# An R Package for Acquiring and Processing Notifiable Infectious Diseases Dataset from the Japan Institute for Health Security

Tomonori Hoshi [1,2,¶], **Erina Ishigaki** [1,3,4], and **Satoshi Kaneko** [1,2,4,5]

1 Institute of Tropical Medicine, Nagasaki University, Japan **2** School of Tropical Medicine and Global Health, Nagasaki University, Japan **3** London School of Hygiene & Tropical Medicine, London, UK **4** Graduate School of Biomedical Sciences, Nagasaki University, Japan **5** DEJIMA Infectious Disease Research Alliance, Nagasaki University, Japan ¶ Corresponding author

Summary

`jpinfect` is an R (R Core Team, 2025) package that provides a set of functions to acquire and process notifiable infectious disease datasets from the Japan Institute for Health Security (Japan Institute for Health Security, 2025). The package helps generate combined datasets of weekly case reports since week 14 of 1999 by prefecture and, where available, sex and suspected location of infection information. In addition to its core functionalities, the package also includes pre-processed built-in datasets. These datasets are ready for immediate analysis, making it easier to utilise officially released public data. The package is designed to streamline epidemiological research, enhance public health response and support educational efforts. Ultimately, `jpinfect` aims to assist researchers and practitioners in responding efficiently to notifiable infectious diseases in Japan. All code is archived on GitHub for extension and adaptation.

Statement of need

The COVID-19 pandemic highlighted the importance of reliable epidemiological dataset for outbreak monitoring and effective disease control (Hoseinpour Dehkordi et al., 2020). Thanks to computer advancements, mathematical modelling methods based on the large dataset could be implemented in many countries for the COVID-19 response (Ferguson et al., 2020; Vynnycky & White, 2011). This experience emphasised that timely and accurate information is crucial for successful disease control (Gulland, 2020). Consequently, public health officers and epidemiologists at government organisations paid particular attention to ensuring that all epidemiological recommendations were grounded in reliable data sources (World Health Organization, 2020).

In real-world settings, officially released datasets from government organisations, while comprehensive, are the most reliable source. However, these are often difficult to process for immediate use due to their complex structure (Vetrò et al., 2016). This barrier hindered rapid response for containing the outbreak (Stoto et al., 2022). Preparing, cleaning and standardising these datasets during non-emergency periods is essential for improving responsiveness in the event of future outbreaks.

The `jpinfect` package was specifically designed to address these challenges by providing a streamlined workflow for acquiring and processing notifiable infectious disease datasets from the Japan Institute for Health Security (Japan Institute for Health Security, 2025). By automating data retrieval, standardisation, and integration into analysable formats, the package reduces

barriers for researchers, public health practitioners and educators, enabling them to focus on actionable insights.

Statement of the field

Epidemiological research and infectious disease modelling have become increasingly critical tools for public health interventions and policy-making (Ferguson et al., 2020). The COVID-19 pandemic demonstrated the global dependency on accurate and timely epidemiological data for forecasting infection trends, evaluating interventions and informing public health strategies. R packages such as epidm and ukbtools prepare health data in the UK (Bhattacharya, 2022; Hanscombe, 2019), while in the United States, epidatr, cdcfluview and covidcast streamline retrieval of epidemiological surveillance datasets via APIs (Arnold et al., 2025; Brooks et al., 2025; Rudis et al., 2022).

In Japan, the government has maintained a national infectious disease monitoring network since week 14 of 1999. However, the raw data are scattered across multiple webpages due to Ministry reorganisations over the past decades. As a result, acquiring the appropriate datasets can be challenging, especially in time-sensitive settings (Guo et al., 2023). No software currently exists to facilitate the raw data acquisition, and the data are often in inconsistent formats, requiring significant preprocessing before analysis. This complexity hinders users' productivity. Notably, these websites lack full English translation and a user-friendly format for non-Japanese speakers. Retrieving and preparing these datasets manually is labour-intensive. It requires downloading and merging 52–53 weekly Excel sheets for each year from 1999 to 2023 for confirmed case data and individual weekly CSV files for provisional reports. These files are scattered across multiple webpages in varying formats, requiring extensive manual preprocessing.

The `jpinfect` package fills a critical gap in the field by offering tools to automate and simplify the preparation of these datasets. By standardising formats, merging sources, and providing built-in datasets, `jpinfect` supports researchers, public health practitioners and educators.

Pipeline Overview

The `jpinfect` package provides an efficient and streamlined pipeline for acquiring and processing infectious disease datasets from the Japan Institute for Health Security. The pipeline consists of the following steps:

1. **Data Acquisition**: Raw data can be acquired using the following functions:
   - `jpinfect_get_confirmed`: Downloads confirmed case reports by year and data type (e.g., sex or place of infection).
   - `jpinfect_get_bullet`: Downloads weekly provisional case reports.

   The downloaded data is stored locally and organised for further processing.

2. **Data Import**: The acquired data can be read into R using:
   - `jpinfect_read_confirmed`: Imports confirmed case reports from local files or directories.
   - `jpinfect_read_bullet`: Imports provisional weekly reports from local directories.

3. **Built-in Datasets**: For immediate analysis, the package provides pre-processed datasets:
   - `sex_prefecture`: Weekly confirmed cases by sex and prefecture.
   - `place_prefecture`: Weekly confirmed cases by place of infection and prefecture.
   - `bullet`: Provisional weekly case reports.

4. **Data Merging**: The imported or built-in datasets can be merged using:

- `jpinfect_merge`: Combines multiple datasets into a single dataset for comprehensive analysis.

5. **Data Transformation**: The merged data can be converted between wide and long formats using:

- `jpinfect_pivot`: Enables users to adjust the data format to suit specific analytical workflows.

6. **Analysis**: The processed data is ready for various epidemiological analyses, such as outbreak monitoring, modelling and reporting.
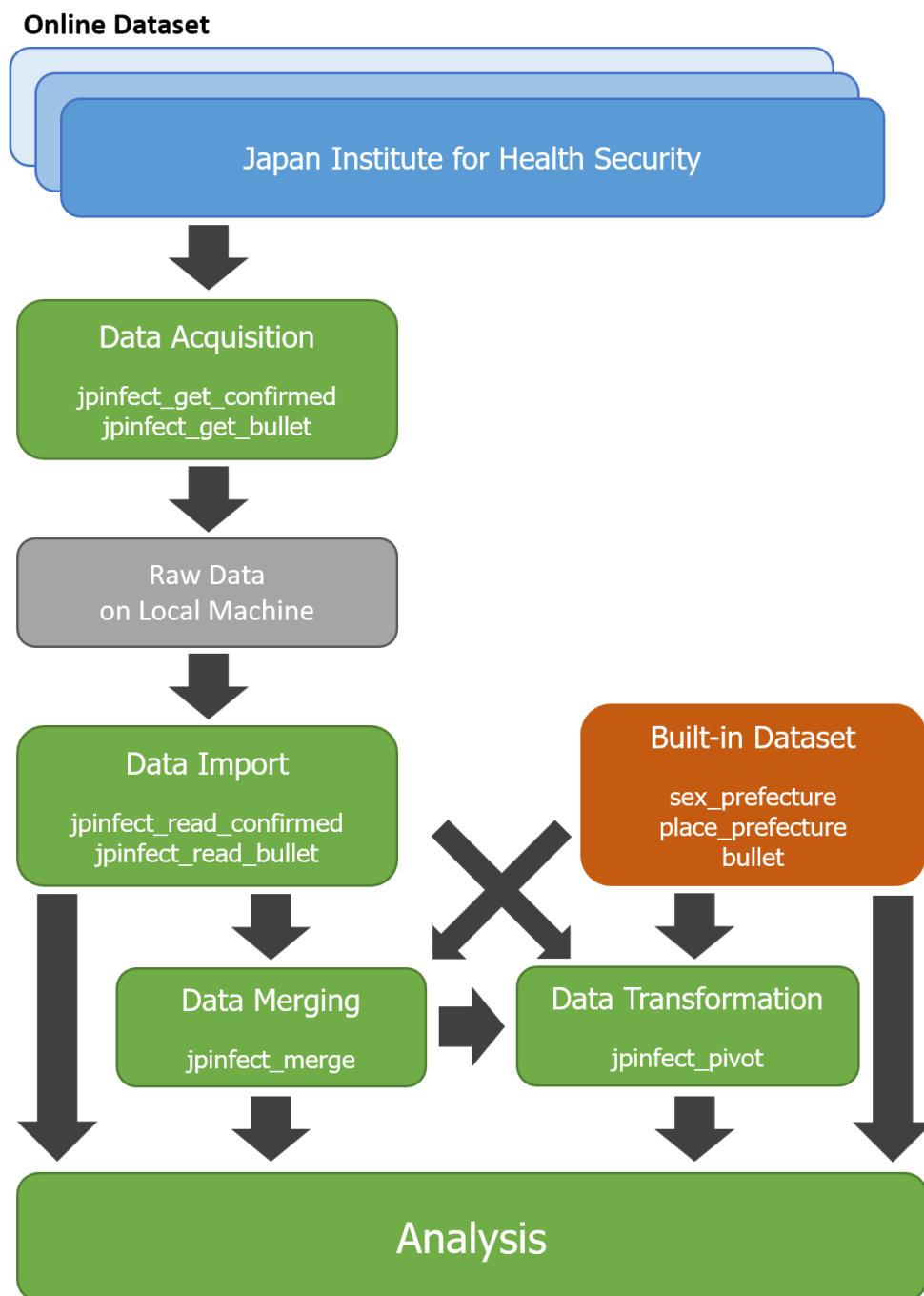
During package development, we used GitHub Copilot within RStudio to assist with coding and Microsoft Copilot to support both coding and debugging.

Acknowledgements

References

Arnold, T., Bien, J., Brooks, L., Colquhoun, S., Farrow, D., Grabman, J., Maynard-Zhang, P., Mazaitis, K., Reinhart, A., & Tibshirani, R. (2025). *Covidcast: Client for delphi's 'COVIDcast epidata' API*. https://doi.org/10.32614/CRAN.package.covidcast

Bhattacharya, A. (2022). *Epidm: UK epidemiological data management*. https://doi.org/10.32614/CRAN.package.epidm

Brooks, L., Shemetov, D., Gratzl, S., Weber, D., DeFries, N., Reinhart, A., McDonald, D. J., Tan, K. M., Townes, W., Haff, G., & Mazaitis, K. (2025). *Epidatr: Client for delphi's 'epidata' API*. https://doi.org/10.32614/CRAN.package.epidatr

Ferguson, N., Laydon, D., Nedjati Gilani, G., Imai, N., Ainslie, K., Baguelin, M., Bhatia, S., Boonyasiri, A., Cucunuba Perez, Z., Cuomo-Dannenburg, G., Dighe, A., Dorigatti, I., Fu, H., Gaythorpe, K., Green, W., Hamlet, A., Hinsley, W., Okell, L., Van Elsland, S., … Ghani, A. (2020). *Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand*. Imperial College London. https://doi.org/10.25561/77482

Gulland, A. (2020). *From maths to maps: The modellers tackling COVID-19*. London School of Hygiene & Tropical Medicine. https://www.lshtm.ac.uk/research/research-action/features/maths-maps-modellers-tackling-covid-19

Guo, M., Wang, Y., Yang, Q., Li, R., Zhao, Y., Li, C., Zhu, M., Cui, Y., Jiang, X., Sheng, S., Li, Q., & Gao, R. (2023). Normal Workflow and Key Strategies for Data Cleaning Toward Real-World Data: Viewpoint. *Interactive Journal of Medical Research*, *12*, e44310. https://doi.org/10.2196/44310

Hanscombe, J. R. I. A. T., Ken B. AND Coleman. (2019). ukbtools: An R package to manage and query UK Biobank data. *PLOS ONE*, *14*(5), 1–6. https://doi.org/10.1371/journal.pone.0214311

Hoseinpour Dehkordi, A., Alizadeh, M., Derakhshan, P., Babazadeh, P., & Jahandideh, A. (2020). Understanding epidemic data and statistics: A case study of COVID-19. *Journal of Medical Virology*, *92*(7), 868–882. https://doi.org/10.1002/jmv.25885

Japan Institute for Health Security. (2025). *Japan Institute for Health Security | The Infectious Disease Information Website*. Japan Institute for Health Security. https://id-info.jihs.go.jp/en/index.html

R Core Team. (2025). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://doi.org/10.32614/r.manuals

Rudis, B., McGowan, C., Chen, J., Meyer, S., Turtle, J., Bates, A., & McGovern, I. (2022). *Cdcfluview: Retrieve flu season data from the united states centers for disease control and prevention ('CDC') 'FluView' portal*. https://github.com/hrbrmstr/cdcfluview

Stoto, M. A., Woolverton, A., Kraemer, J., Barlow, P., & Clarke, M. (2022). COVID-19 data are messy: Analytic methods for rigorous impact analyses with imperfect data. *Globalization and Health*, *18*(1), 2. https://doi.org/10.1186/s12992-021-00795-0

Vetrò, A., Canova, L., Torchiano, M., Minotas, C. O., Iemma, R., & Morando, F. (2016). Open data quality measurement framework: Definition and application to open government data. *Government Information Quarterly*, *33*(2), 325–337. https://doi.org/10.1016/j.giq.2016.02.001

Vynnycky, E., & White, R. G. (2011). *An introduction to infectious disease modelling* (1st ed.). Oxford Univ. Press. ISBN: 978-0-19-856576-5

World Health Organization. (2020). *World health organization data principles*. World Health Organization. https://www.who.int/data/principles