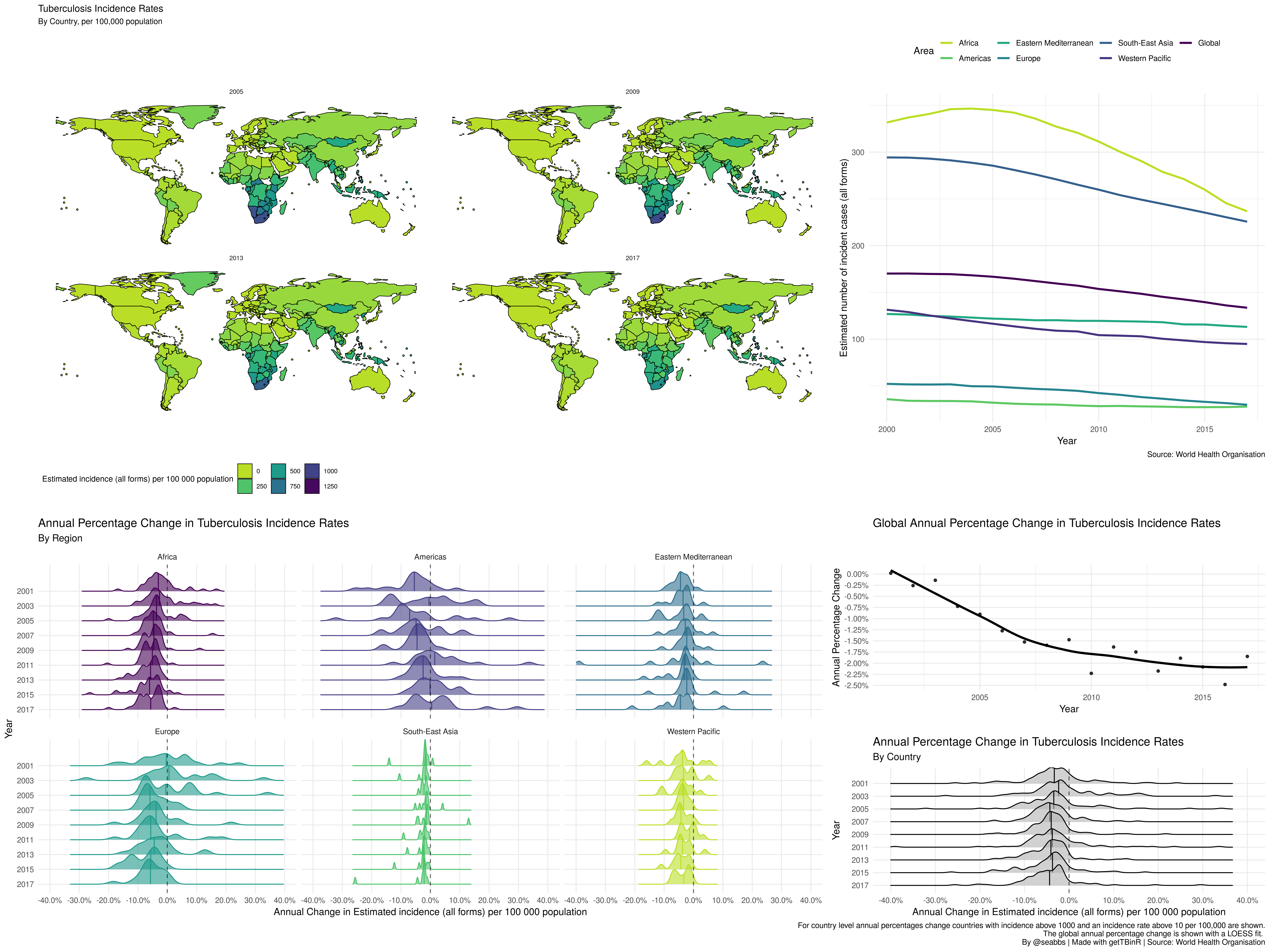
Making global tuberculosis data more accessible: a new R package



Global and regional tuberculosis (TB) incidence rates, as well as trends in the annual percentage change of TB incidence rates. Generated using getTBinR.

## The problem to solve

[Tuberculosis (TB)](https://www.who.int/tb) is one of the oldest human diseases, with recorded cases in ancient Egypt, renaissance Europe, and in the modern day across the globe. It is thought to infect over 1.7 billion people globally, of which 5-15% will develop symptomatic TB in their lifetime (World Health Organisation 2018). Of this number around 10% are likely to die from TB or TB related causes and globally TB remains the leading cause of death from infectious disease. TB is preventable and curable, but the majority of cases occur in less economically developed countries and are never diagnosed. Whilst global incidence rates have decreased year on year since the early 2000’s, global TB incidence remains above 134 per 100,000 people.

The World Health Organization (WHO) makes country-level TB data publicly available in an annual report. These data can be downloaded from the [WHO website](https://www.who.int/tb/country/data/download/en/) in a series of raw .csv (spreadsheet) data files and are a fantastic tool for understanding global trends in TB. Unfortunately, exploring the data requires serveral repetitive steps that need to be repeated for each .csv file. Once these have been completed more work is needed to generate usable summary statistics, data visualisations and maps. This overhead may be too much for many users.

## The package: getTBinR

[getTBinR](https://doi.org/10.21105/joss.01260) (Abbott 2019) is an open-source [R package](.%20https://www.R-project.org/) (R Core Team 2019) that facilitates working with data collected by the [WHO](https://www.who.int/tb/country/data/download/en/) (WHO) on the epidemiology of tuberculosis (TB). All data is freely available from the [WHO](https://www.who.int/tb/country/data/download/en/). The aim of getTBinR is to allow researchers, analysts, and other interested individuals, to quickly and easily gain access to a detailed TB data set and to start using it to derive key insights. It provides a consistent set of tools that can be used to rapidly evaluate hypotheses on a widely used data set before they are explored further using more complex methods or more detailed data.

The functions provided in the package were developed to have sensible defaults to allow those new to the field too quickly gain key insights but also allow sufficient customisation so that experienced users can rapidly prototype new ideas. The package code is archived on [Zenodo](https://doi.org/10.5281/zenodo.2547405) (Abbott 2019) and [Github](https://www.samabbott.co.uk/getTBinR/).

## What data is available?

The data sourced by getBTinR is collected by the WHO from national governments and used to compile the [yearly global TB report](https://www.who.int/tb/country/data/download/en/) (World Health Organisation 2018). These data include TB notifications, TB mortality rates, the proportion of drug resistant cases, case detection rates, treatment rates, interventions, outcomes, and planned intervention budgets.. The [WHO](https://www.who.int/tb/country/data/download/en/) (World Health Organisation 2018) has a complete description of the data and data collection process. These data are used by the WHO, governmental organisations and researchers to summarise country level TB epidemiology, as well as the wider global and regional picture.

## What does the package do?

The package automates downloading, cleaning, and joining the various TB related datasets compiled by the WHO. On top of this, it supplies a data dictionary search tool that makes exploring the data easier. For experianced R users it provides a set of functions to be used within an R session. For those new to R, or new to analyzing public health data, it also provides an [interactive dashboard](https://www.samabbott.co.uk/getTBinR/reference/run_tb_dashboard.html) an [automated country level](https://www.samabbott.co.uk/getTBinR/reference/render_country_report.html) report.

getTBinR provides a range of mapping and visualisation tools. The figures at the top of this post were produced using a small subset of that functionality and are a good example of the kind of information that the package helps to extract. They show the global and regional trends in TB incidence rates and the annual change of those rates. They highlight the fact that TB reductions are no longer increasing annually and that regional trends differ (see [here](https://www.samabbott.co.uk/post/gettbinr-6-0/) for the code to generate these figures).

Another example of a use of the package, to explore estimates of the TB case fatality ratio, can be seen [here](https://www.samabbott.co.uk/post/est-cfr-gettbinr/). See [here](https://www.samabbott.co.uk/getTBinR/) for further examples and the package documentation.

## Why is this package needed?

Developing tools for accessing and exploring data sets benefits the public health research community by enabling multiple data sets to be combined in a consistent manner, increasing their visibility, and providing a framework for exploring key questions. Tooling also reduces barriers to entry, allowing non-specialists to explore data sets that would otherwise be inaccessible. This widening of access may lead to new insights and wider interest for key public health issues.

## References

Abbott, Sam. 2019. “GetTBinR: An R Package for Accessing and Summarising the World Health Organisation Tuberculosis Data.” *Journal of Open Source Software* 4 (34): 1260. <https://doi.org/10.21105/joss.01260>.

Abbott, Sam. 2019. “Seabbs/getTBinR 0.6.0.” *Zenodo*. <https://doi.org/10.5281/zenodo.2547405>.

R Core Team. 2019. “R: A Language and Environment for Statistical Computing.” Vienna, Austria. <https://www.R-project.org/>.

World Health Organisation. 2018. “Global Tuberculosis Report.” <https://www.who.int/tb/country/data/download/en/>.