Reading in and manipulating the list of bacterial human pathogens from Bartlett et al. 2022 in R

Daniel Padfield

03/07/2022

Dataset of bacterial human pathogens

We present a comprehensive list of known bacterial pathogens known to cause infectious symptoms in humans. A summary of all the datasets used and produced are summarised in this document.

The dataset is available on GitHub, but to allow the code to fit cleanly in these walkthroughs we created a shortened URL for the file (https://shorturl.at/hiwy7)

Load in packages

First we will load in the R packages used in the script.

```
# load packages
library(openxlsx)
library(tidyverse)
library(janitor)
```

1. Taylor *et al.*

Taylor *et al.* reported 538 bacterial pathogens in 2001. We took this list as a starting point, and added year of description for each species, removed not validly described species and checked for name changes using the 'List of Prokaryotic names with Standing in Nomenclature' from LPSN. After checking, there were 528 human pathogens that met our definition of a pathogen.

After initial cleaning (see code below), the column names in this spreadsheet are:

- genus: genus of the pathogen
- species: species name of the pathogen
- year: year the pathogen was first described
- status: whether the pathogen is Established or Putative. All pathogens in this list are classified as Established
- original_description_and_or_relevant_clinical_description: reference of where the pathogen was described
- old_name: the old name if it has now been reclassified
- source: source of the data. Here is `taylor`

```
# set url
url <- "https://shorturl.at/hiwy7"

# read in Taylor et al dataset from GitHub
d_taylor <- rio::import(url, sheet = "1 Taylor et al", range = "A1:F529") %>%
    clean_names() %>%
    mutate(source = "taylor") %>%
    rename(original_description_and_or_relevant_clinical_description = x5,
```

```
        status = 4)

# show table without reference column
head(select(d_taylor, -original_description_and_or_relevant_clinical_description))
```

```
##             genus       species year      status old_name source
## 1     Abiotrophia     defectiva 1989 Established     <NA> taylor
## 2   Achromobacter    piechaudii 1986 Established     <NA> taylor
## 3   Achromobacter  xylosoxidans 1971 Established     <NA> taylor
## 4 Acidaminococcus    fermentans 1969 Established     <NA> taylor
## 5   Acinetobacter     baumannii 1986 Established     <NA> taylor
## 6   Acinetobacter calcoaceticus 1911 Established     <NA> taylor
```

2. Munson & Carroll

Three papers published by Munson and Carroll compiled new bacterial species associated with humans described in the years 2012 to 2019, from which all species meeting our definitions were taken. From these papers, 85 species met our definitions of a bacterial human pathogen.

After initial cleaning (see code below), the column names in this spreadsheet are:

- genus: genus of the pathogen
- species: species name of the pathogen
- year: year the pathogen was first described
- status: whether the pathogen is Established or Putative
- original_description_and_or_relevant_clinical_description: reference of where the pathogen was described
- source: source of the data. Here is `munson`

```
# read in Munson & Carroll dataset from GitHub
d_munson <- rio::import(url, sheet = "2 Munson and Carroll", range = "A1:E86") %>%
    janitor::clean_names() %>%
    mutate(source = "munson")

head(select(d_munson, -original_description_and_or_relevant_clinical_description))
```

```
##           genus       species year      status source
## 1 Streptococcus     tigurinus 2012 Established munson
## 2 Streptococcus hongkongensis 2012    Putative munson
## 3   Haemophilus      sputorum 2012 Established munson
## 4 Psychrobacter     sanguinis 2012 Established munson
## 5    Legionella nagasakiensis 2012    Putative munson
## 6      Massilia         oculi 2012    Putative munson
```

3. IJSEM

We screened new species published in the International Journal of Systematic and Evolutionary Microbiology and the bimonthly published series "List of new names and new combinations previously effectively, but not validly, published" in the same journal which lists species described in other journals. We used this approach to find newly described pathogen species in the period 1997-2011 and 2020. Using this approach, 346 species met our definitions of a bacterial human pathogen.

After initial cleaning (see code below), the column names in this spreadsheet are:

- genus: genus of the pathogen
- species: species name of the pathogen
- year: year the pathogen was first described

- status: whether the pathogen is Established or Putative
- original_description_and_or_relevant_clinical_description: reference of where the pathogen was described
- old_name: the old name if it has now been reclassified
- source: source of the data. Here is `ijsem`

```
# read in ijsem dataset from GitHub
d_ijsem <- rio::import(url, sheet = "3 IJSEM", range = "A1:F347") %>%
    janitor::clean_names() %>%
    mutate(source = "ijsem")

head(select(d_ijsem, -original_description_and_or_relevant_clinical_description))
```

```
##           genus       species year      status old_name source
## 1   Acetobacter indonesiensis 2001 Established     <NA>  ijsem
## 2   Acetobacter cibinongensis 2002    Putative     <NA>  ijsem
## 3 Achromobacter    insolitus  2003    Putative     <NA>  ijsem
## 4 Achromobacter       spanius 2003    Putative     <NA>  ijsem
## 5    Acidovorax        oryzae 2009    Putative     <NA>  ijsem
## 6 Acinetobacter  beijerinckii 2009 Established     <NA>  ijsem
```

4. Google Scholar

We performed ad hoc searches using Google Scholar, limiting our searches to peer-reviewed literature in the English language. We did not use mentions of human infection without a primary reference. Using this approach, 260 species met our definitions of a bacterial human pathogen.

After initial cleaning (see code below), the column names in this spreadsheet are:

- genus: genus of the pathogen
- species: species name of the pathogen
- year: year the pathogen was first described
- status: whether the pathogen is Established or Putative
- original_description_and_or_relevant_clinical_description: reference of where the pathogen was described
- source: source of the data. Here is `scholar`

```
# read in google scholar dataset from GitHub
d_scholar <- rio::import(url, sheet = "4 Google Scholar", range = "A1:E261") %>%
    janitor::clean_names() %>%
    mutate(source = "google_scholar")

head(select(d_scholar, -original_description_and_or_relevant_clinical_description))
```

```
##           genus     species year      status         source
## 1   Abiotrophia     elegans 1998    Putative google_scholar
## 2 Achromobacter  mucicolens 2013 Established google_scholar
## 3 Achromobacter    pulmonis 2013 Established google_scholar
## 4 Achromobacter  spiritinus 2013    Putative google_scholar
## 5    Acidomonas methanolica 1989    Putative google_scholar
## 6 Acinetobacter      baylyi 2006 Established google_scholar
```

5. Shaw *et al.*

The pathogen species identified from approaches 1-4 were compared with the list by Shaw *et al.* resulting in an additional 409 species of which 294 met our criteria

After initial cleaning (see code below), the column names in this spreadsheet are:

- genus: genus of the pathogen
- species: species name of the pathogen
- year: year the pathogen was first described
- status: whether the pathogen is Established or Putative
- original_description_and_or_relevant_clinical_description: reference of where the pathogen was described
- source: source of the data. Here is `shaw`

```
# read in google scholar dataset from GitHub
d_shaw <- rio::import(url, sheet = "5 Shaw et al", range = "A1:E295") %>%
    janitor::clean_names() %>%
    mutate(source = "shaw")

head(select(d_shaw, -original_description_and_or_relevant_clinical_description))
```

```
##           genus       species year      status source
## 1  Acholeplasma         oculi 1973    Putative   shaw
## 2 Achromobacter  aegrifaciens 2013 Established   shaw
## 3 Achromobacter       anxifer 2013    Putative   shaw
## 4 Achromobacter denitrificans 1983 Established   shaw
## 5 Achromobacter        dolens 2013 Established   shaw
## 6 Achromobacter      insuavis 2013 Established   shaw
```

This process makes up the whole list, and we have shown how each list can be read into R.

We can bind them all together easily.

```
d_all <- bind_rows(d_taylor, d_munson, d_ijsem, d_scholar, d_shaw)
nrow(d_all)
```

```
## [1] 1513
```

Alternatively we can just read in the sheet with the complete list which also includes the higher taxonomy for each pathogen (derived using taxize)

Overall we found 1513 species that fit our definition of a human pathogen.

After initial cleaning (see code below), the column names in this spreadsheet are:

- superkingdom: kingdom of the pathogen
- phylum: phylum of the pathogen
- class: class of the pathofen
- order: order of the pathogen
- family: family of the pathogen
- genus: genus of the pathogen
- species: species name of the pathogen
- year: year the pathogen was first described
- status: whether the pathogen is Established or Putative
- reference: reference of where the pathogen was described

```
d_all <- rio::import(url, sheet = "Full List", range = "A1:J1514") %>%
    janitor::clean_names()

select(d_all, -reference) %>%
    head()
```

```
##   superkingdom         phylum               class           order
## 1     Bacteria Proteobacteria Gammaproteobacteria Enterobacterales
## 2     Bacteria     Firmicutes             Bacilli       Bacillales
```

```
## 3     Bacteria     Firmicutes          Clostridia    Eubacteriales
## 4     Bacteria   Fusobacteria       Fusobacteriia  Fusobacteriales
## 5     Bacteria Proteobacteria Gammaproteobacteria      Vibrionales
## 6     Bacteria     Firmicutes             Bacilli       Bacillales
##             family        genus    species year      status
## 1     Yersiniaceae     Serratia marcescens 1823 established
## 2      Bacillaceae     Bacillus   subtilis 1835 established
## 3   Clostridiaceae  Clostridium ventriculi 1842 established
## 4 Leptotrichiaceae Leptotrichia   buccalis 1853 established
## 5     Vibrionaceae       Vibrio   cholerae 1854 established
## 6      Bacillaceae     Bacillus  anthracis 1872 established
```

References

- Taylor LH, Latham SM, Woolhouse ME. Risk factors for human disease emergence. Philosophical Transactions of the Royal Society of London Series B: Biological Sciences. 2001;356(1411):983-9.
- Munson E, Carroll KC. What's in a name? New bacterial species and changes to taxonomic status from 2012 through 2015. Journal of clinical microbiology. 2017;55(1):24-42.
- Munson E, Carroll KC. An update on the novel genera and species and revised taxonomic status of bacterial organisms described in 2016 and 2017. Journal of clinical microbiology. 2019;57(2):e01181-18.
- Munson E, Carroll KC. Summary of novel bacterial isolates derived from human clinical specimens and nomenclature revisions published in 2018 and 2019. Journal of clinical microbiology. 2020;59(2):e01309-20.
- Shaw LP, Wang AD, Dylus D, Meier M, Pogacnik G, Dessimoz C, et al. The phylogenetic range of bacterial and viral pathogens of vertebrates. Mol Ecol. 2020;29(17):3361-79.