合统计分析计划等。

* + 1. 请求正式同意

请求数据负责人提供正式的知情同意文件，明确他们同意跨队列研究使用、分享和融合他们的数据。知情同意文件可以是书面协议、合同或许可证书，获取后按照有效期对知情同意文件进行档案保存，知情同意文件包括但不限于以下内容：

* 1. 数据使用目的和范围
  2. 数据分享和融合的具体条款和条件
  3. 数据的保护和隐私保护措施
  4. 同意的有效期限（如果适用）
  5. 同意的签署日期和签名

1. 融合中的规范
   1. 清理拟融合的原始队列数据
      1. 统一变量的赋值方式

对于因单位不一致，赋值编码不一致、调查工具不一致导致的队列数据间相同评估变量的差异，可以通过数据转化的方式进行统一。

* + 1. 识别并确认原始数据问题

识别并确认原始数据的质量问题，如缺失值、异常值、重复值等。

* + 1. 处理缺失数据

可根据研究需求决定在数据融合过程中是否对数据的缺失值进行处理，根据数据的重要性和可用性，考虑以下不同的缺失数据处理方法，包括但不限于以下内容：

* 1. 删除缺失数据行：如果缺失数据占比较小，且对研究问题的影响不大，可以考虑删除包含缺失数据的行。
  2. 插补数据：使用统计方法或机器学习方法来估算缺失数据的值。
  3. 分析缺失数据：缺失数据本身可能包含信息，可以作为一个独立的变量进行分析。
  4. 加权分析：在某些情况下，可以使用加权分析来考虑缺失数据的影响，以降低偏差。
     1. 处理异常数据

可根据研究需求决定在数据融合过程中是否对数据的异常值进行处理，根据异常值的特征和对数据的理解，选择适当的处理策略，包括但不限于以下内容：

* 1. 删除异常值：对于极端异常值，可以考虑删除这些数据点，但需谨慎，以免丢失重要信息。

1. 修复异常值：可以通过插值、取中位数、平均值等方式对异常值进行修复，使其更接近数据集的整体趋势。
2. 分组：将异常值分到适当的分组中，以降低其对分析的影响。
3. 转换数据：应用适当的数学函数对异常值进行转换，以缓解其影响。
   * 1. 处理重复数据

根据重复数据的特征和对数据的理解，选择适当的处理策略，包括但不限于以下内容：

1. 删除重复数据：删除完全重复的记录，保留一条唯一的记录。
2. 合并重复数据：对重复数据进行合并，将重复数据合并为一条记录，保留合并后的唯一记录。
3. 分析重复数据原因：如果重复数据出现的原因不明确，可以分析其产生的原因，修复数据采集或输入过程中的错误。
   * 1. 与研究目标的权衡

在决定如何处理缺失数据时，需要权衡数据的可用性和对研究目标的影响。如果数据的缺失对研究问题非常关键，那么需要采取更谨慎的方法来处理缺失数据，如插补。如果数据的缺失对研究问题影响较小，可以考虑简化的方法，如删除缺失数据行。

* + 1. 记录数据清理过程

记录原始数据的质量问题，使用的数据清理工具、脚本或软件，并记录其参数和设置，缺失值、异常值、重复值的处理方式，以及数据转换，变量的删除或保留及其原因。

* + 1. 数据质量验证

对清理后的数据质量进行验证，确保数据质量满足要求。

* 1. 根据研究需求设定数据融合的规则

制定明确的融合规则和方法，包括数据匹配方式、变量匹配、合并方式等。

* + 1. 选择需要融合的核心变量和次要变量

根据研究目的和问题，选择需要融合的关键变量，并评估这些变量的重要性及局限性。

* + 1. 设定数据质量评估标准

评估标准应包括但不限于数据准确性、完整性、一致性和可靠性。

* + 1. 设定预期满足标准的数据变量的百分比

设定预期满足标准的数据变量的百分比是为了确保在融合过程中，数据质量达到一定的标准。这个百分比应基于评估数据质量的指标和方法来设定。

* 1. 根据设定的数据融合标准评估所需的变量并进行定量评分

根据设定的数据融合标准，对所需的变量进行评估，并根据预设标准进行定量评分。这些评分可以帮助在融合过程中权衡不同变量的重要性和可靠性。

* 1. 验证融合结果

验证融合后数据集的准确性和一致性，确保融合结果符合预期。

1. 融合后的要求
   1. 使用文档记录过程

记录数据融合的所有步骤、规则、参数以及处理过程，以便后续审查和复制研究。

* 1. 确保过程的可追溯性和透明度

确保数据融合的过程具有可追溯性，能够清晰展示每个数据源如何被融合到最终结果中。

* 1. 定期评估和更新融合后的数据集

定期评估数据融合的效果，根据反馈和新的需求更新融合标准，以保持数据融合的质量和适应性。

参 考 文 献

1. Benchimol, E. I., L. Smeeth, A. Guttmann, K. Harron, D. Moher, I. Petersen, H. T. Sørensen, E. von Elm and S. M. Langan (2015). "The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement." PLoS Med 12(10): e1001885.
2. Bernardi, F. A., D. Alves, N. Crepaldi, D. B. Yamada, V. C. Lima and R. Rijo (2023). "Data Quality in Health Research: Integrative Literature Review." J Med Internet Res 25: e41446.
3. Carter, E. D., H. H. Leslie, T. Marchant, A. Amouzou and M. K. Munos (2021). "Methodological considerations for linking household and healthcare provider data for estimating effective coverage: a systematic review." BMJ Open 11(8): e045704.
4. de Kraker, M. E. A., M. Abbas, B. Huttner and S. Harbarth (2017). "Good epidemiological practice: a narrative review of appropriate scientific methods to evaluate the impact of antimicrobial stewardship interventions." Clin Microbiol Infect 23(11): 819-825.
5. de Lusignan, S., S. T. Liaw, P. Krause, V. Curcin, M. T. Vicente, G. Michalakidis, L. Agreus, P. Leysen, N. Shaw and K. Mendis (2011). "Key concepts to assess the readiness of data for international research: data quality, lineage and provenance, extraction and processing errors, traceability, and curation. Contribution of the IMIA Primary Health Care Informatics Working Group." Yearb Med Inform 6: 112-120.
6. Duda, S. N., B. E. Shepherd, C. S. Gadd, D. R. Masys and C. C. McGowan (2012). "Measuring the quality of observational study data in an international HIV research network." PLoS One 7(4): e33908.
7. Dusetzina, S. B., S. Tyree, A. M. Meyer, A. Meyer, L. Green and W. R. Carpenter (2014). AHRQ Methods for Effective Health Care. Linking Data for Health Services Research: A Framework and Instructional Guide. Rockville (MD), Agency for Healthcare Research and Quality (US).
8. Field, K., S. Kosmider, J. Johns, H. Farrugia, I. Hastie, M. Croxford, M. Chapman, M. Harold, N. Murigu and P. Gibbs (2010). "Linking data from hospital and cancer registry databases: should this be standard practice?" Intern Med J 40(8): 566-573.
9. Gassman, J. J., W. W. Owen, T. E. Kuntz, J. P. Martin and W. P. Amoroso (1995). "Data quality assurance, monitoring, and reporting." Control Clin Trials 16(2 Suppl): 104s-136s.
10. Gilbert, R., R. Lafferty, G. Hagger-Johnson, K. Harron, L. C. Zhang, P. Smith, C. Dibben and H. Goldstein (2018). "GUILD: GUidance for Information about Linking Data sets." J Public Health (Oxf) 40(1): 191-198.
11. Harron, K., R. Gilbert, D. Cromwell and J. van der Meulen (2016). "Linking Data for Mothers and Babies in De-Identified Electronic Health Data." PLoS One 11(10): e0164667.
12. Hills, R. A., D. Revere, R. Altamore, N. F. Abernethy and W. B. Lober (2012). "Timeliness and data element completeness of immunization data in Washington State in 2010: a comparison of data exchange methods." AMIA Annu Symp Proc 2012: 340-349.
13. Hoffmann, W., U. Latza, S. E. Baumeister, M. Brünger, N. Buttmann-Schweiger, J. Hardt, V. Hoffmann, A. Karch, A. Richter, C. O. Schmidt, I. Schmidtmann, E. Swart and N. van den Berg (2019). "Guidelines and recommendations for ensuring Good Epidemiological Practice (GEP): a guideline developed by the German Society for Epidemiology." Eur J Epidemiol 34(3): 301-317.
14. Hosking, K., G. Stewart, M. Mobsby, S. Skov, Y. Zhao, J. Y. Su, S. Tong, P. Nihill, J. Davis, C. Connors and J. Davies (2020). "Data linkage and computerised algorithmic coding to enhance individual clinical care for Aboriginal people living with chronic hepatitis B in the Northern Territory of Australia - Is it feasible?" PLoS One 15(4): e0232207.
15. Hosseinzadeh, A., A. Karimpour, R. Kluger and R. Orthober (2022). "Data linkage for crash outcome assessment: Linking police-reported crashes, emergency response data, and trauma registry records." J Safety Res 81: 21-35.
16. Huang, Y., Z. Liu, H. Wang, X. Guan, H. Chen, C. Ma, Q. Li, J. Yan, Y. Yu, C. Kou, X. Xu, J. Lu, Z. Wang, L. Liu, Y. Xu, Y. He, T. Li, W. Guo, H. Tian, G. Xu, X. Xu, S. Lv, L. Wang, L. Wang, Y. Yan, B. Wang, S. Xiao, L. Zhou, L. Li and L. Tan (2016). "The China Mental Health Survey (CMHS): I. background, aims and measures." Soc Psychiatry Psychiatr Epidemiol 51(11): 1559-1569.
17. Kapsner, L. A., J. M. Mang, S. Mate, S. A. Seuchter, A. Vengadeswaran, F. Bathelt, N. Deppenwiese, D. Kadioglu, D. Kraska and H. U. Prokosch (2021). "Linking a Consortium-Wide Data Quality Assessment Tool with the MIRACUM Metadata Repository." Appl Clin Inform 12(4): 826-835.
18. Langseth, H., T. Luostarinen, F. Bray and J. Dillner (2010). "Ensuring quality in studies linking cancer registries and biobanks." Acta Oncol 49(3): 368-377.
19. Liu, Z., Y. Huang, P. Lv, T. Zhang, H. Wang, Q. Li, J. Yan, Y. Yu, C. Kou, X. Xu, J. Lu, Z. Wang, H. Qiu, Y. Xu, Y. He, T. Li, W. Guo, H. Tian, G. Xu, X. Xu, Y. Ma, L. Wang, L. Wang, Y. Yan, B. Wang, S. Xiao, L. Zhou, L. Li, L. Tan, H. Chen and C. Ma (2016). "The China Mental Health Survey: II. Design and field procedures." Soc Psychiatry Psychiatr Epidemiol 51(11): 1547-1557.
20. Nind, T., J. Galloway, G. McAllister, D. Scobbie, W. Bonney, C. Hall, L. Tramma, P. Reel, M. Groves, P. Appleby, A. Doney, B. Guthrie and E. Jefferson (2018). "The research data management platform (RDMP): A novel, process driven, open-source tool for the management of longitudinal cohorts of clinical data." Gigascience 7(7).
21. Parsa, N., M. J. Zibaeenezhad, M. Trevisan, A. Karimi Akhormeh and M. Sayadi (2020). "Magnitude of the Quality Assurance, Quality Control, and Testing in the Shiraz Cohort Heart Study." Biomed Res Int 2020: 8179795.
22. Smith, M. and F. Flack (2021). "Data Linkage in Australia: The First 50 Years." Int J Environ Res Public Health 18(21).
23. Stanfill, B. A., E. S. Nakayasu, L. M. Bramer, A. M. Thompson, C. K. Ansong, T. R. Clauss, M. A. Gritsenko, M. E. Monroe, R. J. Moore, D. J. Orton, P. D. Piehowski, A. A. Schepmoes, R. D. Smith, B. M. Webb-Robertson and T. O. Metz (2018). "Quality Control Analysis in Real-time (QC-ART): A Tool for Real-time Quality Control Assessment of Mass Spectrometry-based Proteomics Data." Mol Cell Proteomics 17(9): 1824-1836.
24. Swart, E., H. Gothe, S. Geyer, J. Jaunzeme, B. Maier, T. G. Grobe and P. Ihle (2015). "[Good Practice of Secondary Data Analysis (GPS): guidelines and recommendations]." Gesundheitswesen 77(2): 120-126.
25. Vandenbroucke, J. P., E. von Elm, D. G. Altman, P. C. Gøtzsche, C. D. Mulrow, S. J. Pocock, C. Poole, J. J. Schlesselman and M. Egger (2007). "Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration." Ann Intern Med 147(8): W163-194.
26. von Elm, E., D. G. Altman, M. Egger, S. J. Pocock, P. C. Gøtzsche and J. P. Vandenbroucke (2008). "The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies." J Clin Epidemiol 61(4): 344-349.
27. Weiskopf, N. G. and C. Weng (2013). "Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research." J Am Med Inform Assoc 20(1): 144-151.
28. Whitney, C. W., B. K. Lind and P. W. Wahl (1998). "Quality assurance and quality control in longitudinal studies." Epidemiol Rev 20(1): 71-80.
29. Willey, B., P. Waiswa, D. Kajjo, M. Munos, J. Akuze, E. Allen and T. Marchant (2018). "Linking data sources for measurement of effective coverage in maternal and newborn health: what do we learn from individual- vs ecological-linking methods?" J Glob Health 8(1): 010601.
30. 岳和欣，湛永乐，边峰，张一方，桂路婷，石英杰，孟耀涵，张娟江，江宇. (2021). "临床队列研究的数据标准与共享." 中华流行病学杂志 42(7).
31. 朱建平，付学敬 (2023). "融合数据标准管理和数据血缘建设的数据治理方法." 上海质量(11): 49-55.