肿瘤登记 R语言 canegtools 2025年12月28日

使用R包作为存储《肿瘤登记年报》历史数据的载体

👨‍💻 陈琼博士（河南省癌症中心） 📪 chenq08@126.com

这篇博文介绍了使用R包作为储存数据的载体这篇博文介绍了使用R包作为储存数据的载体这篇博文介绍了

> 编制和出版《肿瘤登记年报》是肿瘤登记工作的内容之一。我们每年会分析和处理全省各登记处提交的肿瘤发病和死亡数据，按照一定的标准对这些数据进行质量控制、评价和筛选，进而汇总分析全省的恶性肿瘤发病死亡情况。

我们已经连续出版《肿瘤登记年报》长达11年。如何高效保存各登记处历年来提交的原始数据，以及已经发布的肿瘤发病和死亡统计指标数据，使得我们能够方便地进行历年指标的回顾性查询，或基于历史原始数据开展新的统计指标计算，这是我们亟需解决的问题。

> 在 R 语言生态中，我们可以将数据打包进 R 包中，在团队内共享和复用。

在开发 canregtools 包时，我们为肿瘤登记的原始数据设计了两种数据结构：canreg 和 canregs。借助 canregtools，原始数据可以被转换为上述两种统一的数据结构，并以 R 包的形式保存与共享。这样不仅能够规范数据格式，还能实现历史数据的高效复用与便捷调用。

为什么使用 R 包作为载体？

使用 R 包作为数据载体，有一些天生的优势，比如可复用性高、版本控制清洗、可内置说明文档、与分析流程无缝衔接、方便地团队协作利用。

1. 可复用性高：数据以包形式存在，调用时只需`library(teamdata)`，直接加载数据集。团队所有人使用的都是同一份数据，避免了各自维护版本。
2. 版本控制清晰：通过 Git 管理 R 包，可以清楚跟踪数据更新历史。每次数据有调整时，可以打 tag 或版本号（如 v1.0.0 → v1.1.0）。
3. 内置文档：R 包的数据集都可以用`?dataset_name`查看帮助，结构、变量说明一目了然，比 Excel 里散落的注释更专业。
4. 与分析流程无缝衔接：数据直接在 R 环境中可用，避免了频繁导入 CSV/Excel，减少手工操作。
5. 团队协作便利：所有成员都从同一仓库安装，保证数据一致性，尤其适合科研和长期项目。

// 缺点

当然，将数据封装在 R 包中也存在一些天然的局限性。首先，团队成员需要具备一定的 R 包构建与使用基础，否则相比直接分发 Excel 文件，门槛可能更高。其次，R 包并不适合存放超大规模的数据集，但若仅用于团队内部共享中小型数

据集，不提交至 CRAN 或 GitHub，这一问题影响并不大。最后，这种方式依赖 R 生态，如果团队成员中有人不使用 R，直接访问和利用 R 包中的数据会相对不便。

// 实现方法

第一步：准备数据，将数据整理成标准化的 data.frame 或 tibble，例如：人口年龄分布表、疾病分类字典。

第二步： 构建 R 包框架

```
usethis::create_package("teamdata")
```

第三步： 将数据保存到包中

```
usethis::use_data(mydata, overwrite = TRUE)
```

第四步：编写文档，在 R/ 文件夹下为每个数据集写 .R 脚本，用 roxygen2 注释：

```
#' Example dataset
#' A dataset of population structure.
#'
#' @format A data frame with 100 rows and 5 variables:
#' \describe{
#'   \item{year}{Year}
#'   \item{age}{Age group}
#'   \item{sex}{Sex}
#'   \item{population}{Population count}
#' }
"population"
```

第五步：共享给团队

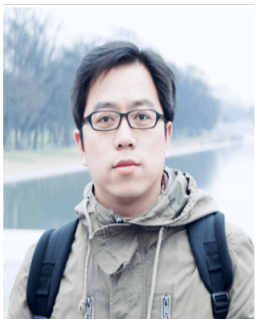
- 可以把包上传到 GitHub，团队成员用 remotes::install_github("org/team-data") 安装。
- 如果团队有私有 GitLab / GitHub，可以作为内部包管理工具。

- 甚至可以投递到 CRAN（如果是公开通用数据）

为什么使用 R

// 缺点缺点

作者简介



陈 琼

博士，副主任医师，就职河南省癌症中心/河南省肿瘤医院，从事肿瘤登记和描述流行病相关工作和研究，担任《河南省肿瘤登记年报》副主编。创建网站（https://www.chenq.site）及微信公众号【普癌新声】，分享肿瘤登记、数据分析、R 语言编程、可视化技巧与可重复性报告解决方案 🚀。

Qsight 博客

微信公众号

扫描二维码，关注公众号，获取更多精彩内容！

💬 欢迎公众号留言交流，期待您的意见和建议！

请点亮 ❤️，点赞 & 分享，一起传播有价值的内容！

P.S. 本文版权归属@陈琼博士所有，如需转载或转发，应以完整 PDF 全文形式进行，不得截取、删改或仅传播部分内容。

使用R包作为存储《肿瘤登记年报》历史数据的载体

肿瘤登记 R语言 canegtools

2025年12月28日