

SCIENTIFIC DATA

OPEN

Data Descriptor: The contemporary distribution of *Trypanosoma cruzi* infection in humans, alternative hosts and vectors

Received: 30 November 2016

Accepted: 13 March 2017

Published: 11 April 2017

Annie J. Browne¹, Carlos A. Guerra², Renato Vieira Alves³, Veruska Maia da Costa³, Anne L. Wilson⁴, David M. Pigott⁵, Simon I. Hay^{1,5}, Steve W. Lindsay⁴, Nick Golding⁶ & Catherine L. Moyes¹

Chagas is a potentially fatal chronic disease affecting large numbers of people across the Americas and exported throughout the world through human population movement. It is caused by the *Trypanosoma cruzi* parasite, which is transmitted by triatomine vectors to humans and a wide range of alternative host species. The database described here was compiled to allow the risk of vectorial transmission to humans to be mapped using geospatial models. The database collates all available records, published since 2003, for prevalence and occurrence of infection in humans, vectors and alternative hosts, and links each record to a defined time and location. A total of 16,802 records of infection have been extracted from the published literature and unpublished sources. The resulting database can be used to improve our understanding of the geographic variation in vector infection prevalence and to estimate the risk of vectorial transmission of *T. cruzi* to humans.

Design Type(s)	data integration objective • epidemiological study • observation design
Measurement Type(s)	parasite epidemiology
Technology Type(s)	data item extraction from journal article
Factor Type(s)	
Sample Characteristic(s)	<i>Trypanosoma cruzi</i>

¹Oxford Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford OX3 7BN, UK. ²Sanaria Institute for Global Health and Tropical Medicine, Rockville, Maryland 20850, USA. ³Secretaria de Vigilância em Saúde, Ministério da Saúde, Brasília, Distrito Federal 70058-900, Brasil. ⁴School of Biosciences, Durham University, Durham DH1 3LE, UK. ⁵Institute for Health Metrics and Evaluation, University of Washington, Seattle, Washington 98121, USA. ⁶School of Biosciences, University of Melbourne, Parkville, Victoria 3010, Australia. Correspondence and requests for materials should be addressed to C.L.M. (email: catherinemoyes@gmail.com).

Background & Summary

Chagas disease is caused by the protozoan *Trypanosoma cruzi*. It is endemic across most of Latin America causing high levels of morbidity and mortality. An estimated 10 million people are affected worldwide^{1,2}, with approximately 8,000 deaths in 2015 (ref. 3). High levels of geographic variation in the burden of disease have been observed^{2,4} with people in poor, rural areas most at risk⁵. Further spatial variation results from heterogeneity in *T. cruzi* vectors^{6,7}, mediated by vector control programmes that are themselves spatially heterogeneous^{8,9}. A clear understanding of the spatial variation in *T. cruzi* transmission and infection is required to enable effective implementation of control measures for Chagas disease.

Our understanding of the geographical distribution of *T. cruzi* infection risk is complicated by the multiple routes of transmission, multiple vector species and multiple diverse reservoir species. The parasite is primarily transmitted through the faeces of triatomine bugs contaminating their own bite wounds⁹. Congenital transmission and transmission through blood transfusion and organ donation also occur. Oral transmission is the primary mode by which wild animals are infected due to frequent ingestion of infected triatomines¹⁰, and outbreaks of orally transmitted *T. cruzi* in humans occur via contaminated food sources, extending the range of vector species that present a risk to humans^{11–13}. Vectorial transmission via wounds or contaminated food is the route by which infections enter human populations and control of this transmission route is essential. The database described here was compiled to support analyses of the risk of vectorial transmission to humans.

Understanding variation in vector infection prevalence alone is not enough to understand the risk to humans from vector-related transmission. Several factors affect the relationship between vector and human infection prevalence. For example, vector species vary in their host preferences and this is affected by factors such as local host availability and house condition^{14,15}. Additionally, vector control programmes reduce the risk to humans and their efficacy is affected by a range of factors^{8,9}. Data on vector infections must therefore be coupled with data on human infections in endemic areas to understand the risk of transmission to humans. Data on human infection is complicated because the parasite can remain in the body for many decades. Acute infections, following a one to two week incubation period, typically last two to three months and during this period parasites can be identified from blood samples but serological testing does not detect infection in the early acute phase¹⁶. In contrast, parasites are rarely detected from the blood during clinical latency, which can last for decades¹⁷. Surveys of infection in humans rely on serology so the duration of infection is unknown at the time of testing, making it hard to generate estimates for the location where a subject was originally infected.

In the absence of data on human and vector infections, data on infections in alternative hosts can provide an indication of pathogen presence. Domestic and wild animal species are important reservoirs of *T. cruzi* and are responsible for maintaining parasite populations over long periods of time¹⁴. Further information can be derived from the known variation in the ability of different host species to sustain and transmit the pathogen¹⁴ and from their proximity to humans.

The World Health Organisation's Global Vector Control Response stresses the importance of conducting surveillance of locally important vectors in addition to monitoring human infection data. Data on infections in human and vector populations thus come from both research studies (usually published) and public health surveillance programmes (often unpublished). We have collated data from both research and surveillance programmes to provide a contemporary database of *T. cruzi* infection in human, vector and alternative host populations across the region. The aim is to provide data for geospatial models that will produce detailed maps for one of Latin America's most important infectious diseases, providing the basis for more effective targeting of resources. To achieve this we extracted data from the published literature and unpublished sources (Fig. 1). Previously published datasets have focused on one endemic country or one data type^{6,18,19}. To our knowledge this is the first database covering the endemic region as a whole and incorporating infection in humans, vectors and alternative host species with each record linked to a defined time and location.

Methods

Identifying and selecting data sources

Potential data sources were identified and then selected for data extraction following steps A to E in Fig. 1. The Web of Science bibliographic database was chosen because it incorporates many relevant databases including the SciELO Citation Index from 1997 onwards (provides access to leading journals from Latin America, Portugal, Spain and South Africa), MEDLINE from 1950 onwards (from the U.S. Library of Medicine), the Data Citation Index from 1993 onwards (provides details of datasets in international data depositories), the BIOSIS Citation Index from 1969 onwards (covers pre-clinical, experimental, and animal research) and the Web of Science's own Core Collection from 1945 onwards. The Web of Science was searched for articles published between 1 January 2003 and 31 December 2015 using the search terms 'Trypanosoma cruzi' and 'Chagas'. This time period was selected to obtain a contemporary dataset while including as many endemic countries as possible. No language restrictions were placed on the search. The initial search yielded 11,633 articles with the term 'Trypanosoma cruzi' or the term 'Chagas' in the title or abstract or key words.

In the first step of the selection process (A), titles and abstracts were scanned by an individual for any indication that the study measured the occurrence, prevalence or seroprevalence of *T. cruzi* in human communities, or in vector or reservoir populations sampled from the field. Articles that gave no

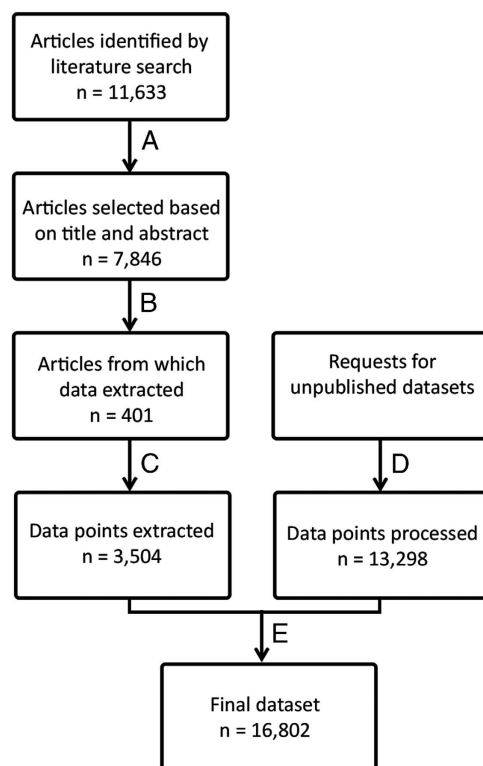


Figure 1. The process from data source identification to data extraction. Steps A-E are described in the Methods section of the main text.

indication that local occurrence or prevalence of infection was measured were discarded during this stage of the process. Typically this excluded studies of, for example, therapeutic options for patients with chronic disease, molecular studies of lab strains of the parasite, mouse model experiments, and so on, unless such studies also incorporated field collected data. This step reduced the number of articles of potential interest to 7,846.

In step B of the process, full text copies were obtained wherever possible and read by an individual. Articles that contained quantitative data on infection prevalence or occurrence in one or more human or vector or alternative host populations, linked to a defined location and time, were retained. If an article contained aggregated data (across multiple locations or multiple collection periods) the authors were contacted and the disaggregated data were requested. Review articles were excluded unless the original data source plus the dates of the original sample collection were provided. This step yielded 401 articles. In addition, if an article mentioned an unpublished dataset or a surveillance programme, the details were noted.

Step C of the process involved the extraction of the data fields listed in the Metadata File from each article identified in step B. First each article was assessed by an individual to determine whether the data met the criteria for infection prevalence data. If so, the relevant fields were extracted and, if not, the fields for infection occurrence were extracted. Seven different data files for infection prevalence and infection occurrence in human, vector and alternative host populations, and for acute infection occurrence in humans, were compiled. The inclusion criteria for each data file, and for the location and time linked to each data point, are detailed below.

Unpublished datasets identified during step B above were requested from the groups managing these data, identified using Google. Step D of the process mirrored step C; each dataset was assessed by an individual to determine whether the data met the criteria for infection prevalence data. If so, the relevant fields were extracted and, if not, the fields for infection occurrence were extracted. Finally, in step E of the process the data extracted from the published articles and from the unpublished datasets were combined into the seven data files listed in the Data Records section.

Infection prevalence in humans

Studies were included if they sampled a cross-section of the local community at a specific time and place (active surveillance). The only exception to this criterion was any study that sampled a restricted age range within the community; these studies were included together with a record of the age range sampled. Studies of prevalence in a non-representative section of the community, for example pregnant women or hospital patients, were excluded.

If a study gave different prevalence values for different diagnostic tests then the value derived from the number of positives in two or more independent serological tests was taken, i.e. the generally recommended diagnostic. In all instances, the combination of diagnostic tests used to derive the prevalence value given was recorded. The number tested and the number positive were extracted, unless these values were not given in which case the prevalence value was recorded. The age range of those sampled was also recorded. If the study stated that the whole community was tested but did not give an age range, this was recorded as 0 to 99.

Infection occurrence in humans

Studies within endemic countries that did not meet the criteria for prevalence were included in the occurrence dataset. These included whole community studies if the prevalence value was missing, infections in a community subset, passive surveillance and case reports. Cases specifically attributed to vertical transmission, blood donation or organ donation, and reports of chronic patients, were excluded. Cases identified as an acute infection were recorded separately. For each record, presence/absence was recorded with the number of individuals tested if available. The combination of diagnostic tests used to confirm each presence or absence was also recorded.

Infections in vectors and alternative hosts

Vectors and reservoir hosts tested were assumed to be a representative sample of the population at that location. The number tested and number positive, or the prevalence, were recorded. If prevalence data were not available then presence/absence was recorded together with the number tested if given. In both instances, the combination of diagnostic or detection tests used was recorded. For each record, the species sampled was recorded using the scientific name, and assigned to one of 19 species groupings.

Sample collection times and locations

All data were disaggregated to a single collection period and collection site wherever possible. If the published article only contained aggregated data (aggregated over time or over locations), the authors were contacted to request the disaggregated data.

Geographical coordinates were assigned to individual sites that could be located precisely within 5×5 km. If samples from more than one site were pooled before they were tested, all site coordinates were linked to the single record. If an area $>25 \text{ km}^2$ was given by the authors or data source (including hospital catchment areas), this was defined as a polygon. Polygons that matched a country's administrative divisions were assigned an identifier from the UN Food and Agriculture Organisation's Global Administrative Unit Layers (GAUL)²⁰.

For all sites $\leq 5 \times 5$ km, geographical coordinates provided in the article or by the data source were converted to decimal degrees. If a map was provided, this was used to assign coordinates to each site. If no coordinates or map were provided, the site name was used together with contextual information such as the district, distance to a border or city and so on, to locate the site in online gazetteers such as Google Maps (<http://maps.google.co.uk>), Geonames (<http://www.geonames.org/>) and Open Street Map (<https://www.openstreetmap.org/>). If more than one potential site was identified, no coordinates were assigned. Where possible the site coordinates were verified in two separate gazetteers. If the site could not be found in any of the gazetteers then further online searches were used to determine the location. If no precise coordinates could be found then the next highest level geography (e.g., a district polygon) was assigned.

Data sources

The data source was recorded and up to two data sources were linked to each data point, for example, if prevalence was recorded in a published article but the date of collection was confirmed through a personal communication then both the article and personal communication were linked to the same data point.

Defining the limits of vectorial transmission of *Trypanosoma cruzi*

The geographical limits of transmission were determined using the published literature and online resources. A mask of areas with no record of vector related human infection was then created using ArcGIS (ESRI 2011. ArcGIS Desktop: Release 10.1. Redlands, CA: Environmental Systems Research Institute).

The continental southern limits of transmission were defined as the Chubut region in Argentina and the Libertador region of Chile. Small numbers of triatomines have been found in Chubut²¹ but all tested negative for *T. cruzi* whilst infections have been reported in the seven northern regions of Chile but not in the south²². Maps from the Pan American Health Organisation (PAHO) and the International Association for Medical Assistance to Travellers (IAMAT) provide clear advice for no risk of vector-related *T. cruzi* transmission south of these regions (<https://www.iamat.org/risks/chagas-disease>, <http://www.paho.org/salud-en-las-americas-2012>).

The northern limits of Chagas are less clearly defined. Autochthonous cases arising from vectorial transmission have been reported in humans in Louisiana and Texas^{23,24} and in mammals and triatomines across the southern USA and as far north as Tennessee^{23–41}. Based on these observations and a map of

states reporting triatomines (www.cdc.gov/parasites/chagas/gen_info/vectors), the northern limit was drawn from the north of California to the north of Pennsylvania.

The World Health Organisation (WHO) and the Centers of Disease and Control and Prevention (CDC) state that vector-related transmission does not occur in the Caribbean (www.who.int/mediacentre/factsheets/fs340; www.cdc.gov/parasites/chagas/gen_info/vectors). The only Caribbean country reporting Chagas cases is Grenada^{42,43} therefore all other Caribbean countries were deemed not at risk. The Galapagos Islands were deemed not at risk because studies have identified an absence of triatomine vectors⁴⁴ and serological testing of dogs and cats showed no evidence of infection⁴⁵. Uruguay is the only continental Latin American country to have interrupted vector-related transmission of *T. cruzi* (www1.paho.org/english/ad/dpc/cd/incosur.htm). Since control of transmission in 1998, there have been no reports in Uruguay and therefore it was also classified as not at risk.

Data Records

The full workflow and numbers of articles contributing to the final dataset is given in Fig. 1. A total of 16,802 data records were extracted and input into seven data files, stored in the Dryad Digital Repository (Data Citation 1).

The seven individual data files within Data Citation 1 are (1) prevalence of infection in humans, (2) occurrence of acute infection in humans, (3) occurrence of infection in humans, (4) prevalence of infection in vectors, (5) occurrence of infection in vectors, (6) prevalence of infection in alternative hosts, and (7) occurrence of infection in alternative hosts. Each data file contains individual records linked to a defined time and location. Locations are defined as either points (an area $\geq 5 \times 5$ km, assigned geographical coordinates in decimal degrees), or polygons (areas $> 5 \times 5$ km with defined boundaries). Times are defined as the start month and year, and the end month and year, for each sample collected. The total number of records is given in Table 1.

All data records were linked to sub-national geographic locations. Data were obtained for 20 of the countries where Chagas is endemic but no data records were identified from Suriname, Guyana or Grenada. Figure 2 highlights states/departments with no available data. Despite some obvious gaps, we provide data on human and/or vector infections for most of the endemic region. At the northern and southern fringes of the regions some areas only have data for alternative hosts. Areas with no publicly available data at all (published since 2003) can be seen throughout the endemic region.

Within the data file for prevalence of infection in vectors, values are recorded for a total of 59 triatomine species. The mean prevalence of infection in the five most commonly sampled species is given in Table 2. These species are all found in Brazil and their status as the most widely tested species largely reflects the high number of locations and volume of data from Brazil. It is interesting to note that although the database has only 162 prevalence records for *T. infestans* populations, a large number of specimens have been tested (11,994, with a mean infection prevalence of 35.4%), reflecting the importance of this vector within its restricted range⁴⁶. Many other species are locally important with higher infection prevalences than those found for the most widely sampled species.

The data files for occurrence and prevalence of infection in alternative hosts encompass 177 species including rodents, bats, non-human primates, other wild mammals, marsupials, domestic mammals and livestock. The mean prevalence of infection in the most sampled species is given in Table 3 and it can be seen that the most widely tested alternative host is the domestic dog, *Canis familiaris*, with over 250 records of infection prevalence.

The aim of this work is to provide contemporary records of infections and only published articles from 2003 onwards were included. Studies of historical trends in the location and prevalence of infection will need to expand the current dataset further by extracting data from older publications. The current data compilation exercise ended in 2016 and, as time passes, increasing amounts of more recent data will also become available so the literature should be monitored for new records by any groups wishing to update the current dataset. Finally, the dataset presented here may not be complete. Potential records of interest may have been missed during each step of the process outlined in Fig. 1 and there are likely to be

Data file	Number of records	Number of locations	Number of countries
Prevalence in humans	1,012	607	15
Acute infections in humans	497	216	8
Occurrence in humans (excluding prevalence data and confirmed acute infections)	328	128	12
Prevalence in vector species	13,798	2,755	15
Occurrence in vector species (excluding prevalence data)	276	220	10
Prevalence in alternative host species	858	338	13
Occurrence in alternative host species (excluding prevalence data)	33	21	7

Table 1. Data volumes for each file within the dataset. A data record is defined as a prevalence value or presence/absence for a defined location and collection period derived from a unique study. The number of locations refers to the number of unique locations with infection data.

unpublished datasets that were not identified or obtained. There is no comprehensive dataset in existence that would allow an assessment of completeness of the current database, however, a georeferenced database of infections in North America has been compiled that contains 669 records from 1936 to 2014 (ref. 6). Here we collated 341 infection records for Mexico and the USA (excluding absence records) from articles published from 2003 to 2015. The average number of records per year is higher in the current dataset, possibly because more data have been generated in more recent years.

Technical Validation

Each data record was extracted by one individual and checked by a second person to ensure accuracy. All site coordinates linked to data records were checked using GIS software (ArcGIS and QGIS) to ensure they fell on land (as defined by GAUL) and in the right country. All point and polygon locations for each data source were plotted to ensure they matched the sampling design described by the data source.

All other fields were checked to confirm that each value fell within the expected range and to identify any missing data. If any information was unclear then authors were contacted for confirmation.

Each full data file was checked to ensure that study records had not been duplicated. The data file for each vector species was also plotted and any outliers geographically distant from the main distribution were investigated to ensure they had been correctly extracted and fell within the species range.

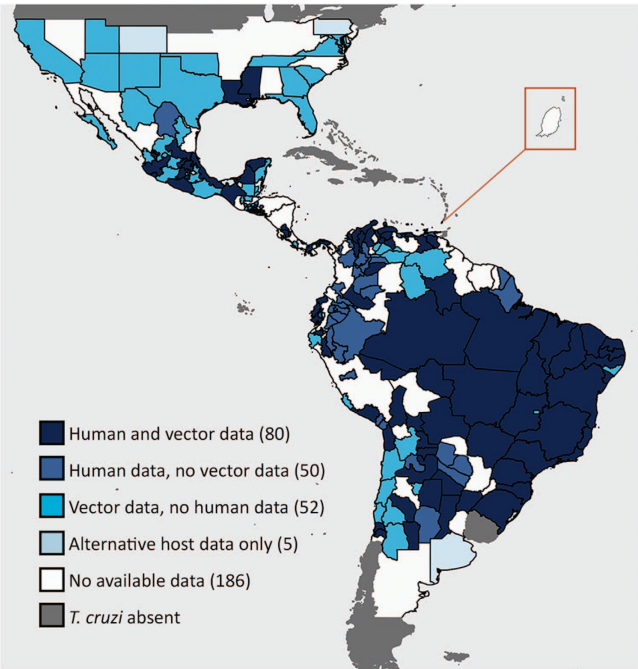


Figure 2. Data availability by first level administrative division. For each first order administrative division (typically called a state or department) the types of data available are shown and the total number of administrative divisions with data is given in parentheses. The combination of human and vector infection data is prioritised and the availability of alternative host infection data is only shown if no other data types are available. An enlarged view of the island nation Grenada is shown.

Vector species	Number of records	Mean prevalence (%)
<i>Triatoma sordida</i>	2,502 (342,791)	1.0 (0–100)
<i>Triatoma pseudomaculata</i>	2,286 (118,064)	4.6 (0–100)
<i>Panstrongylus megistus</i>	1,951 (20,897)	10.2 (0–100)
<i>Triatoma brasiliensis</i>	1,795 (104,851)	4.6 (0–100)
<i>Rhodnius neglectus</i>	1,371 (9,804)	3.7 (0–100)

Table 2. Prevalence of *T. cruzi* infection in the most sampled vector species. A data record is defined as a prevalence value for a defined location and collection period derived from a unique study. The total number of specimens tested is given in parentheses after the number of records, and the full range of prevalence values is given in parentheses after the mean prevalence value.

Host species	Species grouping	Number of records	Mean prevalence (%)
<i>Rattus rattus</i> (black rat)	Rodent	22 (171)	26 (0–100)
<i>Sigmodon hispidus</i> (cotton rat)	Rodent	11 (316)	17.7 (0–50)
<i>Octodon degus</i> (degu)	Rodent	10 (861)	44 (13.3–70.4)
<i>Carollia perspicillata</i> (Seba's short-tailed bat)	Bat	8 (150)	25.9 (0–54.5)
<i>Leontopithecus rosalia</i> (golden lion tamarin)	Non-human primate	5 (607)	33.9 (25.1–43.7)
<i>Nasua nasua</i> (South American coati)	Other wild mammal	11 (343)	52 (0–100)
<i>Procyon lotor</i> (raccoon)	Other wild mammal	11 (302)	11.3 (0–55)
<i>Nasua narica</i> (white-nosed coati)	Other wild mammal	10 (126)	20.2 (0–93.3)
<i>Didelphis marsupialis</i> (common opossum)	Marsupial	20 (111)	48.9 (0–100)
<i>Didelphis albiventris</i> (white-eared opossum)	Marsupial	18 (391)	34.4 (0–88.9)
<i>Monodelphis domestica</i> (gray short-tailed opossum)	Marsupial	12 (227)	19 (0–100)
<i>Canis familiaris</i> (domestic dog)	Domestic mammals	253 (11,506)	24.6 (0–100)
<i>Felis catus</i> (domestic cat)	Domestic mammals	25 (832)	16.3 (0–62.1)
<i>Mus musculus</i> (house mouse)	Domestic mammals	19 (231)	5.5 (0–20)
<i>Sus scrofa domesticus</i> (domestic pig)	Livestock	11 (142)	29.7 (0–100)
<i>Ovis aries</i> (sheep)	Livestock	4 (308)	1.5 (0–3.9)

Table 3. Prevalence of *T. cruzi* infection in the most sampled alternative host species. A data record is defined as a prevalence value for a defined location and collection period derived from a unique study. The total number of specimens tested is given in parentheses after the number of records, and the full range of prevalence values is given in parentheses after the mean prevalence value.

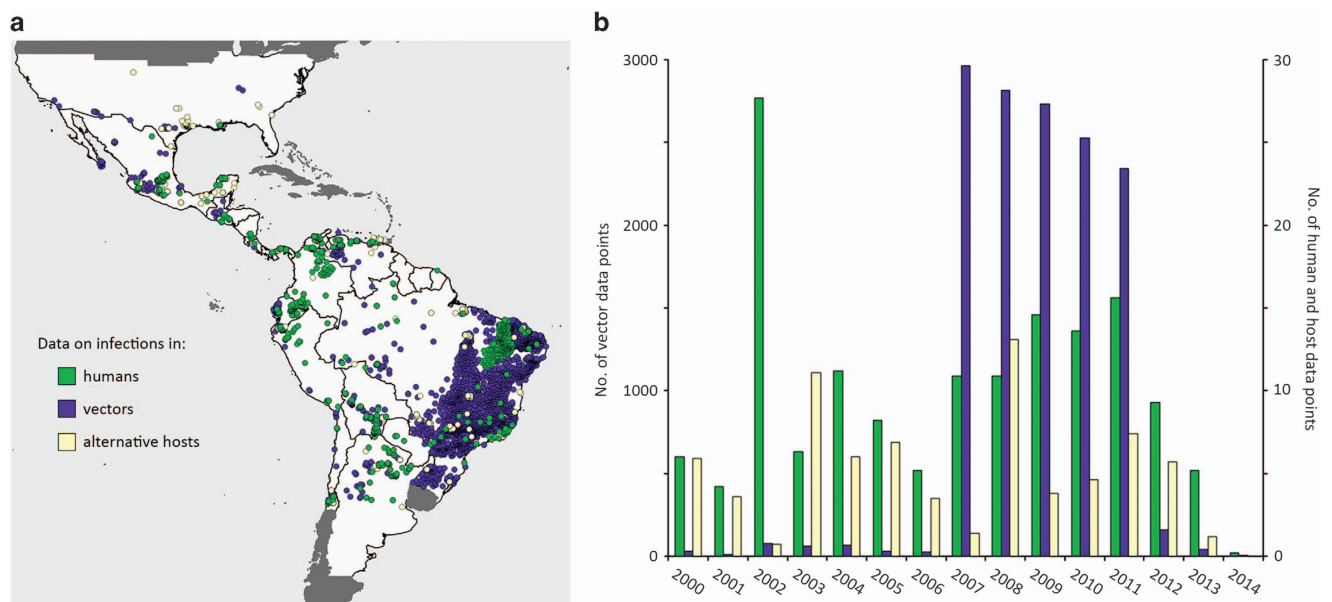


Figure 3. Spatial and temporal clustering in the dataset. (a) The map shows the spatial distribution of the human, vector and alternative host populations sampled. (b) The plot shows the distribution of the data across years for infections in humans, vectors and hosts.

Usage Notes

This dataset does not represent a systematic survey across the Chagas endemic region using a single consistent method. It is a collation of observational data containing inherent biases that should be addressed when the data are used in geospatial models. Figure 3 shows that the distribution of data is not uniform in either time or space and this is in part due to sampling bias. The choice of study locations is biased with a preference for areas of known high infection prevalence or areas linked to risk factors such as poor housing in rural localities⁵. Analysis of these data should therefore make use of methods that have been developed to account for such sampling biases^{47,48}. The full dataset provides the information needed to assess these data appropriately so that users can also take account of potential confounders such as the

combination of diagnostics used, species tested, and age of the subjects sampled. Other important factors contributing to variation in the results recorded, such as the sample size for each record and the number of records for a particular location, are also held within the dataset. This allows modellers to incorporate the strength of evidence associated with each record of disease occurrence or prevalence, for example taking account of the often highly stochastic nature of empirical infection prevalence estimates, especially in vectors.

This database was collated to support analyses that map the risk of vectorial transmission of *T. cruzi* to humans. The inclusion of data on the location and time of each sample collection means that this dataset can be used in combination with spatiotemporal environmental, demographic and socioeconomic covariates to model the risk of human infection at fine spatial resolution across the Americas. Ultimately we hope that these data will be used to improve knowledge of the geographical extent of endemic Chagas disease and disease burden. Understanding the relationship between the prevalence of *T. cruzi* in vectors and the risk of infection to humans will help inform appropriate control measures and highlight areas where resources are required.

References

- Lee, B. Y., Bacon, K. M., Bottazzi, M. E. & Hotez, P. J. Global economic burden of Chagas disease: a computational simulation model. *Lancet Infect. Dis.* **13**, 342–348 (2013).
- Stanaway, J. D. & Roth, G. The burden of Chagas disease: estimates and challenges. *Glob. Heart* **10**, 139–144 (2015).
- Wang, H. *et al.* Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the global burden of disease study 2015. *Lancet* **388**, 1459–1544 (2016).
- Moncayo, A. Chagas disease: current epidemiological trends after the interruption of vectorial and transfusional transmission in the southern cone countries. *Mem. Inst. Oswaldo Cruz* **98**, 577–591 (2003).
- Rassi, A. Jr, Rassi, A. & Marin-Neto, J. A. Chagas disease. *Lancet* **375**, 1388–1402 (2010).
- Ramsey, J. M. *et al.* Atlas of Mexican Triatominae (Reduviidae: Hemiptera) and vector transmission of Chagas disease. *Mem. Inst. Oswaldo Cruz* **110**, 339–352 (2015).
- Waleckx, E., Gourbiere, S. & Dumonteil, E. Intrusive versus domiciliated triatomines and the challenge of adapting vector control practices against Chagas disease. *Mem. Inst. Oswaldo Cruz* **110**, 324–338 (2015).
- Quinde-Calderon, L., Rios-Quitizaca, P., Solorzano, L. & Dumonteil, E. Ten years (2004–2014) of Chagas disease surveillance and vector control in Ecuador: successes and challenges. *Trop. Med. Int. Health* **21**, 84–92 (2016).
- Sosa-Estani, S. & Segura, E. L. Integrated control of Chagas disease for its elimination as public health problem—a review. *Mem. Inst. Oswaldo Cruz* **110**, 289–298 (2015).
- Coura, J. R. The main sceneries of Chagas disease transmission. The vectors, blood and oral transmissions—a comprehensive review. *Mem. Inst. Oswaldo Cruz* **110**, 277–282 (2015).
- Aglaër, A. N. *et al.* Oral transmission of chagas disease by consumption of açai palm fruit, Brazil. *Emerg. Infect. Dis.* **15**, 653 (2009).
- Dario, M. A. *et al.* Ecological scenario and *Trypanosoma cruzi* DTU characterization of a fatal acute Chagas disease case transmitted orally (Espírito Santo state, Brazil). *Parasites & Vectors* **9**, 477 (2016).
- de Noya, B. A. & Gonzalez, O. N. An ecological overview on the factors that drives to *Trypanosoma cruzi* oral transmission. *Acta Trop.* **151**, 94–102 (2015).
- Gurtler, R. E. & Cardinal, M. V. Reservoir host competence and the role of domestic and commensal hosts in the transmission of *Trypanosoma cruzi*. *Acta Trop.* **151**, 32–50 (2015).
- Schofield, C. J. Control of Chagas-disease vectors. *Br. Med. Bull.* **41**, 187–194 (1985).
- Sathler-Avelar, R. *et al.* Phenotypic features of peripheral blood leucocytes during early stages of human infection with *Trypanosoma cruzi*. *Scand. J. Immunol.* **58**, 655–663 (2003).
- Prata, A. Clinical and epidemiological aspects of Chagas disease. *Lancet Infect. Dis.* **1**, 92–100 (2001).
- Cruz-Reyes, A. & Pickering-Lopez, J. M. Chagas disease in Mexico: an analysis of geographical distribution during the past 76 years—a review. *Mem. Inst. Oswaldo Cruz* **101**, 345–354 (2006).
- Mischler, P. *et al.* Environmental and socio-economic risk modelling for Chagas disease in Bolivia. *Geospat Health* **6**, S59–S66 (2012).
- Food and Agricultural Organization of the United Nations. *The global administrative unit layers (GAUL): technical aspects*. EC-FAO Food Security Program (2008).
- Wisnivesky-Colli, C., Vezzani, D., Pietrokovsky, S. M., Scurti, H. & Iriarte, J. Ecological characteristics of *Triatoma patagonica* at the southern limit of its distribution (Chubut, Argentina). *Mem. Inst. Oswaldo Cruz* **98**, 1077–1081 (2003).
- Schenone, H. *et al.* Panorama general de la epidemiología de la enfermedad de Chagas en Chile. *Bol. Chil. Parasitol.* **46**, 19–30 (1991).
- Dorn, P. L. *et al.* Autochthonous transmission of *Trypanosoma cruzi*, Louisiana. *Emerg. Infect. Dis.* **13**, 605–607 (2007).
- Klotz, S. A., Dorn, P. L., Mosbacher, M. & Schmidt, J. O. Kissing bugs in the United States: risk for vector-borne disease in humans. *Environ. Health Insights* **8**, 49–59 (2014).
- Rosypal, A. C. *et al.* Serologic survey of antibodies to *Trypanosoma cruzi* in coyotes and red foxes from Pennsylvania and Tennessee. *J. Zoo. Wildl. Med.* **45**, 991–993 (2014).
- Wozniak, E. J. *et al.* The biology of the triatomine bugs native to south central Texas and assessment of the risk they pose for autochthonous Chagas disease exposure. *J. Parasitol.* **101**, 520–528 (2015).
- Newsome, A. L. & McGhee, C. R. *Trypanosoma cruzi* in triatomines from an urban and a domestic setting in middle Tennessee. *J. Ten. Acad. Sci.* **81**, 62–65 (2006).
- Cesa, K., Caillouet, K. A., Dorn, P. L. & Wesson, D. M. High *Trypanosoma cruzi* (Kinetoplastida: Trypanosomatidae) prevalence in *Triatoma sanguisuga* (Hemiptera: Reduviidae) in southeastern Louisiana. *J. Med. Entomol.* **48**, 1091–1094 (2011).
- Buhaya, M. H., Galvan, S. & Maldonado, R. A. Incidence of *Trypanosoma cruzi* infection in triatomines collected at Indio Mountains Research Station. *Acta Trop.* **150**, 97–99 (2015).
- Herrera, C. P., Licon, M. H., Nation, C. S., Jameson, S. B. & Wesson, D. M. Genotype diversity of *Trypanosoma cruzi* in small rodents and *Triatoma sanguisuga* from a rural area in New Orleans, Louisiana. *Parasites & Vectors* **8**, 123 (2015).
- Waleckx, E., Suarez, J., Richards, B. & Dorn, P. L. *Triatoma sanguisuga* blood meals and potential for Chagas disease, Louisiana, USA. *Emerg. Infect. Dis.* **20**, 2141–2143 (2014).
- McPhatter, L. *et al.* Vector surveillance to determine species composition and occurrence of *Trypanosoma cruzi* at three military installations in San Antonio, Texas. *US Army. Med. Dep. J.* 12–21 (2012).

33. de la Rua, N., Cesa, K., Perniciaro, L., Wesson, D. & Dorn, P. L. Genetic structure of a highly *Trypanosoma cruzi*-infected population of *Triatoma sanguisuga* in New Orleans, Louisiana, USA. *Am. J. Trop. Med. Hyg.* **79**, 248–249 (2008).
34. Hall, C. A., Polizzi, C., Yabsley, M. J. & Norton, T. M. *Trypanosoma cruzi* prevalence and epidemiologic trends in lemurs on St. Catherines Island, Georgia. *J. Parasitol.* **93**, 93–96 (2007).
35. Shadomy, S. V., Waring, S. C. & Chappell, C. L. Combined use of enzyme-linked immunosorbent assay and flow cytometry to detect antibodies to *Trypanosoma cruzi* in domestic canines in Texas. *Clin. Diagn. Lab. Immunol.* **11**, 313–319 (2004).
36. Parrish, E. A. & Mead, A. J. Determining the prevalence of *Trypanosoma cruzi* in road-killed opossums (*Didelphis virginiana*) from Baldwin county, Georgia, using polymerase chain reaction. *Ga. J. Sci.* **68**, 132–139 (2010).
37. Gates, M. *et al.* Parasitology, virology, and serology of free-ranging coyotes (*Canis latrans*) from central Georgia, USA. *J. Wildl. Dis.* **50**, 896–901 (2014).
38. Tenney, T. D., Curtis-Robles, R., Snowden, K. F. & Hamer, S. A. Shelter dogs as sentinels for *Trypanosoma cruzi* transmission across Texas, USA. *Emerg. Infect. Dis.* **20**, 1323–1326 (2014).
39. Reisenman, C. E. *et al.* Infection of kissing bugs with *Trypanosoma cruzi*, Tucson, Arizona, USA. *Emerg. Infect. Dis.* **16**, 400–405 (2010).
40. Hwang, W. S., Zhang, G., Maslov, D. & Weirauch, C. Short report: infection rates of *Triatoma protracta* (Uhler) with *Trypanosoma cruzi* in southern California and molecular identification of trypanosomes. *Am. J. Trop. Med. Hyg.* **83**, 1020–1022 (2010).
41. Charles, R. A., Kjos, S., Ellis, A. E., Barnes, J. C. & Yabsley, M. J. Southern plains woodrats (*Neotoma micropus*) from southern Texas are important reservoirs of two genotypes of *Trypanosoma cruzi* and host of a putative novel *Trypanosoma* species. *Vector Borne Zoonotic Dis* **13**, 22–30 (2013).
42. Rosypal, A. C. *et al.* Seroprevalence of canine leishmaniasis and American trypanosomiasis in dogs from Grenada, West Indies. *J. Parasitol.* **96**, 228–229 (2010).
43. Chikweto, A. *et al.* Seroprevalence of *Trypanosoma cruzi* in stray and pet dogs in Grenada, West Indies. *Trop. Biomed.* **31**, 347–350 (2014).
44. Jasicki, S., Herman, D., M. G. & Ocaña, S. Household infestation rates of Chagas disease vectors and *Trypanosoma cruzi* prevalence in southern Ecuador. Available at <http://digital.library.wisc.edu/1793/54759> (2011).
45. Levy, J. K. *et al.* Infectious diseases of dogs and cats on Isabela Island, Galapagos. *J. Vet. Intern. Med.* **22**, 60–65 (2008).
46. Schofield, C. J., Jannin, J. & Salvatella, R. The future of Chagas disease control. *Trends in Parasitology* **22**, 583–588 (2006).
47. Diggle, P. J., Menezes, R. & Su, T. L. Geostatistical inference under preferential sampling. *J. R. Stat. Soc. Ser. C Appl. Stat* **59**, 191–232 (2010).
48. Ho, L. P. & Stoyan, D. Modelling marked point patterns by intensity-marked Cox processes. *Stat. Probab. Lett.* **78**, 1194–1199 (2008).

Data Citation

1. Browne, A. J. *et al.* Dryad Digital Repository <http://dx.doi.org/10.5061/dryad.93mn0> (2017).

Acknowledgements

The authors are grateful to the following people who provided unpublished data: Mariolga Berrizbeita, Laurent Brutus, Jean-Philippe Chippaux and Jose Alejandro Martínez Ibarra, in addition to our co-authors Veruska Maia Costa and Renato Vieira Alves. We are also grateful to the authors who confirmed details of their published work and who are cited in the relevant published datasets linked to this data descriptor. We are also grateful to Maria Devine who obtained the published articles that fed into this work and proof read the manuscript. This work was funded by grants from the Bill & Melinda Gates Foundation [OPP1093011; OPP1053338]. C.L.M. had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Author Contributions

C.L.M. and N.G. drafted the data collation protocol. C.A.G. and A.J.B. extracted the data and C.A.G. requested unpublished data. V.M.C. and R.V.A. processed and provided unpublished datasets from Brazil. C.L.M. checked all extracted data. A.J.B. and C.L.M. wrote the first draft and all authors contributed to the manuscript.

Additional Information

Competing interests: The authors declare no competing financial interests.

How to cite this article: Browne, A. J. *et al.* The contemporary distribution of *Trypanosoma cruzi* infection in humans, alternative hosts and vectors. *Sci. Data* **4**:170050 doi: 10.1038/sdata.2017.50 (2017).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0>

Metadata associated with this Data Descriptor is available at <http://www.nature.com/sdata/> and is released under the CC0 waiver to maximize reuse.

© The Author(s) 2017