# Metadata concerning the diseases data stored under the directory “Human CSI”, where CSI means “Climate Sensitive Infections”.

Thierfelder T.[1], Berggren C.[1], Omazic A.[2], Evengård B.[3], 2019-01-25

*1: Dept. of Energy & Technology, Swedish University of Agricultural Sciences, Uppsala, Sweden*
*2: Dept. of Chemistry, Environment, and Feed Safety, National Veterinary Institute, Uppsala, Sweden*
*3: Dept. of Clinical Microbiology, Umeå University, Umeå, Sweden*

## Contact person and citations

For any questions concerning the published data, please make contact with Tomas Thierfelder at Tomas.Thierfelder@slu.se.

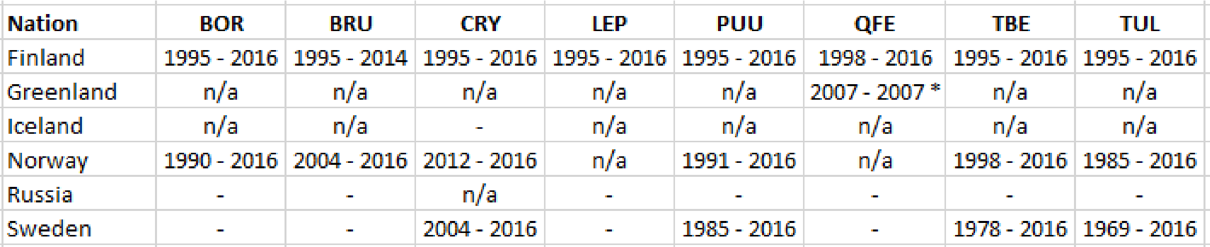
Kindly refer to the title, date, and list of authors provided above in all publications that utilise CLINF human diseases data.

## Diseases

With most diseases addressed by CLINF being zoonotic, meaning that they may occur in humans as well as in (other) animals, the term “Human CSI” indicates that the corresponding diseases have been observed in humans and reported via human diseases report systems. In the CLINF repository, human CSI data are stored per disease, where the respective diseases data are stored in separate directories that have been given abbreviated names of the associated diseases. The names of directories and diseases are:

BOR = Borreliosis, BRU = Brucellosis, CRY = Cryptosporidiosis, LEP = Leptospirosis, PUU = Haemorrhagic fever with renal syndrome (Puumala virus infection), QFE = Q fever, TBE = Tick-borne encephalitis, TUL = Tularaemia.

With human diseases data being reported case-by-case, individually reported cases may have been supplemented with information regarding gender and age of the patient depending on routines that vary with nation, disease, and period of time, as illustrated in *Table 1*.



| Nation | BOR | BRU | CRY | LEP | PUU | QFE | TBE | TUL |
|---|---|---|---|---|---|---|---|---|
| Finland | 1995 - 2016 | 1995 - 2014 | 1995 - 2016 | 1995 - 2016 | 1995 - 2016 | 1998 - 2016 | 1995 - 2016 | 1995 - 2016 |
| Greenland | n/a | n/a | n/a | n/a | n/a | 2007 - 2007 * | n/a | n/a |
| Iceland | n/a | n/a | - | n/a | n/a | n/a | n/a | n/a |
| Norway | 1990 - 2016 | 2004 - 2016 | 2012 - 2016 | n/a | 1991 - 2016 | n/a | 1998 - 2016 | 1985 - 2016 |
| Russia | - | - | n/a | - | - | - | - | - |
| Sweden | - | - | 2004 - 2016 | - | 1985 - 2016 | - | 1978 - 2016 | 1969 - 2016 |

*Table 1: Coverage of supplementary information concerning gender and age per nation and disease. Where not applicable (n/a), the diseases have not been reported. A bar (-) annotates the lack of supplementary information despite reported diseases. * = A single case of QFE reported in Greenland 2007.*

Human diseases data are either reported clinically or via laboratories, in written form or via digital report systems. This introduces a potential of overlay error since single cases of diseases may be reported twice, and since older written report systems may overlap with the implementation of digital dittos. CLINF has spent much effort into reducing such sources of error to their minimum.

## Spatial resolution

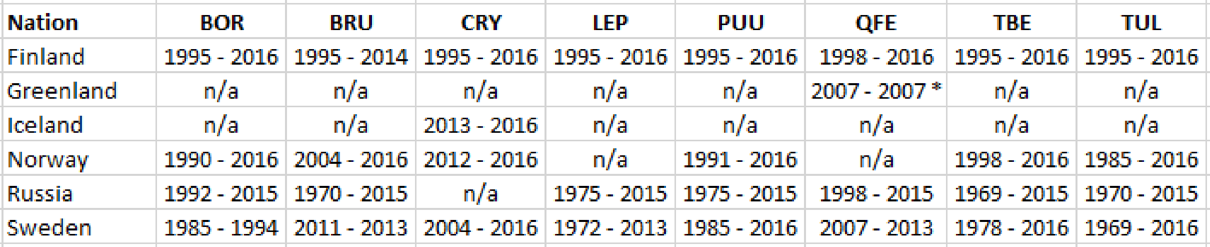
The data covering human diseases are constituted by empirical observations concerning individual cases reported to authorities in the six nations of:

GRE = Greenland, ICE = Iceland, NOR = Norway, SWE = Sweden, FIN = Finland, RUS = Russia.

The national administration of reported diseases is managed per report district, where the size of a typical report district approximately equals the size of counties everywhere except in Russia, where diseases mainly are reported per oblast, republic, or autonomous region. Hence, the smallest possible spatial resolution of CLINF Human CSI data is report district – see *Appendix A* for a list of CLINF report districts. In Greenland and Iceland, the entire nations constitute one report district, although confined to their respective coastal regions. In Greenland, human diseases are predominantly reported in the southern and western coastal regions.

## Temporal resolution

In the temporal domain, an attempt has been made to cover the 30-year climate reference period with diseases data. This ambition has been more or less successful depending on national differences regarding the inclusion of different diseases in their respective lists of reportable diseases (diseases that should be reported by law). As a consequence, the time-period covered by CLINF human CSI data varies with diseases and nation, in accordance with *Table 2*.



| Nation | BOR | BRU | CRY | LEP | PUU | QFE | TBE | TUL |
|---|---|---|---|---|---|---|---|---|
| Finland | 1995 - 2016 | 1995 - 2014 | 1995 - 2016 | 1995 - 2016 | 1995 - 2016 | 1998 - 2016 | 1995 - 2016 | 1995 - 2016 |
| Greenland | n/a | n/a | n/a | n/a | n/a | 2007 - 2007 * | n/a | n/a |
| Iceland | n/a | n/a | 2013 - 2016 | n/a | n/a | n/a | n/a | n/a |
| Norway | 1990 - 2016 | 2004 - 2016 | 2012 - 2016 | n/a | 1991 - 2016 | n/a | 1998 - 2016 | 1985 - 2016 |
| Russia | 1992 - 2015 | 1970 - 2015 | n/a | 1975 - 2015 | 1975 - 2015 | 1998 - 2015 | 1969 - 2015 | 1970 - 2015 |
| Sweden | 1985 - 1994 | 2011 - 2013 | 2004 - 2016 | 1972 - 2013 | 1985 - 2016 | 2007 - 2013 | 1978 - 2016 | 1969 - 2016 |

*Table 2: Temporal coverage of CLINF human CSI data per nation and disease. Where not applicable (n/a), the diseases have not been reported. * = A single case of QFE reported in Greenland 2007.*

With diseases reported case-by-case, the exact reporting dates of every single case are available in the CLINF diseases database, although not publicly disseminated. This temporal case-by-case resolution may be scaled up into any other temporal resolution such as monthly or annual cumulations of cases. The CLINF temporal standard resolution is *the number of annually cumulated cases per 100,000 report-district inhabitants*, which we call "diseases incidence". Such incidences are what CLINF disseminates via its standard repository of human diseases data.

## Data disposition

In any of the disease's directories, four different files are to be found: XXX_inci_ann_tot, XXX_inci_list_all, XXX_map_inci_prim, and XXX_map_inci_ord, where XXX is the abbreviated name of the disease.

### XXX_inci_ann_tot

In XXX_inci_ann_tot, the annual incidences of the XXX disease are stored per nation and report district, where Distr_code is the official abbreviated identity of report districts. These data are dispositioned with one column per year, where Ave_inci contains the average annual incidence across the observed period of time, and where Bin_inci contains a binary code which indicates whether or not the XXX disease is ever reported in the respective report districts (trough the observed period of time).

### XXX_inci_list_all

In XXX_inci_list_all, information regarding gender and age is included in the data table, and data has been dispositioned into one factor per column (which is ideal for formal statistical inference). The following factors/variables are included in the CLINF standard disposition of human diseases data:

CSI = [BOR, BRU, CRY, LEP, PUU, QFE, TBE, TUL]

Nation = [GRE, ICE, NOR, SWE, FIN, RUS]

District = Diseases report districts, see *Appendix A*.

Distr_code = Official abbreviated identities of diseases report district, see *Appendix A*.

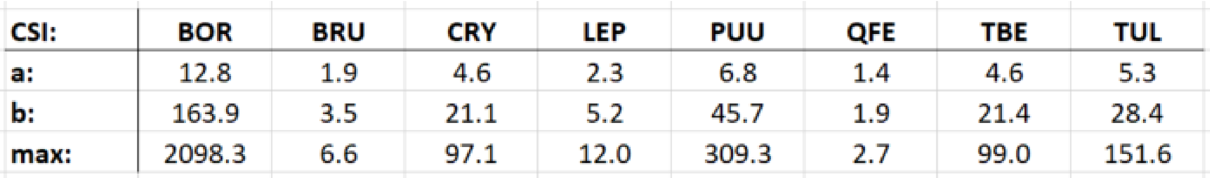
Year = Year of annually cumulated diseases incidences.

Pop_count = Number of inhabitants per diseases report district and year.

Inc_prim = Primary diseases incidences per report district and year. With incidences calculated as the number of annually cumulated cases multiplied with factor 100,000 and subdivided with Pop_count.

Inc_nonzero = Inc_prim with zero incidences removed. There are several motivations for the removal of zero incidences, with one simply reflecting the general interest of focusing on the geographic regions (report districts) where diseases actually occur. Another reason is strictly statistical, since the presence of abundantly many zeros in the lower tail of a probability density function is difficult to transform into an approximate normal distribution.

Inc_ord = A transformation of ratio-scale Inc_prim into the four ordinal categories [0, 1, 2, 3], where 0: Inc_prim = 0 (min), 1: Inc_prim <= a, 2: Inc_prim <= b, 3: Inc_prim <=max. In order to categorise Inc_prim, Inc_nozero was Box-Cox transformed into an approximately normal distributed variable that was subdivided into three equidistant classes, and where a and b were calculated by retransforming the equidistant subdivision into the original distribution (basically a log – antilog operation). When applied to the respective diseases, through the respective matrices of empirical diseases data, a, b, and max differ from one disease to another as illustrated in *Table 3*:



| CSI: | BOR | BRU | CRY | LEP | PUU | QFE | TBE | TUL |
|---|---|---|---|---|---|---|---|---|
| a: | 12.8 | 1.9 | 4.6 | 2.3 | 6.8 | 1.4 | 4.6 | 5.3 |
| b: | 163.9 | 3.5 | 21.1 | 5.2 | 45.7 | 1.9 | 21.4 | 28.4 |
| max: | 2098.3 | 6.6 | 97.1 | 12.0 | 309.3 | 2.7 | 99.0 | 151.6 |

*Table 3: Categorisation of Inc_ord = [0, 1, 2, 3] per human disease, where 0 = 0 (min), 1 <= a, 2 <= b, 3 <= max. The numbers are given in the unit of primary incidences (Prim_inci).*

Inc_bin = [0, 1] is a binary categorisation that indicates whether or not the XXX disease is reported across the respective combinations of years and report districts.

Inc_prim, Inc_nonzero, Inc_ord, and Inc_bin have very different statistical characteristics. With a small amount of noise introduced to Inc_prim with the overlay errors discussed above, Inc_ord is relatively robust to such errors, and Inc_bin is free of such errors. In addition, where a Box_Cox transformation of Inc_nozero may be considered as being approximately normal distributed, Inc_ord may be considered as being basically Poisson distributed while retaining all zero incidences (which is a huge advantage). And Inc_bin may, of course, be considered as being binomially distributed while retaining all zero incidences. When used as statistical model responses, a Box-Cox transformation of Inc_nozero would require a general modelling approach, whereas Inc_ord and Inc_bin would require generalised modelling approaches. A generalised Poisson regression approach to Inc_ord is the method of statistical inference that is recommended by CLINF.

Gender = [F, M, F+M] is a nominal factor that categorises cases of the XXX disease into the female (F) and male (M) classes of gender, and where F+M summarises cases across the two primary classes of gender.

Age = [Child, Youth, Young, Middle, Old, Tot], where Child <= 14 years, Youth <= 24, Young <= 44, Middle <= 69, Old > 69, and where Tot summarises cases across all categories of Age. Age is a nominal or ordinal factor depending on how it is being used. It is categorised across all diseases XXX, and hence across all the CLINF human CSI data, using AI-algorithms that identifies natural breaks.

Long and Lat provides the WGS84 longitudes and latitudes of the unweighted centroid of diseases report districts.

### XXX_map_inci_prim

A map in pdf-format that depicts the average (across the entire time-period of observation) primary incidences of diseases XXX (basically variable Ave_inci in file XXX_inci_ann_tot).

### XXX_map_inci_ord

The ordinal version of XXX_map_inci_prim, where individual diseases XXX have been categorised in accordance with *Table 3*.

## Appendix 1

CLINF diseases report districts (simply click the hyperlink).