DOCTORAL THESIS

# Toward global estimates of spatial and temporal transmission of Chagas disease

Julia V. C. Ledien

Submitted for the degree of Doctor of Philosophy

University of Sussex

September 2022

# Declaration

I hereby declare that this thesis has not been and will not be submitted in whole or in part to another University for the award of any degree.

Signature:

Julia Ledien

# Thesis summary

Estimating spatiotemporal variation in disease exposure is critical to developing cost-effective and targeted strategies to reduce the burden of the disease. As The WHO is expecting to eliminate Chagas disease as a public health problem by 2030, being able to measure the local burden of the disease and the progress made thus far is critical. However, obtaining such information when there is no dedicated surveillance system set up to monitor incidence can be challenging.

Cross-sectional prevalence studies provide information on past exposure but cannot be used directly to evaluate the epidemiological situation, especially for long-lasting diseases such as Chagas disease. However, the Force-of-Infection (FoI), i.e., the yearly per-susceptible rate of disease acquisition, can be estimated using age prevalence data and provide insight into the local temporal pattern of the disease. Such methodology, relying on localised surveys, can inform the dynamic of transmission locally, but extrapolating FoI estimates from them to assess the burden across the country requires robust statistical methods. In this thesis, we develop and implement a modelling process to, first, predict FoI in space and time from cross-sectional studies and, estimate the burden of the disease, while appropriately propagating uncertainties.

Where such an approach has been used, typically mean or median FoI estimates are used as a dependent variable, ignoring the uncertain nature of such FoI being estimates rather than observations. Therefore, the first objective of this thesis was to account for such uncertainty both when fitting models and evaluating their predictive ability. We implemented a set of comprehensive analyses to assess the impact of this uncertainty on performance, by characterising the ability to estimate accurately the central trends while correctly characterising the level of uncertainty. We, then, compare the implementation and performance of this framework to Machine Learning methods to optimise the methodology. Finally, we, propose a modelling process where the predicted FoIs at a fine spatial resolution are used to estimate the burden of Chagas disease.

The process, applied to the 76 serosurveys conducted in Colombia, showed a substantial risk of overconfidence when using median estimates to fit and evaluate models, instead of accounting for the uncertain nature of estimated FoI. Machine Learning methods provided a more flexible and reproducible framework while recentring the uncertainty and are thus better suited to provide good burden estimates. Implementing the modelling pipeline, we estimated that the FoI varied considerably across Colombia, but temporal changes were less marked. Relying on predicted current and past exposure, 506,000 (95%CrI: 395,000-648,000) people were estimated to be infected by *T. cruzi* in Colombia in 2020, representing a 1.0% (95%CrI: 0.8%-1.3%) prevalence in the general population and leading to an estimated 2,400 (95%CrI: 1,900-3,400) deaths. We estimated a substantial increase in the burden of Chagas disease over time, resulting from the interplay between exposure and demography: a slight decrease in exposure was overcompensated by the large increase in population size and the gradual ageing of the population.

# Acknowledgements

This journey would not have been possible without the considerable support I have received throughout the last four years.

I'm particularly grateful to my main supervisor, Professor Pierre Nouvellet, for his patience and continuous involvement. I'm aware I question anything and call for explanations on everything while not being a genius in statistics. Thank you for all the teaching you've employed, I've learnt a lot. I would like to thank you as well for your constant support even if you had other responsibilities and for your invaluable guidance and help.

I would also like to thank my collaborators from the Chagas modelling team, Prof María-Gloria Basáñez, Dr Zulma Cucunubá, Dr Gabriel Parra-Henao, Ms Eliana Rodríguez-Monguí, Prof Andrew Dobson and Dr Susana Adamo. None of this work could have been possible without you. Thank you for generously providing your expertise and guidance.

Thanks to my co-supervisor, Prof Fiona Mathews, and the members of my thesis committee, Prof Paul Graham, Prof Jorn Scharlemann and Dr Tai Yang, for their insightful advice.

My thanks also go to Dr Luis Gerardo Castellanos and Dr Alpha Forna for valuable discussions.

A special thanks to Dr Arnaud Tarantola for prodding me to undertake a PhD and for his guidance and support when I was looking for the right position.

Thanks to Dr Julien Cappelle, Dr Yves Froehlich, Dr Philippe Vignoles and Prof Marie-Laure Darde for their support throughout my career.

I would also like to thank the School of Life Sciences at the University of Sussex for generously covering the costs of tuition and living expenses as well as travel to international conferences.

Finally, I would like to thank my friends and family for their endless support.

Un grand merci à tous ceux qui ont partagé ma vie pendant ces quatre années, ses hauts et ses bas, les éclats de rire et les épaules réconfortantes.

D'abord, la meilleure cohorte EBE des 100 dernières années, Julie, Elias, Lizzie et Jon. Compagnons de souffrance, vous garderez une place spéciale dans mon cœur. J'aurais aimé qui nous puissions célébrer notre graduation tous ensemble. Cette photo aurait eu sa place sur la cheminée.

Non moins importants, les adeptes du sofa du 5B23 et des bancs de l'Open House, Yanet, Cannelle et Dani, Chris, Heena, Yujin, Andrès, Priesha, Rosie et le gang des Grecques. Merci pour cet environnement de travail chaleureux et les longues soirées sur la plage, juste inoubliable.

Également, ma Dream Team d'épidémiologistes, Louisa et Alex, Camille, Clara et Juan.

Encore quelques mots pour celui qui a sans doute été le plus vaillant, Pierre. Merci d'avoir patiemment enduré toutes mes questionnements méthodologiques et apporté ton soutient pédagogique, ainsi que pour tes analyses aiguisées et tes précieux conseils. Ce fut un plaisir de travailler avec toi.

Dans un registre plus personnel, J'aimerais remercier mes proches pour avoir cru en mon potentiel avant que j'ai moi-même pris conscience des portes qui pouvait s'ouvrir à moi, mes parents, Varta, Dorianne, Estève et Soline, Vincent et Héléna.

Tout particulièrement, merci à mon père pour m'avoir forgé un mental de guerrière, pour son ouverture d'esprits et sa soif de connaissance qu'il a su me transmettre.

Merci à ma mère pour m'avoir appris qu'on peut être doux mais fort et que le monde est fait de nuances qu'on peut discuter. Son goût pour les analyses critiques et son sens du détail, que je partage désormais, sont des outils indispensables dans mon travail.

Héléna, je suis heureuse de t'avoir à nouveau dans ma vie. Merci pour toutes les heures passées au téléphone dans les bons et les mauvais moments.

Varta, Dorianne, merci d'être des super sœurettes. C'est bon de savoir qu'on trouvera toujours une oreille attentive et réconfortante au bout du fil.

Merci également à Matou pour m'avoir prouvé que les choses que l'on prend pour établies peuvent être bousculées si on accepte d'y croire et d'y engager les efforts nécessaires. Essayer en vaut toujours la peine, non pas pour le résultat mais pour le voyage lui-même et ce qu'il a à nous apprendre.

Spéciale dédicace à Pumpkin comme c'est la tradition dans l'équipe EBE Epi de Sussex.

Enfin, merci à Awsten, Geoff and Otto pour avoir écrit la bande son accompagnant toutes ces heures de travail.

# Preface

## Scientific papers included in this thesis

*"Spatiotemporal variations in exposure: Chagas disease in Colombia as a case study"*

Julia Ledien, Zulma M. Cucunubá, Gabriel Parra-Henao, Eliana Rodríguez-Monguí, Andrew P. Dobson, María-Gloria Basáñez, Pierre Nouvellet

Published by BMC Medical Research Methodology in January 2022.

Contribution: statistical data analyses (predictors data cleaning, conceptualization, and programming), interpretations and communication of the results (design and production of the figures, writing of the first draft of the paper)

*"Linear and Machine Learning Modelling for Spatiotemporal Disease Predictions: Force-of-Infection of Chagas Disease"*

Julia Ledien, Zulma M. Cucunubá, Gabriel Parra-Henao, Eliana Rodríguez-Monguí, Andrew P. Dobson, Susana B. Adamo, María-Gloria Basáñez, Pierre Nouvellet

Under review with PLoS NTD since February 2022.

Contribution: statistical data analyses (predictors data cleaning, conceptualization, and programming), interpretations and communication of the results (design and production of the figures, writing of the first draft of the paper)

*"From local prevalence surveys to national burden of Chagas disease, a modelling pipeline"*

Julia Ledien, Zulma M. Cucunubá, Gabriel Parra-Henao, Eliana Rodríguez-Monguí, Andrew P. Dobson, Susana B. Adamo, María-Gloria Basáñez, Pierre Nouvellet

In preparation (being reviewed by the co-authors as early as March 2022).

Contribution: statistical data analyses of the predictive models (predictors data cleaning, conceptualization, and programming), optimization of the burden model, interpretations, and communication of the results (design and production of the figures, writing of the first draft of the paper)

## Scientific papers not included in this thesis

As part of the wider consortium described above, throughout my PhD. I significantly contributed to other analyses relevant to my PhD thesis. Below is a list of outputs that resulted from those activities.

*"Global Trends of Seroprevalence and Universal Screening Policy for Chagas Disease in Donors: a systematic review and meta-analysis"*

Kim JYH, Ledien J, Rodriguez-Monguí E, Dobson A, Basáñez M-G, Cucunubá ZM.

medRxiv; 2019 [cited 2022 Feb 8]. p. 2019.12.25.19015776. Available from: https://www.medrxiv.org/content/10.1101/2019.12.25.19015776v1

This paper is a systematic review and meta-analyses of the seroprevalence captured in blood banks. I participated in the literature review and discussion of the paper.

*"Seroprevalence trends for Chagas disease in Chile: a DICTUM study / Tendencias de la seroprevalencia para enfermedad de Chagas en Chile: un estudio DICTUM"*

E Rodríguez-Monguí, A Oyarce Fierro, L Valderrama Pérez, J Valdebenito Pino, MI Jercic, A Parra Garcés, J Halder, J Ledien, R Salvatella, A Dobson, P Nouvellet, M-G Basáñez, L-G Castellanos, ZM Cucunubá.

Under review by the co-authors.

This paper describes the Force-of-infection estimates obtained for Chile. I participated in the figure design and discussion.

*"Thirty-five years of Chagas disease, population based seroprevalence in Paraguay: a DICTUM Study / Treinta y cinco años de estudios de seroprevalencia poblacional para enfermedad de Chagas en Paraguay: un estudio DICTUM"*

E Rodríguez-Monguí, H Giménez, D Giménez, C Villalba, V Lesmo, J Halder, **J Ledien**, T Barán, R Salvatella, A Dobson, P Nouvellet, M-G Basáñez, L-G Castellanos, ZM. Cucunubá

Under review by the co-authors.

This paper describes the Force-of-infection estimates obtained for Paraguay. I participated in the figure design and discussion.

## Other knowledge transfer activities

The work realized for this thesis has been presented through three poster presentations and three oral presentations to an audience ranging from medical staff, and students with various backgrounds to researchers specialized in disease modelling.

### *Poster presentation*

**Julia Ledien**, Zulma M. Cucunubá, Eliana Rodriguez Monguí, Gabriel Jaime Parra Henao, Maria-Gloria Basáñez, Pierre Nouvellet. **Chagas Disease Force of Infection and disease burden estimation at the departmental level in Colombia.** 7th International Conference on Infectious Disease Dynamics (Epidemics7), Charleston, USA, December 3-6, 2019.

**Julia Ledien**, Zulma M. Cucunubá, Eliana Rodriguez Monguí, Gabriel Jaime Parra Henao, Maria-Gloria Basáñez, Pierre Nouvellet. **From local serosurveys to global burden estimates of Chagas disease. Colombia as a case study.** Life Sciences PhD Careers Symposium, University of Sussex, ONLINE, 30th of June 2021.

**Julia Ledien**, Zulma M. Cucunubá, Eliana Rodriguez Monguí, Gabriel Jaime Parra Henao, Maria-Gloria Basáñez, Pierre Nouvellet. **Spatiotemporal variations in exposure: Chagas disease in Colombia as a case study.** 8th International Conference on Infectious Disease Dynamics (Epidemics8), ONLINE, 30th of November to 3rd of December 2021.

### *Oral presentation*

Julia Ledien, **From local serosurveys to global burden estimates of Chagas disease. Colombia as a case study**, Measuring progress and challenges for Chagas disease control in the Americas, ASTMH 2020, 15th–19th of November 2020.

Julia Ledien, **Force-of-infection models and machine learning methods applied to Chagas disease**, International Course in Outbreak Analysis and Modelling for Public Health, Colombia - Peru 2021, 6th of July 2021.

Julia Ledien, **Estimating spatiotemporal variations of exposure to Chagas disease across Colombia**, EBE seminars, University of Sussex, 25th of November 2021.

# Contents

Declaration ..... 2

Thesis summary ..... 3

Acknowledgements ..... 5

Preface ..... 8

Scientific papers included in this thesis ..... 8

Scientific paper not included in this thesis ..... 9

Other knowledge transfer activities ..... 10

Contents ..... 12

List of figures ..... 15

List of Tables ..... 17

List of abbreviations and acronyms ..... 18

Chapter 1: Introduction ..... 19

Chagas disease causes, symptoms, and treatment ..... 19

Chagas disease control ..... 22

Chagas disease elimination challenges ..... 24

Opportunity raised by modelling ..... 25

Chagas disease in Colombia ..... 27

Objectives and contribution to knowledge ..... 28

Chapter 2: Spatiotemporal variations in exposure: Chagas disease in Colombia as a case study..... ..... 31

Abstract ..... 31

Significance statement ..... 32

Introduction ..... 32

Methods ..... 34

1. General approach ..... 34
2. Data input ..... 35

Chagas Disease Force-of-Infection ..... 35

Potential explanatory variables ..... 35

3. Model selection strategy ..... 36
4. Modelling approaches and predictive performance ..... 37
5. FoI prediction for the entire country ..... 38
6. Comparing observations and predictions across serosurveys ..... 38

7. Spatial correlation and spatial heterogeneity tests ....39

8. Availability of data and materials ....39

Results....39

1. Importance of accounting for the uncertainty in FoI ....39

2. Spatial and temporal predictions of FoI in Colombia .... 44

3. Main predictors of Trypanosoma cruzi exposure ....47

Discussion ....48

**Chapter 3: Linear and Machine Learning Modelling for Spatiotemporal Disease Predictions: Force-of-Infection of Chagas Disease ....52**

Abstract ....52

Author Summary....53

Introduction....53

Methods ....55

1. Data sources ....55

2. Linear Model (LM) framework ....56

3. Machine Learning (ML) framework....56

4. Models' workflow ....57

5. Indicators used to compare LM with ML frameworks....60

6. Availability of data and materials .... 61

Results.... 61

1. Comparison of the performance of LM and ML frameworks .... 61

2. Comparison of the influence of predictor variables ....62

3. Comparison of spatial trends in predictions....64

4. Comparison of temporal trends in predictions across serosurveys....66

5. Residual spatial and temporal correlation ....68

6. Computational aspects....69

Discussion ....69

Concluding Remarks ....72

**Chapter 4: From Serological Surveys to Disease Burden: A Modelling Pipeline for Chagas Disease....73**

Summary....73

Research in context ....74

Introduction....75

Material and Methods ....76

1. The DICTUM platform....76

2. Step 1: the FoI at the serosurvey level (catalytic model) ... 79

3. Step 2: Predictive model for the FoI at the municipal level ... 79

Predictors included ... 79

Model definition ... 80

4. Step 3: from FoI to burden ... 81

5. Data sharing ... 83

Results ... 83

1. FoIs at the serosurvey level ... 83

2. FoI Predicted yearly FoI at the municipal level ... 83

3. Burden model over 1985-2020 ... 87

Discussion ... 90

Chapter 5: Discussion ... 94

Summary of findings ... 94

Future work and limitations ... 96

1. Extension of the modelling pipeline to Chile and Paraguay ... 96

2. Validation of the models ... 97

3. Development of Models built at a very fine scale ... 98

4. Consideration for the effect of the treatment in the model ... 99

5. Consideration for the specificity of large cities in the models ... 100

6. Development of burden model for the indigenous setting ... 101

Implications of research ... 102

Challenges ... 103

Conclusions ... 105

References ... 106

Appendix ... 122

Supplementary Tables ... 124

Supplementary Figures ... 132

Supplementary method ... 144

# List of Figures

## Chapter 1

Figure 1. 1: Tripasonoma cruzi. among red blood cells. ........ 19

Figure 1. 2: Triatoma sanguisuga, one of the 138 Triatomine species. ........ 19

Figure 1. 3: Trypanosoma cruzi life cycle according to CDC (extracted from (1)). ........ 21

Figure 1. 4: Chagas disease, transmission by the principal vector ........ 25

Figure 1. 5: Map of Colombia ........ 28

## Chapter 2

Figure 2. 1: Comparison of the predictive ability of the best-fit models for the three approaches investigated ........ 41

Figure 2. 2: Goodness-of-fit of the model averaging of the 3 modelling approaches for all serosurveys ........ 42

Figure 2. 3: Force-of-Infection of Chagas disease in urban, rural and indigenous settings, Colombia, 1990 ........ 43

Figure 2. 4: Spatiotemporal trends in Chagas disease Force-of-Infection, Colombia, 1980–2010 ........ 46

Figure 2. 5: Predictors included in the model averaging of the FoI of Chagas disease in Colombia ........ 48

## Chapter 3

Figure 3. 1: Graphical representation of the two modelling approaches used for each of the frameworks tested ........ 58

Figure 3. 2: Description of the modelling workflow for the Linear Models (LM) and the Machine Learning (ML) frameworks. ........ 59

Figure 3. 3: Spatial distribution of the Force-of-Infection of Chagas Disease (per year and per susceptible individual), in Colombia ........ 65

Figure 3. 4: Comparison of predicted Chagas disease Force-of-Infection (FoI) values for urban or rural settings at municipality level, in Colombia for the year 1990. ........ 66

Figure 3. 5: Standardised comparisons of 'observed' and predicted distributions across serosurveys and by setting type. ........ 67

Chapter 4

Figure 4. 1: Modelling pipeline, from local serosurveys to National burden of diseases' estimates ....78

Figure 4. 2: Chagas disease burden model ....82

Figure 4. 3: Catalytic model fit for the survey with, panel a: the largest sample size (n=1,680); and panel b: the smallest sample size (n=30) ....83

Figure 4. 4: Spatial distribution of the Force-of-Infection of Chagas Disease (per year and susceptible individual), in Colombia at the level of municipalities in 1995 and 2020. ....86

Figure 4. 5: Spatial, temporal and age class distribution of the prevalence of Chagas disease in Colombia. ....87

# List of Tables

Chapter 2

Table 2. 1: Advantages and disadvantages of the three approaches investigated. Approach 1: (A1) models fitted with median FoI estimates and selected based on predictive R2; Approach 2 (A2): models fitted with median FoI estimates and selected based on predictive R2 and overlap; Approach 3 (A3): models fitted with the full posterior distribution of FoI estimates and selected based on the predictive R2 and overlap. ........40

Table 2. 2: Predicted FoI averaged across all Colombian municipalities in 1980, 1990 and 2010, the percentage of decrease between 1980 and 2010 (trend) for each setting and the spatial clustering effect given by the Moran's I statistic for the test under randomisation in 1980, 1990, 2000 and 2010 (n=1065 municipalities). ........45

Chapter 3

Table 3. 1: Median cross-validation performance values for the two Machine Learning modelling methods investigated. ........62

Table 3. 2: Standardized relative influence, importance and rank of the predictors included in Boosted Regression Trees (BRT) and Random Forest (RF) Machine Learning models and normalised number of times the predictors were used in the linear model (LM) framework and their rank when using the full posterior distribution of FoI estimates. ........63

Table 3. 3: Uncertainty across Chagas disease Force-of-Infection predictions for the three frameworks under comparison. ........68

Table 3. 4: Spatial and temporal correlation test statistics and statistical significance of the spatial correlation test applied to the cross-validation residuals for the two Machine Learning (BRT, RF) and the Linear Model (LM) methods under consideration. ........69

Table 3. 5: Comparison of computational aspects for the Machine Learning (Boosted Regression Trees (BRT), Random Forest (RF)) and Linear Model (LM) methods investigated. ........69

Chapter 4

Table 4. 1: Importance of the predictors used in the Random Forest model used to predict the FoI of Chagas disease in Colombia at the municipal level. ........85

Table 4. 2: Burden of Chagas disease in 1995 and 2020. ........89

# List of abbreviations and acronyms

BRT: Boosted Regression Trees

CI: confidence Interval

CDC: Centers for Disease Control and Prevention

CrI: Bayesian Credible Interval

CV: Cross-validation

DANE: Colombia's Department of Statistics

DICTUM: Decreasing the Impact of Chagas Disease Through Modelling

DW: Durbin-Watson (statistic)

ELISA: enzyme-linked immunosorbent assay

FoI: Force-of-Infection

FullPostFoi: full posterior distribution of FoI estimates

GADM: Database of Global Administrative Areas

GBM: Global Burden Model

*I*: Moran's I (statistic)

*Ind*: performance indicator

LM: Linear Model

MAD: Median Absolute Deviation

MAD-CV: Median Absolute Deviation Coefficient of variation

MedFoI: Median FoI

ML: Machine Learning

NDVI: Normalized Difference Vegetation Index

NTD: Neglected Tropical Disease

PAHO: Pan-American Health Organization

PCR: Polymerase chain reaction

$R^2$: coefficient of determination

RF: Random Forest.

WHO: World Health Organisation

# Chapter 1: Introduction

## Chagas disease causes, symptoms, and treatment

Chagas disease is a Neglected Tropical Disease (NTD) caused by a protozoan parasite, *Trypanosoma cruzi*, and transmitted to humans by hematophagous vectors of the subfamily of the Triatominae insect (Figures 1.1 and 1.2) (1,2). *T. cruzi* has a high genetic diversity and is classified into six near-clades. This diversity is reflected by the large geographical extent of the parasite, i.e., it has colonised different types of habitats in Latin America through infection of a variety of triatomine species and mammal hosts, with some differences in pathogenicity and symptoms (2).

Triatomines become infected by the parasite through blood meals on infected animals, including humans. The parasites develop in the triatomine's digestive system and are excreted with faeces, which typically occur while the triatomine is feeding. The parasites, in the infected faeces, may be introduced into the body of the new host when he scratches the site of the bite (Figure 1.3) (2). Aside from this main mode of transmission, other transmission routes are possible and contribute to the spread of the disease. As the parasites circulate in the blood of the host, vertical transmission, from mother to foetus, is possible, as well as via organ transplant or blood transfusion. Finally, new infections via direct ingestion of the parasite, from contaminated food, has also been identified (3).

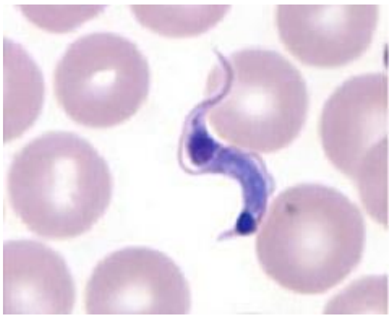


*Figure 1. 1: Tripasonoma cruzi. among red blood cells.* *Credit CDC (extracted from (1))*

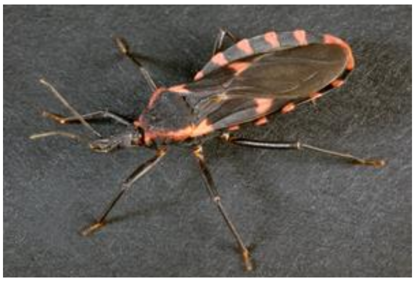


*Figure 1. 2: Triatoma sanguisuga, one of the 138 Triatomine species.* *Credit: CDC, courtesy of James Gathany (extracted from (1))*

In humans, the disease develops and can be characterised into phases. Shortly after infection, the acute phase starts and can last 4 to 8 weeks (4). Often asymptomatic, some mild symptoms can be observed like fever, inflammation or oedema around the bite location. Severe symptoms occur in less than 5% of cases and are characterised by cardiac manifestations (acute myocarditis, pericardial effusion) or meningoencephalitis (2). The acute phase seems to be more severe when the transmission has occurred by ingestion of the parasite (oral transmission route), e.g., when food or beverages have been contaminated by the infected triatomine or triatomines' faeces. For severe acute cases contaminated orally, the risk of mortality may increase from 5-10% to 8-30% which is most probably related to the number of parasites entering the body, i.e. more than 600,000 units through the oral route while it is estimated between 3,000 and 4,000 units through the vectorial transmission route (5). The second stage of the disease is the indeterminate phase, which is asymptomatic, and chronic. It has traditionally been defined as a complete absence of clinical manifestations that can last 10-30 years but recent studies have found cardiac or digestive abnormalities that are more complicated to detect and an excess of mortality compared to non-infected people (2,6). The third phase is the chronic symptomatic stage and the progression rate to this phase is 1.5%-5.0% per year (2,7). Among all infected persons, cardiovascular symptoms leading to chronic heart failure or sudden deaths are observed in 25% of cases, digestive symptoms (mega-colon and mega-oesophagus) in 6% of cases and peripheral nervous involvement in 3% of cases (4). Cardiac symptoms are the most frequent and the most serious Chagas disease symptoms. Patients at an early stage (Chronic mild stage) will suffer from fatigue, dizziness, palpitations, shortness of breath that might have a limited impact on their daily life (2,8). When patients reach the chronic severe phase, the cardiac symptoms, i.e., dyspnoea, left ventricular dysfunction and congestive heart failure, are causing large morbidity as daily physical activity is compromised and medical care is required (2,8). While the physiological causes of the symptoms of the acute phase are well understood as they mainly rely on the inflammation processes, the physiological process involved in the indeterminate and chronic phases have not been fully elucidated. After entering the body, the parasite is nesting in cardiac, skeletal and smooth muscles tissues creating an inflammatory response and damages in these tissues. Recent consensus states that the progression of the disease might be related to a balance between human immune response and the parasite persistence in the tissues which create a chronic inflammation (2).

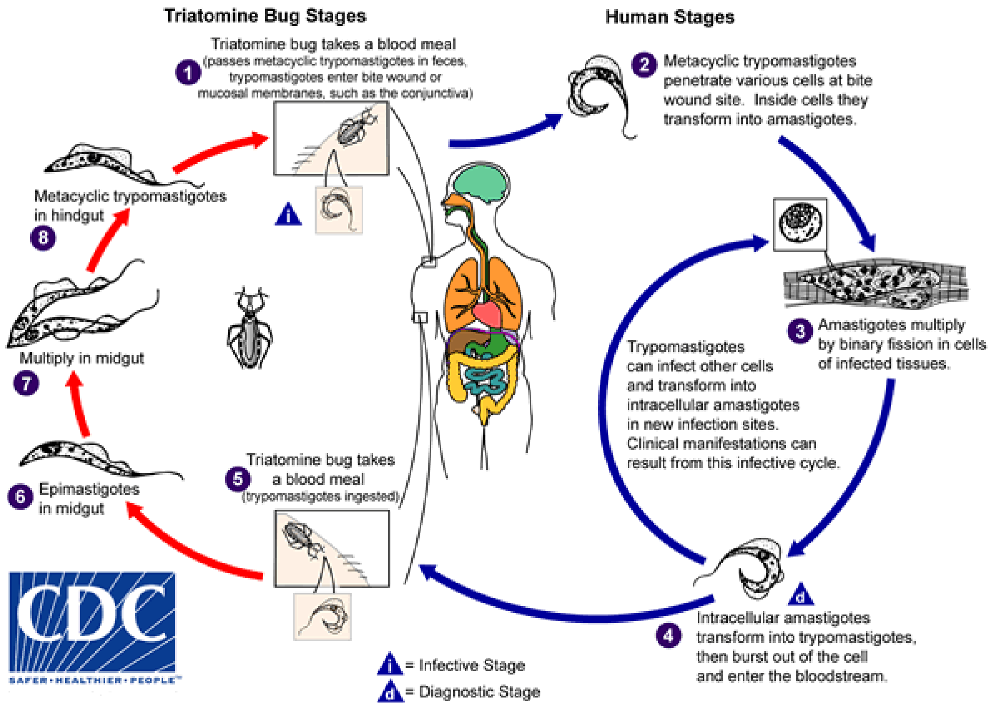

Triatomine Bug Stages
Human Stages
1 Triatomine bug takes a blood meal (passes metacyclic trypomastigotes in feces, trypomastigotes enter bite wound or mucosal membranes, such as the conjunctiva)
2 Metacyclic trypomastigotes penetrate various cells at bite wound site. Inside cells they transform into amastigotes.
3 Amastigotes multiply by binary fission in cells of infected tissues.
Trypomastigotes can infect other cells and transform into intracellular amastigotes in new infection sites. Clinical manifestations can result from this infective cycle.
4 Intracellular amastigotes transform into trypomastigotes, then burst out of the cell and enter the bloodstream.
5 Triatomine bug takes a blood meal (trypomastigotes ingested)
6 Epimastigotes in midgut
7 Multiply in midgut
8 Metacyclic trypomastigotes in hindgut
i
d
i = Infective Stage
d = Diagnostic Stage
CDC
SAFER·HEALTHIER·PEOPLE™

*Figure 1. 3: Trypanosoma cruzi life cycle according to CDC (extracted from (1)). "An infected triatomine insect vector (or "kissing bug") takes a blood meal and releases trypomastigotes in its faeces near the site of the bite wound. Trypomastigotes enter the host through the wound or through intact mucosal membranes, such as the conjunctiva* ❶. *Common triatomine vector species for trypanosomiasis belong to the genera Triatoma, Rhodnius, and Panstrongylus. Inside the host, the trypomastigotes invade cells near the site of inoculation, where they differentiate into intracellular amastigotes* ❷. *The amastigotes multiply by binary fission* ❸*and differentiate into trypomastigotes, and then are released into the circulation as bloodstream trypomastigotes* ❹. *Trypomastigotes infect cells from a variety of tissues and transform into intracellular amastigotes in new infection sites. Clinical manifestations can result from this infective cycle. The bloodstream trypomastigotes do not replicate (different from the African trypanosomes). Replication resumes only when the parasites enter another cell or are ingested by another vector. The "kissing bug" becomes infected by feeding on human or animal blood that contains circulating parasites* ❺. *The ingested trypomastigotes transform into epimastigotes in the vector's midgut* ❻. *The parasites multiply and differentiate in the midgut* ❼*and differentiate into infective metacyclic trypomastigotes in the hindgut* ❽ *." (1)*

While all phases of the diseases are associated with excess mortality (6), most of the infected persons have no access to diagnosis and treatment (9). Moreover, once diagnosed, access and adherence to the treatment is a real challenge: treatment is long, require many visits to the health facilities, and  is associated with many side effects (3). Research on Chagas disease treatment has been slow over the last five decades and only two antiparasitic drugs are currently available (benznidazole and nifurtimox). While these medicines have shown their effectiveness to avoid vertical transmission, i.e. from mother to child during pregnancy, when treating women of child-bearing age, they do not appear to significantly improve the conditions of patients with mild and severe cardiomyopathies (2). The efficacy of the treatment seems to decrease with time from the infection. The duration of the treatment is between 60 and 90 days for an adult, and associated side effects include anorexia, digestive intolerance, hypersensitivity, or psychological disorders (irritability, insomnia, disorientation). These side effects occurred in 43-96% of the cases with nifurtimox but are a little bit lower with benznidazole, and represent a major cause for treatment discontinuations, i.e. 15-75% of the cases for nifurtimox and 9-29% of the cases with benznidazole (2).

Regarding Chagas disease diagnostic, direct observation by microscopy is only possible during the acute phase (10). During the indeterminate and chronic phases, serological tests are preferred as PCR suffer from a low sensitivity ranging between 50% and 70% (10). ELISA, indirect immunofluorescence, and indirect haemagglutination are the tests the  most often used to detect the parasite among asymptomatic with a screening process typically involving multiple testing for confirmation (10).

## Chagas disease control

Chagas disease etiology and epidemiology have been first described in 1909 by Carlos Chagas (11). He rapidly identified that prevention relying on vector control was the best strategy to break the epidemiological cycle of the disease (12). Some triatomine vectors, e.g., *T. infestans*, are very efficient in colonizing human dwellings, hiding and nesting in house cracks and roofs during the day and feeding on humans during the night. Poor housing conditions were then identified as a major risk of house infestation and thus of Chagas disease infection (12). As a result, most of the efforts made to fight the disease have focused on entomological research with a first map of the triatomine species distribution across Latin America realized in 1919 (12).

It appeared that only three genera of triatomines can transmit *T. cruzi* (2) and that most of them are sylvatic, i.e. living in the forest while others species may live in peri-domiciliated areas, mainly in palm trees (13), and few are 'domiciliated'. The peri-domiciliated vectors do not colonize houses but can enter the house at night to feed. Historically, in term of transmission to humans, the most problematic species are the domiciliated, i.e., able to colonize human dwellings, such as *Triatoma infestans*, and, to a lesser extent, *Triatoma dimidiata* and *Rhodnius prolixus*.

Currently, Chagas disease control programs mainly rely on house improvement and insecticide spraying in houses, annexes and other types of building to eradicate the vector in domicile and peri-domicile areas (4,14). This strategy has shown its effectiveness, especially in contexts where the vectors are domiciliated, i.e. the vectors are living inside the buildings (4), but less for peri-domiciliated vectors (15). Improving awareness and engagement with communities affected has also been part of the programs in 12 of the endemic countries (4). Human socioeconomic factors play an important role to explain the risk of infection  as house infestation is more likely to happen in poorly built dwellings (16). This has led to an association between Chagas disease, poverty, and rurality, and, altogether, contributes to the stigmatisation of people having their house infested or being themselves infected by the parasite. The rural nature of the disease is however being questioned, with recent research in Peru showing that the vectors have started to settle in urban areas (17). In addition, human movements between rural and urban locations (short and long-term), as well as progressive urbanisation have contributed and continue to contribute to significant level of infection in urban environment, leading to continued bloodborne and congenital transmission, and substantial level of infection observed during blood screening in blood banks (screening being compulsory in most Latin American countries) (4,18).

As the epidemiology of the disease and the clinical tools available are evolving and progressing (e.g., success of past vector-control program, improved diagnostics), the relative balance in cost-effectiveness between prevention and improved diagnostics and treatment is likely changing. In the next phase of the fight against Chagas disease, the identification of cases before they reach the symptomatic stages of the disease is likely to become critical to treat before permanent damages are caused by the parasites. Asymptomatic cases are therefore a key population as up to 70% of them can stay asymptomatic for more than 30 years (19). Identifying those populations would help reducing instance of bloodborne and congenital transmissions. However, progresses on the treatment therapies available to cure Chagas disease still have to be made to ensure treatment access and adherence by the patients.

## Chagas disease elimination challenges

In 2012, the World Health Organization (WHO) set the elimination of Chagas disease transmission as a goal in its first neglected tropical disease roadmap (20). Currently, between 5 and 18 million persons are estimated to be infected and the disease causes an estimated 10,000 deaths per year in the 21 endemic countries in Latin America (Figure 1.4) (9,14). While scarce and highly uncertain, reliable estimates of the spatial and temporal pattern of the burden are essential for the governments and Health organisations coordinating the fight against the disease as it can inform targeted and cost-effective vector-control and screening interventions, as well as help monitor their needs for diagnosis and treatment supplies (21). The key challenges in estimating the burden of the disease include the lack of estimates of incidence trends due to the lengthy asymptomatic phase (14), the high level of spatial heterogeneities, as well as the potential for rapid changes in temporal patterns.

The surveillance systems for the disease, when present, suffer from severe underreporting (14) e.g. in Colombia in 2021, 306 chronic and 172 acute cases have been reported, with only 170 and 14 of them being confirmed respectively (22,23). Meanwhile, estimations of the number of cases from WHO, Global Burden Model (GBM) and other authors ranged widely between 186,000 and 438,000 for the 2005-2010 period (4,8,24,25).

In the absence of incidence data, evidence of past exposure through seroprevalence surveys are often used to quantify the epidemiological situation. However, as Chagas disease is a long-lasting disease, prevalence data are not reflecting the current epidemiological situation, with high prevalence potentially reflecting a high level of past transmission but little to no current circulation of the parasite. To overtake this challenge, the use of prevalence in children under 5 years old have been identify as a strategy to observe the current transmission (21). However, it requires a large number of participants as the prevalence in younger age classes is expected to be lower, e.g., they had less time to get infected.

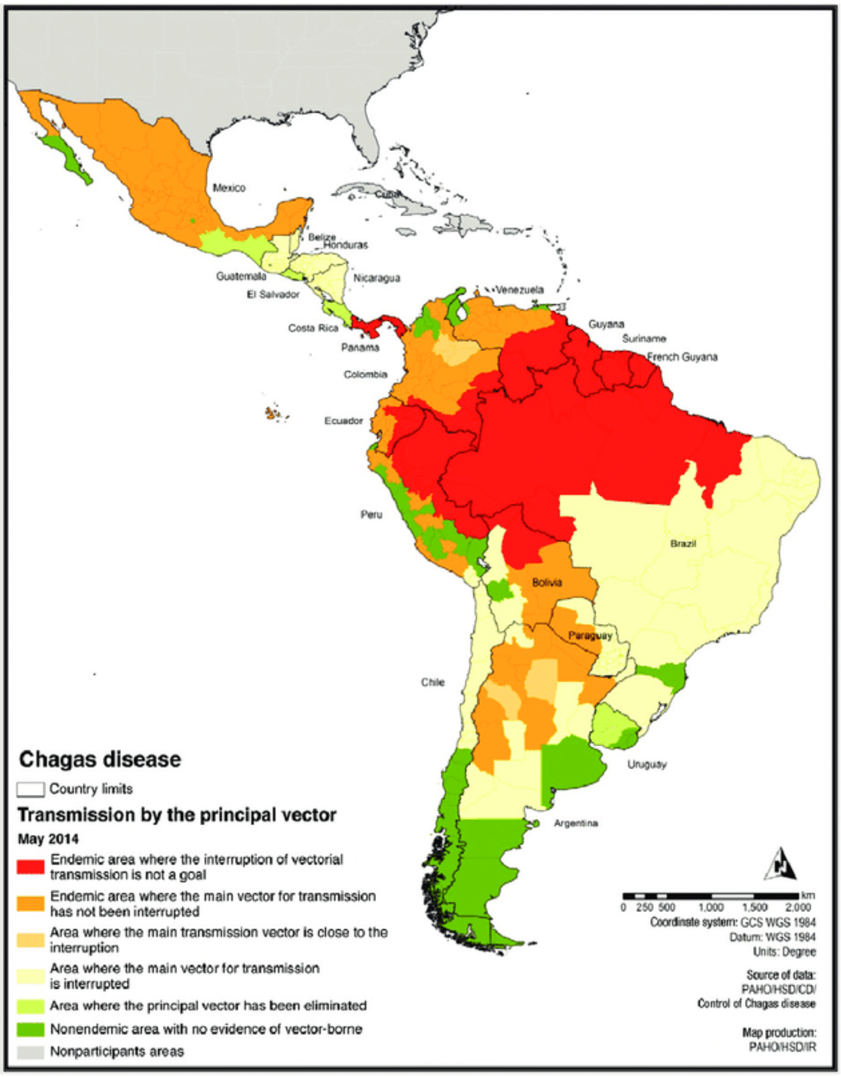

Mexico
Cuba
Belize
Honduras
Guatemala
El Salvador
Nicaragua
Costa Rica
Panama
Venezuela
Guyana
Suriname
French Guyana
Colombia
Ecuador
Peru
Brazil
Bolivia
Paraguay
Chile
Uruguay
Argentina
Chagas disease
Country limits
Transmission by the principal vector
May 2014
Endemic area where the interruption of vectorial transmission is not a goal
Endemic area where the main vector for transmission has not been interrupted
Area where the main transmission vector is close to the interruption
Area where the main vector for transmission is interrupted
Area where the principal vector has been eliminated
Nonendemic area with no evidence of vector-borne
Nonparticipants areas
N
0 250 500 1.000 1.500 2.000 km
Coordinate system: GCS WGS 1984
Datum: WGS 1984
Units: Degree
Source of data:
PAHO/HSD/CD/
Control of Chagas disease
Map production:
PAHO/HSD/IR

*Figure 1. 4: Chagas disease, transmission by the principal vector, credit: PAHO, 2014. The colours on the map characterise the status of the vectorial transmission in the area: in red, areas where the vectorial transmission can occur and interruption of the transmission is not a goal; in dark orange, areas where the vectorial transmission occur but elimination effort are realised; in light orange, areas that are close to elimination of the vectorial transmission route; in yellow, areas where the vectorial transmission have been interrupted; in light green, areas where the principal vector have been eliminated; and in dark green, non-endemic areas.*

## Opportunity raised by modelling

The presence of the vectors has been traditionally used to assess the risk of Chagas disease. As a result, most of the studies interested in the spread and distribution of Chagas disease modelled Triatomines ((16,26–31) or (32)). Environmental variables also play an important role as they define the ecological niche of the vector and can as well impact the infectiousness of the parasite itself (33). They were used to identify the potential geographical distribution of the triatomine species of interest (15,27,34–48), but also to assess the impact of climate change (29,49–53), urbanisation (17,54,55), deforestation (56), and other environmental changes (28,55,57–60) on the risk of Chagas disease transmission.

When modelling the transmission dynamics of an infectious disease, compartmental models, such as Susceptible and Infected (SI) models are typically used and this also applied to Chagas disease with models explicitly characterising the dynamics within and between hosts and vector populations (32). However, these models rely on ecological parameters, such as triatomine feeding behaviour, transmission rates by blood meal or survival of the parasite in the host, that needed to be estimated and as research on Chagas disease is limited these models might lack robustness (32). These parameters are then associated with different levels of uncertainty that need to be considered. First, the lack of research to characterise these parameters values increase uncertainty. Then, the diversity in potential host and vector creates high heterogeneities due to context-specific dynamics. These aspects hindered the widespread use of compartmental model to study Chagas disease transmission and make them informative about general dynamics but not the most suitable to inform policies.

Another source of information used by modelling studies to inform prevention and control are seroprevalence surveys organised by government and NGOs. Those are typically conducted to estimate the local prevalence by age class, and represent, in our view, an under-exploited resource for the modelling of Chagas disease. They have already proved useful to reconstruct past and present incidence patterns of Chagas disease (21). Using a catalytic model fitted to age-structured seroprevalence data, a seroprevalence survey is used to estimate the local Force-of-Infection (FoI), i.e., the rate of parasite acquisition, and its temporal heterogeneities (61–79). However, even if several serological surveys are available, the FoI estimates remain limited in space, making spatially resolved national estimates difficult to obtain. To a lesser extent, estimates of FoI from serosurveys also remain limited in time as they can only inform on exposure up to the time of the serosurvey and back to the time of birth of the oldest participants. As a result, such methods are typically used to characterise local dynamics but not to draw predictions at large spatial and temporal scales.

To be able to obtain FoI estimates at the national scale, predictive models can be applied to local FoI estimates from serosurveys. Such methods have already been used in several contexts such as dengue (66,80) or Yellow Fever (61). However, important assumptions have been made on the FoI to use it as a response variable in a predictive model. The FoI, and therefore exposure, is often assumed to be constant over time and age classes (66,78,80). If for Chagas disease the intensity of the transmission is not known to be influenced by age, the transmission intensity is likely changing over time, e.g., linked to the implementations of vector control interventions. Additionally, the uncertainty surrounding the FoI estimation by the catalytic model is generally ignored and only the average estimated FoI's are used to fit the predictive models (61,80). This assumption has not yet been challenged in the literature as the incorporation of uncertainty in the input of predictive models raises technical challenges.

## Chagas disease in Colombia

In Colombia, the goal of elimination of vectorial transmission was agreed on by 1997 (4). Blood bank screenings were made mandatory in 1995 but became fully effective in 2003 (8). The vector control program has been decentralised at the departmental level with the most affected department being Arauca, Casanare and Santander (4). *Rhodnius prolixus* and *Triatoma dimidiata* are the two main vectors of Chagas disease in Colombia (37). While most of the country is defined as endemic by the WHO (Figure 1.4), the screening coverage, i.e., proportion of the at-risk population having received a test, was estimated at 1.2%. Access to treatment is also limited with only 0.3%-0.4% of infected people having received anti-parasitic treatment (8).

A recent meta-analyses found a pooled prevalence of 2.0% (95%CI: 1.0%-4.0%) based on 12 published studies (19). The highest prevalence was observed in the Orinoco region (in the department of Casanare, Yopal being the capital city (Figure 1.5)) with 7.0% (95% CI: 2.2%–12.6%) but large spatial heterogeneity at the municipal level is observed (19).

In Colombia, 109 serosurveys were conducted over 19 years (1995-2014) which represent around 6 serosurveys by year, 768 persons tested by serosurvey and a geographic coverage of 15 of the 32 departments (21). However, on the scale of the country, this gives a relatively sparse picture of the prevalence level. Furthermore, only 75 of these serosurveys provide age prevalence data at the municipality level, covering 35 of 1122 municipalities.

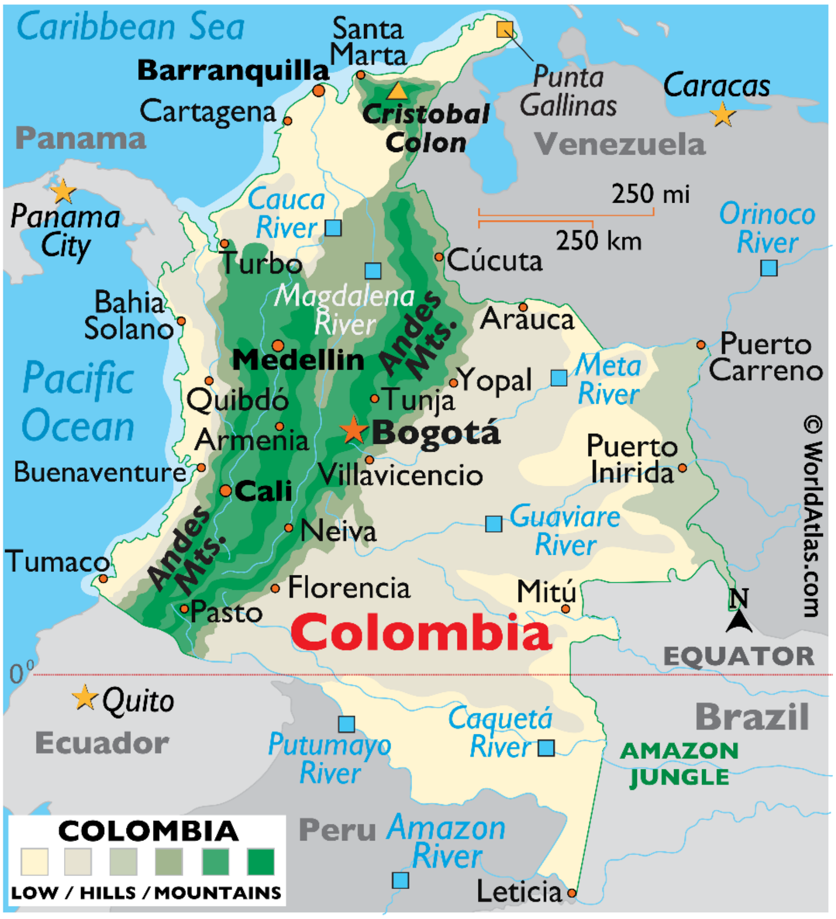


*Figure 1. 5: Map of Colombia*, *credit: World Atlas*

## Objectives and contribution to knowledge

Even if national and international goals have been set and a large-scale intervention implemented, there is a crucial need to measure and monitor the progress made and assess where new efforts should be targeted. In contexts where the vectorial transmission is becoming under control, understanding the impact of interventions as well as local epidemiological patterns is necessary to adjust the fighting efforts to new challenges. In addition, having a better view of the burden of disease can help coordinate medical care and reach people that need them more efficiently.

The general aim of the project was to support governments in their fight against Chagas disease and is part of a larger research consortium that have already collaborated for over 6 years with researchers from the university of Sussex, Imperial College London (UK), Princeton University (USA), the National Institute of Health in Colombia and Columbia University (USA) and with support from the Pan American Health Organization (PAHO). The consortium aims to use disease modelling methods to bring innovative and cost-effective solutions to guide the Chagas disease fight. In this context, information useful for Chagas disease research has been gathered and standardised. These data include entomological surveillance, prevalence recorded in blood banks and prevalence estimates measured through serosurveys. Thanks to cooperation with Colombian, Chilian and Paraguayan authorities, published and unpublished data are being processed. Colombia was the first participating country and has thus been used as a case study in this thesis. My work built on the results and progresses obtained by the consortium and in particular that of Cucunubá *et al.*, where they used serosurvey data to estimate the local FoI.

The aim of my thesis was to develop global geostatistical models of the risk of Chagas disease transmission using the Force-of-Infection as an input along with a set of relevant environmental and socioeconomic predictors to ultimately estimate the spatial and temporal heterogeneities in the burden of Chagas disease.

In particular, the following questions needed to be addressed:

> *What challenges are raised when using the FoI as a response variable and how they can be addressed?*
>
> *What types of biases are present in serosurvey data and what is their impact on the predictions?*
>
> *Are more flexible methods, such as Machine Learning, better suited to handle the challenges raised by the FoI?*
>
> *Is it possible to reliably estimate burden of disease estimates at an operational scale?*

In order to answer these questions, the specific objectives of this thesis project were:

- To develop a methodological framework to use the FoI as a response variable in predictive models. In particular, find how to integrate information, and propagate uncertainty, in the model and how to evaluate model performances.
- To assess the potential of Machine Learning methods in this context.
- To develop a predictive model that can be applied to other countries.
- To optimize the compartmental burden of disease model developed by Cucunubá *et al.* and modify it to work at the municipal level.
- To write programming code and communicate our results to stakeholders with various expertise, e.g., knowledge transfer and scientific collaboration.

The thesis has been then organised into five chapters. The first chapter presents the context of the research. The second chapter aims to improve our understanding of the FoI when used on predictive models. Based on linear model methodologies, a modelling framework has been progressively built to face each of the challenges raised by the use of the FoI as a response variable. It results in a clearer view of the data, including its strengths, weaknesses, and biases. In the third chapter, the framework is optimized using advanced modelling techniques relying on Machine Learning. Results are confronted to assess the benefits and weaknesses of each method. The fourth chapter, relying on previous chapters, present a modelling pipeline that is effective in Colombia and can be adapted without major change to other countries. Finally, the fifth chapter present future work and the challenges that remain.

# Chapter 2: Spatiotemporal variations in exposure: Chagas disease in Colombia as a case study

*Published in BMC Medical Research Methodology on the 13th of January 2022 (DOI: 10.1186/s12874-021-01477-6).*

## Abstract

Age-stratified serosurvey data are often used to understand spatiotemporal trends in disease incidence and exposure through estimating the Force-of-Infection (FoI). Typically, median or mean FoI estimates are used as the response variable in predictive models, often overlooking the uncertainty in estimated FoI values when fitting models and evaluating their predictive ability. To assess how this uncertainty impacts predictions, we compared three approaches with three levels of uncertainty integration. We propose a performance indicator to assess how predictions reflect initial uncertainty.

In Colombia, 76 serosurveys (1980–2014) conducted at the municipality level provided age-stratified Chagas disease prevalence data. The yearly FoI was estimated at the serosurvey level using a time-varying catalytic model. Environmental, demographic and entomological predictors were used to fit and predict the FoI at the municipality level from 1980 to 2010 across Colombia.

A stratified bootstrap method was used to fit the models without temporal autocorrelation at the serosurvey level. The predictive ability of each model was evaluated to select the best-fit models within urban, rural and (Amerindian) indigenous settings. Model averaging, with the 10 best-fit models identified, was used to generate predictions.

Our analysis shows a risk of overconfidence in model predictions when median estimates of FoI alone are used to fit and evaluate models, failing to account for uncertainty in FoI estimates. Our proposed methodology fully propagates uncertainty in the estimated FoI onto the generated predictions, providing realistic assessments of both central tendency and current uncertainty surrounding exposure to Chagas disease.

## Significance statement

Estimating spatiotemporal variation in disease exposure is critical to developing cost-effective strategies to reduce disease burden. However, where there is no well-established surveillance system, it might be challenging to obtain such information. Serosurveys provide information on past exposure at a certain location but do not reflect the current situation, particularly for long-lasting diseases such as Chagas disease. The FoI provides insight into the temporal patterns of the disease and is particularly relevant for assessing spatiotemporal heterogeneities and interventions' impacts. However, assessing incidence over countries and decades, when seroprevalence information remains limited, requires robust statistical methods. We developed a modelling framework that predicts FoI in space and time from serosurveys able to propagate uncertainties using Colombia as a case study.

## Introduction

Between 5 and 18 million persons are estimated to be currently infected by *Trypanosoma cruzi*, the protozoan parasite causing Chagas disease, and between 4,200 and 33,000 per year are estimated to die in the 21 endemic countries in Latin America (14,81). These figures give a coarse picture of the epidemiological situation, which is problematic as reliable estimates of the spatial and temporal patterns of the disease burden are essential for governments and health organisations to assess progress towards control or elimination goals. Indeed, spatial estimates of exposure are critical to target vector control activities. Additionally, the current clinical burden depends on past exposure as people infected by *T. cruzi* may develop a chronic form of the disease, requiring long-term care. Temporal estimates of exposure to *T. cruzi* are essential to monitor diagnostic and treatment needs (21), and ultimately to coordinate intervention strategies (e.g. targeted vector control and screening interventions). Finally, temporal patterns in exposure can also be used to evaluate past control interventions and guide future planning.

Estimating the burden of Chagas disease is challenging; there are no reliable measures of incidence, for example, in Colombia, only an estimated 1.2% of the at-risk population received a screening test in 2008–2014 (3). The low level of detection is partly linked to the unspecific nature of early symptoms and the long-lasting asymptomatic period, i.e. asymptomatic or unspecific symptoms can last for over 10 years and around 50% of those infected may never reach the chronic phase (14). Moreover, the disease affects disproportionately poorer populations with limited access to the health system (12).

As demonstrated for other infectious diseases with a relatively low proportion of symptomatic cases, burden estimates typically rely on exposure estimates, particularly the Force-of-Infection (FoI), i.e. the per-susceptible rate of parasite acquisition (21). Seroprevalence surveys are typically used to reconstruct past and present incidence patterns in various locations and a geostatistical model smooths the estimated FoI over space (61,80).

Where this framework has been applied, given the complexity of the inference and relative scarcity of ground-truth data, it is common to assume that exposure has been constant over time. Although this may hold for FoI estimates for dengue (66,78,80), yellow fever (61), rubella (71,75) and malaria (67), it is more challenging for Chagas disease, as its protracted nature and substantial spatial and temporal heterogeneities in the implementation of control measures lead to temporal and spatial heterogeneities in exposure.

Additionally, predicting FoI spatial patterns relies upon point estimates of FoI, with geostatistical models smoothing the central estimates (61,66,80), often neglecting their uncertainty. This may generate over-confidence in FoI estimates and ultimately burden. Generating FoI and disease burden estimates that robustly incorporate uncertainty is essential to inform policy-relevant questions, from affected communities to stakeholders and policy-makers (82).

Here, we propose a framework to predict spatial as well as temporal variations in FoI that fully account for uncertainties at various levels, particularly, the uncertainty in estimated FoI. The framework is applied to 76 *T. cruzi* serosurveys in Colombia to obtain estimates of exposure across Colombia from 1980 to 2014 at the municipality level. The importance of propagating uncertainty in estimated FoI and its impact on model selection and prediction was then quantified.

# Methods

## 1. General approach

Our general aim is to predict the FoI at the municipality level across Colombia using data from 76 serosurveys (27 urban, 36 rural, 5 indigenous and 8 mixed as defined by the Colombian government) conducted between 1980 and 2014 (Supp. Figure 1 and Supp. Figure 2). Environmental, demographic and entomological predictors were available for each location. For each serosurvey, the full posterior distributions of the FoI were obtained using a catalytic model (21). As a serosurvey reflects exposure since the birth of the oldest participant, estimated FoIs include past and contemporary (to the serosurvey) estimates of FoI. The potential predictors included in the models were selected based on expert knowledge and preliminary analyses (Supp. Table 1 presents the full list of predictors considered). Log-linear models were fitted using a combination of these predictors. Due to temporal autocorrelation, a stratified bootstrapping was applied to fit the models using single-year FoI estimates (randomly chosen at each iteration). To avoid overfitting, a repeated random sub-sampling validation was applied by selecting multiple times and randomly using half of the serosurveys for either training or validation. The predictive ability of each model (i.e., central estimate across the out-of-sample sets) was then evaluated to select the best models within urban, rural and indigenous settings. Finally, model averaging, with the 10 best models identified in the 3 different settings studied, was used to produce predictions of FoI as described by (83).

Typically median, or mean, FoI estimates are used as the dependent variable (61,66,80); however, ideally, the uncertainty in estimated FoIs should be accounted for when fitting the models and evaluating their predictive ability. To assess how this uncertainty impacted predictions, we compared three approaches incorporating different levels of uncertainty:

A1. Central estimates of FoI are used, i.e., no uncertainty is accounted for as commonly used in the literature. The selection of the best model is based on the central trends.

A2. Uncertainty in estimated FoI is used to quantify the model's predictive ability but not for fitting. For a given model, the predictions remain the same. This approach potentially changes which models are selected as the best ones based on a more realistic measure of predictive ability.

A3. Uncertainty in estimated FoI is used for both fitting and quantifying the model's predictive ability. Models are fitted and evaluated repeatedly using samples of the FoI posterior distribution leading to changes in both the predictions for a given model and which models are selected as the best.

The uncertainty of the predictions was characterised using a coefficient of variation based on the Median Absolute Deviation (MAD-CV) accounting for the non-normality of the FoI distribution (84). A3, although computationally more intensive, appropriately propagates the uncertainty in FoI estimates in both the predictions and the model selection processes.

### 2. Data input

*Chagas Disease Force-of-Infection*

From the 112 Chagas disease serosurveys conducted in Colombia, only 76 serosurveys were selected, where the catchment area was smaller than the municipality level. Indeed, serosurveys having a catchment area at the departmental level have been excluded to be able to run analyses at the municipality level. The Force-of-Infection (FoI) is the per-susceptible rate of parasite acquisition (21) and had been estimated using Bayesian inference (to account for diagnostic uncertainty) for all those 76 age-stratified serosurveys (21). Thus, for each serosurvey, we extracted the full posterior distribution of the estimated annual FoI from the year of birth of the oldest participant up to the year the serosurvey was conducted. The median and the 95% Bayesian Credible Intervals (CrI) were then extracted from the posterior distribution. The methodology used to calculate the FoI has been described elsewhere (21) and relies on estimating time-varying FoI based on catalytic models (85) (see SI for more details).

*Potential explanatory variables*

For each covariate, the geographical scale of interest was the municipality (ADM2) level when available or the departmental (ADM1) level, otherwise. The pool of variables tested related to both human population and environmental conditions (Supp. Table 1).

The *Trypanosoma cruzi* seroprevalence in public blood banks by year and department was provided by the Pan American Health Organization (PAHO). The presence of *Triatoma dimidiata* and *Rhodnius prolixus* at the municipality level was obtained after combining records from a national surveillance report of 2013 (86) and data from (37,46). We also extracted data on presence/absence of these two vector species, from which the proportion of municipalities infested for each department was calculated. Data on vector control interventions implemented in Colombia (1998–2014) were extracted from (87). Census data were obtained from the Colombia's Department of Statistics (DANE) website (88). Climate variables were extracted from the Köppen-Geiger climate classification maps at a 1-km resolution (89). Finally, the map layer used was obtained from the Database of Global Administrative Areas (GADM) (https://gadm.org/ (90)).

Other covariates included the setting of the survey (urban, rural, indigenous, or mixed population (including urban, rural and unknown settings); the year when the survey was conducted; an effect for years and decades (full details in Supp. Table 1). Indigenous settings comprised those with Amerindian populations mostly following traditional lifestyles as described in (21). Definitions for urban and rural populations followed the Colombian government criteria (88).

### 3. Model selection strategy

Due to temporal autocorrelation in estimated FoI, a stratified bootstrapping was applied to fit log-linear models using single-year FoI estimates (randomly chosen at each iteration).

To avoid overfitting, the method of Leave-p-out cross-validation (with p=50%), while ultimately ideal, was unpractical given the computational cost. Instead, we used a repeated random sub-sampling validation by selecting multiple times and randomly half of the serosurveys for either training or validation. As the number of random splits increases, the repeated random sub-sampling validation results approach the exhaustive Leave-p-out cross-validation. We used 10,000 splits to ensure convergence. The variation in the first and second 5,000 out-of-sample predictive $R^2$ values for the 10 best models varied by less than 1% in rural and urban settings (3% for indigenous settings).

A total of 464 models, combining 27 covariates (including some 2-ways interactions), were evaluated using the above procedure. The combinations of variables have been built to be plausible and to reduce predictor collinearity. Thus, predictors representing the same information in different forms (ie, presence of *Rhodnius prolixus* and proportion of municipalities where *R prolixus* is present) were not included together in the same model. For each model, the parameters were estimated using data from all settings (urban, rural, indigenous), but predictive performance (see below) was evaluated separately for each setting. For each setting, a model-averaging method (83) was used to account for structural uncertainties based on the 10 best models in each setting. Models' weights based on predictive performance (see below) were used to obtain model-averaged predictions and maps.

### 4. Modelling approaches and predictive performance

We used 2 predictive performance indicators:

- The standard predictive (out-of-sample) $R^2$ (91) (Eq. 1),

$$Predictive\ R^2 = 1 - \frac{\sum(y_i - \hat{y_i})^2}{\sum(y_i - \bar{y})^2} \qquad \text{(Eq. 1)}$$

- An overlap indicator estimating the percentage overlap between observed and predicted distributions (using the R-package *overlap* 1.5.4. (92)).

The predictive $R^2$ compares the central estimate of the prediction against observations. The overlap indicator compares the full distribution of the predictions against the full distribution of the observations. Therefore, while the overlap indicator quantifies well the predicted uncertainty, the predictive $R^2$ focuses on the central trend in observations and predictions. Model selection relied on an average of both indicators and models' weights were adapted from (83) (Eq. 2),

$$w_i = \frac{e^{(-\frac{1}{2}(\max(Ind) - Ind_i))}}{\sum_{r=1}^{R} e^{(-\frac{1}{2}(\max(Ind) - Ind_r))}} \qquad \text{(Eq. 2)}$$

With $R$ being the total number of candidate models (here 10) and $Ind$ the performance indicator.

Three modelling approaches were compared which differ in how much uncertainty in estimated FoI is accounted for while i) fitting the model and ii) assessing its predictive performance (for model selection). The 3 approaches were:

- Approach 1 (A1): only the median FoI estimates were used as the response variable, i.e., no uncertainty is used (a common approach in the literature). In this approach, as the response variable is characterised by its median, only the predictive *R*² was used to select the best models. For comparison, the overlap indicator for each model was retrospectively estimated but not used.
- Approach 2 (A2): Uncertainty in estimated FoI is used to quantify the model's predictive ability but not for fitting. In this approach, while only the median FoI is used for fitting, both the predictive *R*² and the overlap indicator are used (averaged) to select the best models.
- Approach 3 (A3): Uncertainty in estimated FoI is used when both fitting and quantifying model's predictive ability. Each model is repeatedly fitted to the posterior samples of the estimated FoI, and the predictive *R*² and the overlap indicator are used (averaged) to select the best models.

### 5. FoI prediction for the entire country

The model average built for each setting was then used to generate FoI estimates in each municipality of Colombia for the years 1980, 1990, 2000 and 2010. The median FoI and its uncertainty were extracted. The uncertainty was characterised using a standardised coefficient of variation calculated using the standardised Median Absolute Deviation (MAD-CV) because the FoI values were not normally distributed (84).

### 6. Comparing observations and predictions across serosurveys

For each serosurvey, we compared, across years, the median and 95%CI (Confidence Interval) of the predicted FoI against the median and 95%CrI of the originally estimated FoI (21) (i.e. the dependent variable or 'observed' FoI).

For each quantile of interest $q_x$ (i.e., median, 2.5%, and 97.5% percentiles, denoted $q_m$, $q_l$ and $q_u$ respectively), we computed a distance between the 'observed' and predicted quantile ($\delta_{q_x}$). This distance was standardised by the interval between the observed median and observed upper (or lower) 95% CrI,

$$\begin{cases} \delta_{q_x} = \frac{q_x(\hat{y}) - q_x(y)}{q_x(y) - q_l(y)} & if\ q_x(\hat{y}) < q_x(y) \\ \delta_{q_x} = \frac{q_x(\hat{y}) - q_x(y)}{q_u(y) - q_x(y)} & if\ q_x(\hat{y}) > q_x(y) \end{cases} \qquad \text{(Eq. 3)}$$

When the predicted and 'observed' medians are equal, we expect $\boldsymbol{\delta_{q_m} = 0}$. If the predicted median was equal to the upper (or lower) 95%CrI of the 'observed' FoIs, then we would have $\delta_{q_m} = 1$ ($\delta_{q_m} = -1$).

If the predicted and 'observed' upper (or lower) 95% CI/CrI were equal, then we expect $\boldsymbol{\delta_{q_u}} = 1$ ($\delta_{q_u} = -1$). A value $\delta_{q_u} = 2$ would indicate that the interval between the median and upper CI in the prediction is twice as wide as the interval between the median and upper CrI in the observations.

The change in the denominator reflects the non-symmetrical nature of the 95%CI.

As it is rescaled, this measure of bias allows an assessment of the predictive ability of our approaches across serosurveys. For each year, we estimated the median and interquartile range in the bias. This was also done by setting.

### 7. Spatial correlation and spatial heterogeneity tests

The Spatial Correlation Diagnostic test from the PrevMap R-package (based on a permutation of locations (93)) and the Moran's I test from spdep R-package (based on neighbourhood values (94)) were used to assess spatial auto-correlation for the best and second-best model in each setting. To analyse the spatial correlation independently from the temporal one, the tests were bootstrapped 200 times with stratification on the location (one value for each municipality by iteration).

In order to assess the spatial heterogeneity among predictions, the Moran's I test under randomisation from spdep R-package (94) was undertaken in each setting on the predicted FoI values at the municipality level.

### 8. Availability of data and materials

The datasets supporting the conclusions of this article are available in the repository in (95).

## Results

### 1. Importance of accounting for the uncertainty in FoI

When using only the central FoI estimates (A1), we obtained higher predictive R$^2$ but the overlap between the predicted and estimated distribution was lower (Figure 2.1 and Supp.

Table 2). This is reflected in the 95% credible intervals (95%CrI) of the predicted FoI values being smaller than the 95%CrI in the original FoI estimates, indicative of substantial overconfidence in the models' predictions (Figure 2.2). This overconfidence in predictions is likely propagated to municipalities where we do not have estimates of FoI, leading to widespread overconfidence nationally (Figure 2.2 and Figure 2.3). This simple approach also leads to reduced heterogeneity in both space and time (Figure 2.3).

In contrast, when using the full estimated distribution of FoI for both fitting and model selection (A3), we observed a lower predictive $R^2$ but a greater overlap between observations and predictions, indicating that both the central FoI estimates and their uncertainties are well characterised (Figure 2.1 and Figure 2.2). This is reflected in the 95%CrI of the predicted FoIs being much closer to the 95%CrI in the originally estimated FoIs (Figure 2.2 and Supp. Figure 3). A3 did not, however, lead to higher uncertainty across municipalities, even where serological surveys have not been conducted. Using A3, we estimated that the MAD-CV in FoI predictions was greater than 2 in 25% of municipalities (compared to 31% and 27% in 2010 for A1 and A2, respectively) (Supp. Table 3). Furthermore, the number of extreme MAD-CV values (above 5) is reduced in A3 (39, 81, 17 municipalities with MAD-CV above 5 in 1990 for A1, 2 and 3, respectively). In municipalities where serosurveys had been conducted, the median MAD-CV was higher with A3 (median MAD-CV= 1.29, 1.28 and 1.33 with approaches A1, A2 and A3, respectively), but the maximum was lower (maximum MAD-CV= 2.76 for A3, 4.06 for A1 and 4.56 for A2) (Supp. Table 4).

Table 2.1 summarised the advantages and disadvantages of each of the three approaches. Approach 1 and 2 are easy to implement and require a limited computational effort but are not able to fully represent the uncertainty around prediction and thus provide an over-optimistic prediction.

*Table 2. 1: Advantages and disadvantages of the three approaches investigated. Approach 1: (A1) models fitted with median FoI estimates and selected based on predictive R2; Approach 2 (A2): models fitted with median FoI estimates and selected based on predictive R2 and overlap; Approach 3 (A3): models fitted with the full posterior distribution of FoI estimates and selected based on the predictive R2 and overlap.*

| Advantages | Disadvantages |
|---|---|

| | | |
|---|---|---|
| A1 | Low computational time<br>Low computational space<br>Give good central estimates | Under-estimate uncertainty<br>Reduce heterogeneity |
| A2 | Low computational time<br>Low computational space<br>Give good central estimates | Under-estimate uncertainty<br>Reduce heterogeneity |
| A3 | Give good central estimates<br>Give a good representation of the uncertainty<br>Maintain heterogeneity | Computationally demanding |

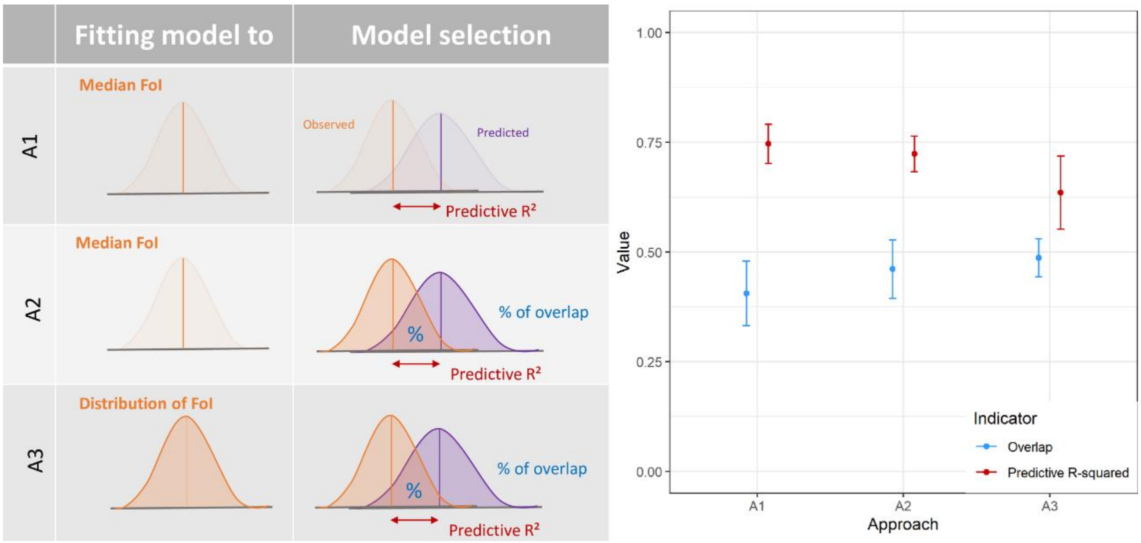


*Figure 2. 1: Comparison of the predictive ability of the best-fit models for the three approaches investigated.* *Approach 1: (A1) models fitted with median FoI estimates and selected based on predictive $R^2$; Approach 2 (A2): models fitted with median FoI estimates and selected base on predictive $R^2$ and overlap; Approach 3 (A3): models fitted with the full posterior distribution of FoI estimates and selected based on the predictive $R^2$ and overlap. Note: The overlap obtained for A1 is presented for comparison purpose and has been calculated using the same methodology as A2 but is never taken into consideration for the model selection*

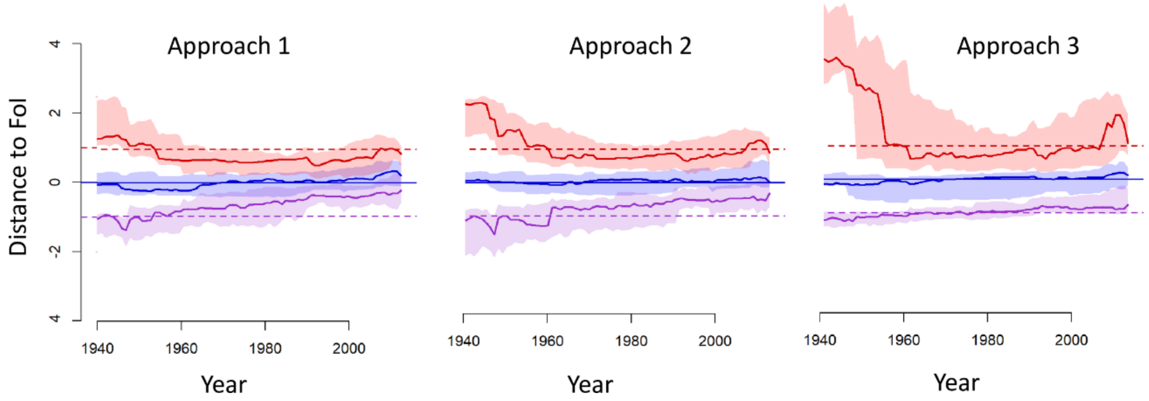


*Figure 2. 2: Goodness-of-fit of the model averaging of the 3 modelling approaches for all serosurveys. The solid lines and envelopes show standardised distances between observations and predictions' median (blue), and 95%CrI (upper bound in red and lower bound in purple). A perfect fit would translate in all coloured solid lines overlapping with the correspondingly-coloured dotted lines. A blue solid line overlapping the blue dotted line, together with a red and purple solid lines at 2 and -2 respectively would reflect a good central prediction with CrI in predictions twice as large as the CrI in the 'observed' FoI. Approach 1: models fitted with median FoI estimates and selected based on predictive $R^2$; Approach 2: models fitted with median FoI estimates and selected based on predictive $R^2$ and overlap; Approach 3: models fitted with the full posterior distribution of FoI estimates and selected based on the predictive $R^2$ and overlap.*

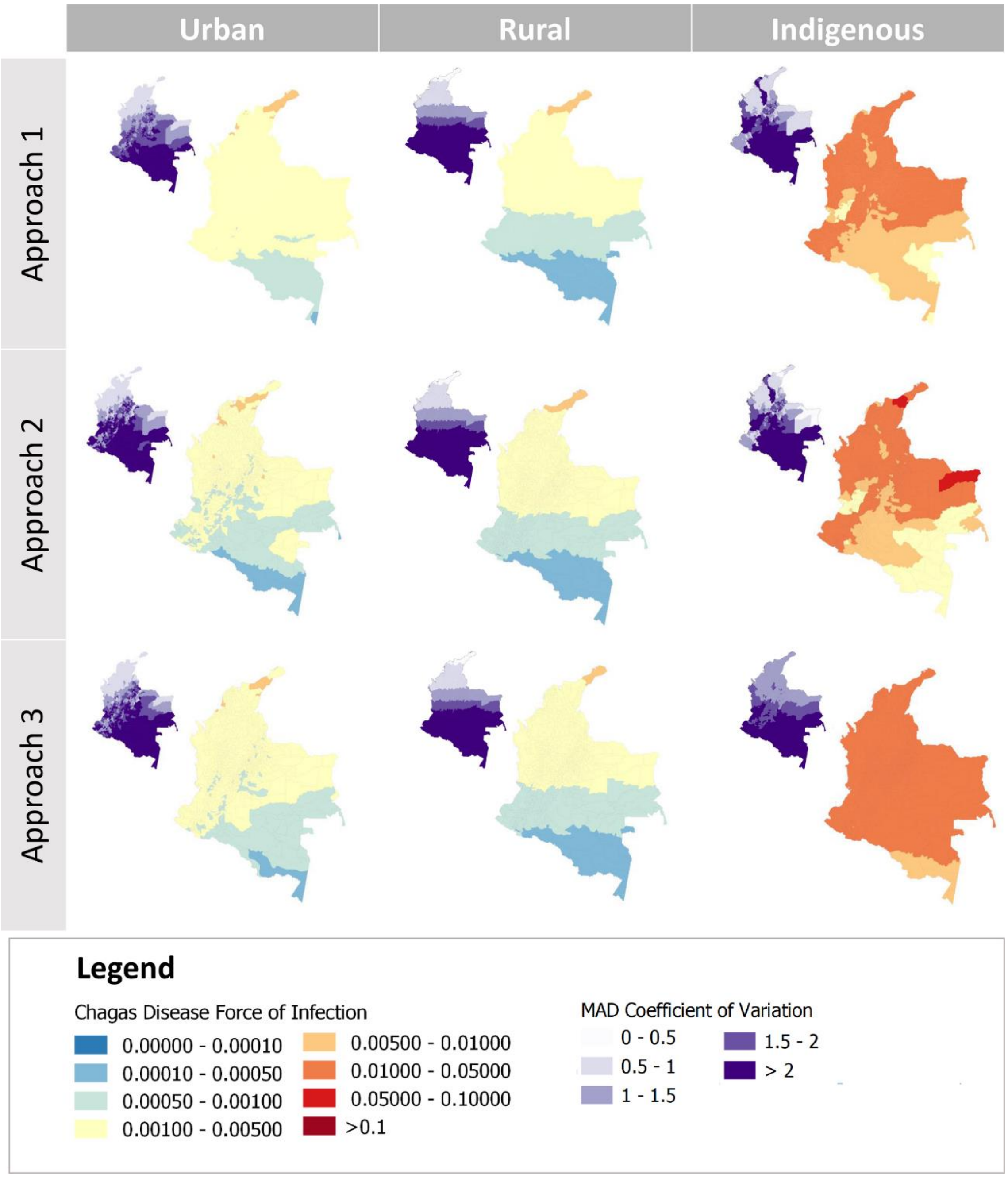


*Figure 2. 3: Force-of-Infection of Chagas disease in urban, rural and indigenous settings, Colombia, 1990. Main map, predictions per year and per susceptible individual; small map, Median Absolute Deviation (MAD) Coefficient of Variation) (n=1065 municipalities). Rows correspond to the 3 modelling approaches. Maps show model-averaged estimates (across the 10 best setting-specific models). Approach 1: models fitted using the median FoI estimates and selected based on predictive $R^2$; Approach 2: models fitted with median FoI estimates and selected based on predictive $R^2$ and overlap; Approach 3: models fitted with the full posterior distribution of FoI estimates and selected based on the predictive $R^2$ and overlap.*

## 2. Spatial and temporal predictions of FoI in Colombia

In the following, we present results from Approach 3 (unless otherwise stated); this leads to a more accurate assessment of the variations in FoI and its uncertainty. No residual spatial autocorrelation in the FoI estimates was found for any of the models as assessed by methods developed in (93,94); therefore, municipalities' predictions were obtained directly from estimated models' parameters and sets of predictors.

The FoI varied significantly by setting, with overall FoI predicted to be 9.1 and 11.8 times lower in urban and rural settings than in indigenous settings (respectively, FoI values of 2.2 x $10^{-3}$, 1.7 x $10^{-3}$ and 2.0 x $10^{-3}$ per year and per susceptible individual).

Between 1980 and 2010, the predicted FoIs showed a decreasing trend, with relative decreases of 23%, 0.07% and 7% in urban, rural and indigenous settings respectively. The decrease in predicted FoIs was statistically significant in urban and indigenous settings (Table 2.1 and Supp. Table 1), but not in rural settings.

Spatially, rural FoIs showed a clear north-south gradient, with estimated FoI values per year reaching 0.05-0.01 in the north compared to 0.0001 in most southern municipalities (Figure 2.4). In all settings, the uncertainty estimated was higher in most southern municipalities. In 1990, the Moran's I test under randomisation shows that there was spatial clustering in the predicted FoIs. The heterogeneity in predicted FoI was higher in urban settings (Moran's I statistic value of 0.82) than in rural settings (Moran's I statistic value of 0.93). In addition, the clustering effect seemed to decrease over time in urban settings, but not in rural ones (Moran's I statistic in urban settings in 1980 is 0.82 while it is 0.78 in 2010).

*Table 2. 2: Predicted FoI averaged across all Colombian municipalities in 1980, 1990 and 2010, the percentage of decrease between 1980 and 2010 (trend) for each setting and the spatial clustering effect given by the Moran's I statistic for the test under randomisation in 1980, 1990, 2000 and 2010 (n=1065 municipalities).*

| | Predicted FoI values | | | | Moran's I statistic | | | |
|---|---|---|---|---|---|---|---|---|
| | 1980 | 1990 | 2010 | trend | 1980 | 1990 | 2000 | 2010 |
| | mean<br>(sd) | mean<br>(sd) | mean<br>(sd) | % | | | | |
| Urban | $2.2 \times 10^{-3}$<br>($1.1 \times 10^{-3}$) | $2.1 \times 10^{-3}$<br>($1.1 \times 10^{-3}$) | $1.7 \times 10^{-3}$<br>($9.9 \times 10^{-4}$) | -23* | 0.82 | 0.82 | 0.79 | 0.78 |
| Rural | $1.7 \times 10^{-3}$ ($1.0 \times 10^{-3}$) | $1.7 \times 10^{-3}$<br>($1.0 \times 10^{-3}$) | $1.7 \times 10^{-3}$<br>($1.0 \times 10^{-3}$) | -0.07 | 0.93 | 0.93 | 0.93 | 0.93 |
| Indigenous | $2.0 \times 10^{-2}$<br>($4.5 \times 10^{-3}$) | $2.0 \times 10^{-2}$<br>($4.5 \times 10^{-3}$) | $1.8 \times 10^{-2}$<br>($4.4 \times 10^{-3}$) | -7* | 0.91 | 0.91 | 0.90 | 0.90 |

*Statistically significant at a 5% significance level according to Student's *t*-test comparing FoI values between 1980 and 2010

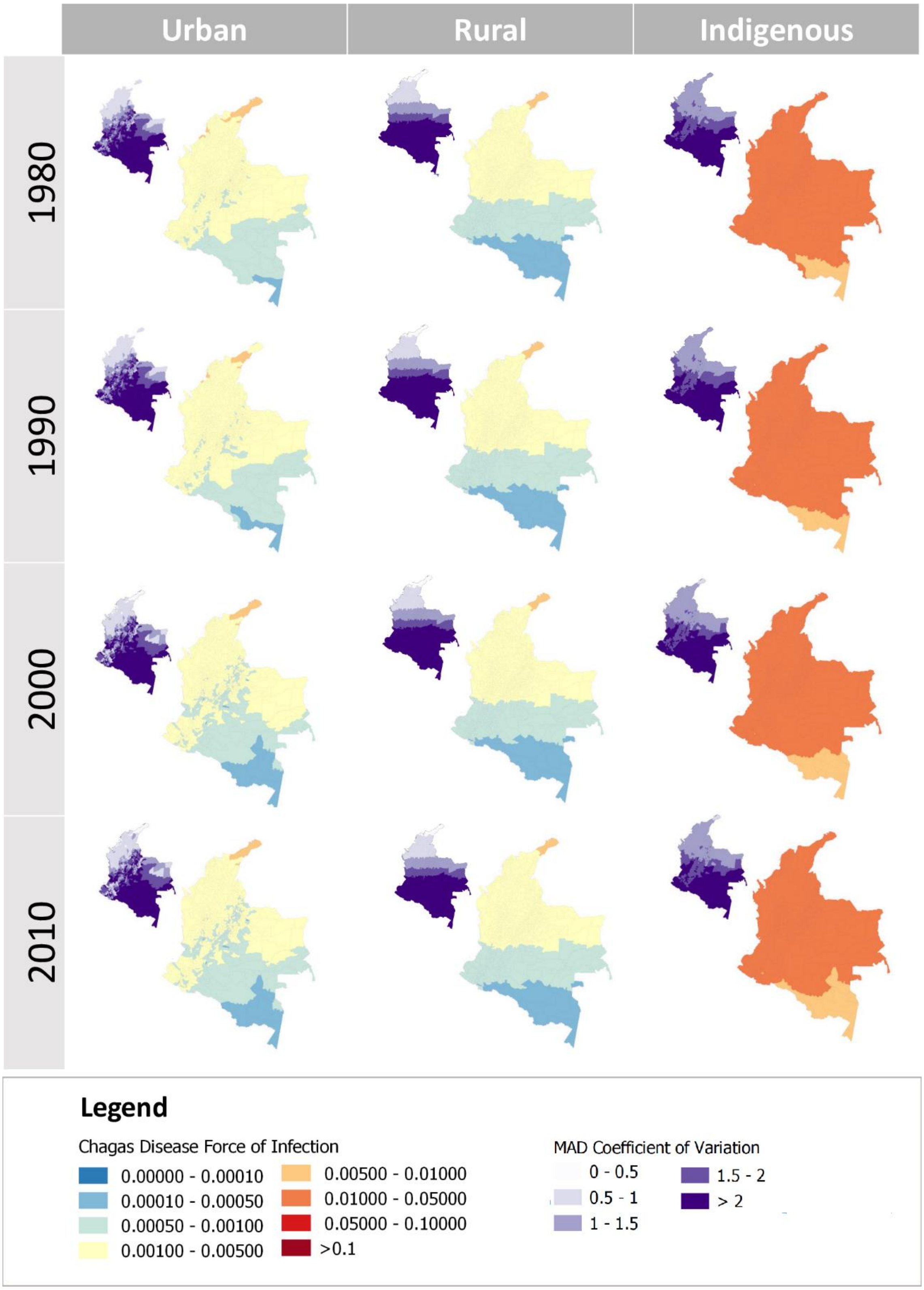


*Figure 2. 4: Spatiotemporal trends in Chagas disease Force-of-Infection, Colombia, 1980–2010. Main maps, predictions per year using approach 3 and model averaging; small maps, MAD Coefficient of Variation (n=1065 municipalities)*

### 3. Main predictors of Trypanosoma cruzi exposure

Model complexity was similar across settings, with the number of predictors included in the 10 best-fit models varying from 10–14 in urban settings, 7–13 in rural settings and 6–12 in indigenous settings (Figure 2.5).

In urban and rural areas, the predictors selected in each of the 5 best-fit models were consistent, with small changes from one model to another; while in indigenous settings, models were more distinct.

The urban-setting models always included the setting of the survey (urban, rural and indigenous) (S01), as well as its latitude (S05). Seroprevalence in blood banks and climate variables were included in 4 out of the 5 models. The level of poverty (D02) was selected and positively correlated with FoI in 3 models out of the 5 models. The interaction between the prevalence in blood banks and tropical climate (X05) was selected in 4 of the models. The year and the interaction between the amount of vector control interventions and the proportion of municipalities infested by *Triatoma dimidiata* were both included in one of the models.

The rural-setting models always included the year when the serosurvey was conducted (S01), as well as the setting (urban, rural or indigenous) (S02) and its latitude (S05). Four out of the 5 models included a climate variable. Blood bank and vector variables were only included once. Demographic, vector interventions and time variables were never selected in rural models, not even as interaction terms. Only two interactions were included; the interaction between prevalence in blood banks and tropical climate (X05), and the proportion of municipalities infested by *Rhodnius prolixus* and longitude (X11).

The indigenous-setting models were far more varied. The year when the serosurveys were conducted (S01) was included in one model. The setting was always included (S02 and S03/S04) but one of the models used the indigenous setting (S03) and the urban setting (S04) against the others as risk factors. The effect of latitude (S05) was not as clear as for urban and rural settings. Poverty (D02) was the only demographic variable included directly, but the population density was included in interaction terms with the prevalence in blood banks (X03). Vector variables played an important role in the three models. These predictors were also included as interaction terms in X11 (the proportion of municipalities infested by *R. prolixus* and longitude) and X14 (*T. dimidiata* density and vector-control interventions).

While all the best-fit models selected for prediction in rural settings included a predictor specifying the year when the serosurvey was conducted (S01, Figure 2.5), this variable was not included in any of the best models for predictions in urban settings and was included in only one of the models for indigenous settings. Consistently, for a given year and municipality, the predicted FoI values from older serosurveys were higher than those of more contemporary serosurveys (Supp. Figure 4). The inclusion of the year of the survey as a predictor for rural settings highlights potentially a bias in sampling, with older serosurveys being less representative and biased toward municipalities with higher FoI (Supp. Table 6 and Supp. Figure 4).

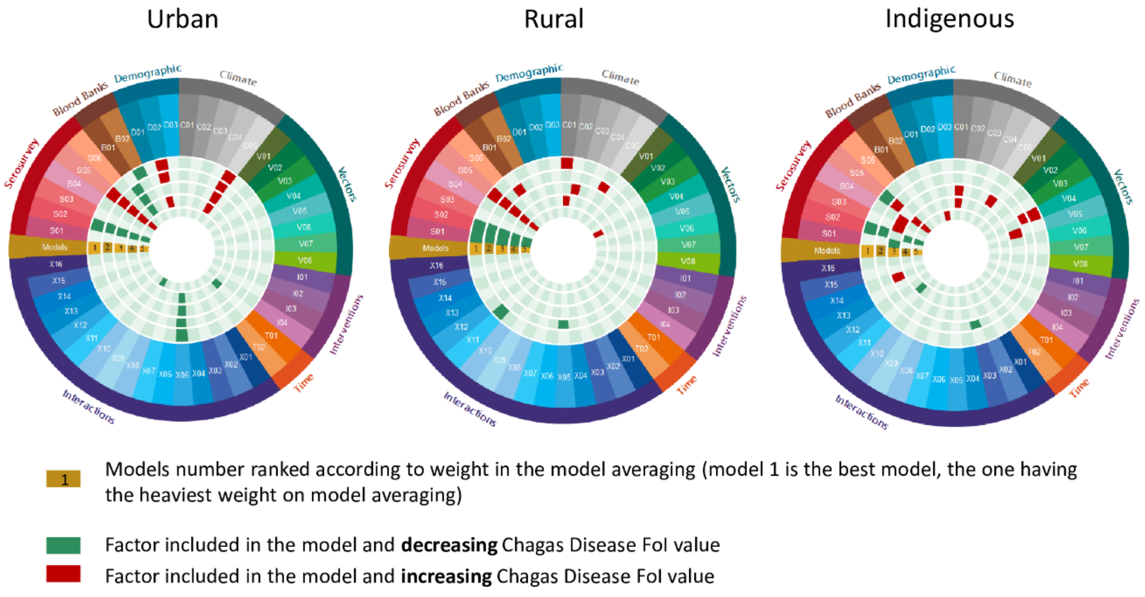


*Figure 2. 5: Predictors included in the model averaging of the FoI of Chagas disease in Colombia. Models fitted with the full posterior distribution of FoI estimates and selected based on predictive R$^2$ and overlap. For the full set of predictor variables see Supp. Table 1.*

## Discussion

We predicted spatial and temporal variations in FoI across Colombia based on estimated FoI from 76 serosurveys conducted between 1980 and 2014. Our analysis highlights the importance of accounting for the uncertain nature of the estimated FoI by demonstrating a substantial risk of overconfidence when using median estimates of FoI to fit and evaluate models, as typically done in the literature (61,66,80). We propose a novel methodology to fully propagate uncertainty from the estimated FoI onto the predicted one, giving a realistic assessment of both the central tendency and uncertainty surrounding past and current exposure to Chagas disease across Colombia.

Accounting for and communicating uncertainty in FoI estimates is critical to better inform public health and clinician stakeholders (82). It allows a better assessment of where information is missing, rather than giving a false sense of certainty. Our framework offers the opportunity to prioritise areas where serosurveys would be needed. In addition, where uncertainty is low, the models identified areas where we can be confident that populations have experienced, or are experiencing, high exposure to *T. cruzi*, which is critical to better inform focused interventions for patient diagnosis and care.

The performances of the models obtained were good, with performance indicators measuring the predictive ability of both central trends and uncertainty, estimated to vary between 0.46 and 0.67 for the five best-fit models (Supp. Table 2). When predicting FoI in new areas (where serosurveys have not yet been conducted), the uncertainty, characterised by the CV, can become much larger, while the median remains consistent across settings (in 1990, urban: median MAD-CV=1.48, range MAD-CV=0.32–8.19; rural: median MAD-CV=1.50, range MAD-CV=0.24–11.00; indigenous: median MAD-CV=1.50, range MAD-CV=1.07–3.52). In contrast, Garske *et al*. obtained FoI predictions of yellow fever with a MAD-CV ranging from 0 to 3 using central estimates of the FoI to fit their model. Using the same methodology (i.e., Approach 1), our results showed similar median uncertainty (urban: median MAD-CV=1.48, range=0.34–6.05; rural: median MAD-CV=1.51, range=0.23–11.98). To some extent, the relatively smaller uncertainty obtained in the context of yellow fever by Garske *et al*. might also be explained by their assumption of a constant FoI over time, rather than the time-varying FoI we used in this work for Chagas disease. Given the demographic and public health changes that have occurred in Colombia over the past decades (considerable rural-to-urban migration, housing improvements, scaled-up vector control, and more efficient diagnostic protocols), we believe that accounting for temporal variations in Chagas disease FoI is critical for our analysis, even at the ‘cost’ of increased uncertainty.

At first glance, our analysis highlights some unexpected results. The effect of time was relatively weak, i.e., with FoI not showing a significant decrease in rural settings; as was the effect of rural vs. urban settings. Such results contrast with previous evidence, which showed a strong temporal trend (21,32,96,97), and increased exposure in rural settings where vectorial transmission is much more prevalent (21,32,97). In terms of temporal trends, our final models always include time-varying variables, such as poverty levels and vector density, which have decreased over time, due to intervention implementation and general improvement of living conditions in the country. However, we showed that the year when the serosurvey was conducted impacted the estimated FoI, with older serosurveys biased toward high-risk areas (Supp. Figure 4). Regarding the lack of substantial differences in the level of exposure between rural and urban settings, the great population migration trends observed across the country are likely blurring this effect. Considerable rural to urban migration has taken place in Colombia, with one-third of the rural population aged below 40 in 1951 having migrated to urban settings by 1964, mostly to find better employment opportunities (98). More recently, it has been estimated that more than 3.5 million people had migrated to urban centres to escape violence in rural areas (99). Having lived for an extended period of time in rural settings, these migrants may well have been exposed to *T. cruzi* in rural areas but now account for the estimated FoI in urban settings. Unfortunately, the participants' migration history was not recorded (or available) in the serosurveys used. Similar dynamics of migrations have been shown to explain a substantial burden of Chagas disease in both endemic (e.g. in Arequipa, Peru (17)) and non-endemic settings (100).

Another spatial challenge is the scale at which the analyses have been conducted. Indeed, we demonstrate small-scale spatial heterogeneity in Chagas disease exposure between the municipalities within a department. And, while our approach was designed to be conservative by excluding serosurveys providing information only at the departmental level, we acknowledge that further small-scale heterogeneity may exist, i.e., differences could occur between villages of the same municipality. However, the municipal level is the operational level in the control of Chagas disease and is, therefore, the most useful level to characterize exposure in a way that actionable information can be extracted. Also, we found that most of the important variables for predictions were available at the municipality level (poverty indicator, vector density), but not disaggregated further. Thus, even if a small-scale analysis could provide some insights, technically and operationally, the municipal level remains the most relevant one.

While serosurveys provide invaluable information on exposure, our analysis highlights the importance of appropriate sampling strategies. Sampling decisions taken to collect the data have a clear impact on our ability to provide representative predictions over large spatial and temporal scales. One issue linked to sample representativeness is the location of the serosurveys. Indeed, the likely past focus on estimating exposure in high-risk populations may have created a selection bias that cannot be easily handled when modelling the data. In Colombia, this seemed especially true in rural settings (Supp. Figure 4). This bias likely explains much of the temporal trends that have been reported in previous studies (e.g. (21)). This highlights the problem of relying on surveys that were not designed to provide a representative sample but rather organised to confirm and quantify incidence in high-risk areas. Extrapolation to areas where no serosurveys have been conducted is then made more uncertain and needs to be interpreted accordingly. Another issue linked to sample representativeness is the targeted age groups of the surveys. In 2012, the World Health Organization set the elimination of (intradomicilary) Chagas disease transmission as a goal in its first neglected tropical disease roadmap; one of the indicators used to monitor progress towards this goal was the seroprevalence among under-five children, aiming to measure active transmission as opposed to past exposure (21,101). Unfortunately, such (narrow age-range) sampling scheme hampers obtaining valuable information about past exposure, which for a chronic illness, such as Chagas disease, is crucial to target diagnosis and treatment. We argue that organising representative serosurveys and covering a broader age range is essential to obtain a reliable picture of the epidemiological situation and the impact of control interventions in endemic countries, particularly for infectious diseases that use serosurveys for the purposes of surveillance.

# Chapter 3: Linear and Machine Learning Modelling for Spatiotemporal Disease Predictions: Force-of-Infection of Chagas Disease

*Published in PLOS Neglected Tropical Diseases on the 19th of July 2022 (DOI: 10.1371/journal.pntd.0010594).*

## Abstract

Background: Chagas disease is a long-lasting disease with a prolonged asymptomatic period. Cumulative indices of infection such as prevalence do not shed light on the current epidemiological situation, as they integrate infection over long periods. Instead, metrics such as the Force-of-Infection (FoI) provide information about the rate at which susceptible people become infected and permit sharper inference about temporal changes in infection rates. FoI is estimated by fitting (catalytic) models to available age-stratified serological (ground-truth) data. Predictive FoI modelling frameworks are then used to understand spatial and temporal trends indicative of heterogeneity in transmission and changes effected by control interventions. Ideally, these frameworks should be able to propagate uncertainty and handle spatiotemporal issues.

Methodology/Principal findings: We compare three methods in their ability to propagate uncertainty and provide reliable estimates of FoI for Chagas disease in Colombia as a case study: two Machine Learning (ML) methods (Boosted Regression Trees (BRT) and Random Forest (RF)), and a Linear Model (LM) framework that we had developed previously. Our analyses show consistent results between the three modelling methods under scrutiny. The predictors (explanatory variables) selected, as well as the location of the most uncertain FoI values, were coherent across frameworks. RF was faster than BRT and LM, and provided estimates with fewer extreme values when extrapolating to areas where no ground-truth data were available. However, BRT and RF were less efficient at propagating uncertainty.

Conclusions/Significance: The choice of FoI predictive models will depend on the objectives of the analysis. ML methods will help characterise the mean behaviour of the estimates, while LM will provide insight into the uncertainty surrounding such estimates. Our approach can be extended to the modelling of FoI patterns in other Chagas disease-endemic countries and to other infectious diseases for which serosurveys are regularly conducted for surveillance.

## Author Summary

Metrics such as the per susceptible rate of infection acquisition (Force-of-Infection) are crucial to understand the current epidemiological situation and the impact of control interventions for long-lasting diseases in which the infection event might have occurred many years previously, such as Chagas disease. FoI values are estimated from serological age profiles, often obtained in a few locations. However, when using predictive models to estimate the FoI over time and space (including areas where serosurveys had not been conducted), methods able to handle and propagate uncertainty must be implemented; otherwise, overconfident predictions may be obtained. Although Machine Learning (ML) methods are powerful tools, they may not be able to entirely handle this challenge. Therefore, the use of ML must be considered in relation to the aims of the analyses. ML will be more relevant to characterise the central trends of the estimates while Linear Models will help identify areas where further serosurveys should be conducted to improve the reliability of the predictions. Our approaches can be used to generate FoI predictions in other Chagas disease-endemic countries as well as in other diseases for which serological surveillance data are collected.

## Introduction

Chagas disease is a neglected tropical disease estimated to affect between 6 and 7 million persons worldwide. While only endemic in 21 countries in Latin America, the number of Chagas disease cases detected in Europe, North America, and the Far East has greatly increased, due to migration of infected populations (102). Being able to identify how the cases are distributed in space and whether the control interventions implemented have been successful is critical to identifying how resources should be allocated to eliminate the disease as a public health problem in the 2021–2030 time horizon (20). As a long-lasting and chronic disease, analyses based solely on the current prevalence of infection (typically measured as seroprevalence) has limited scope. Indeed, the prevalence recorded at a given time does not reflect the current epidemiological situation, as infection may have occurred in the past. The Force-of-Infection (FoI), i.e. the rate at which susceptible individuals become infected, is a modelling-derived metric that can be used to understand changes in incidence in space and time as a result of deliberate control interventions and/or secular changes, including environmental change (85). However, the use of FoI raises its own challenges, particularly those regarding quantification and propagation of uncertainty when used as a response variable in predictive models. A catalytic model (fitted to age-structured seroprevalence data, often using Bayesian methods) has been used to obtain the FoI and thus, the FoI values for each serosurvey and each year are posterior distributions and not only single values (21). When the derived FoI is used to fit predictive models, the mean or median values of FoI are predominantly used, neglecting the uncertainty surrounding the estimated values (61,66,80). Furthermore, when a non-constant (e.g. a yearly-varying) FoI is assumed, each serosurvey analysed becomes a temporal series at a certain location, requiring specific and computationally-intensive methods to be included into predictive models (103). Machine Learning methods could represent a faster and more flexible framework to implement such models.

Machine Learning (ML) methods are computational processes based on probabilities and algorithms that use prior knowledge to produce predictions. ML can handle non-linear and non-parametric models that are able to flout the linearity, normality (Gaussian distribution) and equal variance assumptions of statistical models (104). Essentially, ML methods make no assumptions about the statistical distribution of the data (104).

These methods have previously been used in contexts in which those assumptions are challenged, such as spatial, temporal and spatiotemporal analyses of infectious diseases, e.g. mapping of human leptospirosis (105,106), severe fever with thrombocytopenia syndrome (107), lymphatic filariasis (108), or to identify individuals with a higher risk of HIV infection based on socio-behavioural-driven data (109).

Two types of ML models have been extensively used in the context of infectious disease epidemiology, namely, Boosted Regression Trees (BRT) and Random Forest (RF). Although they are not spatial approaches (as data locations and sampling patterns are ignored to produce estimates), they have shown potential in spatial modelling (110), in particular, when used with appropriate sampling strategies (111). Specifically, BRT and RF have been used to study the spatial spread of numerous infectious diseases, including epidemics among swine farms in the USA (112), Ebola case-fatality ratio (113), risk factors for visceral leishmaniasis (114,115), African swine fever (116), scrub typhus (117), dengue incidence (118), and dengue FoI (80). RF also proved its potential in modelling epidemics in a spatiotemporal framework, outperforming time series models (112).

This paper aims to compare the performance of two ML methods, namely, BRT and RF, with a Linear Model (LM) framework we have previously developed (103) in their ability to predict the FoI of Chagas disease across space and time. We use detailed data from Colombia as a case study and describe the advantages and disadvantages of using such Machine Learning methods compared to Linear Model frameworks, specifically focussing on their ability to handle uncertainty on the response variable.

## Methods

### 1. Data sources

Current and past exposure to Chagas disease can be characterised by estimating the (time-varying) Force-of-Infection (FoI), i.e. temporal changes in the per susceptible rate of parasite acquisition (21,85). Using results of 76 age-stratified serosurveys conducted at municipal level in Colombia between 1998 and 2014 (Supp. Figure 2), yearly-varying FoI values were estimated, for each serosurvey, by fitting a catalytic model to age-stratified seropositivity data (see (21) for details). For each serosurvey, FoI estimates, for the period ranging from the birth of the oldest participants to the year when the serosurvey was conducted, were obtained using a Bayesian framework to fit the catalytic model to data, thus allowing for extraction of the full joint posterior distribution of the yearly FoI estimates. We refer to those municipalities where at least one serosurvey was conducted as municipalities 'in catchment areas', whereas those municipalities for which serosurveys were not conducted, not available, or not used in our analyses, are referred as 'out of catchment areas'. Supp. Figure 2 in the Supporting Information file depicts the geographical distribution of the available serosurveys ('ground-truth' data).

The predictors used in these analyses included demographic, entomological and climatic variables (recorded at the municipality level), contextual information about the serosurveys (location, year conducted and setting, i.e. urban, rural, mixed and indigenous (as defined in (21)), and information from public blood banks on the prevalence of Chagas disease and number of blood units tested (available at the departmental level). A full list and description of the predictors is available in Supp. Table 1 in the Appendix.

### 2. Linear Model (LM) framework

The LM framework relied on a list of plausible linear combinations of predictors that were then integrated into an ensemble model using model averaging with weights based on the performance indicators of each individual linear model. The 10 best models for each setting type (urban, rural and indigenous) were averaged and used to obtain FoI predictions. The LM framework has been fully described in (103).

### 3. Machine Learning (ML) framework

Both ML methods tested in this paper are based on decision trees. A decision tree is an intuitive process that builds an algorithm by generating a step-by-step tree, whereby the dataset is repeatedly split to make a decision at each node. The splitting relies on optimising a variable-specific threshold that best discriminates the data into two branches at each node. Sequentially, the entire dataset is divided by defining new variable-specific thresholds defining the nodes in the decision tree.

The size of the tree, its complexity (reflecting predictors' interactions), the number of observations in the terminal nodes and the criteria to stop the process are defined as model hyperparameters and form the basis of more complex designs.

*Boosted Regression Trees (BRT)*

Boosting Regression Trees (BRT) or Gradient Boosting Trees (GBT) are based on the building of a large number of small decision trees. The boosting aspect refers to fitting repeatedly very simple and basic classifiers, in which a different subset of the data is used for fitting at each iteration (104). The Gradient technique is used to reduce the variance in the model; sequentially, each new tree added to the model is fitted to explain the remaining variance from the previous trees.

While BRT is considered a robust ML method, including its use for spatiotemporal analyses (114,115,117,119), it is known as having a tendency to overfit, unless a very large amount of data is available (120).

*Random Forest (RF)*

Random Forest (RF), first described by Breiman in 2001 (121), consists of a large collection of decision trees (104). To grow an RF tree, random inputs and predictors are selected at each node (121), and this randomness is thought to reduce overfitting. RF is also considered a robust ML method that can handle outliers and noise while being faster than bagging- and boosting-based methods (121).

RF is not explicitly designed to explore spatial observations (110), and is known to produce suboptimal prediction when sampling is spatially biased and/or in the presence of strong spatial correlation (110). However, spatiotemporal resampling strategies and variable selection processes have been developed to overcome this challenge (111,122).

### 4. Models' workflow

In order to assess the importance of integrating uncertainty on the response variable, we implemented two approaches The former relies on generating and assessing model predictions using the ***median*** estimates of the FoI for each serosurvey as an outcome (referred to as MedFoI). The latter seeks to propagate the uncertainty linked to the catalytic model-derived 'observations' by accounting for the ***full posterior distribution*** of the FoI estimates (referred to as FullPostFoI).

With the MedFoI approach, models are fitted to the median FoI estimates and the performance indicator, the predictive $R^2$, is based on central tendencies only. With the FullPostFoI approach, models are fitted on the full posterior distribution of FoI estimates and the performance indicator is based both on central tendency and on the amount of overlap between the 'observed' and predicted distribution of the outcome. This allowed us to quantify the ability of the models to match the uncertainty surrounding the FoI estimates (i.e. the outcome) inherited from the catalytic model (Figure 3.1). The percentage of overlap was obtained using the "overlap" function from the Overlapping R-package (123) and provided the proportion of the area of two kernel density estimations that overlap (124).

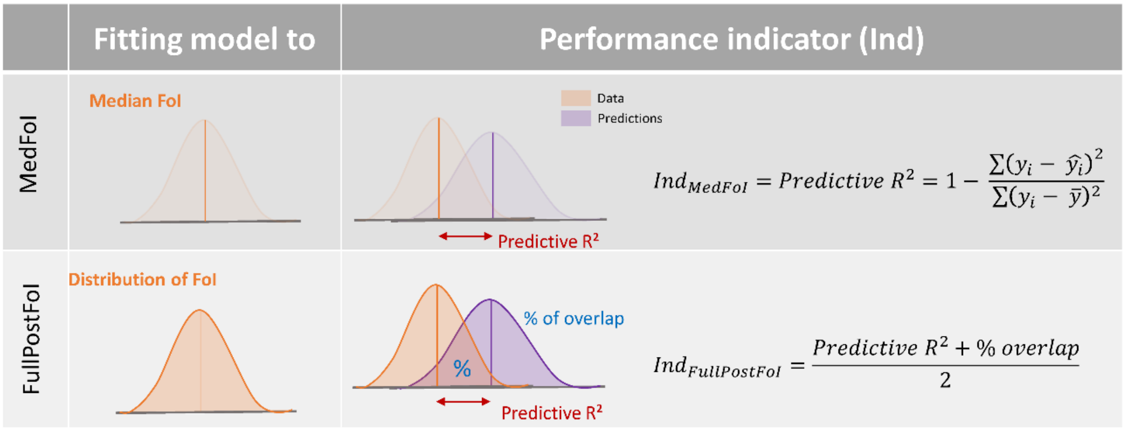

Fitting model to
Performance indicator (Ind)
MedFoI
Median FoI
Data
Predictions
Predictive R²
$Ind_{MedFoI} = Predictive\ R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$
FullPostFoI
Distribution of FoI
% of overlap
%
Predictive R²
$Ind_{FullPostFoI} = \frac{Predictive\ R^2 + \%\ overlap}{2}$

*Figure 3. 1: Graphical representation of the two modelling approaches used for each of the frameworks tested. The upper panel corresponds to MedFoI (fitted on median FoI); the lower panel depicts the FullPostFoI (fitted on full posterior FoI). The predictive $R^2$ values are calculated on cross-validation sets for both approaches (see Figure 3.2). In the upper panel, the performance indicator, Ind, is the $R^2$, based on central tendency alone; in the lower panel, both central tendency and percentage of overlap enter into the calculation of the performance indicator, Ind (as the arithmetic mean between $R^2$ and percentage overlap). The percentage of overlap (% of overlap) represents the proportion of the 'observed' and predicted distributions that overlap.*

For the two approaches, we defined six different coefficients of determination ($R^2$) linked to the sampling strategy. An $R^2$ was estimated based directly on the data used to fit the models; a predictive $R^2$ was estimated based on a proportion of the dataset that was not used to fit the models, i.e. the cross-validation (CV) set. In addition, both urban- and rural-specific predictive $R^2$ were estimated based on the urban/rural data from the CV set. Finally, in the ML frameworks, a resample $R^2$ was estimated based on out-of-sample data for each resample iterations (see Figure 3.2 for a schematic representation of these approaches).

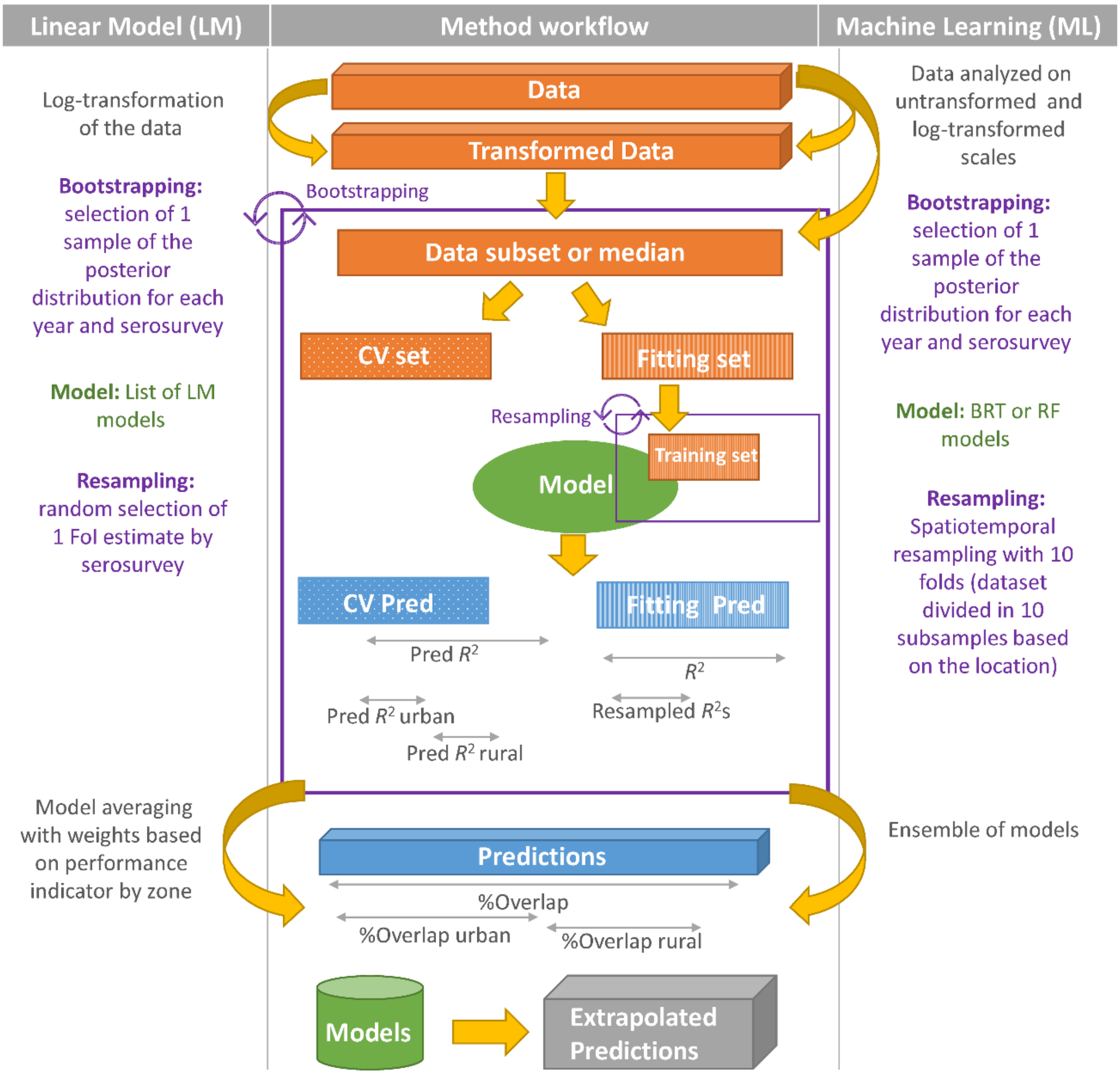

Linear Model (LM)
Method workflow
Machine Learning (ML)
Log-transformation of the data
Bootstrapping: selection of 1 sample of the posterior distribution for each year and serosurvey
Model: List of LM models
Resampling: random selection of 1 FoI estimate by serosurvey
Model averaging with weights based on performance indicator by zone
Data
Transformed Data
Bootstrapping
Data subset or median
CV set
Fitting set
Resampling
Training set
Model
CV Pred
Fitting Pred
Pred $R^2$
$R^2$
Pred $R^2$ urban
Resampled $R^2$s
Pred $R^2$ rural
Predictions
%Overlap
%Overlap urban
%Overlap rural
Models
Extrapolated Predictions
Data analyzed on untransformed and log-transformed scales
Bootstrapping: selection of 1 sample of the posterior distribution for each year and serosurvey
Model: BRT or RF models
Resampling: Spatiotemporal resampling with 10 folds (dataset divided in 10 subsamples based on the location)
Ensemble of models

*Figure 3. 2: Description of the modelling workflow for the Linear Models (LM) and the Machine Learning (ML) frameworks. ML framework include Boosted Regression Trees (BRT) and Random Forest (RF) methods). CV denotes cross-validation; Pred $R^2$ urban and Pred $R^2$ rural denote urban- and rural-specific predictive $R^2$ values that were estimated based on the urban/rural data from the CV set; %Overlap indicates the proportion of the 'observed' and predicted distributions that overlap (see Figure 3.1), assessed over all settings and for urban and rural settings separately.*

While the LM framework necessitated transformation of the data to normalise them, ML methods should, in principle, be able to handle data without requiring normalisation (i.e. without requiring that their distribution is Gaussian). However, this process can help improve the performance of the model and was, therefore, tested (i.e., ML approaches were used to predict the FoI values both on non-transformed and log-transformed scales).

While the LM framework relies on a list of plausible and pre-defined linear models including interactions between factors (predictors), the ML framework is implemented in two steps, to be fitted only on the ten most important variables. At first, ML models were fitted using all the predictors available, then the importance/influence of each predictor was assessed, and the 10 most influential factors were used in the second step, during which the models were fitted again, and predictions extracted.

Finally, ML requires a tuning step, during which the best hyperparameters are selected. A detailed description of the tuning of hyperparameters and the comparison of several resampling strategies is available in Supp. Method 2, including details about the tuning of hyperparameters and the comparison of several resampling strategies.

### 5. Indicators used to compare LM with ML frameworks

The best ML models obtained were then compared with the LM framework previously developed (103) in terms of their performance indicators, predictions, and ability to propagate uncertainty. In addition to these aspects, the models' ability to deal with temporal and spatial correlation, as well as their different computational aspects entered the comparison.

To allow comparison of predictive ability across multiple serosurveys, the distributions of predictions were standardised to the 'observations', allowing us to visualise whether the median and confidence intervals of the predictions matched those (median and credible intervals) of the catalytic model-derived FoI 'observations'. This process was performed at the serosurvey level to assess how much of the uncertainty inherited from the FoI calculation (via catalytic model fitting) was propagated into the predictions (see Supp. Method 3).

The uncertainty in the predictions was quantified using the Coefficient of Variation based on the standardised Median Absolute Deviation (MAD-CV), as FoI values were not normally distributed (84). (Note that MAD-CV refers to coefficient of variation, whilst CV denotes cross validation.)

The residual spatial correlation was assessed using the Moran's *I* heterogeneity test from the "spdep" R package (94). For the LM framework, the Moran's *I* test was applied on all the residuals (originating from the cross validation (CV) and fitting sets) excluding those from the rural–urban mixed settings (as LM model selection was based on setting type and no model was explored, selected or averaged for mixed settings, and thus no predictions were produced for the 'observed' FoI values corresponding to such settings). For ML models, the Moran's *I* test was applied to the residuals of the CV set. Residuals for a single year were used to exclude potential temporal autocorrelation, and for presentation purposes, we selected 2005 as the year with the largest number of independent FoI 'observations'.

The residual temporal correlation was tested using a Durbin-Watson test (DW) (125) (see Eq. 4 for the DW statistic). In order to capture the residual correlation inherited from the estimation of the FoI values through fitting the catalytic model, the residuals being compared were always from the same serosurvey and for consecutive years. Thus, the DW statistic provided the residual serial correlation for a lag of one year,

$$DW = \frac{\sum_{i=0}^{n} \left(r_i - lag(r_i)\right)^2}{\sum_{i=0}^{n} r_i^2} \qquad \text{Eq. 4}$$

where $r$ denotes the residuals for $i$ serosurveys, $lag$ is one year (for consecutive, yearly series of serosurveys), and $n$ the number of 'observations' tested.

### 6. Availability of data and materials

All ML analyses were run under the mlr3 framework (an object-oriented machine learning framework in R) (126) using R-4.0.3 software (127). The datasets used for these analyses are available in the repository of (95).

## Results

### 1. Comparison of the performance of LM and ML frameworks

The predictive $R^2$ values for the LM framework obtained, on average, for its 5 best-fitting models, were 77% and 70%, with %overlap of 54% and 39% for urban and rural settings, respectively (103). For the ML frameworks, the MedFoI approach yielded substantially better predictive $R^2$ values (ranging between 90% and 98%), but the degree of overlap between the distributions of the FoI 'observations' and the predictions was substantially lower (19%–25%), reflective of a tighter distribution around the central estimates and thus indicating over-confidence in the predictions when using such a simple approach (i.e. an approach that ignores the uncertainty linked to the outcome). The FullPostFoI approach gave a more balanced performance indicator, with predictive $R^2$ values ranging between 39% and 70%, and %overlap between 22% and 42% (Table 3.1). For both BRT and RF methods, the use of log-transformation to normalise the distribution of the FoI 'observations' consistently led to improved results (Table 3.1), with predictive $R^2$ values ranging between 59% and 70%, and %overlap between 34% and 42%.

Nested resampling, tested on the RF method with log-transformation, did not substantially improve model performance. Thus, the following subsections focus on the results obtained by fitting the frameworks on the full posterior distribution of the log-transformed FoI.

*Table 3. 1: Median cross-validation performance values for the two Machine Learning modelling methods investigated.*

| | BRT | | | | | | RF | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | All settings | | Urban | | Rural | | All settings | | Urban | | Rural | |
| | non | log | non | log | non | log | non | log | non | log | non | log |
| *MedFoI* | | | | | | | | | | | | |
| $R^2$ (%) | 98 | 95 | 95 | 96 | 94 | 90 | 98 | 96 | 90 | 98 | 93 | 93 |
| Overlap | 23 | 19 | 21 | 19 | 25 | 19 | 22 | 21 | 20 | 21 | 25 | 21 |
| *FullPostFoI* | | | | | | | | | | | | |
| $R^2$ (%) | 60 | 58 | 53 | 68 | 39 | 68 | 63 | 59 | 63 | 68 | 69 | 70 |
| Overlap | 25 | 36 | 24 | 34 | 22 | 36 | 40 | 42 | 42 | 42 | 40 | 42 |
| Indicator | 43 | 48 | 37 | 52 | 25 | 52 | 51 | 50 | 50 | 55 | 52 | 55 |

*MedFoI*: models fitted on median FoI; *FullPostFoI*: models fitted on full posterior FoI, without (non) or with log-transformation (log) of Force-of-infection (FoI) 'observations' (generated by fitting catalytic models to age-stratified serological surveys for Chagas disease in Colombia, with yearly-varying FoI (103)).

BRT: Boosted Regression Trees; RF: Random Forest methods; performance indicators are reported for either all settings (urban, rural, indigenous and mixed), urban, or rural settings separately. The predictive $R^2$ values were calculated on cross-validation datasets and are expressed as percentage.

Overlap: proportion (expressed as percentage) of 'observed' and predicted distributions that overlaps (reflective of the degree of dispersion around central tendency). For *MedFoI* models, the performance indicator is the value of $R^2$ alone. Therefore % overlap is presented for comparison but was not used in fitting or selecting models. For *FullPostFoI*, the performance indicator is the arithmetic mean between $R^2$ and % of overlap (see Figure 3.1).

### 2. Comparison of the influence of predictor variables

The factors selected for the ML models were consistent with those that had been selected for the LM framework (Table 3.2); a Spearman correlation test showed that there was substantial rank correlation of the predictors included among the three models investigated (with

Spearman rank correlation coefficient, $r_S$, between LM and BRT = 0.50; between LM and RF = 0.54, and between BRT and RF = 0.64, all p-values <0.05).

The type of setting was the most important factor for both the LM and BRT. Latitude, year when the serosurvey was conducted, and population density also played an important role. Poverty, climatic and entomological features had a moderate role.

For the ML frameworks, blood-bank and intervention-related features were less influential than for the LM framework. Generally, the year (temporal trend) seemed to play a greater role in the ML models.

*Table 3. 2: Standardized relative influence, importance and rank of the predictors included in Boosted Regression Trees (BRT) and Random Forest (RF) Machine Learning models and normalised number of times the predictors were used in the linear model (LM) framework and their rank when using the full posterior distribution of FoI estimates.*

| | Predictors | BRT | | RF | | LM | |
|---|---|---|---|---|---|---|---|
| Code | Name | Influence | Rank | Importance | Rank | Used | Rank |
| | Serosurvey characteristics: | | | | | | |
| S01 | Year of the survey | 0.20 | 2 | 0.17 | 1 | 0.05 | 4 |
| S02 | Rural setting | 0.03 | 11 | 0.03 | 11 | 0.14 | 2 |
| S03 | Urban setting | 0.03 | 10 | 0.04 | 7 | 0.15 | 1 |
| S04 | Indigenous setting | 0.20 | 1 | 0.12 | 4 | 0.15 | 1 |
| S05 | Latitude | 0.14 | 3 | 0.16 | 2 | 0.14 | 2 |
| S06 | Longitude | 0.04 | 8 | 0.09 | 5 | 0.02 | 7 |
| | Blood-bank data: | | | | | | |
| B01 | Seroprevalence | 0.00 | NU | 0.03 | 14 | 0.04 | 5 |
| B02 | Proportion of blood units screened | 0.00 | NU | 0.03 | 10 | 0.01 | 8 |
| | Demography: | | | | | | |
| D01 | Population density | 0.10 | 4 | 0.13 | 3 | 0.01 | 8 |
| D02 | Poverty | 0.07 | 6 | 0.01 | 15 | 0.04 | 5 |
| D03 | Rural Indigenous Population size | 0.00 | NU | 0.03 | 9 | 0.00 | NU |
| | Climate: | | | | | | |
| | *Continuous* | | | | | | |
| C01 | Polar climate frequency | 0.03 | 12 | 0.00 | 18 | 0.04 | 5 |
| C02 | Tropical climate frequency | 0.03 | 9 | 0.04 | 8 | 0.01 | 8 |
| C03 | Temperate climate frequency | 0.04 | 7 | 0.00 | 17 | 0.00 | NU |
| C04 | Arid climate frequency | 0.00 | NU | 0.00 | 21 | 0.00 | NU |
| | *Categorical* | | | | | | |
| C05 | Tropical climate categorised | NT | | | | 0.06 | 3 |
| C06 | Polar Climate Presence | 0.00 | NU | 0.00 | NU | 0.00 | NU |
| | Entomological data: | | | | | | |
| | *At departmental level* | | | | | | |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| V01 | *R. prolixus* geographical extent | 0.00 | NU | 0.06 | 6 | 0.02 | 7 |
| V02 | *T. dimidiata* geographical extent | 0.00 | NU | 0.03 | 12 | 0.01 | 8 |
| V03 | *R. prolixus* presence | 0.00 | NU | 0.00 | NU | 0.00 | NU |
| V04 | *T. dimidiata* presence | 0.00 | NU | 0.00 | NU | 0.03 | 6 |
| | *At municipality level* | | | | | | |
| V05 | *R. prolixus* density | 0.00 | 13 | 0.00 | NU | 0.01 | 8 |
| V06 | *T. dimidiata* density | 0.00 | 15 | 0.00 | 16 | 0.01 | 8 |
| V07 | *R. prolixus* presence | 0.00 | NU | 0.00 | NU | 0.00 | NU |
| V08 | *T. dimidiata* presence | 0.00 | NU | 0.00 | 20 | 0.00 | NU |
| | Interventions: | | | | | | |
| | *At municipality level* | | | | | | |
| I01 | Intervention intensity | 0.00 | 14 | 0.00 | NU | 0.00 | NU |
| I02 | Intervention category | NT | | NT | | 0.01 | 8 |
| | *At household level* | | | | | | |
| I03 | Household intervention | 0.00 | NU | 0.00 | NU | 0.00 | NU |
| I04 | Household intervention category | NT | | NT | | 0.01 | 8 |
| | Temporal factors: | | | | | | |
| T01 | Year | 0.09 | 5 | 0.03 | 13 | 0.01 | 8 |
| T02 | Decade | 0.00 | NU | 0.00 | NU | 0.00 | NU |

NU: Not used in the model; NT: not tested in the model.

*R. prolixus: Rhodnius prolixus; T. dimidiata: Triatoma dimidiata.*

The shade of green is associated to the rank of the predictors with darker green predictors having more importance.

### 3. Comparison of spatial trends in predictions

All methods showed generally similar spatial trends and comparable levels of uncertainty (Figure 3.3) for FoI prediction across Colombia (using the year 1990 for illustration as the pattern is consistent in time). Generally, FoI estimates were higher in northern and eastern municipalities and lower in the south of the country, with the latter being associated with higher uncertainty (Figure 3.3). The BRT framework predictions showed increased spatial heterogeneity, while predictions from the LM framework resulted in more spatially uniform predictions.

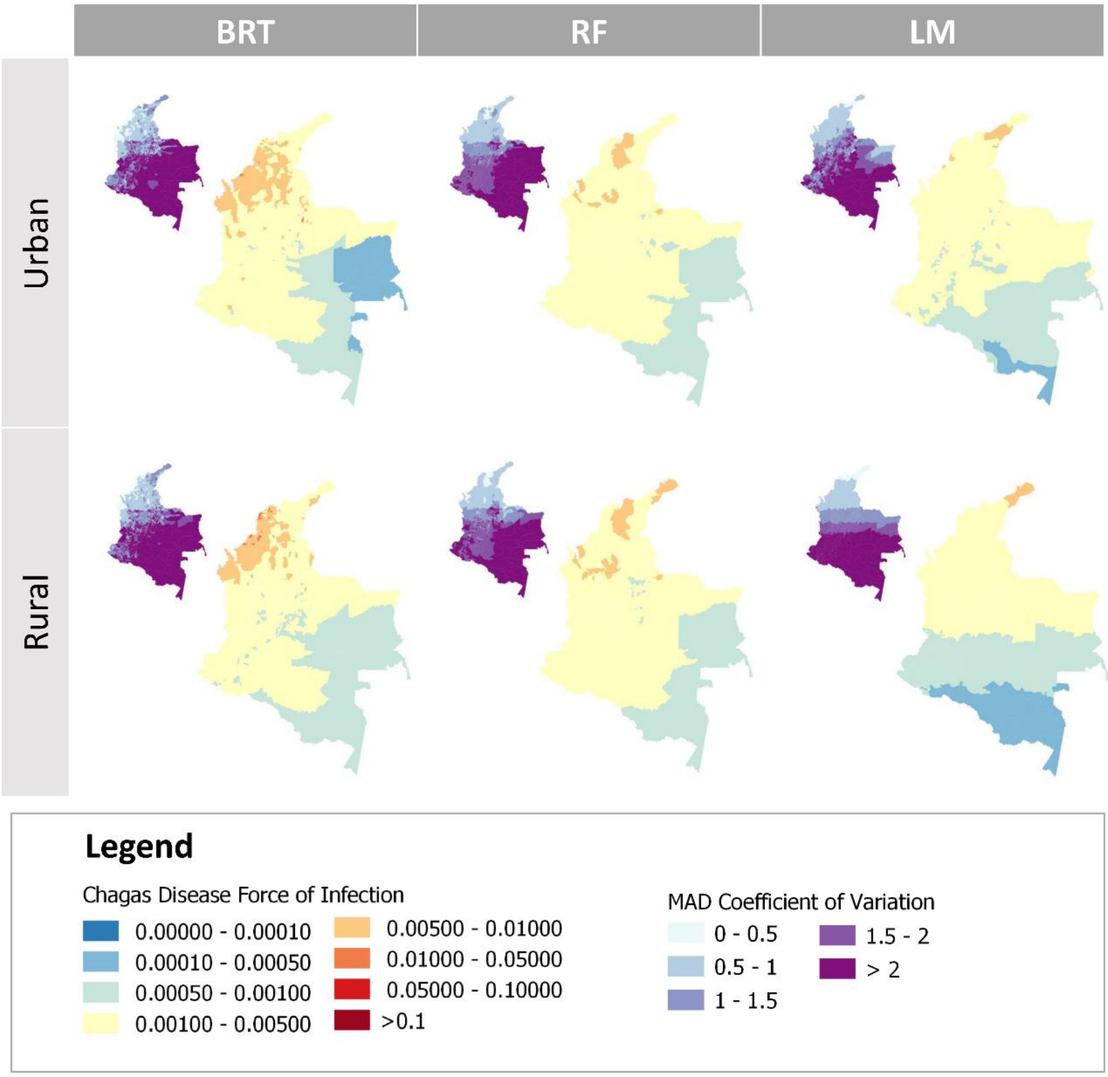


*Figure 3. 3: Spatial distribution of the Force-of-Infection of Chagas Disease (per year and per susceptible individual), in Colombia. The predicted distribution was generated using two Machine Learning (Boosted Regression Trees (BRT) and Random Forest (RF)) methods and a Linear Model (LM) framework (main maps); the associated uncertainty (small map insets) presents the Median Absolute Deviation (MAD) Coefficient of Variation (MAD-CV). Predictions were obtained at the municipality level for urban and rural settings, in 1990.*

When comparing FoI predictions directly across the three methods, for urban and rural settings (Figure 3.4), we found good agreement between all of them, particularly between RF and LM. Generally, the BRT tended to predict higher FoI values in both settings. The patterns observed in the entire country seemed to follow what was observed in the catchment areas (municipalities where at least one serosurvey was conducted).

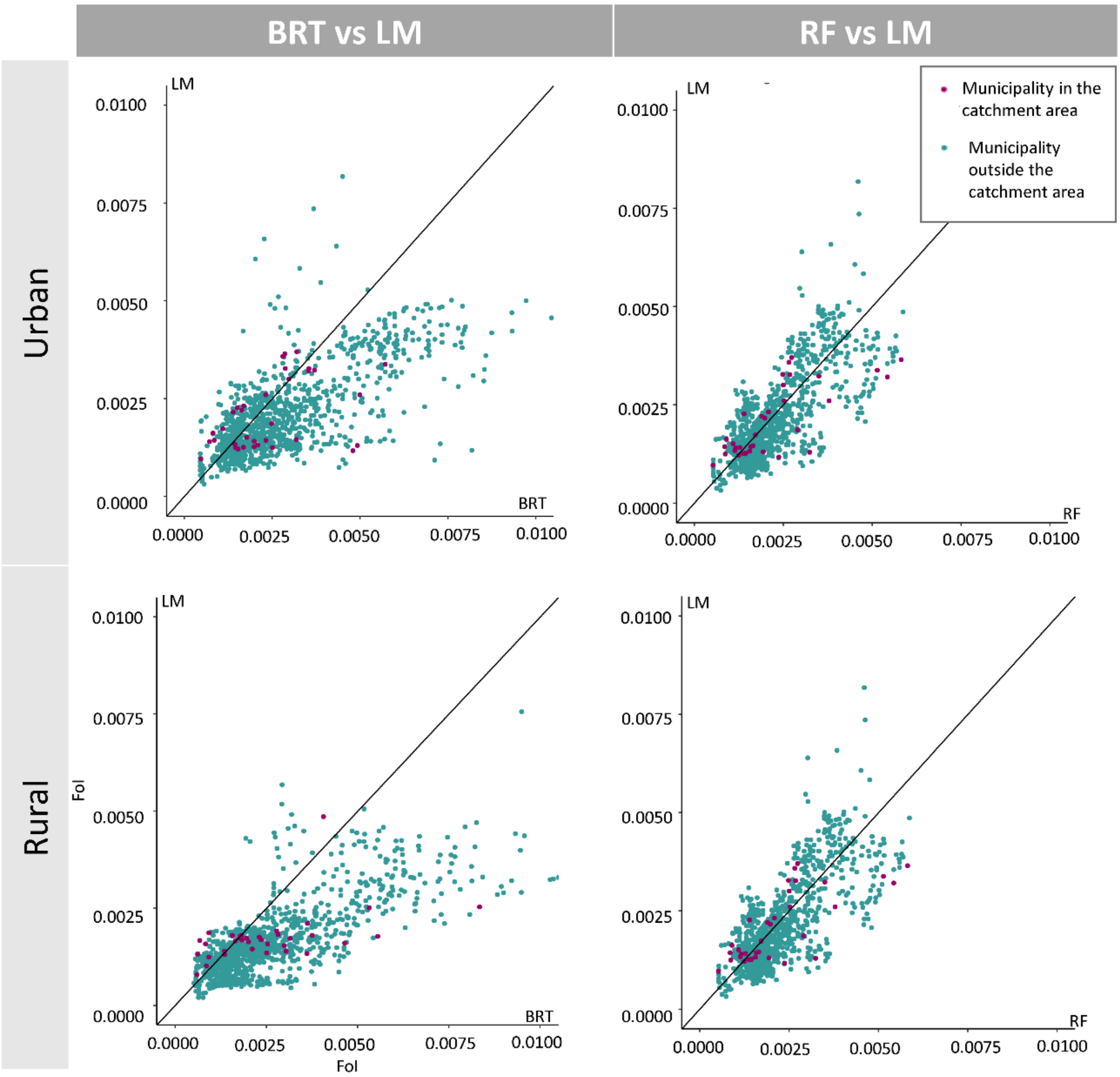


*Figure 3. 4: Comparison of predicted Chagas disease Force-of-Infection (FoI) values for urban or rural settings at municipality level, in Colombia for the year 1990. The values were obtained by two Machine Learning (Boosted Regression Trees (BRT) and Random Forest (RF)) methods and a Linear Model (LM) framework using log-transformed FoI estimates from the FullPostFoI approach (see Models' workflow subsection in Methods and Figure 3.1 for a description of this approach). The upper panel presents the results for urban settings; the lower panel presents the results for rural settings. Purple-coloured dots denote municipalities where at least one serosurvey had been conducted ('in catchment area'); teal-coloured dots denote municipalities where no serosurveys had been conducted or were not included in our analyses ('outside catchment area'). The black solid diagonal line represents perfect agreement between the two frameworks being compared.*

## 4. Comparison of temporal trends in predictions across serosurveys

When comparing 'observations' with predictions over time, all methods performed well regarding their ability to capture central trends (Figure 3.5). However, the LM framework was better at capturing uncertainty, as the confidence bounds of the predictions mirrored more closely the credible intervals (CrI) of the 'observations'.

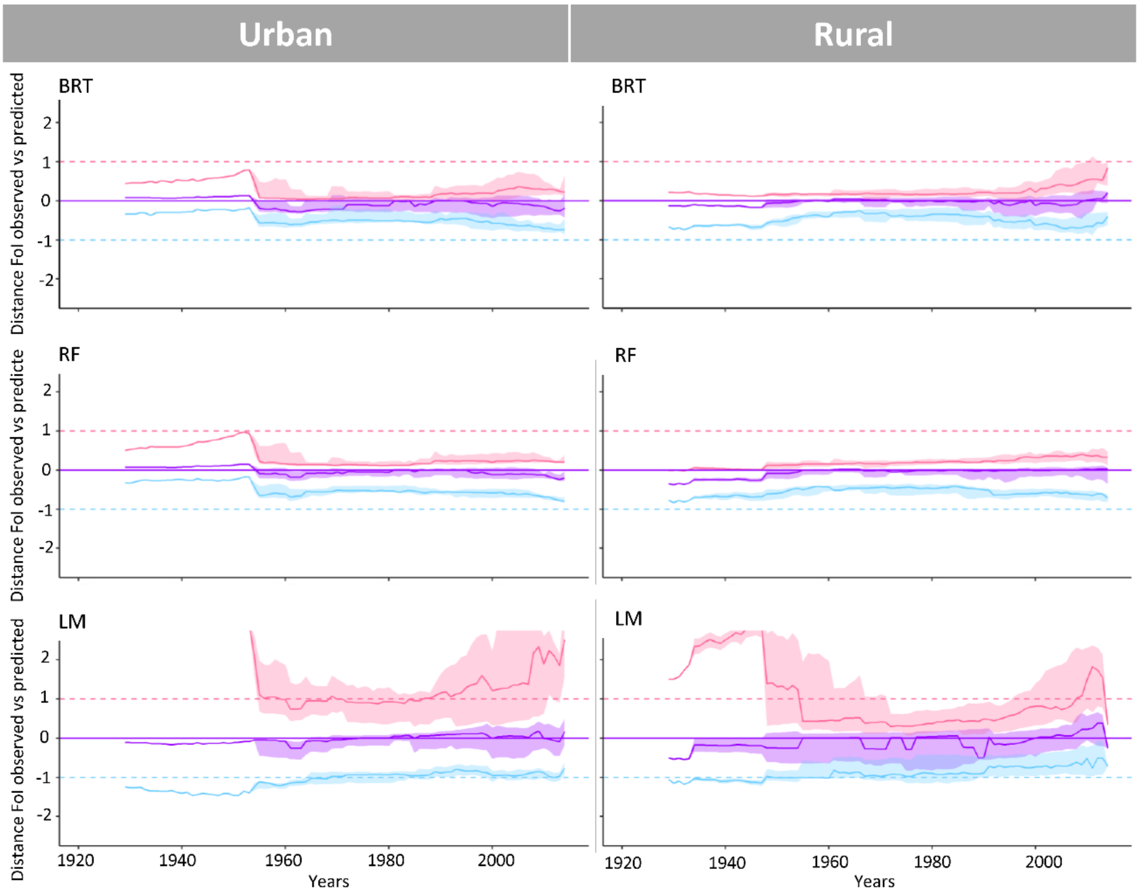


*Figure 3. 5: Standardised comparisons of 'observed' and predicted distributions across serosurveys and by setting type. Comparisons were made for two Machine Learning (Boosted Regression Trees (BRT) and Random Forest (RF)) methods (upper and middle panels) and a Linear Model (LM) framework (lower panel) using log-transformed (log) Force-of-Infection (FoI) estimates from the FullPostFoI approach for urban and rural Chagas disease settings in Colombia across 9 decades. The solid lines and envelopes show standardised distances between FoI 'observations' and predictions, with purple-colour lines representing the median, and the pink and blue lines representing, respectively, the upper and lower bounds of the 95%CrI. If medians and 95% confidence bounds of the predictions matched exactly the corresponding measures for all the 'observations' across serosurveys, then the solid and dashed lines would fully overlap.*

The median uncertainty across municipalities (Table 3.3) was comparable using any of the methods and restricting the assessment to 'in catchment area' only (i.e., municipalities where at least one serosurvey had been conducted) or not ('out of catchment area'). However, for some municipalities, the uncertainty associated with the LM framework increased dramatically.

Comparatively, the RF method produced more uniform uncertainty across predictions, with median and range similar to those yielded by the other (BRT and LM) methods, but with fewer municipalities with substantial uncertainty (defined as MAD-CV>2), and only a moderate number of municipalities with extreme uncertainty (defined as MAD-CV>5).

*Table 3. 3: Uncertainty across Chagas disease Force-of-Infection predictions for the three frameworks under comparison. The uncertainty was estimated using the Median Absolute Deviation Coefficient of Variation (MAD-CV) of the predictions for Colombia in 1990, in (urban and rural) areas where at least one serosurvey had been conducted ('in catchment area') and where no data were available or used in the analyses ('out of catchment area'). The number of municipalities where MAD-CV is greater than 2 (substantial uncertainty) or greater than 5 (extreme uncertainty) is also included.*

| | MAD CV values (range) | | | | Number of municipalities MAD CV> 2 | | Number of municipalities MAD CV> 5 | |
|---|---|---|---|---|---|---|---|---|
| | In catchment area | | Out of catchment area | | | | | |
| | Urban | Rural | Urban | Rural | Urban | Rural | Urban | Rural |
| BRT | 1.45<br>(0.31-7.16) | 1.54<br>(0.39-5.40) | 1.48<br>(0.31-7.41) | 1.48<br>(0.17-6.32) | 338 | 335 | 25 | 24 |
| RF | 1.47<br>(0.48-5.28) | 1.45<br>(0.40-5.29) | 1.48<br>(0.47-5.24) | 1.49<br>(0.44-5.22) | 145 | 198 | 10 | 8 |
| LM | 1.60<br>(0.70-2.73) | 1.29<br>(0.44-2.76) | 1.48<br>(0.32-8.19) | 1.50<br>(0.24-11.00) | 284 | 266 | 6 | 11 |

BRT: Boosted Regression Trees; RF: Random Forest; LM: Linear Model.

### 5. Residual spatial and temporal correlation

While the ML-based methods did not show any significant spatial correlation in their residuals, this was not the case with the LM framework (Table 3.4). For all models, the DW test's statistic (see Eq.4) showed a significant residual temporal correlation between residuals from the same serosurvey, with a stronger effect for serosurveys conducted in indigenous settings (Supp. Figure 5 in Appendix).

*Table 3. 4: Spatial and temporal correlation test statistics and statistical significance of the spatial correlation test applied to the cross-validation residuals for the two Machine Learning (BRT, RF) and the Linear Model (LM) methods under consideration.*

| | BRT | RF | LM† |
|---|---|---|---|
| Spatial correlation test: | | | |
| Moran's *I* statistic | 0.00 | 0.00 | 0.06* |
| | | | |
| Temporal correlation test: | | | |
| DW statistic | 0.06* | 0.04* | 0.00* |

BRT: Boosted Regression Trees; RF: Random Forest; LM: Linear Model.

DW: Durbin-Watson statistic (see Eq. 1).

†See methods for calculation of the LM residual correlation.

*p-values significant et 5%.

### 6. Computational aspects

Computationally, RF and BRT required the least effort (31 and 42 hrs respectively, on standard laptop, with an i7-8565U processor and 16.0 GB RAM) (Table 3.5). By contrast, although implementation of LM required far fewer R-packages than the ML framework, it took over twice the time to run when compared to RF (72 hr). Also, the computer hard-drive space that was required to store 'objects' and model outputs was about 20 times higher for LM than for the ML framework. Finally, the overall implementation of the models was substantially simpler for the ML framework; particularly to make adjustments and updates.

***Table 3. 5: Comparison of computational aspects for the Machine Learning (Boosted Regression Trees (BRT), Random Forest (RF)) and Linear Model (LM) methods investigated.*** *The methods under comparison used log-transformed FoI values from the FullPostFoI approach*

| | BRT | RF | LM |
|---|---|---|---|
| Number of R packages needed | 20 | 20 | 6 |
| Time required for models to run (hr) | 42.5 | 31.0 | 72.0 |
| Hard-drive space requirements (MB) | 149 | 114 | 2,048 |

BRT: Boosted Regression Trees; RF: Random Forest; LM: Linear Model.

hr: hours; MB: Megabytes.

## Discussion

Our comparative analyses indicated generally consistent results among the three modelling methods investigated to generate Chagas disease FoI predictions, namely, the linear model

(LM) framework we previously developed (103), and the two Machine Learning (ML) methods explored here (Boosted Regression Trees (BRT) and Random Forest (RF)). The predictors that were selected, as well as the location of the most uncertain FoI values were coherent and generally consistent among the three methods (Table 3.2, Figure 3.3, Figure 3.4). Not entirely surprising, RF was faster to run than BRT and LM (121) (Table 3.5).

Based on the performance indicators used, RF performed best (Table 3.1) but did less well when considering the propagation of uncertainty in the FoI inherited from the catalytic model (Figure 3.5). Also, RF generated fewer municipality-level predicted values with substantial or extreme uncertainty (Table 3.3). All methods, when fitted on the median FoI alone (MedFoI approach), were unable to capture the uncertainty in the response variable (the FoI 'observations' generated by fitting the catalytic model to the age-stratified serosurveys), leading to overconfident predictions (with high predictive $R^2$ values but smaller % of overlap values). This highlights an important issue not fully addressed in the literature, as most publications using FoI data to infer spatiotemporal patterns of infectious disease incidence tend to use the central FoI estimates alone to fit predictive models (i.e., using what we labelled here as the MedFoI approach). We argue that neglecting to appreciate and propagate the uncertainty inherent in their estimation (61,66,80) may therefore lead to significant over-confidence in predictions. This issue, already highlighted in our previous LM work (103), is not mitigated by implementing ML frameworks, and deserves careful consideration, not only from a methodological perspective, but importantly, when the results are applied in policy-relevant contexts (82).

Indeed, quantifying and communicating uncertainty in FoI appropriately is critical when the results of predictive models are used to inform stakeholders and public health programme managers on the level of certitude associated with exposure risk or number of cases. Thus areas/populations for which exposure has been certainly high or low can be differentiated from those with exposure levels or number of cases that necessitate further investigation due to highly uncertain estimation.

Even when the three methods showed good performance and generally good agreement at the serosurvey level (Figure 3.4), the residuals remained correlated in time (Table 3.4). Thus, the correlation inherited from the FoI calculation was not fully accounted for in any of our methods, i.e., none of the predictors included was able to account for the full extent of this correlation.

While the final ML models showed no evidence of residual spatial correlation (Table 3.4), the spatial extrapolation shown (Figure 3.3) should be interpreted with caution, as the (ground-truth) serosurveys available had only been conducted in a relatively small number of municipalities and tended to be aggregated in the same area (Supp. Figure 2). When using RF, the degree of uncertainty inside and outside 'catchment areas' was consistent, suggesting reliable extrapolation. This contrasted with the LM framework, which predicted large uncertainty in some municipalities.

Most of the earliest serosurveys (up to early 2000) seemed to have targeted high-risk populations (14), presumably because the perceived risk of Chagas disease transmission in those areas was higher and required improved situational awareness. However, using only such information to make predictions across Colombia would have led to higher predicted FoI in areas where no ground-truth data had been collected. By contrast, the most recent serosurveys (2010-2014) seem to have been conducted on more representative samples of the population, presumably motivated by providing a more realistic assessment of the epidemiological situation nationally and demonstrating progress in reducing vectorial transmission. We, therefore, included the year when the serosurvey took place to account for this bias and, in fact, this variable appeared to be one of the most influential in all three methods and particularly for BRT and RF (Table 3.2). This demonstrates the crucial importance of understanding the motivation behind the implementation of serosurveys in order to assess the sampling strategy and ultimately quantify potential biases that may interfere with the representativeness of FoI estimates. Indeed, in context where resources are limited, rational and cost-effective decisions have to be made to reach primary objectives of the survey even at the detriment of statistical representativeness of the sample of the population targeted.

Finally, and regarding computational aspects, the LM framework required substantial user-input to prepare the data for model fitting (including data transformation; choice of predictors included in each model; tests for spatial and temporal correlation, etc.). In contrast, ML frameworks were faster (particularly RF) and required less pre-processing of the data and hard-drive space (Table 3.5). These features render the ML models more flexible, more readily updatable, and thus easier and simpler to be extended to other Chagas disease-endemic countries, and potentially to other infectious diseases, including neglected tropical diseases, for which serological surveys are regularly conducted as surveillance tools to assess epidemiological situation, incidence, and impact of control interventions across spatial and temporal scales (61,62,80).

## Concluding Remarks

ML methods are increasingly used to derive computationally efficient algorithms for data analysis that are agnostic to the distributional properties of such data. They represent an attractive modelling tool for the generation of predictive maps of important infectious disease epidemiological metrics, such as the FoI. Most published literature on the subject use measures of FoI central tendency, neglecting to quantify, propagate and ultimately communicate the uncertainty appropriately. We show that the choice of modelling framework requires careful consideration according to the ultimate objectives of the modelling endeavour. If the aim is, for instance, to use the predicted FoI patterns to provide numbers of cases and estimates of the associated disease burden, ML framework (and particularly RF) would indeed be an optimal choice, as capturing the median (central tendency behaviour) may be sufficient and computationally efficient. However, if the objective is to identify areas where serological surveillance surveys are scarce and should be conducted to improve the reliability of FoI estimates and provide ground-truth data, we conclude that the LM framework, albeit more time-consuming and computationally intensive, would provide a better indication of where uncertainty is greatest. Although in this paper we focused on Chagas disease in Colombia as a case study, the modelling frameworks compared here can be applied to other Chagas disease-endemic countries and to infectious diseases (including neglected tropical diseases) for which age-stratified serological data are regularly collected.

# Chapter 4: From Serological Surveys to Disease Burden: A Modelling Pipeline for Chagas Disease

*In preparation for submission to the 'Challenges in the fight against Neglected Tropical Diseases' issue of Philosophical Transactions B.*

## Summary

### *Background*

In 2012, the World Health Organization (WHO) set the elimination of Chagas disease intradomicilary vectorial transmission as a goal in its first neglected tropical diseases (NTDs) roadmap. After a decade, progress has been made, but the new 2021–2030 WHO roadmap on NTDs has set even more ambitious targets. The challenges raised require innovative and robust modelling methods to monitor progress towards these goals.

### *Methods*

We have developed a modelling pipeline using local prevalence data to obtain national burden estimates at the municipality level while propagating uncertainty in ways that are consistent when aggregated across different locations to give a broader scale perspective. initially, local seroprevalence information is used to estimate the local trend in temporal exposure (quantified by the force-of-infection (FoI). Exposure estimates from such surveys are then used to predict spatiotemporal trends across larger geographical areas. Finally, large-scale predicted exposure estimates (based on the fine spatial resolution), are used to estimate disease burden based on a disease progression model.

### *Findings*

Using 76 serosurveys conducted in Colombia between 1990 and 2020, we estimated that the number of infected people would reach an estimated 506,000 (95% CrI: 395,000-648,000) in 2020 with a 1.0% (95% CrI: 0.8%-1.3%) prevalence in the general population and 2,400 (95% CrI: 1,900-3,400) deaths (~0.5% of those infected). Temporally, the interplay between a slight

decrease in exposure was overcompensated by the large increase in population size and the gradual ageing of the population, leading to a substantial increase in the burden of Chagas disease over time.

### *Interpretation*

The modelling pipeline has been initially built with Colombian data but can be used on other Chagas disease endemic countries or even on other long-lasting infectious diseases for which serosurveys are conducted.

## Research in context

### *Evidence before this study*

Chagas disease, caused by infection with *Trypanosoma cruzi*, is a long-lasting disease. Therefore, current prevalence data do not truly reflect the epidemiological situation. For instance, high prevalence potentially reflects a high level of past transmission rather than current exposure. Using mathematical modelling, seroprevalence studies can be used to reconstruct temporal trends in the force-of-infection (FoI, the per-susceptible rate of parasite acquisition.)

Therefore, age-stratified seroprevalence studies have the potential to provide a largely untapped resource to predict spatiotemporal trends in Chagas disease incidence, which can, in turn, be used to predict the burden of Chagas disease over time and space at a resolution much finer than that available from current national estimates.

### *Added value of this study*

We present a modelling pipeline able to estimate the incidence and burden of disease. We collated information from 76 seroprevalence studies in Colombia, from published and unpublished sources between 1990 and 2020. Those studies were used to estimate local temporal trends in the FoI. Spatiotemporal predictive models were used to obtain FoI estimates over the last 7 decades at the municipality level across Colombia. Finally, those estimates were used in age-specific compartmental models linking infection to disease to estimate the burden of Chagas disease and its spatial and temporal heterogeneities.

*Implications of all the available evidence*

Our study highlights the benefit of using currently available but largely under-utilised seroprevalence studies to inform the burden of Chagas disease. Our modelling pipeline relies on robust statistical modelling which propagates the various uncertainties at each step, providing a more realistic assessment of the past and current epidemiological situation.

By providing age-specific and spatially resolved estimates of disease burden, we hope to assist public health professionals in the targeting of specific interventions (e.g. those targeting vectorial transmission vs. those targeting improvement in diagnosis and treatment).

## Introduction

Chagas disease is a Neglected Tropical Disease (NTD) caused by the protozoan parasite, *Trypanosoma cruzi*. Vectorial transmission (by reduviid, triatomine bugs) is the main (but not exclusive) transmission route. While Chagas disease is endemic in 21 Latin American countries, population migration has resulted in its globalisation. Infections can remain asymptomatic for many years, with 20-35%% of those infected eventually developing clinical manifestations and requiring medical interventions (4). Such interventions (including treatment) aim at alleviating symptoms and/or reducing disease progression when possible. Disease control efforts have mainly focused on infection prevention (e.g., through vector control, education, and housing improvement) and on testing for prompt identification of asymptomatic cases (14).

In 2012, the World Health Organization (WHO) set the elimination of intradomiciliated vectorial transmission in the Americas by 2020 as a goal in its first NTD roadmap (128). After a decade, progress has been made but the new 2021–2030 WHO roadmap on NTDs is even more ambitious, proposing that all routes of transmission be interrupted in nearly 40% of the endemic countries by 2030 (20). The application of innovative and robust statistical methods can help monitor the epidemiological situation and the progress to be made to meet this challenge. To this end, estimating the spatiotemporal variations in disease exposure is critical, but this is hampered by weak surveillance (14) (e.g. in Colombia in 2021, 306 chronic and 172 acute cases were reported, with only 170 and 14 of them being confirmed, respectively (129,130)). By contrast, estimations of the number of cases for the country from WHO, Global

Burden Model (GBM) and others ranged between 186,000 and 438,000 for the 2005-2010 period (4,8,24,25).

Cross-sectional, age-stratified serological surveys have been used to estimate past trends in exposure in the context of Chagas disease (21), dengue (66,72,76,80), malaria (70), schistosomiasis (69) and yellow fever (61). Provided enough surveys are available, predictive models can be used to estimate spatiotemporal trends in exposure for Chagas disease (103) and other infections (61,65,66,80). The crucial next step is that of linking such trends with models of disease progression so robust estimates of disease burden can be obtained to better target the necessary interventions (8). However, in some of the applications above mentioned, estimates of the FoI have been assumed to be constant over time (66,75,80) or only average FoI values have been used to fit predictive models (61,72,75,80), substantially neglecting the associated uncertainty. Therefore, appropriately propagating the uncertainty surrounding each step is essential for reliable estimation of disease burden.

In this paper, we propose a modelling pipeline to produce robust municipal, departmental, and national estimates of the disease burden of Chagas disease, acknowledging the uncertainties associated with each step of the process, using Colombia as a case study. In Colombia, seroprevalence surveys have been conducted by governmental and non-governmental organisations over time, providing a rich source of information to showcase our approach. We proceed to discuss its applications for estimating disease burden across the remaining Chagas disease endemic countries in the Americas, as well as for other infectious diseases for which seroprevalence surveys are conducted and models linking infection and disease can be formulated.

## Material and Methods

### 1. The DICTUM platform

With Pan-American Health Organisation (PAHO) support, the Decreasing the Impact of Chagas Disease Through Modelling (DICTUM) platform has been created to collate, standardise, and communicate data relevant to Chagas disease epidemiology, including information on serosurveys, vector surveillance and blood-banks screening. A key aim of the platform is to then exploit such a database to inform public health professionals on key relevant aspects of its current epidemiology, e.g., to obtain estimates of the number of asymptomatic, chronic, and severe cases by age classes allowing targeting of diagnostics and treatments activities.

The process of estimating the burden of Chagas disease using local serosurveys (Figure 4.1) involves three steps:

1) Local seroprevalence information is used to estimate the local trend in temporal exposure (quantified by the Force-of-Infection (FoI).
2) The estimates of exposure from various surveys are used to predict spatiotemporal trends across larger geographical areas.
3) Large-scale predicted exposure estimates, at a fine spatial resolution, are used to predict the burden based on a disease progression model.

Special attention is given to propagating all uncertainty between steps and at different spatial and temporal scales (Supp. Method 5).

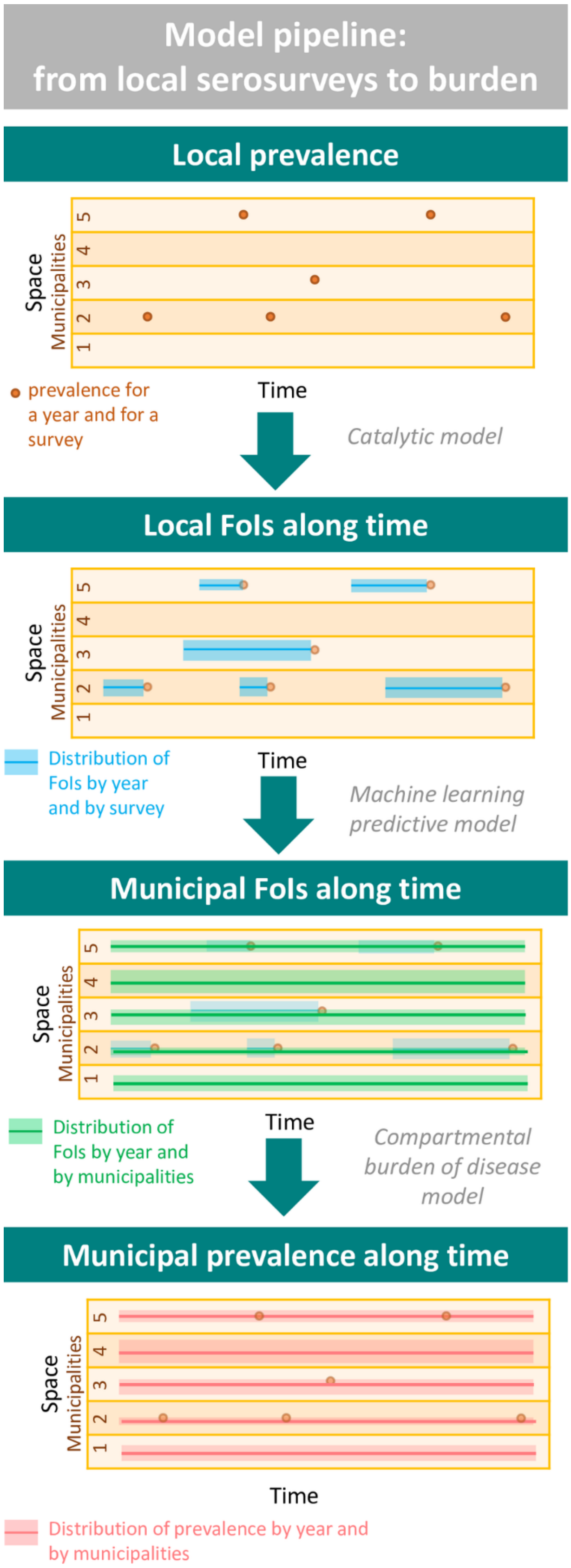

Model pipeline:
from local serosurveys to burden
Local prevalence
Space
Municipalities
5
4
3
2
1
Time
prevalence for a year and for a survey
Catalytic model
Local FoIs along time
Space
Municipalities
5
4
3
2
1
Time
Distribution of FoIs by year and by survey
Machine learning predictive model
Municipal FoIs along time
Space
Municipalities
5
4
3
2
1
Time
Distribution of FoIs by year and by municipalities
Compartmental burden of disease model
Municipal prevalence along time
Space
Municipalities
5
4
3
2
1
Time
Distribution of prevalence by year and by municipalities

*Figure 4. 1: Modelling pipeline, from local serosurveys to National burden of diseases' estimates. Using local prevalence data, the modelling pipeline use three consecutive models (FoI catalytic model, FoI predictive model and* *compartmental burden of disease model) to obtain prevalence and burden of disease information across the country at the municipal level while propagating uncertainty from one model to another. The Catalytic model gets as input the prevalence by age class for each serosurvey and produced as output the FoI by year for each serosurvey. The Machine learning predictive model receives as input the FoI by year for each serosurvey and a set of environmental and demographic predictors, these are used to produce as output the FoI by year at the municipality level. The Compartmental burden of disease model then receives as input the FoI by year and the population size at the municipal level and produces as output the prevalence by year and by age class of different stages of the diseases (including the death) at the municipal level.*

## 2. Step 1: the FoI at the serosurvey level (catalytic model)

We relied on 76 serological surveys (serosurveys) conducted in Colombia and organised at the municipality level (Supp. Figure 1). These age-specific seroprevalences were used to fit a Bayesian catalytic model and provide yearly estimates of the local FoI (i.e., in the catchment areas) from the birth of the oldest participants to the year the survey was conducted. For each serosurvey, the model estimates smoothed profile of the FoI relying on an autoregressive process, which allowed us to obtain yearly samples of the joint posterior distribution of the local FoI (see (21) and Supp. Method 1 for more detail).

## 3. Step 2: Predictive model for the FoI at the municipal level

*Predictors included*

The second step aimed to predict the FoI in areas where no serosurveys have been conducted (i.e., outside of the catchment areas), to obtain yearly FoI predictions across the entire country, at the municipality level.

Informed by previous studies (103,131) and in order to build a pipeline that could be applied in other countries, the predictors selected were available across Latin America and included characteristics of the serosurveys, as well as spatiotemporal environmental and demographic predictors.

The setting where the serosurvey was conducted was included and defined as urban, rural, indigenous or mixed settings (composed of urban and rural settings). The urban/rural definition followed the government's one (88).  None of the large urban centres with a population of over 100,000 inhabitants in 1985 was included in the catchment area of the serosurveys, therefore the FoI predictions made relate to small to medium-size cities.

The year when the serosurvey was conducted was included to correct for a selection bias in the early serosurveys, i.e. serosurveys organised before 2000 were more focused on high-risk populations (103), especially in rural settings.

We allowed a temporal trend by including the 'midpoint'  'years' as a covariate but assume other predictors would account for spatial heterogeneities and therefore latitude and longitude were not included (111).

In terms of environmental predictors, we focussed on climatic variables and indicators of the triatomine vectors. BioClim for Colombia were collated between 1979 and 2013 on a 1km$^2$ scale. Following the triatomine niche modelling literature (29–31,35,36,40,42,43,47,50,53,132), we focussed on relative day-night temperatures differences (Bio03), median minimum temperature (Bio06), and seasonality of precipitation (Bio15). We included additional predictors based on available literature that included median municipality elevation (27,42,43,47) and vegetation index (i.e. NDVI)(27,31,39,42,57).

The year when a municipality has been certified free from infestation was used as a predictor of the model. Our analysis did not use further vector indicators 1) to keep the pipeline flexible in terms of country-specific information available, and 2) implicitly assuming that environmental variables included above would encapsulate such information.

Finally, population sizes and proportions of the municipal population living in urban settings were also included along with an IPUMS indicator characterising the proportion of houses with unfinished floors (133), a proxy for poverty with relevance to Chagas disease (housing improvement being highly correlated with vector intra-domiciliary infestation (12)).

A full description of the predictors is given in Supp. Table 8 and Supp. Method 4.

### Model definition

We used the available collated data to predict the spatiotemporal trends in the FoI between 1950 and 2020, at the municipality level across Colombia using a random forest regression model (126). Following previous work (131), a nested resampling has been applied for model tuning with a spatial resampling strategy. Cross-validation (CV) was used to assess model performance and included half of the data, which were excluded from the fitting process. To propagate the uncertainty inherited from the calculation of the FoI, the fitting and performance evaluation was repeated with 100 bootstrap samples from the posterior distribution of the FoI.

A composite performance indicator was used to assess performance and estimated as the mean of the coefficient of variation ($R^2$) calculated among the cross-validation (CV) set and the percentage of overlap between predicted and "observed" distributions of the FoI using the function "overlap" from package overlapping (123), as in (131). This performance indicator ensured that predictions reflected the central tendency, while also accounting for the uncertainty in the response variable. More detail on the modelling process is available in Supp. Method 5.

As previously mentioned, serosurveys were not available for large cities, therefore our predictions of 'urban' exposure were representative of small to medium cities, and not of large cities. For the cities having a population size bigger than 100,000 in 1985 (based on DANE estimates (88)), the prevalence observed in blood banks was used to estimate a constant FoI.

### 4. Step 3: from FoI to burden

To estimate the spatiotemporal trends in the burden of Chagas disease in Colombia, we developed a model of disease progression, reflecting the life history of infections (Figure 4.2).

The progression, or burden, model consists of an age-specific compartmental model that estimates the prevalence at each stage of the disease for each age class. This model use parameters that describe the disease progression and mortality (see Supp. Method 6).

In the progression model, individuals may acquire the parasite at a rate specified by the municipality specific predicted FoI for a given year. Immediately after infection, while some may present no or mild symptoms, while a given proportion may develop acute symptoms, i.e., acute phase of Chagas disease, with symptoms including cardiomyopathy. Then cases who displayed mild or no symptoms will transition to the indeterminate phase, during which they will remain largely asymptomatic. Cases in the indeterminate and severe acute phase can then progress to the mild and thereafter severe chronic phases.

Individuals in the acute, chronic mild and chronic severe phases contribute most to the mortality associated with Chagas disease. Progression to mild and severe chronic phases (i.e. with mild or severe cardiomyopathy) may be linked to co-morbidity rather than T. cruzi infection itself. We, therefore, allowed all infected, and non-infected, to transition to those phases regardless of infection status. Digestive forms of Chagas disease (e.g. megaoesophagus or megacolon) were not included in this model as they are uncommon in Colombia (8).

From this progression model, we tracked for each cohort the proportion of individuals in each stage, as well as the yearly proportion of the cohort deaths that were directly attributable to *T. cruzi* infection. Cohort size over time was informed by census data, while yearly mortality per cohort was informed by death line-listing.

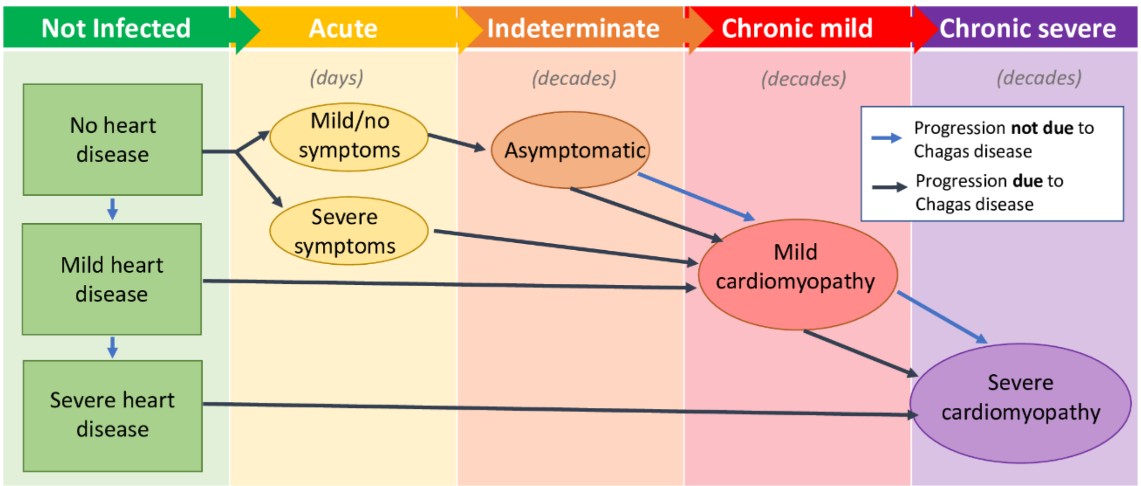


*Figure 4. 2: Chagas disease burden model. Schematic representation of the compartmental model used to obtain the burden of disease. The model is taking into account comorbidity; indeed, patients can have heart diseases before getting infected by Chagas disease. For each compartment in the model, the prevalence by age class was calculated as long as the mortality caused or not caused by Chagas disease depending on if the progression was or not due to the disease. Details on the progression rates and compartments used are provided in Supp. Method 6.*

All the analyses have been realised on R studio (134).

### 5. Data Sharing

The data will be made available on a repository upon publication.

## Results

### 1. FoIs at the serosurvey level

The seroprevalence data by age class were used to back-calculate the FoI using a catalytic model. Figure 4.3 presents a subset of the fitting process for the serosurvey with the largest and smallest sample sizes. The full 76 serosurveys are presented in Supp. Figure 6.

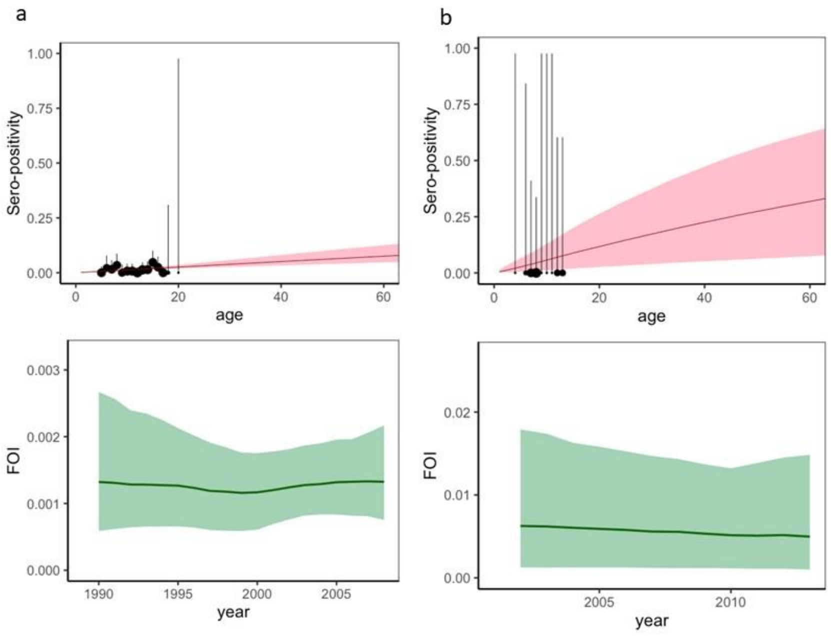


*Figure 4. 3: Catalytic model fit for the survey with, panel a: the largest sample size (n=1,680); and panel b: the smallest sample size (n=30). The upper panels represent the prevalence by age class observed as black dots and the inferred prevalence in pink. The lower panels represent the FoI along time in green.*

### 2. FoI Predicted yearly FoI at the municipal level

The predictive model of the FoI showed good performances with a coefficient of variation ($R^2$) on the cross-validation set of 64% in urban and 71% in rural areas. Model uncertainty was well propagated with predicted and observed FoI distributions showing a 59% overlap (Supp. Table 9).

To accurately predict FoI's, including "the year when the serosurvey was conducted" was, in term of importance (which represent how helpful the predictor has been to the models) reaching 99.2 (table 4.1). Accounting for settings was also key, especially distinguishing between indigenous vs. non-indigenous settings (importance of 87.7). Further accounting for differences between urban and rural settings had more marginal importance (around 11). All environmental (excluding NDVI) and demographic predictors substantially improve the fit, with importance ranging from 19.0 to 43.7. Both year and NDVI were associated with more limited improvement (approximately 10).

Predicted FoI across Colombia showed similar patterns in urban and rural settings (Figure 4.4), with high spatial heterogeneity and a substantial decrease in exposure over time, especially in the Andean and northeast. The median of the municipal FoI ranged from $2.0\text{x}10^{-4}$ to $2.8\text{x}10^{-3}$ in 1995 and $1.6\text{x}10^{-4}$ to $2.7\text{x}10^{-3}$ in 2020 in urban settings and from $2.1\text{x}10^{-4}$ to $2.8\text{x}10^{-3}$ in 1995 to $1.6\text{x}10^{-4}$ and $1.8\text{x}10^{-4}$ to $2.8\text{x}10^{-3}$ in 2020 in rural settings. Uncertainty gradually decreased over time (with MAD-CV being greater than 2 for 9% and 8% of the municipality in 1995 and 2020 across urban and rural settings) and was larger in the south (where fewer surveys were available).

*Table 4. 1: Importance of the predictors used in the Random Forest model used to predict the FoI of Chagas disease in Colombia at the municipal level.*

| Predictor | Importance |
|---|---|
| Serosurvey characteristics: | |
| Year when the serosurvey was conducted | 99.2 |
| Setting type: | |
| Urban | 11.1 |
| Rural | 10.9 |
| Indigenous | 87.7 |
| Temporal Coordinates: | |
| Year | 11.7 |
| Environmental predictors: | |
| Year of certification for intradomiciliated vector elimination | 34.7 |
| Isothermality (Bio3) | 43.7 |
| Minimum temperature of the coldest month (Bio6) | 26.8 |
| Seasonality of precipitation (Bio15) | 36.4 |
| Normalized Difference Vegetation Index (NDVI) | 9.9 |
| Elevation | 29.3 |
| Demographic predictors: | |
| Population size | 36.5 |
| Proportion of urban population | 29 |
| Proportion of households with unfinished floor materials | 19 |

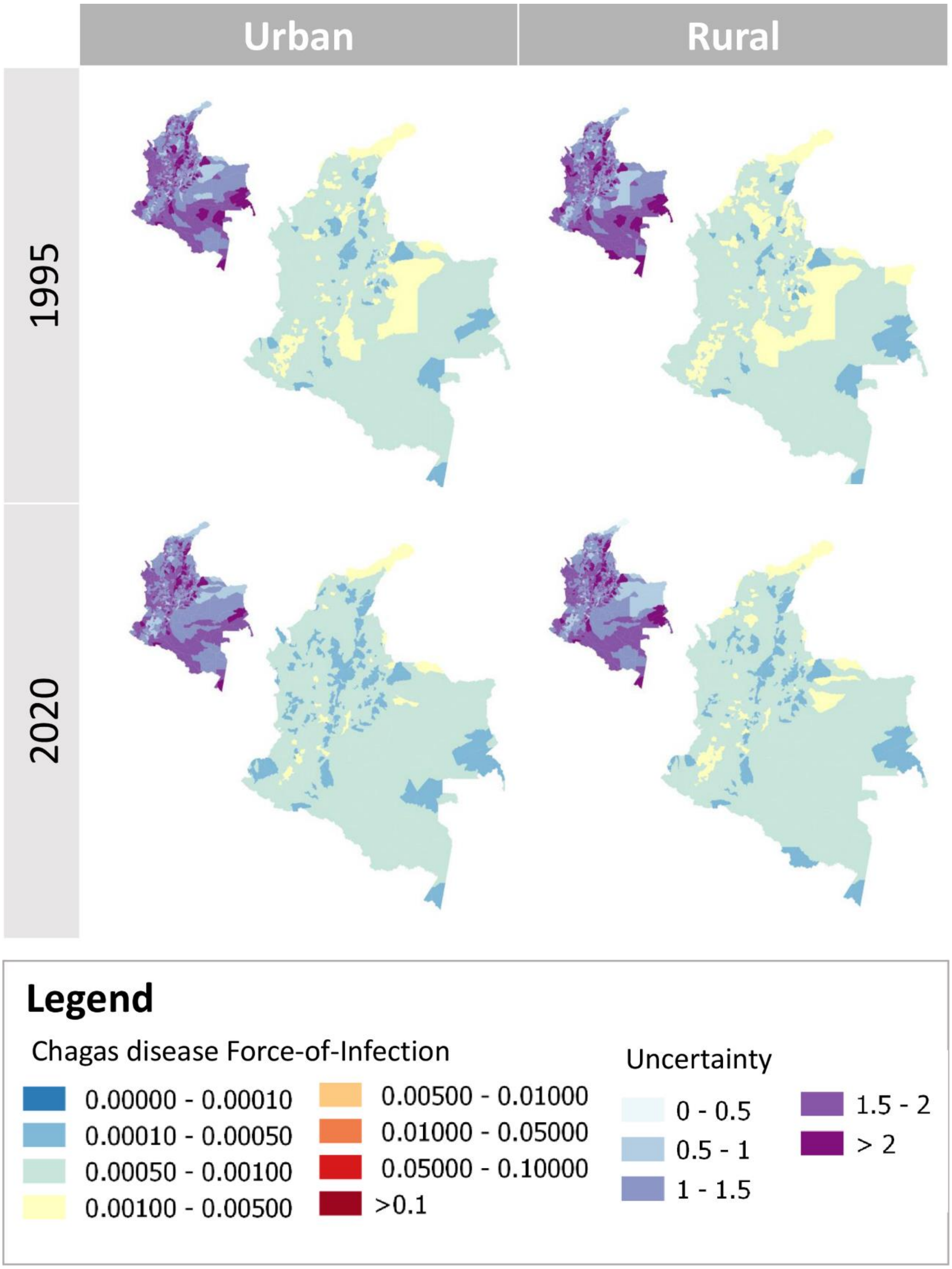


*Figure 4. 4: Spatial distribution of the Force-of-Infection of Chagas Disease (per year and susceptible individual), in Colombia at the level of municipalities in 1995 and 2020. The predicted distribution was generated using a Random Forest model (main maps); the associated uncertainty (small map insets) presents the Median Absolute Deviation (MAD) Coefficient of Variation (MAD-CV).*

### 3. Burden model over 1985-2020

Chagas disease prevalence showed some spatial disparities with the northern part of the country generally being more impacted (Figure 4.5). The prevalence was higher in rural areas with median and Interquartile Range (IQR) prevalence across municipalities in 2020 of 1.26% [1.10%;1.45%] for urban and 1.36% [1.15%;1.57%] and rural settings. However, given the higher population size in urban settings (76% of the total population in 2020), 61% of predicted cases belonged to urban settings in 2020 (Supp. Figure 7 and Supp. Figure 8).

While FoI showed an overall decreasing trend between 1995 and 2020, the prevalence was relatively stable with the national median prevalence and IQR across urban and rural settings being 1.0% (0.8%-1.2%) in 1995 and 1.0% (0.8%-1.3%) in 2020. A 6% decline in the prevalence in the acute stage was observed between 1995 And 2020, but this was compensated by a 13% increase in the predicted prevalence in the severe stage.

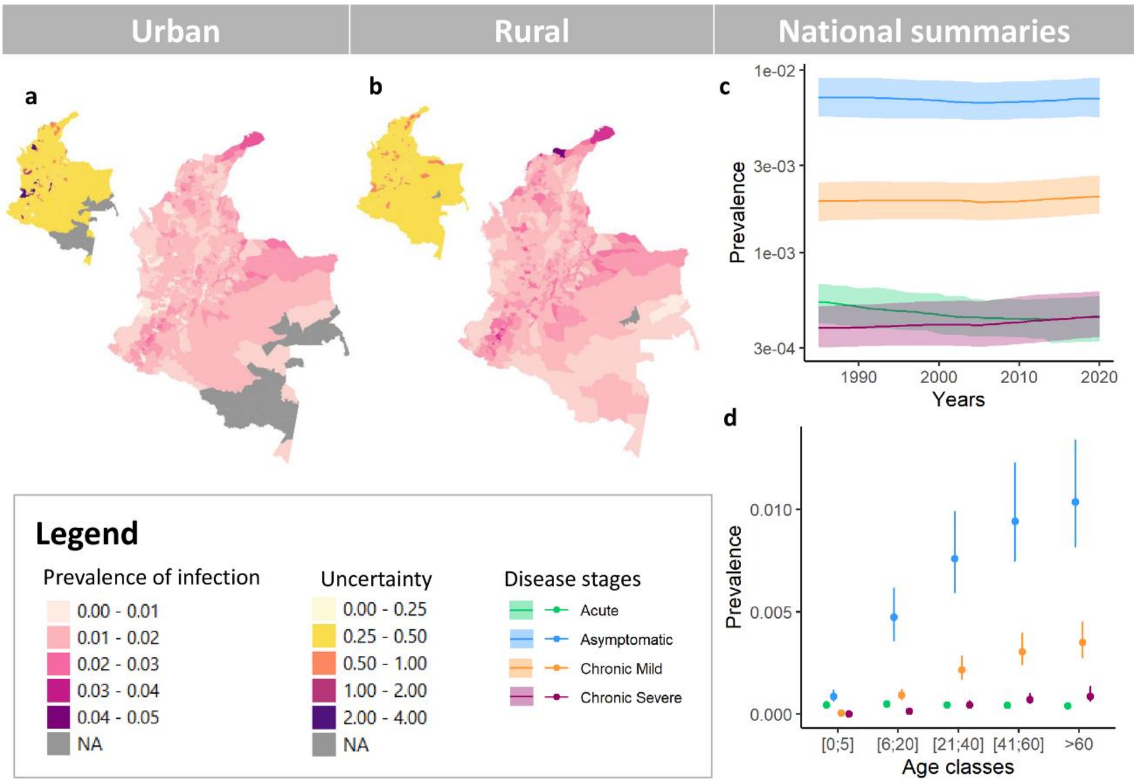


*Figure 4. 5: Spatial, temporal and age class distribution of the prevalence of Chagas disease in Colombia. a and b: Municipal prevalence in 2020 in urban and rural areas (main maps); the associated uncertainty (small map insets) represent the interquartile range divided by the median. c: Yearly national prevalence (median, solid line and interquartile, ribbon) from 1985 to 2020; each colour corresponds to a different disease stage. d: National prevalence by age class (median, point and interquartile, error bar) in 2020; each colour corresponds to a different disease stage.*

As the prevalence remain largely stable but the population increased by 39% between 1995 and 2020 (Supp. Figure 9), the overall burden of Chagas disease was predicted to have significantly increased (Table 4.2).  We estimated the total number of cases across Colombia had increased by 43% between 1995 and 2020 reaching half a million cases. Between 1995 and 2020, even larger increases of 57% and 79% applied to cases with severe cardiomyopathy and deaths attributable to Chagas disease respectively. These were driven by an increase in both population size and the gradual ageing of the population (Supp. Figure 9 and Supp. Figure 10).

Large spatial heterogeneity and clustering in the burden were observed, with the three departments having the heaviest number of deaths attributable to Chagas disease in 1995 Bogotá DC, Cundinamarca and Santander accounting for 31% of the deaths but only 25% of the total population.

*Table 4. 2: Burden of Chagas disease in 1995 and 2020.*

| Median (95%Credible Interval) | | |
|---|---|---|
| | 1995 | 2020 |
| Number of cases | | |
| total | 355,000 | 506,000 |
| | (278,000-451,000) | (395,000-648,000) |
| chronic mild | 70,000 | 102,000 |
| | (55,000-88,000) | (82,000-133,000) |
| chronic severe | 14,000 | 22,000 |
| | (11,000-19,000) | (17,000-31,000) |
| Prevalence % | | |
| total | 1.0 | 1.0 |
| | (0.8-1.2) | (0.8-1.3) |
| urban | 0.7 | 0.8 |
| | (0.6-0.9) | (0.6-1.1) |
| rural | 1.6 | 1.6 |
| | (1.1-2.1) | (1.2-2.2) |
| Children [0-5] | 0.1 | 0.1 |
| | (0.1-0.2) | (0.1-0.2) |
| Older [>60} | 1.8 | 1.5 |
| | (1.4-2.3) | (1.2-2.0) |
| Number of deaths | | |
| total | 1,400 | 2,400 |
| | (1,100-1,900) | (1,900-3,400) |
| Bogotá..D.C. | 160 | 238 |
| | (64-443) | (93-700) |
| Cundinamarca | 142 | 240 |
| | (101-200) | (164-341) |
| Santander | 129 | 210 |
| | (87-173) | (142-248) |

## Discussion

We have developed a modelling pipeline that uses local prevalence data to obtain national burden estimates at the municipality level. From the 76 serosurveys conducted in Colombia, we estimated that the number of infected people would reach an estimated 506,000 (95% CrI: 395,000-648,000) in 2020 with a 1.0% (95% CrI: 0.8%-1.3%) prevalence in the general population and 2,400 (95% CrI: 1,900-3,400) deaths. Temporally, the interplay between a slight decrease in exposure was overcompensated by the large increase in population size and the gradual ageing of the population, both of which lead to a substantial increase in the burden of Chagas disease over time. The complex dynamics observed in the burden reflect the long progression of the disease. Large spatial heterogeneities in the burden indicated that spatial targeting of interventions could improve the cost-effectiveness of resource allocation. Our results could help such targeting, as we can predict locations where the impact of the reduction in vector transmission (exposure) would be optimal; and these locations may differ if the interventions considered relate to diagnosis and treatment.

The performances of the FoI predictive models were good with cross-validation performances that would suggest limited overfitting (Supp. Table 9). However, the validation of the burden model with external data was more complex as sources are limited. Indeed, the surveillance system for Chagas disease cannot be used to validate our estimates as the reporting is extremely sparse, although, we observe that cases are detected in almost all departments of Colombia (Supp. Table 10). The four departments having reported confirmed chronic cases in 2019 (Arauca, Santander, Cesar, and Boyacá) are representing only 14% of the severe cases in our results. These departments are endemic areas that most probably have infrastructures to make the diagnostic as well as staff and population more aware of the disease which is not the case in all of the country. In Colombia, in the general population, the prevalence of heart conditions has been estimated at around 11% (135). In 2019, a total of 41,848 deaths related to heart diseases have been reported across the country (88). If, as suggested by our models, 2,400 of these deaths have been caused by Chagas disease, then Chagas disease would have accounted for 6% of the heart diseases related deaths in 2019, in Colombia, in line with Brazilian data, where Chagas disease was estimated between 1% and 21% of the in-patients with heart failure (136). Also, our model estimates 88,226 deaths among people suffering from heart disease and 23,158 among people suffering from a severe heart condition.

Our burden estimates compared well with estimates from the WHO (438 thousand cases estimated in 2010 and a 1.0% prevalence) (25) and Moncayo et al. (436 thousand cases estimated in 2005) (4). Disconcertingly, our estimated increase in the burden contrasted with a sharp decrease in burden predicted by WHO and Moncayo. Estimates from the GBD project (24) showed a similar temporal increase in burden in terms of deaths (143% increase between 1995 and 2019, compared to our estimate, 71% for the deaths), but a much lower absolute burden with 170 (74-283) deaths predicted in Colombia (compared to 2,400 deaths for our central estimate). However, in term of prevalence and number of cases, the GBM predict a stable number of cases between 1995 and 2019 with 123,000 (106,000-144,000) cases and a decreasing prevalence from 0.34% (0.29%-0.39%) to 0.26% (0.22%-0.31%) between 1995 and 2019, which is again far below our estimates. The estimates from the GBM as well as ours relied on demographic data and the demography in Colombia has shown a sharp increase during this period as well as changes in the shape of the age classes distribution with a population that is getting older (Supp. Figure 10). Indeed, the GBM at the global scale was estimating a 3% decrease in Chagas disease-related deaths for the same period with a decrease in Brazil and Argentina, but not in the other endemic countries (24). The lower estimation from the GBM might be explained by the methodology used but also, our study is the first that comprehensively incorporated published and unpublished data.

Our modelling pipeline is the first attempt to estimate the burden of Chagas disease in Colombia over a number of years at the municipal level. Serosurveys with a geographical extent that was higher than the municipality level were excluded from our analyses to be able to account for the small spatial heterogeneity in Chagas disease transmission and be able to provide estimates at the most operational level. Given the complex modelling tasks, major simplifications were made which could influence our results. First, while using information from serological surveys in indigenous populations to inform spatiotemporal trends in exposure, we did not include indigenous populations in our burden estimates. Since 1980, only 3 serosurveys have been conducted in indigenous settings at the municipality level, and we felt this was wholly insufficient to map exposure in indigenous settings across the country. We believe our overall estimates of burden would remain largely unaffected as only 0.6% of the population is estimated to live in indigenous settings, i.e., in traditional houses (88). However, improving our understanding of the spatial distribution of the burden in this setting would be critical as prevalences of up to 48.7% (95% CI 42.6% to 51.6%). have been measured in this setting in 2012 (21).

We developed our pipeline to be flexible, robust, and transparent, with most variables used to predict FoI and estimate the burden being available across Latin America. However, our model currently doesn't account for the influence of digestive morbidity and mortality. Indeed, digestive symptoms (mega-colon and mega-oesophagus) can be present in 6% of the cases (4) but are only present in the Southern Cone of Latin America (2). Currently, our model only considers the vectorial transmission route; in contexts where the vectorial transmission has been interrupted, the model will need to integrate the other transmission routes, from mother to child during pregnancy, i.e., congenital, and by organ transplant or blood transfusion. While a significant reduction in the risk related to blood transfusion and organ transplant has been observed, it remains the main transmission route in non-endemic countries (18). Crucially in the case of congenital transmission, even if the diagnosis is made during the pregnancy, access to the treatment is often delayed after the delivery. Systematic screening of pregnant women has not been implemented yet and remains an important challenge in the fight against Chagas disease. Given the current low access to treatment, the model did not consider medical improvements that might help reduce mortality due to cardiomyopathy. Finally, there were important population movements in Colombia that are not accounted for in our model and thus people who tested positive in a municipality could have been infected in another place. This is probably explaining why our model showed FoI patterns that were quite similar in urban and rural areas while the vectorial transmission is expected to be stronger in rural areas.

Getting additional prevalence data would help the validation of the modelling pipeline. However, sampling bias introduced by the sampling strategy chosen has to be considered, minimized or at least well documented. Indeed, we highlighted the importance of the year when the serosurvey was conducted as sampling tended to be directed toward at-risk populations for serosurveys before 2000. Also, sampling should be made in areas where the model suggests higher uncertainty so the model would be generally improved. If rural areas are well represented, this is not the case for urban ones. Indeed, serosurveys conducted in areas defined as urban are not representative of densely populated cities. Organising serosurveys in large cities would help improve the estimates of the burden and are crucial as the population is migrating to cities; these large cities already represent 27% of the population (88) (Supp. Figure 9). At the moment, the modelling pipeline use prevalence information from blood banks for large cities, these cities account for 20% of the cases in 2020 while it was accounting for 17% of the cases in 1995.

We are describing here this serial modelling process, using Colombia as a case study. However, the platform is hosting data from an increasing number of countries and the modelling process is ready to be implemented in other contexts. The inclusion criteria for new studies should only rely on the availability of the information necessary for the implementation of the model, and where possible include consideration of quality to ensure as much study as possible are included. Indeed, sampling strategies impact the results of the model in one direction or another depending on the population group or geographic area that is targeted. Thus, by pulling together surveys with different sampling designs, we might be able to cover a larger portion of the population and capture more variation. Currently, the pipeline includes the year when the serosurvey has been conducted to account for the sampling bias observed in Colombia which reflects the reason why the surveys have been conducted at different moments in time. However, in other countries, the use of some quality criteria could be relevant and used to associate a weight to each serosurvey depending on objective criteria defined to assess the level of quality associated with the survey. This could help reducing the impact of serosurveys' bias on the overall models.

# Chapter 5: Discussion

## Summary of findings

This project aimed to develop global geostatistical models of the risk of Chagas disease transmission using the Force-of-Infection as an input along with a set of environmental and socioeconomic predictors in order to estimate the burden at a small scale over an entire country using Colombia as a case study. Age-stratified prevalence data obtained from serological surveys are often used to estimate the historical trends in the Force-of-Infection (FoI) for a given location (21). This metric help understand spatiotemporal trends in disease incidence and exposure (61–80). However, when the FoI is used to fit another model, the uncertainty surrounding this metric is often neglected, i.e. central estimates are used to fit predictive models instead of the full distribution of values (61,80). The first objective was then to develop and evaluate a methodology that accounted for the uncertainty in the FoI inherited from the catalytic model estimation. Methodologies, based on bootstrapping and model ensemble, were proposed to integrate the uncertainty, with new performance indicators developed to assess how well the uncertainty was propagated from one prediction to the other.

Relying on the general framework developed, two modelling strategies were compared, the first based on linear models and model averaging, and the other based on Machine Learning algorithms. The two strategies confirmed that using central tendencies (mean or median) of the FoI, i.e., neglecting the uncertainty, leads to overconfident predictions.

Among all the predictive modelling approaches tested, the drivers of transmission were consistent. Environmental factors related to the presence of vectors, as well as socio-economic data and information on vector control interventions, demonstrated their usefulness to characterise Chagas disease exposure. Furthermore, serosurvey features, such as the setting (urban, rural, indigenous) and the year when the serosurvey was conducted were consistently found to be the most influential factors independently of the modelling method used. A much higher prevalence and FoI were observed in people living in the indigenous setting. Also, the year when the serosurvey was conducted was included to control for a sampling bias, i.e., older serosurvey focusing on high-risk populations. This highlights the paramount importance of the quality of the data (e.g., representativeness) above the complexity of the model.

We also demonstrated that the two modelling approaches, Machine Learning and Linear models, can be used in conjunction depending on the objectives of the analyses. The Linear Model framework performed better to identify areas where the uncertainty was high and such a framework would therefore be recommended to guide the implementation of new serosurveys. On the other hand, the Machine Learning framework showed the best performance to obtain good central estimates and a general picture of the current epidemiological situation.

Based on the results obtained, a new Machine Learning model using the Random Forest method and a set of predictors available across Latin America was developed to estimate the burden of Chagas disease. A standardised modelling pipeline was then proposed to estimate the burden of Chagas disease from the prevalence data obtained locally. First, local seroprevalence information is used to estimate the local trend in temporal exposure (quantified by the FoI). Then, the estimates of exposure from various surveys are used to predict spatiotemporal trends across larger geographical areas. Finally, large-scale predicted exposure estimates, at a fine spatial resolution, are used to predict the burden based on a disease progression model. While the modelling pipeline was initially built and implemented for and with Colombian data, its flexible nature means that it could easily be applied to other Chagas disease endemic countries or even for other long-lasting diseases where serosurveys are conducted.

Relying on predicted current and past exposure, 506,000 (95%CrI:395,000-648,000) people were estimated to be infected by *T. cruzi* in Colombia in 2020, representing a 1.0% (95%CrI:0.8%-1.3%) prevalence in the general population and leading to an estimated 2,400 (95%CrI:1,900-3,400) deaths this year. Temporally we observed a substantial increase in the burden of Chagas disease over time, suggesting a complex interplay between a slight decrease in exposure being overcompensated by the large increase in population size and the gradual ageing of the population.

# Future work and limitations

## 1. Extension of the modelling pipeline to Chile and Paraguay

Within the Decreasing the Impact of Chagas Disease Through Modelling (DICTUM) platform and in collaboration with PAHO and the governments, published and unpublished data have been gathered for Chile (245 serosurveys extracted) and Paraguay (433 serosurveys extracted). The local FoIs have already been estimated (manuscripts in preparation) and, as the predictors used in the modelling of the FoI at the municipality level in Colombia are also available in these countries, the Machine Learning model is ready to be applied in these two new contexts. Future efforts will therefore concentrate on applying the models to these two countries and comparing models' performance and relevance in new contexts. Being able to rely on the results from three different countries will then help us showcase the usefulness of our approach in Colombia, Chile and Paraguay, but also in other countries that might be interested to join the effort.

However, in applying our framework to other countries in Latin America, the disease progression model associated with the burden estimates may require some adjustments. Currently, it does not account for the symptoms, morbidity and mortality associated with digestive aetiology as they are rare in Colombia. However, these symptoms have been observed in Chile and can represent up to 21% of chronic patients, so they will need to be considered (2). A preliminary literature search revealed that the morbidity associated with these symptoms is different from the cardiac ones and the mortality is mainly related to an increased risk of cancer (2). In addition, cardiac and digestive symptoms co-occurrence have been observed in 5% to 20% of the patients with cardiac involvement (2). Gathering robust estimates of progression and excess mortality associated with digestive symptoms will be challenging, the literature seems extremely scarce and mainly based on migrant populations in non-endemic countries, which may suffer from a selection bias (137–140).

On the other hand, including in the framework countries with a much lower level of vectorial transmission, e.g., Chile, would increase the relative importance of correctly characterising vertical transmission, i.e., transmission from mother to child during pregnancy. Indeed, congenital transmission can occur in endemic and non-endemic countries. Even if most of the infected newborns are asymptomatic, symptoms such as respiratory distress syndrome, myocarditis, meningoencephalitis can happen and the neonatal death rate can be as high as 2% (2). Obtaining estimates of disease progression in this context is challenging as research efforts have been focused on the other transmission routes. However, new studies have emerged in the last few years raising awareness of this upcoming issue so more data might be available in the future (141–143).

Finally, including an increasing number of surveys from different countries may necessitate reviewing inclusion criteria. Indeed, political and practical choices made when designing serosurveys can generate sampling bias and increase the risk of having serosurvey of various levels of quality (e.g. representativness). A conservative approach should be preferred to be able to exploit all of the information available. However, a weight system could be implemented based on a list of objective quality criteria, giving more weight to the serosurveys of the highest quality, ie with rigorous sampling strategies and diagnostic workflow.

### 2. Validation of the models

Our analyses concluded that serosurveys conducted before the '00s were more focused on high-risk populations. Ideally, to validate our results, new data from representative serosurveys (i.e., that have been conducted without sampling bias) would be used to assess the robustness of our work. Having representative serosurveys conducted in the general population without assumptions on the expected prevalence would help us better understand the impact of sampling biases in previous serosurveys. The sampling biases observed in our data are linked to the focus on populations that were known to be more exposed to Chagas disease risk, i.e., the presence of vectors in houses, in the early stages of the control program's implementation. Consequently, it caused a temporal and spatial issue when all the serosurveys were aggregated. Even if we have addressed and quantified the temporal issue by adding a specific predictor in the model, one potential remaining problem is that this variable is likely to be "absorbing" a part of the true temporal trend. Indeed, in the FoI estimates calculated by Cucunubá *et al*., a sharp decrease in FoI over years was observed while this is not the case in our models (21).

Regarding the limited geographical extent of the catchment area, this issue might create overfitting in the predictive model and complicate the extrapolation of the model to areas where there is no information. This is why special attention was given to the predictors' ranges within and outside the catchment areas and a spatial resampling strategy was applied. Indeed, predictors ranges in the catchment areas cover well overall trends in Colombia except for the population sizes. High population density areas observed in the large cities are not represented in the catchment areas. Therefore, the FoI in these large cities has been treated slightly differently in the burden model.

Another avenue for model validation could involve being more inclusive in the use of serosurveys, and in particular making use of those that only provide geographical information at the departmental level, rather than the municipal level. Currently, those were excluded from the analyses. As some may be representative enough to reflect the prevalence across entirely new departments (i.e., a department for which we currently have no single municipal level survey), they could prove very useful for cross-validation.  Finding a way to include information available at the departmental level would help improve geographic representativeness. This might involve contacting the serosurvey's project manager to obtain more details about their catchment area and thus retrospectively collect more information about the surveys. This approach could prove to be much cheaper than organising new serosurveys.

### 3. Development of Models built at a very fine scale

Another interesting improvement would be to produce models at a smaller scale. The main benefit of using a 1-10 km$^2$ spatial scale instead of the municipal one would be to dissociate urban and rural areas within the municipalities. The definition of urban and rural often represents a technical issue as the administrative definition is not matching the operational and ecological ones. In Colombia, there is one urban area for each municipality, meaning that some areas defined as 'urban' would have been defined as a village in an ecological context. A modelling approach based on a fine spatial grid could resolve this issue. The setting type of the square, i.e., based on ground truth delimitation, could be used to attribute the "observed" FoI within the square, i.e., attributing the FoI "observed" in urban parts of the municipality to squares defined as urban. Then, predictors, at the square level as well, could help discriminate between urban and rural squares based on environmental and demographic criteria and give a better understanding of the factors that impact the transmission independently of the setting type.

Working at a smaller scale could also help understand whether the somehow counter-intuitive consistency observed in the estimated exposure between urban and rural areas is caused by this definition issue.  Indeed, the vectors have not settled in urban areas in Colombia (37,86). Another potential explanation for the lack of observed differences in exposure between urban and rural settings could lie in migration or the increasing level of urbanization and the long-lasting nature of Chagas disease. Areas defined as urban nowadays might have been defined as rural 10 or 20 years ago, thus, some infections currently recorded in an urban setting might have occurred in a rural setting many years ago. Similarly, infections recorded in an urban setting might have occurred in a rural setting many years ago due to mass migration from rural to urban settings.

Exploring such hypotheses would be insightful but is hindered by the availability and quality of the data needed for the models, or rather lack of thereof. The predictors used in the FoI predictive models can be found at small spatial scales (1-25 km$^2$), however, the demographic data are raising several challenges. First, the quality of the censuses is not consistent in time with older censuses having less robust methodologies. In addition, population data estimated from these censuses are not accounting for these inconsistencies and are not providing estimates with uncertainty either.  Regarding the number of deaths, a line-listing was used in the burden model (88). However, the data were not available for the entire period of interest (1985-2020), being available only from 1993. Also, the reliability of these data might need to be questioned. Comparing these data to the population's estimates could help understand if there is a large underreporting and if it is localized in certain areas or not.

### 4. Consideration for the effect of the treatment in the model

Within the modelling pipeline, some enhancements can be realised. In particular, while the FoI predictive models were accounting for the vector control interventions, the FoI predictive and burden models did not consider the impact of treatment and medical care on the epidemiology and burden of disease. The treatment is an important aspect of the Chagas disease fight and its cost was estimated to reach, in Colombia, between $46 and $8,000 per patient depending on the stage and the severity of the disease (144). The recovery of the patients from Chagas disease has also been neglected. Nevertheless, treatments have been distributed to persons diagnosed with Chagas disease. The treatments distributed in the 90s might be partly responsible for the reduction of the FoI observed in serosurveys conducted in the same areas but at a more recent time. However, no data can help validate this assumption as there are no records of the number of persons successfully treated. As treatment becomes hopefully much more accessible in the future, adding this aspect to the model will become critical.

Also, incorporating the treatment to the burden model could help understand the impact of this intervention on the epidemiology of the disease and be able to assess at which stage of the disease we should aim to treat infections to keep the best cost-effective ratio. Indeed some cost-effectiveness studies found that treating 5% of the cases yearly would significantly reduce the health and economic burden of the disease (145).  However, this is, at the moment, not the most important factor in disease burden reduction as access to treatment is still complicated and about 20% of chronic patients will never seek care (3,144). In addition, the drugs available are causing several side effects that lead to low medical adherence (2). From a technical point of view, there are very little data available to assess how many treatments have been distributed and where. Also, medical improvements to manage patients with acute and severe Chagas disease symptoms have been implemented (12) but, again, the data are scarce, leading to challenges in quantifying their impact.

### 5. Consideration for the specificity of large cities in the models

Understanding Chagas disease epidemiology in cities or highly urbanized areas remains a challenge that will need to be addressed, especially as the population in Latin America is becoming increasingly urban (98).

At the moment, our models use the prevalence observed in blood banks to estimate the FoI in large cities, i.e., cities with a population size above 100,000 in 1985. However, data from blood banks may not be representative of the entire population. They are typically biased toward healthy individuals, a selection bias called the healthy donor effect (146). In Colombia, the blood should have been screened for Chagas disease since 1995 and a questionnaire is filled out by the applicant-donors to reduce the risk of collecting blood from people infected by *T. cruzi* (147). Given such pre-selection of donors, the prevalence observed in blood banks could be underestimating the true prevalence in the general population.

While using blood banks prevalence is a strong assumption, no other data were available in large cities. Across Latin America, studies on Chagas disease in urban areas are scarce and unbiased prevalence studies in highly urbanized areas are even scarcer. A recent study in Rio de Janeiro (Brazil) has demonstrated the insight that studies in highly urbanised areas can provide (148). They observed that a large portion of the infections was carried by people over 65 years old and even more interesting, that the geographical origins of the infected persons are different depending on their age indicating that interventions have been successful (148). This type of study, which explicitly account for migration patterns, can give insight into the past exposure patterns across the country. Given this long-lasting nature, a representative survey conducted in 2022 could inform about the level of exposure in the ‘80s, i.e., relying on prevalence among those 40 and over. Therefore, unlike fast progressing diseases, we still have an opportunity to quantify the relatively long-term exposure pattern for Chagas disease. However, estimations of historical FoI trends need to be adjusted on the life expectancy of the targeted population back in time to account for the survival bias. Indeed, people infected with Chagas disease have a lower chance to reach elder age as they could have died from the disease. Also, the disease is mostly affecting vulnerable and poor populations that have a lower life expectancy. On a more technical side, there is a possibility that diagnostic tests are waning as time from the infection is going by. Identification of people that had been infected a long time ago might be more challenging but more research in this area is required to assess the probabilities associated.

Finally, in some context, vectorial transmission has been observed in urban settings (148–152) which represent a new threat that will need to be monitored. In contexts where no vectorial transmission has been observed in cities, the FoI predicted will need to reflect migration, urbanization, mother-to-child transmission and blood-borne transmission. Integrating these aspects into the modelling pipeline would represent a substantial improvement but this is, again, hindered by the availability and quality of the data that would be required.

### 6. Development of burden model for the indigenous setting

In Colombia, only a few studies have been conducted in indigenous settings (21). Thus, extrapolation of the exposure and estimation of the burden of disease was not possible in this setting. In addition, the characteristics of their exposure may differ substantially from the rural and urban populations, possibly requiring different predictors. However in Colombia, prevalence as high as 48.7% (95% CI: 42.6% to 51.6%) have been recorded in indigenous populations in 2012 (21), likely linked to increased exposure to vectors due to their traditional housing and proximity with wildlife habitat. Including these populations in the modelling pipeline, would therefore be critical. However, again, more serosurveys would need to be organised in these settings to unveil the real burden of disease in areas.

## Implications of research

Obtaining reliable burden estimates, at the relevant operational scale is crucial for disease control programs. This can improve the visibility of the progress made as well as identify the locations where more efforts are needed, and thus guide targeted and cost-effective interventions. When traditional surveillance systems suffer from severe underreporting, new approaches have to be developed to inform such interventions. This might involve surveillance based on symptoms (153), hospital admissions (154,155), sewage water analyses (156), animal surveillance (157,158) or even using serological surveys, i.e. serosurveillance (159).

We proposed a modelling pipeline that uses seroprevalence data to obtain burden estimates while also providing information on the locations where more serosurveys should be conducted. The level of analysis, the municipal level, is the most operational level for control programs (86). Also, we identified areas with an estimated high burden of disease that were not previously identified as being endemic. Such municipalities may lack experience with Chagas disease (e.g., prevention, diagnostic and treatment) and are probably not as well prepared as the ones known to be endemic to detect cases and implement the relevant interventions. Indeed, while positive blood samples in blood banks have been reported from every department and notifications are coming from all departments in the country, the surveillance bulletins suggested that most of the confirmed cases are reported from departments known to be at risk, e.g., 70% of the confirmed cases are reported from Santander, Casanare, Boyacá and Arauca (Supp. Table 10)(129,130). This seems to suggest that locations that were traditionally less impacted might not be confirming cases because of a lack of infrastructure. They might also be more likely to misdiagnose a case as the medical staff is not considering the disease. Our analyses suggest that dedicated investigations might be required in these areas to confirm the epidemiological situation.

Finally, the modelling pipeline presented here could be applied to other diseases that suffer from a weak surveillance system and require innovative methods to disentangle the underlying epidemiological situations and estimate their burden. Serosurveillance is routinely used for malaria and Hepatitis B, C and D (159). Also, cross-sectional and outbreak-based serosurveys are conducted for diseases such as Chagas disease but also Nipah (160), Japanese Encephalitis and Dengue fever (161). These types of surveys can inform on the prevalence of asymptomatic disease in the population as well as the exposure level for multiple diseases all at once (159).

## Challenges

Chagas disease is a Neglected Tropical Disease (NTD), and as for other NTDs, many epidemiological challenges but also technical and political ones are hampering the fight against the disease.

From an epidemiological point of view, the diversity in triatomine species as well as their ability to colonise different habitats, being domiciliated, peri-domiciliated or even sylvatic means that the entomological surveillance cannot be relaxed. Indeed, triatomines are able to colonise urban areas (17,148–150,152,162) and, on the other hand, deforestation is creating new opportunities for sylvatic species to evolve into new ecological niches in peri-domiciliated areas (13,38,56,163). This phenomenon would not be critical if humans were the only host for *T. cruzi*. Eliminating the parasite in the human population would be achievable. Unfortunately, *T. cruzi* has a large animal reservoir in the wildlife as well as in domesticated animals (164–167). Such multi-host dynamics make *T. cruzi* eradication out of reach. It also means that substantial efforts need to be maintained over time to handle the burden of disease in humans and avoid upsurges, and those might need to be re-evaluated and adapted to face upcoming changes in the dynamics and behaviour of vectors.

From a technical point of view, triatomine resistance to insecticide might represent an increasing problem (2). Also, entomological surveillance has low sensitivity, with staff requiring training and experience to increase the chance of finding a bug in a house (4,163). New research is required to assess the most effective vector control strategy, in particular, improving house insecticide spaying strategies (168), test new approaches (169) and work within a framework that brings together all of the vector control programs to increase efficiency and save resources (170). On another note, the drugs available to treat *T. cruzi* infection are not very effective when the patients have reached the mild and severe forms of the disease (2). This implies that the care of the patients suffering from these forms of disease is becoming a growing issue. Again, research is necessary to find new treatment strategies.

From a political point of view, the main challenges are related to resource allocation, organization of care and lobbying to obtain better treatment options. First of all, Chagas disease is not and will not become a priority in the upcoming years, therefore, resources allocated will remain limited. However, there are other areas where substantial improvement could be achieved. In particular, the access to the diagnostic and treatment could be simplified. Currently, in Colombia, the proportion of the population at-risk that has been tested for Chagas disease is about 1.2% (3). Also, only 0.3%-0.4% of the infected persons have received an antiparasitic drug (3). Advances would require governments' buy-in to better organise and communicate the availability of medical services for Chagas disease patients but also, with support from PAHO and WHO to stimulate the pharmaceutical industry to develop new treatments with fewer side effects. Also, as progress has been made and certificates distributed to attest to the elimination of intradomiciliated transmission, fewer serosurveys are organised. If the cost-benefit of these surveys become minor as the prevalence is reducing, new strategies will need to be found to continue the surveillances.

COVID-19 will also probably represent another important challenge for Chagas disease patients, researchers and stakeholders.  First, new research is necessary to understand how COVID-19 affect people with Chagas disease. Indeed, people suffering from the severe form of the disease could be more at risk of the severe outcome due to COVID-19 but also, as Chagas disease is a multisystemic disorder, COVID-19 might have a strong impact on infected people at any stage of the disease (171). Some studies have looked at the risk of hospitalization for patients coinfected by SARS-CoV2 and *T. cruzi* and observed a worsening of the symptoms for patients with severe Chagas disease but the analyses still lack robustness (171–174). In addition, and perhaps more critically, the pandemic has and continues to disrupt health services and probably reduced the chances of diagnosis and the speed at which treatment can be delivered as several consultations are necessary.

# Conclusions

The elimination of Chagas disease transmission has been set as a goal by the World Health Organization within its first neglected tropical disease roadmap in 2012 (128). Considerable efforts have been implemented in Colombia and other Latin American countries and their impacts are visible with a reduction in Chagas disease prevalence in some areas and elimination of the vector of main concern (4), but long-term efforts are required and highlight the need to improve the sustainability and cost-effectiveness of control programs. On the other hand, the long-term nature of the disease can be harnessed. Serosurveys that will be conducted over the next years will still be able to characterise the exposure pattern of the last 6-7 decades. If they are conducted based on the needs of research, i.e., with an emphasis on representativeness and collecting information on the individual potential exposure and migration pattern, we may still be able to effectively characterise and monitor the progress that has been made in the last decades, and which interventions would be key to reduce further the burden of Chagas disease. Random large-scale sampling strategies might seem unpractical where resources are limited. However, creative and adaptative design can help approximate those strategies. Indeed, the two main points in the representativeness required in the context of Chagas disease are the age range of the participants that need to be varied and the geographical extent of the survey that needs to be maximised, ie a larger number of municipalities covered. An new approach could be to sample persons (patients or not) in emergency unit in major hospitals. Asking for current residence and migration history would help covering a larger number of municipalities and all ages could be included. Again, a sampling bias would be created, ie only person able to go to hospital could be included, but increasing the diversity of study design would help capturate more information.

# References

1. Centers for Disease Control and Prevention. Parasites - American Trypanosomiasis (also known as Chagas Disease) [Internet]. CDC. Available from: https://www.cdc.gov/parasites/chagas/

2. Pérez-Molina JA, Molina I. Chagas disease. The Lancet. 2018 Jan 6;391(10115):82–94.

3. Cucunubá ZM, Manne-Goehler JM, Díaz D, Nouvellet P, Bernal O, Marchiol A, et al. How universal is coverage and access to diagnosis and treatment for Chagas disease in Colombia? A health systems analysis. Soc Sci Med. 2017 Feb;175:187–98.

4. Moncayo A, Silveira AC. Current epidemiological trends for Chagas disease in Latin America and future challenges in epidemiology, surveillance and health policy. Mem Inst Oswaldo Cruz. 2009 Jul;104 Suppl 1:17–30.

5. Franco-Paredes C, Villamil-Gómez WE, Schultz J, Henao-Martínez AF, Parra-Henao G, Rassi A, et al. A deadly feast: Elucidating the burden of orally acquired acute Chagas disease in Latin America – Public health and travel medicine importance. Travel Med Infect Dis. 2020 Jul 1;36:101565.

6. Cucunubá ZM, Okuwoga O, Basáñez M-G, Nouvellet P. Increased mortality attributed to Chagas disease: a systematic review and meta-analysis. Parasit Vectors. 2016 Jan 27;9:42.

7. Chadalawada S, Sillau S, Archuleta S, Mundo W, Bandali M, Parra-Henao G, et al. Risk of Chronic Cardiomyopathy Among Patients With the Acute Phase or Indeterminate Form of Chagas Disease: A Systematic Review and Meta-analysis. JAMA Netw Open. 2020 Aug 31;3(8):e2015072.

8. Cucunubá ZM. Modelling the epidemiology and healthcare burden of Chagas disease in Colombia. Imperial College of London; 2017.

9. Pinheiro E, Brum-Soares L, Reis R, Cubides J-C. Chagas disease: review of needs, neglect, and obstacles to treatment access in Latin America. Rev Soc Bras Med Trop. 2017 Jun;50(3):296–300.

10. Pérez-Molina JA, Perez AM, Norman FF, Monge-Maillo B, López-Vélez R. Old and new challenges in Chagas disease. Lancet Infect Dis. 2015 Nov 1;15(11):1347–56.

11. Lidani KCF, Andrade FA, Bavia L, Damasceno FS, Beltrame MH, Messias-Reason IJ, et al. Chagas Disease: From Discovery to a Worldwide Health Problem. Front Public Health. 2019 Jul 2;7:166.

12. Dias JCP. Evolution of Chagas disease screening programs and control programs: historical perspective. Glob Heart. 2015 Sep;10(3):193–202.

13. Abad-Franch F, Monteiro FA, Jaramillo O N, Gurgel-Gonçalves R, Dias FBS, Diotaiuti L. Ecology, evolution, and the long-term surveillance of vector-borne Chagas disease: a multi-scale appraisal of the tribe Rhodniini (Triatominae). Acta Trop. 2009 Jun;110(2–3):159–77.

14. Stanaway JD, Roth G. The burden of Chagas disease: estimates and challenges. Glob Heart. 2015 Sep;10(3):139–44.

15. Dumonteil E, Nouvellet P, Rosecrans K, Ramirez-Sierra MJ, Gamboa-León R, Cruz-Chan V, et al. Eco-Bio-Social Determinants for House Infestation by Non-domiciliated Triatoma dimidiata in the Yucatan Peninsula, Mexico. PLoS Negl Trop Dis. 2013 Sep 26;7(9):e2466.

16. Brito RN, Gorla DE, Diotaiuti L, Gomes ACF, Souza RCM, Abad-Franch F. Drivers of house invasion by sylvatic Chagas disease vectors in the Amazon-Cerrado transition: A multi-year, state-wide assessment of municipality-aggregated surveillance data. PLoS Negl Trop Dis. 2017 Nov;11(11):e0006035.

17. Levy MZ, Barbu CM, Castillo-Neyra R, Quispe-Machaca VR, Ancca-Juarez J, Escalante-Mejia P, et al. Urbanization, land tenure security and vector-borne Chagas disease. Proc Biol Sci. 2014 Aug 22;281(1789):20141003.

18. Kim JYH, Ledien J, Rodriguez-Monguí E, Dobson A, Basáñez M-G, Cucunubá ZM. Global Trends of Seroprevalence and Universal Screening Policy for Chagas Disease in Donors: a systematic review and meta-analysis [Internet]. medRxiv; 2019 [cited 2022 Feb 8]. p. 2019.12.25.19015776. Available from: https://www.medrxiv.org/content/10.1101/2019.12.25.19015776v1

19. Olivera MJ, Fory JA, Porras JF, Buitrago G. Prevalence of Chagas disease in Colombia: A systematic review and meta-analysis. PLoS ONE [Internet]. 2019 Jan 7 [cited 2021 Apr 6];14(1). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6322748/

20. World Health Organization. Ending the neglect to attain the sustainable development goals: a road map for neglected tropical diseases 2021–2030 [Internet]. Geneva: World Health Organization; 2020 [cited 2022 Feb 1]. 177 p. Available from: https://apps.who.int/iris/handle/10665/338565

21. Cucunubá ZM, Nouvellet P, Conteh L, Vera MJ, Angulo VM, Dib JC, et al. Modelling historical changes in the force-of-infection of Chagas disease to inform control and elimination programmes: application in Colombia. BMJ Glob Health. 2017;2(3):e000345.

22. CHAGAS AGUDO PE XII 2021.pdf [Internet]. [cited 2022 Feb 17]. Available from: http://www.ins.gov.co/buscador-eventos/Informesdeevento/CHAGAS%20AGUDO%20PE%20XII%202021.pdf

23. CHAGAS CRÓNICO PE XII 2021.pdf [Internet]. [cited 2022 Feb 17]. Available from: http://www.ins.gov.co/buscador-eventos/Informesdeevento/CHAGAS%20CR%C3%93NICO%20PE%20XII%202021.pdf

24. Vos T, Allen C, Arora M, Barber RM, Bhutta ZA, Brown A, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. The Lancet. 2016 Oct;388(10053):1545–602.

25. World Health Organization. Chagas diseases in Latin America: an epidemiological update based on 2010 estimates. Wkly Epidemiol Rec. 2015;No. 6,(90):33–4.

26. Cantillo-Barraza O, Chaverra D, Marcet P, Arboleda-Sánchez S, Triana-Chávez O. Trypanosoma cruzi transmission in a Colombian Caribbean region suggests that secondary vectors play an important epidemiological role. Parasit Vectors. 2014 Aug 20;7:381.

27. Carbajal de la Fuente AL, Porcasi X, Noireau F, Diotaiuti L, Gorla DE. The association between the geographic distribution of Triatoma pseudomaculata and Triatoma wygodzinskyi (Hemiptera: Reduviidae) with environmental variables recorded by remote sensors. Infect Genet Evol J Mol Epidemiol Evol Genet Infect Dis. 2009 Jan;9(1):54–61.

28. Cordovez JM, Guhl F. The impact of landscape transformation on the reinfestation rates of Rhodnius prolixus in the Orinoco Region, Colombia. Acta Trop. 2015 Nov;151:73–9.

29. Carmona-Castro O, Moo-Llanes DA, Ramsey JM. Impact of climate change on vector transmission of Trypanosoma cruzi (Chagas, 1909) in North America. Med Vet Entomol. 2018;32(1):84–101.

30. DE LA Vega GJ, Schilman PE. Ecological and physiological thermal niches to understand distribution of Chagas disease vectors in Latin America. Med Vet Entomol. 2018;32(1):1–13.

31. Cavallo MJ, Amelotti I, Gorla DE. Invasion of rural houses by wild Triatominae in the arid Chaco. J Vector Ecol J Soc Vector Ecol. 2016;41(1):97–102.

32. Nouvellet P, Cucunubá ZM, Gourbière S. Ecology, evolution and control of Chagas disease: A century of neglected modelling and a promising future. In: Mathematical Models for Neglected Tropical Diseases: Essential Tools for Control and Elimination, R M Anderson, MG Basáñez [Internet]. Advances in Parasitology. Academic Press; 2015 [cited 2018 Nov 15]. p. 135–91. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0065308X14000050

33. de Fuentes-Vicente JA, Gutiérrez-Cabrera AE, Flores-Villegas AL, Lowenberger C, Benelli G, Salazar-Schettino PM, et al. What makes an effective Chagas disease vector? Factors underlying Trypanosoma cruzi-triatomine interactions. Acta Trop. 2018 Jul;183:23–31.

34. Ibarra-Cerdeña CN, Valiente-Banuet L, Sánchez-Cordero V, Stephens CR, Ramsey JM. Trypanosoma cruzi reservoir-triatomine vector co-occurrence networks reveal meta-community effects by synanthropic mammals on geographic dispersal. PeerJ. 2017;5:e3152.

35. Ferro E Silva AM, Sobral-Souza T, Vancine MH, Muylaert RL, de Abreu AP, Pelloso SM, et al. Spatial prediction of risk areas for vector transmission of Trypanosoma cruzi in the State of Paraná, southern Brazil. PLoS Negl Trop Dis. 2018 Oct 26;12(10):e0006907.

36. Gurgel-Goncalves R, Cuba CAC. Predicting the potential geographical distribution of Rhodnius neglectus (Hemiptera, Reduviidae) based on ecological niche modeling. J Med Entomol. 2009 Jul;46(4):952–60.

37. Parra-Henao G, Suárez-Escudero LC, González-Caro S. Potential distribution of Chagas disease vectors (Hemiptera, Reduviidae, Triatominae) in Colombia, based on ecological niche modeling. J Trop Med. 2016;2016:1439090.

38. Santana K de SO, Bavia ME, Lima AD, Guimarães ICS, Soares ES, Silva MMN, et al. Spatial distribution of triatomines (Reduviidae: Triatominae) in urban areas of the city of Salvador, Bahia, Brazil. Geospatial Health. 2011 May;5(2):199–203.

39. Dias JVL, Queiroz DRM, Martins HR, Gorla DE, Pires HHR, Diotaiuti L. Spatial distribution of triatomines in domiciles of an urban area of the Brazilian Southeast Region. Mem Inst Oswaldo Cruz. 2016 Jan;111(1):43–50.

40. Espinoza Echeverria J, Rodriguez AN, Cortez MR, Diotaiuti LG, Gorla DE. Spatial and temporal distribution of house infestation by Triatoma infestans in the Toro Toro municipality, Potosi, Bolivia. Parasit Vectors. 2017 02;10(1):58.

41. Sousa A da S, Palácios VR da CM, Miranda C do S, Costa RJF da, Catete CP, Chagasteles EJ, et al. Space-temporal analysis of Chagas disease and its environmental and demographic risk factors in the municipality of Barcarena, Pará, Brazil. Rev Bras Epidemiol Braz J Epidemiol. 2017 Dec;20(4):742–55.

42. Hernández J, Núñez I, Bacigalupo A, Cattan PE. Modeling the spatial distribution of Chagas disease vectors using environmental variables and people´s knowledge. Int J Health Geogr. 2013 May 31;12:29.

43. Parra-Henao G, Angulo VM, Osorio L, Jaramillo-O N. Geographic Distribution and Ecology of Triatoma dimidiata (Hemiptera: Reduviidae) in Colombia. J Med Entomol. 2016 Jan 1;53(1):122–9.

44. Altamiranda-Saavedra M, Osorio-Olvera L, Yáñez-Arenas C, Marín-Ortiz JC, Parra-Henao G. Geographic abundance patterns explained by niche centrality hypothesis in two Chagas disease vectors in Latin America. PLOS ONE. 2020 Nov 4;15(11):e0241710.

45. Rossi JCN, Duarte EC, Gurgel-Gonçalves R. Factors associated with the occurrence of Triatoma sordida (Hemiptera: Reduviidae) in rural localities of Central-West Brazil. Mem Inst Oswaldo Cruz. 2015 Apr;110(2):192–200.

46. Parra-Henao G, Quirós-Gómez O, Jaramillo-O N, Segura-Cardona Á. Environmental determinants of the distribution of Chagas disease vector Triatoma dimidiata in Colombia. Am J Trop Med Hyg. 2016 Apr;94(4):767–74.

47. Mischler P, Kearney M, McCarroll JC, Scholte RGC, Vounatsou P, Malone JB. Environmental and socio-economic risk modelling for Chagas disease in Bolivia. Geospatial Health. 2012 Sep;6(3):S59-66.

48. Grijalva MJ, Villacís AG, Moncayo AL, Ocaña-Mayorga S, Yumiseva CA, Baus EG. Distribution of triatomine species in domestic and peridomestic environments in central coastal Ecuador. PLoS Negl Trop Dis. 2017 Oct;11(10):e0005970.

49. Cordovez JM, Rendon LM, Gonzalez C, Guhl F. Using the basic reproduction number to assess the effects of climate change in the risk of Chagas disease transmission in Colombia. Acta Trop. 2014 Jan;129:74–82.

50. Medone P, Ceccarelli S, Parham PE, Figuera A, Rabinovich JE. The impact of climate change on the geographical distribution of two vectors of Chagas disease: implications for the force of infection. Philos Trans R Soc Lond B Biol Sci. 2015 Apr 5;370(1665).

51. Tamayo LD, Guhl F, Vallejo GA, Ramírez JD. The effect of temperature increase on the development of Rhodnius prolixus and the course of Trypanosoma cruzi metacyclogenesis. PLoS Negl Trop Dis. 2018 Aug;12(8):e0006735.

52. Garza M, Feria Arroyo TP, Casillas EA, Sanchez-Cordero V, Rivaldi C-L, Sarkar S. Projected future distributions of vectors of Trypanosoma cruzi in North America under climate change scenarios. PLoS Negl Trop Dis. 2014 May;8(5):e2818.

53. Tapia-Garay V, Figueroa DP, Maldonado A, Frías-Laserre D, Gonzalez CR, Parra A, et al. Assessing the risk zones of Chagas' disease in Chile, in a world marked by global climatic change. Mem Inst Oswaldo Cruz. 2018 Jan;113(1):24–9.

54. Levy MZ, Malaga Chavez FS, Cornejo Del Carpio JG, Vilhena DA, McKenzie FE, Plotkin JB. Rational spatio-temporal strategies for controlling a Chagas disease vector in urban environments. J R Soc Interface. 2010 Jul 6;7(48):1061–70.

55. Gottdenker NL, Calzada JE, Saldaña A, Carroll CR. Association of anthropogenic land use change and increased abundance of the Chagas disease vector Rhodnius pallescens in a rural landscape of Panama. Am J Trop Med Hyg. 2011 Jan;84(1):70–7.

56. Santos WS, Gurgel-Gonçalves R, Garcez LM, Abad-Franch F. Deforestation effects on Attalea palms and their resident Rhodnius, vectors of Chagas disease, in eastern Amazonia. PLOS ONE. 2021 May 20;16(5):e0252071.

57. Gottdenker NL, Chaves LF, Calzada JE, Saldaña A, Carroll CR. Host life history strategy, species diversity, and habitat influence Trypanosoma cruzi vector infection in Changing landscapes. PLoS Negl Trop Dis. 2012;6(11):e1884.

58. Moreno ML, Hoyos L, Cabido M, Catalá SS, Gorla DE. Exploring the association between Trypanosoma cruzi infection in rural communities and environmental changes in the southern Gran Chaco. Mem Inst Oswaldo Cruz. 2012 Mar;107(2):231–7.

59. Grijalva MJ, Terán D, Dangles O. Dynamics of sylvatic Chagas disease vectors in coastal Ecuador is driven by changes in land cover. PLoS Negl Trop Dis. 2014 Jun;8(6):e2960.

60. Vianna EN, Souza E Guimarães RJ de P, Souza CR, Gorla D, Diotaiuti L. Chagas disease ecoepidemiology and environmental changes in northern Minas Gerais state, Brazil. Mem Inst Oswaldo Cruz. 2017 Nov;112(11):760–8.

61. Garske T, Van Kerkhove MD, Yactayo S, Ronveaux O, Lewis RF, Staples JE, et al. Yellow Fever in Africa: Estimating the burden of disease and impact of mass vaccination from outbreak and serological data. Hay SI, editor. PLoS Med. 2014 May 6;11(5):e1001638.

62. Pinsent A, Solomon AW, Bailey RL, Bid R, Cama A, Dean D, et al. The utility of serology for elimination surveillance of trachoma. Nat Commun. 2018 Dec 21;9(1):5444.

63. Ross A, Koepfli C, Schoepflin S, Timinao L, Siba P, Smith T, et al. The Incidence and Differential Seasonal Patterns of Plasmodium vivax Primary Infections and Relapses in a Cohort of Children in Papua New Guinea. PLoS Negl Trop Dis [Internet]. 2016 May 4 [cited 2020 Sep 14];10(5). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4856325/

64. Caicedo E-Y, Charniga K, Rueda A, Dorigatti I, Mendez Y, Hamlet A, et al. The epidemiology of Mayaro virus in the Americas: A systematic review and key parameter estimates for outbreak modelling. PLoS Negl Trop Dis. 2021 Jun 3;15(6):e0009418.

65. Moore SM. The current burden of Japanese encephalitis and the estimated impacts of vaccination: Combining estimates of the spatial distribution and transmission intensity of a zoonotic pathogen. PLoS Negl Trop Dis. 2021 Oct 13;15(10):e0009385.

66. O'Driscoll M, Imai N, Ferguson NM, Hadinegoro SR, Satari HI, Tam CC, et al. Spatiotemporal variability in dengue transmission intensity in Jakarta, Indonesia. PLoS Negl Trop Dis. 2020 Mar 6;14(3):e0008102.

67. Corran P, Coleman P, Riley E, Drakeley C. Serology: a robust indicator of malaria transmission intensity? Trends Parasitol. 2007 Dec 1;23(12):575–82.

68. Lewnard JA, Lopman BA, Parashar UD, Bennett A, Bar-Zeev N, Cunliffe NA, et al. Heterogeneous susceptibility to rotavirus infection and gastroenteritis in two birth cohort studies: Parameter estimation and epidemiological implications. PLoS Comput Biol [Internet]. 2019 Jul 26 [cited 2020 Sep 14];15(7). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6690553/

69. Arnold BF, Kanyi H, Njenga SM, Rawago FO, Priest JW, Secor WE, et al. Fine-scale heterogeneity in Schistosoma mansoni force of infection measured through antibody response. Proc Natl Acad Sci. 2020 Sep 15;117(37):23174–81.

70. Aleshnick M, Ganusov VV, Nasir G, Yenokyan G, Sinnis P. Experimental determination of the force of malaria infection reveals a non-linear relationship to mosquito sporozoite loads. PLOS Pathog. 2020 May 26;16(5):e1008181.

71. Hachiya M, Miyano S, Mori Y, Vynnycky E, Keungsaneth P, Vongphrachanh P, et al. Evaluation of nationwide supplementary immunization in Lao People's Democratic Republic: Population-based seroprevalence survey of anti-measles and anti-rubella IgG in children and adults, mathematical modelling and a stability testing of the vaccine. PLoS ONE. 2018 Mar 29;13(3):e0194931.

72. Lim JK, Carabali M, Edwards T, Barro A, Lee J-S, Dahourou D, et al. Estimating the Force of Infection for Dengue Virus Using Repeated Serosurveys, Ouagadougou, Burkina Faso. Emerg Infect Dis. 2021 Jan;27(1):130–9.

73. Rees EM, Waterlow NR, Centre for the Mathematical Modelling of Infectious Diseases COVID-19 Working Group, Lowe R, Kucharski AJ. Estimating the duration of seropositivity of human seasonal coronaviruses using seroprevalence studies. Wellcome Open Res. 2021;6:138.

74. Flasche S, Lipsitch M, Ojal J, Pinsent A. Estimating the contribution of different age strata to vaccine serotype pneumococcal transmission in the pre vaccine era: a modelling study. BMC Med [Internet]. 2020 Jun 10 [cited 2020 Jun 17];18. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7285529/

75. Alleman MM, Wannemuehler KA, Hao L, Perelygina L, Icenogle JP, Vynnycky E, et al. Estimating the burden of rubella virus infection and congenital rubella syndrome through a rubella immunity assessment among pregnant women in the Democratic Republic of the Congo: Potential impact on vaccination policy. Vaccine. 2016 Dec 12;34(51):6502–11.

76. Biggs JR, Sy AK, Sherratt K, Brady OJ, Kucharski AJ, Funk S, et al. Estimating the annual dengue force of infection from the age of reporting primary infections across urban centres in endemic countries. BMC Med. 2021 Sep 30;19(1):217.

77. Arnold BF, Martin DL, Juma J, Mkocha H, Ochieng JB, Cooley GM, et al. Enteropathogen antibody dynamics and force of infection among children in low-resource settings. eLife [Internet]. [cited 2020 Sep 14];8. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6746552/

78. Nealon J, Bouckenooghe A, Cortes M, Coudeville L, Frago C, Macina D, et al. Dengue endemicity, force of infection and variation in transmission intensity in 13 endemic countries. J Infect Dis [Internet]. [cited 2020 Apr 1]; Available from: https://academic.oup.com/jid/advance-article/doi/10.1093/infdis/jiaa132/5811404

79. Li S, Ma C, Hao L, Su Q, An Z, Ma F, et al. Demographic transition and the dynamics of measles in six provinces in China: A modeling study. PLoS Med [Internet]. 2017 Apr 4 [cited 2020 Sep 14];14(4). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5380361/

80. Cattarino L, Rodriguez-Barraquer I, Imai N, Cummings DAT, Ferguson NM. Mapping global variation in dengue transmission intensity. Sci Transl Med. 2020 Jan 29;12(528):eaax4144.

81. GBD 2013 Mortality and Causes of Death Collaborators. Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. Lancet. 2015 Jan 10;385(9963):117–71.

82. Behrend MR, Basáñez M-G, Hamley JID, Porco TC, Stolk WA, Walker M, et al. Modelling for policy: The five principles of the Neglected Tropical Diseases Modelling Consortium. PLoS Negl Trop Dis. 2020 Apr 9;14(4):e0008033.

83. Symonds MRE, Moussalli A. A brief guide to model selection, multimodel inference and model averaging in behavioural ecology using Akaike's information criterion. Behav Ecol Sociobiol. 2011 Jan 1;65(1):13–21.

84. Rousseeuw PJ, Croux C. Alternatives to the Median Absolute Deviation. J Am Stat Assoc. 1993 Dec 1;88(424):1273–83.

85. Muench H. Catalytic Models in Epidemiology [Internet]. Harvard University Press; 2013. Available from: https://doi.org/10.4159/harvard.9780674428928

86. Parra-Henao GJ, Flórez Martínez M, Angulo Silva VM. Vigilancia de Triatominae (Hemiptera: Reduviidae) en Colombia. In: Red Chagas Colombia. 1era Edition., p. 127. Bucaramanga Colombia: Sic Editorial Ltda; 2013. (Memorias del Curso de Capacitación Métodos Básicos en Epidemiología y Redacción Científica. Septiembre 23 a 27 de 2013. Bogotá D.C. Colombia.).

87. Parra-Henao G, Angulo V, Cucunubá Z. Colombian Chagas Network. Final report, project 1. 2015.

88. Departamento Administrativo Nacional de Estadística (DANE): www.dane.gov.co [Internet]. [cited 2022 Jan 25]. Available from: https://www.dane.gov.co/index.php/estadisticas-por-tema/demografia-y-poblacion

89. Beck HE, Zimmermann NE, McVicar TR, Vergopolan N, Berg A, Wood EF. Present and future Köppen-Geiger climate classification maps at 1-km resolution. Sci Data. 2018 Oct 30;5:180214.

90. Database of Global Administrative Areas (GADM): https://gadm.org/ [Internet]. [cited 2020 May 25]. Available from: https://gadm.org/

91. Montgomery DC, Peck EA, Vining GG. Introduction to Linear Regression Analysis. John Wiley & Sons; 2012. 678 p.

92. Ridout MS, Linkie M. Estimating overlap of daily activity patterns from camera trap data. J Agric Biol Environ Stat. 2009 Sep 1;14(3):322–37.

93. Giorgi E, Diggle PJ. PrevMap : An R Package for Prevalence Mapping. J Stat Softw [Internet]. 2017 [cited 2020 May 28];78(8). Available from: http://www.jstatsoft.org/v78/i08/

94. Bivand RS, Wong DWS. Comparing implementations of global and local indicators of spatial association. TEST. 2018 Sep;27(3):716–48.

95. GitHub repository for Chagas disease FoI with Linear Models [Internet]. [cited 2021 Oct 27]. Available from: https://github.com/jledien/Chagas-disease-FoI-with-Linear-Models.git

96. Massad E. The elimination of Chagas' disease from Brazil. Epidemiol Infect. 2008 Sep;136(9):1153–64.

97. Feliciangeli MD, Campbell-Lendrum D, Martinez C, Gonzalez D, Coleman P, Davies C. Chagas disease control in Venezuela: lessons for the Andean region and beyond. Trends Parasitol. 2003 Jan 1;19(1):44–9.

98. Schultz TP. Rural-urban migration in Colombia. Rev Econ Stat. 1971;53(2):157–63.

99. Álvarez-Berríos NL, Parés-Ramos IK, Aide TM. Contrasting patterns of urban expansion in Colombia, Ecuador, Peru, and Bolivia between 1992 and 2009. AMBIO. 2013 Feb 1;42(1):29–40.

100. Gascon J, Bern C, Pinazo M-J. Chagas disease in Spain, the United States and other non-endemic countries. Acta Trop. 2010 Jul 1;115(1):22–7.

101. Oficina Sanitaria Panamericana, Encuentro Continental de Educación Médica. Educación, práctica medica y necesidades sociales: una nueva visión de calidad [Internet]. Estados Unidos: OPS (Organización Panamericana de la Salud); 1995. Available from: https://iris.paho.org/bitstream/handle/10665.2/51648/9789275121528-spa.pdf?sequence=7&isAllowed=y&ua=1

102. WHO. Chagas disease [Internet]. [cited 2021 Dec 14]. Available from: https://www.who.int/news-room/fact-sheets/detail/chagas-disease-(american-trypanosomiasis)

103. Ledien J, Cucunubá ZM, Parra-Henao G, Rodríguez-Monguí E, Dobson AP, Basáñez M-G, et al. Spatiotemporal variations in exposure: Chagas disease in Colombia as a case study. BMC Med Res Methodol. 2022 Jan 13;22(1):13.

104. Malley JD, Malley KG, Pajevic S. Statistical learning for biomedical data. Cambridge: Cambridge University Press; 2011. 285 p. (Practical guides to biostatistics and epidemiology).

105. Ledien J, Sorn S, Hem S, Huy R, Buchy P, Tarantola A, et al. Assessing the performance of remotely-sensed flooding indicators and their potential contribution to early warning for leptospirosis in Cambodia. PLOS ONE. 2017 Jul 13;12(7):e0181044.

106. Mohammadinia A, Saeidian B, Pradhan B, Ghaemi Z. Prediction mapping of human leptospirosis using ANN, GWR, SVM and GLM approaches. BMC Infect Dis [Internet]. 2019 Nov 13 [cited 2020 Feb 12];19. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6854714/

107. Miao D, Dai K, Zhao G-P, Li X-L, Shi W-Q, Zhang JS, et al. Mapping the global potential transmission hotspots for severe fever with thrombocytopenia syndrome by machine learning methods. Emerg Microbes Infect. 2020 Dec;9(1):817–26.

108. Eneanya OA, Fronterre C, Anagbogu I, Okoronkwo C, Garske T, Cano J, et al. Mapping the baseline prevalence of lymphatic filariasis across Nigeria. Parasit Vectors. 2019 Sep 16;12(1):440.

109. Mutai CK, McSharry PE, Ngaruye I, Musabanganji E. Use of machine learning techniques to identify HIV predictors for screening in sub-Saharan Africa. BMC Med Res Methodol. 2021 Jul 31;21(1):159.

110. Hengl T, Nussbaum M, Wright MN, Heuvelink GBM, Gräler B. Random forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ. 2018 Aug 29;6:e5518.

111. Meyer H, Reudenbach C, Wöllauer S, Nauss T. Importance of spatial predictor variable selection in machine learning applications – Moving from data reproduction to spatial prediction. Ecol Model. 2019 Nov 1;411:108815.

112. Augusta C, Deardon R, Taylor G. Deep learning for supervised classification of spatial epidemics. Spat Spatio-Temporal Epidemiol [Internet]. 2018 Aug 29; Available from: http://www.sciencedirect.com/science/article/pii/S1877584517301636

113. Forna A, Nouvellet P, Dorigatti I, Donnelly CA. Case Fatality Ratio Estimates for the 2013–2016 West African Ebola Epidemic: Application of Boosted Regression Trees for Imputation. Clin Infect Dis. 2020 Jun 10;70(12):2476–83.

114. Jiang D, Ma T, Hao M, Qian Y, Chen S, Meng Z, et al. Spatiotemporal patterns and spatial risk factors for visceral leishmaniasis from 2007 to 2017 in Western and Central China: A modelling analysis. Sci Total Environ. 2021 Apr 10;764:144275.

115. Ding F, Wang Q, Fu J, Chen S, Hao M, Ma T, et al. Risk factors and predicted distribution of visceral leishmaniasis in the Xinjiang Uygur Autonomous Region, China, 2005–2015. Parasit Vectors. 2019 Nov 8;12(1):528.

116. Andraud M, Bougeard S, Chesnoiu T, Rose N. Spatiotemporal clustering and Random Forest models to identify risk factors of African swine fever outbreak in Romania in 2018–2019. Sci Rep. 2021 Jan 22;11(1):2098.

117. Yao H, Wang Y, Mi X, Sun Y, Liu K, Li X, et al. The scrub typhus in mainland China: spatiotemporal expansion and risk prediction underpinned by complex factors. Emerg Microbes Infect. 2019 Jun 24;8(1):909–19.

118. Ashby J, Moreno-Madriñán MJ, Yiannoutsos CT, Stanforth A. Niche Modeling of Dengue Fever Using Remotely Sensed Environmental Factors and Boosted Regression Trees. Remote Sens. 2017 Apr;9(4):328.

119. Forna A, Dorigatti I, Nouvellet P, Donnelly CA. Spatiotemporal variability in case fatality ratios for the 2013-2016 Ebola epidemic in West Africa. Int J Infect Dis IJID Off Publ Int Soc Infect Dis. 2020 Apr;93:48–55.

120. Elith J, Leathwick JR, Hastie T. A working guide to boosted regression trees. J Anim Ecol. 2008 juillet;77(4):802–13.

121. Breiman L. Random Forests. Mach Learn. 2001 Oct 1;45(1):5–32.

122. Meyer H, Reudenbach C, Hengl T, Katurji M, Nauss T. Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. Environ Model Softw. 2018 Mar 1;101:1–9.

123. Pastore M. Overlapping: a R package for Estimating Overlapping in Empirical Distributions. J Open Source Softw. 2018 Dec 5;3(32):1023.

124. Pastore M, Calcagnì A. Measuring Distribution Similarities Between Samples: A Distribution-Free Overlapping Index. Front Psychol. 2019 May 21;10:1089.

125. Durbin J, Watson GS. Testing Fot Serial Correlation In Least Squares Regression. II. Biometrika. 1951 Jun 1;38(1–2):159–78.

126. Lang M, Binder M, Richter J, Schratz P, Pfisterer F, Coors S, et al. mlr3: A modern object-oriented machine learning framework in R. J Open Source Softw. 2019 Dec 11;4(44):1903.

127. R Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2020. Available from: https://www.R-project.org/

128. World Health Organization. Investing to overcome the global impact of neglected tropical diseases: third WHO report on neglected diseases 2015 [Internet]. Geneva: World Health Organization; 2015 [cited 2022 Mar 31]. 191 p. Available from: https://apps.who.int/iris/handle/10665/152781

129. Instituto Nacional de Salud. Informe de invento. Chagas crónico, periodo epidemiológico XII. Colombia 2021 [Internet]. [cited 2022 Feb 17]. Available from: http://www.ins.gov.co/buscador-eventos/Informesdeevento/CHAGAS%20CR%C3%93NICO%20PE%20XII%202021.pdf

130. Instituto Nacional de Salud. Informe de invento. Chagas agudo, periodo epidemiológico XII. Colombia 2021 [Internet]. [cited 2022 Feb 17]. Available from: http://www.ins.gov.co/buscador-eventos/Informesdeevento/CHAGAS%20AGUDO%20PE%20XII%202021.pdf

131. Ledien J, Cucunubá ZM, Parra-Henao G, Rodríguez-Monguí E, Dobson AP, Adamo SB, et al. Linear and Machine Learning Modelling for Spatiotemporal Disease Predictions: Force-of-Infection of Chagas Disease. Rev PLOS NTD.

132. Pereira JM, Almeida PS de, Sousa AV de, Paula AM de, Machado RB, Gurgel-Gonçalves R. Climatic factors influencing triatomine occurrence in Central-West Brazil. Mem Inst Oswaldo Cruz. 2013 May;108(3).

133. Minnesota Population Center. Integrated Public Use Microdata Series, International: Version 7.2 [Internet]. Minneapolis, MN: IPUMS; 2019 [cited 2022 Feb 25]. Available from: https://www.ipums.org/projects/ipums-international/d020.V7.2

134. R Studio. RStudio – RStudio [Internet]. [cited 2017 Feb 15]. Available from: https://www.rstudio.com/products/rstudio/

135. Stevens B, Verdian L, Tomlinson J, Zegenhagen S, Pezzullo L. PM021 The Economic Burden of Heart Diseases in Colombia. Glob Heart. 2016 Jun;11(2):e73–4.

136. Bocchi EA, Arias A, Verdejo H, Diez M, Gómez E, Castro P. The Reality of Heart Failure in Latin America. J Am Coll Cardiol. 2013 Sep;62(11):949–58.

137. de Oliveira RB, Troncon LEA, Dantas RO, Meneghelli UG. Gastrointestinal Manifestations of Chagas' Disease. Off J Am Coll Gastroenterol ACG. 1998 Jun;93(6):884–9.

138. Matsuda NM, Miller SM, Evora PRB. The Chronic Gastrointestinal Manifestations of Chagas Disease. Clinics. 2009 Dec;64(12):1219–24.

139. Pérez-Ayala A, Pérez-Molina JA, Norman F, Navarro M, Monge-Maillo B, Díaz-Menéndez M, et al. Chagas disease in Latin American migrants: a Spanish challenge. Clin Microbiol Infect Off Publ Eur Soc Clin Microbiol Infect Dis. 2011 Jul;17(7):1108–13.

140. Coura J, Abreu LL de, Pereira JB, Willcox HP. [Morbidity in Chagas' disease. IV. Longitudinal study of 10 years in Pains and Iguatama, Minas Gerais, Brazil]. Mem Inst Oswaldo Cruz. 1985;

141. Norman FF, López-Vélez R. Mother-to-child transmission of Trypanosoma cruzi infection (Chagas disease): a neglected problem. Trans R Soc Trop Med Hyg. 2014 Jul;108(7):388–90.

142. Howard EJ, Xiong X, Carlier Y, Sosa-Estani S, Buekens P. Frequency of the congenital transmission of Trypanosoma cruzi: a systematic review and meta-analysis. BJOG Int J Obstet Gynaecol. 2014 Jan;121(1):22–33.

143. Edwards MS, Montgomery SP. Congenital Chagas disease: progress toward implementation of pregnancy-based screening. Curr Opin Infect Dis. 2021 Oct 1;34(5):538–45.

144. Castillo-Riquelme M, Guhl F, Turriago B, Pinto N, Rosas F, Martínez MF, et al. The Costs of Preventing and Treating Chagas Disease in Colombia. PLoS Negl Trop Dis. 2008 Nov 18;2(11):e336.

145. Bartsch SM, Avelis CM, Asti L, Hertenstein DL, Ndeffo-Mbah M, Galvani A, et al. The economic value of identifying and treating Chagas disease patients earlier and the

impact on Trypanosoma cruzi transmission. PLoS Negl Trop Dis. 2018 Nov 5;12(11):e0006809.

146. van den Hurk K, Zalpuri S, Prinsze FJ, Merz E-M, de Kort WLAM. Associations of health status with subsequent blood donor behavior—An alternative perspective on the Healthy Donor Effect from Donor InSight. PLoS ONE. 2017 Oct 19;12(10):e0186662.

147. Beltrán M, Bermúdez MI, Forero MC, Ayala M, Rodríguez MJ. Control of Trypanosoma cruzi infection in blood donors in Colombia, 2003. Biomédica. 2005 Dec 1;25(4):527–32.

148. Vizzoni AG, Varela MC, Sangenis LHC, Hasslocher-Moreno AM, do Brasil PEAA, Saraiva RM. Ageing with Chagas disease: an overview of an urban Brazilian cohort in Rio de Janeiro. Parasit Vectors. 2018 Jun 19;11(1):354.

149. Fraser B. Controlling Chagas' disease in urban Peru. The Lancet. 2008 Jul 5;372(9632):16–7.

150. Levy MZ, Bowman NM, Kawai V, Waller LA, Carpio JC del, Benzaquen EC, et al. Periurban Trypanosoma cruzi–infected Triatoma infestans, Arequipa, Peru. Emerg Infect Dis. 2006 Sep 1;12(9):1345–52.

151. Foley EA, Khatchikian CE, Hwang J, Ancca-Juárez J, Borrini-Mayori K, Quıspe-Machaca VR, et al. Population structure of the Chagas disease vector, Triatoma infestans, at the urban–rural interface. Mol Ecol. 2013;22(20):5162–71.

152. Provecho YM, Fernández M del P, Salvá L, Meli S, Cano F, Sartor P, et al. Urban infestation by Triatoma infestans (Hemiptera: Reduviidae), an overlooked phenomena for Chagas disease in Argentina. Mem Inst Oswaldo Cruz. 2021 Jun 2;116:e210056.

153. Mandl KD, Overhage JM, Wagner MM, Lober WB, Sebastiani P, Mostashari F, et al. Implementing Syndromic Surveillance: A Practical Guide Informed by the Early Experience. J Am Med Inform Assoc JAMIA. 2004;11(2):141–50.

154. Reis BY, Mandl KD. Time series modeling for syndromic surveillance. BMC Med Inform Decis Mak. 2003 Jan 23;3(1):2.

155. Buckeridge DL, Burkom H, Campbell M, Hogan WR, Moore AW. Algorithms for rapid outbreak detection: a research synthesis. J Biomed Inform. 2005 Apr;38(2):99–113.

156. Nakamura T, Hamasaki M, Yoshitomi H, Ishibashi T, Yoshiyama C, Maeda E, et al. Environmental Surveillance of Poliovirus in Sewage Water around the Introduction Period for Inactivated Polio Vaccine in Japan. Appl Environ Microbiol. 2015 Mar;81(5):1859–64.

157. Boden LA, Auty H, Reeves A, Rydevik G, Bessell P, McKendrick IJ. Animal Health Surveillance in Scotland in 2030: Using Scenario Planning to Develop Strategies in the Context of "Brexit." Front Vet Sci. 2017;4:201.

158. Castillo-Neyra R, Chou Chu L, Quispe-Machaca V, Ancca-Juarez J, Malaga Chavez FS, Bastos Mazuelos M, et al. The potential of canine sentinels for reemerging Trypanosoma cruzi transmission. Prev Vet Med. 2015 Jul 1;120(3–4):349–56.

159. Murray J, Cohen AL. Infectious Disease Surveillance. Int Encycl Public Health. 2017;222–9.

160. Yong M-Y, Lee S-C, Ngui R, Lim YA-L, Phipps ME, Chang L-Y. Seroprevalence of Nipah Virus Infection in Peninsular Malaysia. J Infect Dis. 2020 May 11;221(Suppl 4):S370–4.

161. Conlan JV, Vongxay K, Khamlome B, Jarman RG, Gibbons RV, Fenwick SG, et al. Patterns of Flavivirus Seroprevalence in the Human Population of Northern Laos. Am J Trop Med Hyg. 2015 Nov 4;93(5):1010–3.

162. Berry ASF, Salazar-Sánchez R, Castillo-Neyra R, Borrini-Mayorí K, Arevalo-Nieto C, Chipana-Ramos C, et al. Dispersal patterns of Trypanosoma cruzi in Arequipa, Peru. Buscaglia CA, editor. PLoS Negl Trop Dis. 2020 Mar 9;14(3):e0007910.

163. Abad-Franch F, Ferraz G, Campos C, Palomeque FS, Grijalva MJ, Aguilar HM, et al. Modeling Disease Vector Occurrence when Detection Is Imperfect: Infestation of Amazonian Palm Trees by Triatomine Bugs at Three Spatial Scales. PLoS Negl Trop Dis. 2010 Mar 2;4(3):e620.

164. Jansen AM, Xavier SC das C, Roque ALR. Trypanosoma cruzi transmission in the wild and its most important reservoir hosts in Brazil. Parasit Vectors. 2018 Sep 6;11:502.

165. Orozco MM, Enriquez GF, Alvarado-Otegui JA, Cardinal MV, Schijman AG, Kitron U, et al. New sylvatic hosts of Trypanosoma cruzi and their reservoir competence in the humid Chaco of Argentina: a longitudinal study. Am J Trop Med Hyg. 2013 May;88(5):872–82.

166. Fujita O, Sanabria L, Inchaustti A, De Arias AR, Tomizawa Y, Oku Y. Animal reservoirs for Trypanosoma cruzi infection in an endemic area in Paraguay. J Vet Med Sci. 1994 Apr;56(2):305–8.

167. Jiménez-Coello M, Acosta-Viana KY, Guzman-Marin E, Ortega-Pacheco A. American trypanosomiasis infection in fattening pigs from the south-east of Mexico. Zoonoses Public Health. 2012 Sep;59 Suppl 2:166–9.

168. Gonçalves R, Logan RAE, Ismail HM, Paine MJI, Bern C, Courtenay O. Indoor residual spraying practices against Triatoma infestans in the Bolivian Chaco: contributing factors to suboptimal insecticide delivery to treated households. Parasit Vectors. 2021 Jun 16;14(1):327.

169. Loza A, Talaga A, Herbas G, Canaviri RJ, Cahuasiri T, Luck L, et al. Systemic insecticide treatment of the canine reservoir of Trypanosoma cruzi induces high levels of lethality in Triatoma infestans, a principal vector of Chagas disease. Parasit Vectors. 2017 Jul 19;10(1):344.

170. Organisation mondiale de la Santé. Rapport de la 13e réunion du Groupe consultatif pour la lutte antivectorielle de l'OMS : rapport de réunion, 7-10 décembre 2020, réunion virtuelle [Internet]. Genève: Organisation mondiale de la Santé; 2021 [cited 2022 Feb 1]. Available from: https://apps.who.int/iris/handle/10665/343108

171. Molina I, Marcolino MS, Pires MC, Ramos LEF, Silva RT, Guimarães-Júnior MH, et al. Chagas disease and SARS-CoV-2 coinfection does not lead to worse in-hospital outcomes. Sci Rep. 2021 Oct 13;11(1):20289.

172. Alberca RW, Yendo TM, Leuzzi Ramos YÁ, Fernandes IG, Oliveira L de M, Teixeira FME, et al. Case Report: COVID-19 and Chagas Disease in Two Coinfected Patients. Am J Trop Med Hyg. 2020 Dec;103(6):2353–6.

173. Zaidel EJ, Forsyth CJ, Novick G, Marcus R, Ribeiro ALP, Pinazo M-J, et al. COVID-19: Implications for People with Chagas Disease. Glob Heart. 15(1):69.

174. Diaz-Hernandez A, Gonzalez-Vazquez MC, Arce-Fonseca M, Rodriguez-Morales O, Cedilllo-Ramirez ML, Carabarin-Lima A. Risk of COVID-19 in Chagas Disease Patients: What Happen with Cardiac Affectations? Biology. 2021 May 6;10(5):411.

175. PAHO. Misión Internacional de Evaluación de la situación de la interrupción de la transmisión vectorial domiciliaria de Trypanosoma cruzi por Rhodnius prolixus en 24 municipios del centro oriente de Colombia [Internet]. 2017 [cited 2022 Feb 2]. Available from: https://www.minsalud.gov.co/sites/rid/Lists/BibliotecaDigital/RIDE/VS/PP/ET/informe-verificacion-interrupcion-transmision-vectorial-chagas-2017.pdf?ID=20760

176. Karger DN, Conrad O, Böhner J, Kawohl T, Kreft H, Soria-Auza RW, et al. Climatologies at high resolution for the earth's land surface areas. Sci Data. 2017 Sep 5;4(1):170122.

177. Goodman S, BenYishay A, Lv Z, Runfola D. GeoQuery: Integrating HPC systems and public web-based geospatial data tools. Comput Geosci. 2019 Jan;122:103–12.

178. Amatulli G, Domisch S, Tuanmu M-N, Parmentier B, Ranipeta A, Malczyk J, et al. A suite of global, cross-scale topographic variables for environmental and biodiversity modeling. Sci Data. 2018 Dec 18;5(1):180040.

179. Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, et al. *Stan* : A Probabilistic Programming Language. J Stat Softw [Internet]. 2017 [cited 2022 Mar 31];76(1). Available from: http://www.jstatsoft.org/v76/i01/

180. Vehtari A, Gelman A, Gabry J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Stat Comput. 2017 Sep;27(5):1413–32.

181. DANE. Proyecciones de población [Internet]. 2015. [cited 2017 Feb 24]. Available from: https://www.dane.gov.co/index.php/estadisticas-por-tema/demografia-y-poblacion/proyecciones-de-poblacion

182. Benziger CP, do Carmo GAL, Ribeiro ALP. Chagas Cardiomyopathy: Clinical Presentation and Management in the Americas. Cardiol Clin. 2017;35(1):31–47.

# Appendix

## Contents

### Supplementary Tables

Supp. Table 2 :Variables tested as factors in the geospatial analyses of Chagas disease in Colombia ..... 124

Supp. Table 3: Final performances of the 5 best models for each of the three settings investigated for the 3 different approaches ..... 127

Supp. Table 4: Number of municipalities where the MAD Coefficient of Variation of the predictions of the model averaging is above 2 for each of the 3 approaches and for predictions in 1980 and 2010. ..... 127

Supp. Table 5: Median values of the MAD Coefficient of Variation (MAD-CV) of the predictions of the model averaging in areas where serosurveys have been conducted (in catchment area) and where no data were available (out of catchment area) and number of municipalities where the MAD-CV is greater than 5 ..... 127

Supp. Table 6: Predicted FoI averaged across all Colombian municipalities and among municipalities where serosurveys have been conducted and used in the analyses in 1980, 1990 and 2010, the percentage of decrease between 1980 and 2010 (trend) ..... 128

Supp. Table 7: Comparison of the observed FoI of Chagas Disease for serosurveys organised before and after 2005 in urban, rural, indigenous and mixed settings, Colombia, 1998-2014. ..... 128

Supp. Table 8: predictors used in Random Forest predictive model of the Force-of-Infection of Chagas disease at the municipal level in Colombia. ..... 129

Supp. Table 9: Performances of the Random Forest models ..... 130

Supp. Table 10: Comparison of the number of cases at the departmental level obtain through the modelling pipeline or the surveillance system in 2019 ..... 130

### Supplementary Figures

Supp. Figure 1: Locations and sample sizes of Chagas disease serosurveys conducted in Colombia with information on the location at the departmental level, from 1998 to 2014. ..... 133

Supp. Figure 2: Locations and sample sizes of Chagas disease serosurveys conducted in Colombia with information on the location at the municipality level, from 1998 to 2014. ..... 134

Supp. Figure 3: Goodness of fit of the model averaging of the 3 modelling approaches for each setting (urban, rural and indigenous). ..... 135

Supp. Figure 4 :Averaged observed FoI values by year for serosurveys conducted before 2010 in blue (n=39) and after or in 2010 in pink (n=40) for a) urban, b) rural and c) indigenous settings. ........ 136

Supp. Figure 5: Comparison of residual serial correlation for Boosted Regression Trees (BRT), Random Forest (RF) and Linear Model (LM) models built on the log scale based on the FullPostFoI approach. ........ 136

Supp. Figure 6: Observed and predicted prevalence using the catalytic model. The figure numbering corresponds to the code in the table A. ........ 137

Supp. Figure 7: Spatial, temporal and age class distribution of the number of Chagas disease cases (per year and per individual), in Colombia. ........ 140

Supp. Figure 8: Spatial distribution of the number of Chagas disease chronic cases (per year and per individual), in Colombia. ........ 141

Supp. Figure 9: Population size by year between 1985 and 2020 at the national level in Colombia. Urban population in blue, large cities population in red (i.e., cities with above 100,000 inhabitants in 1985) and total population in black ........ 142

Supp. Figure 10: Population size by age classes in 1990 and 2020 at the national level in Colombia. ........ 143

## Supplementary Methods

Supp. Method 1: Models used to estimate the Force-of-Infection ........ 144

Supp. Method 2: Machine Learning tuning ........ 146

i. Resampling strategies ........ 146

ii. Hyperparameters tuning ........ 148

iii. Results of tuning ........ 149

Supp. Method 3: Comparing observations and predictions across serosurveys ........ 149

Supp. Method 4: Predictive model for the FoI at the municipal level ........ 150

Supp. Method 5: Models used to estimate the FoI across Colombia ........ 166

Supp. Method 6: Models used to estimate the Burden of disease ........ 168

# Supplementary Tables

*Supp. Table 1 :Variables tested as factors in the geospatial analyses of Chagas disease in Colombia*

| Name | Description | Spatial scale | Data Source |
|---|---|---|---|
| **Serosurvey characteristics:** | | | |
| Year of the survey | Year when the serosurvey was conducted | – | (21) |
| Setting | Setting of the serosurvey: urban, rural, indigenous or mixed | – | (21) |
| Latitude | latitude of the centroid of the catchment area of the serosurvey | – | (21) |
| Longitude | longitude of the catchment area of the serosurvey | – | (21) |
| **Blood banks data:** | | | |
| Seroprevalence | Number of blood units positive for *Trypanosoma cruzi* divided by the number of blood units tested. Data aggregated for 1993–2010 by department | Department | PAHO |
| Proportion of blood units screened | Number of blood units tested for *T. cruzi* divided by the number of blood units received. Data aggregated for 1993–2010 by department | Department | PAHO |
| **Demography:** | | | |
| Population density | Estimates of the annual population size at municipality level from the government divided by the surface of the municipality in km$^2$ extracted from GDAM shapefiles | Municipality | (88) |
| Poverty | Proportion of households with deficit from 1993 census for 1950–1999 and from 2005 census for 2000–2014 | Municipality | (88) |
| Rural Indigenous Population size | Population size of the indigenous communities living in rural areas from 2005 census | Department | (88) |
| **Climate:** | | | |
| *Continuous* | | | |
| Polar climate frequency | Number of pixels defined as polar climate divided by the total number of pixels in the municipality | Municipality | (89) |
| Tropical climate frequency | Number of pixels defined as tropical climate divided by the total number of pixels in the municipality | Municipality | (89) |

| | | | |
|---|---|---|---|
| Temperate climate frequency | Number of pixels defined as temperate climate divided by the total number of pixels in the municipality | Municipality | (89) |
| Arid climate frequency | Number of pixels defined as arid climate divided by the total number of pixels in the municipality | Municipality | (89) |
| *Categorical* | | | |
| Tropical climate categorized | Tropical climate frequency categorized as follows: low (<10%), medium (10%–60%), large (60%–90%) and extra–large (>90%) | Municipality | (89) |
| **Entomological data:** | | | |
| *At Departmental level* | | | |
| *R. prolixus* geographical extent | Number of municipalities where *Rhodnius prolixus* is present divided by the number of municipalities in the department. Combined data from National report of 2013 and more recent data from Parra-Henao *et al*. | Department | (37,46,86) |
| *T. dimidiata* geographical extent | Number of municipalities where *Triatoma dimidiata* is present divided by the number of municipalities in the department. Combined data from National report of 2013 and more recent data from Parra-Henao *et al.* | Department | (37,46,86) |
| *R. prolixus* presence | Presence of *R. prolixus* in the department (yes/no). Combined data from National report of 2013 and more recent data from Parra-Henao *et al*. | Department | (37,46,86) |
| *T. dimidiata* presence | Presence of *T. dimidiata* in the department (yes/no). Combined data from National report of 2013 and more recent data from Parra-Henao *et al.* | Department | (37,46,86) |
| *At Municipality level* | | | |
| *R. prolixus* density | Number of *R. prolixus* specimens found divided by the number of households in the municipality. Data extracted from the National report of 2013 | Municipality | (86) |
| *T. dimidiata* density | Number of *T. dimidiata* specimens found divided by the number of households in the municipality. Data extracted from the National report of 2013 | Municipality | (86) |
| *R. prolixus* presence | Presence of *R. prolixus* in the municipality (yes/no). Combined data from National report of 2013 and more recent data from Parra-Henao *et al*. | Municipality | (37,46,86) |

| | | | |
|---|---|---|---|
| *T. dimidiata* presence | Presence of *T. dimidiata* in the municipality (yes/no). Combined data from National report of 2013 and more recent data from Parra-Henao *et al*. | Municipality | (37,46,86) |
| **Interventions:** | | | |
| *At Municipality level* | | | |
| Intervention intensity | Number of municipalities where interventions were organized divided by the number of municipalities in the department for the following time periods: before 1996, 1996–2000, 2001–2010 and 2011–2014. | Department | (87) |
| Intervention category | Municipality-level intervention intensity categorized as follows: no intervention (0%), low (0%–25%), medium (25%–50%), high (50%–75%), very high (>75%) | Department | (87) |
| *At Household level* | | | |
| Household intervention | Number of households having received interventions divided by the total number of households in the department for the following time periods: before 1996, 1996–2000, 2001–2010 and 2011–2014 | Department | (87) |
| Household intervention category | Household-level intervention intensity categorized as follows: no intervention (0%), medium (0%–10%), high (>10%) | Department | (87) |
| **Time:** | | | |
| Year | Year of the FoI value | – | – |
| Decade | Decade of the FoI value defined as follows: 1[1900–1909], 2[1910–1919], 3[1920–1929], 4[1930–1939], 5[1940–1949], 6[1950–1959], 7[1960–1969], 8[1970–1979], 9[1980–1989], 10[1990–1999], 11[2000–2009], 12[2010–2014] | – | – |

*Supp. Table 2: Final performances of the 5 best models for each of the three settings investigated for the 3 different approaches*

| | Approach 1 | | | Approach 2 | | | Approach 3 | | |
|---|---|---|---|---|---|---|---|---|---|
| | Urban | Rural | Ind. | Urban | Rural | Ind. | Urban | Rural | Ind. |
| 1st models | | | | | | | | | |
| Indicator | 0.827 | 0.781 | 0.672 | 0.638 | 0.541 | 0.637 | 0.670 | 0.5475 | 0.526 |
| Predictive $R^2$ | 0.827 | 0.781 | 0.672 | 0.718 | 0.768 | 0.672 | 0.776 | 0.708 | 0.505 |
| Overlap | - | - | - | 0.558 | 0.314 | 0.601 | 0.563 | 0.387 | 0.546 |
| 2nd models | | | | | | | | | |
| Indicator | 0.821 | 0.781 | 0.649 | 0.631 | 0.536 | 0.619 | 0.667 | 0.5430 | 0.524 |
| Predictive $R^2$ | 0.821 | 0.781 | 0.649 | 0.827 | 0.685 | 0.642 | 0.777 | 0.599 | 0.500 |
| Overlap | - | - | - | 0.435 | 0.387 | 0.595 | 0.556 | 0.487 | 0.548 |
| 3rd models | | | | | | | | | |
| Indicator | 0.821 | 0.778 | 0.642 | 0.630 | 0.535 | 0.606 | 0.654 | 0.5425 | 0.464 |
| Predictive $R^2$ | 0.821 | 0.778 | 0.642 | 0.821 | 0.771 | 0.617 | 0.771 | 0.721 | 0.402 |
| Overlap | - | - | - | 0.439 | 0.298 | 0.595 | 0.537 | 0.364 | 0.526 |
| 4th models | | | | | | | | | |
| Indicator | 0.819 | 0.776 | 0.623 | 0.626 | 0.534 | 0.599 | 0.652 | 0.5425 | 0.460 |
| Predictive $R^2$ | 0.819 | 0.776 | 0.623 | 0.802 | 0.771 | 0.615 | 0.741 | 0.718 | 0.404 |
| Overlap | - | - | - | 0.449 | 0.297 | 0.582 | 0.562 | 0.367 | 0.515 |
| 5th models | | | | | | | | | |
| Indicator | 0.818 | 0.771 | 0.617 | 0.625 | 0.533 | 0.597 | 0.628 | 0.5415 | 0.456 |
| Predictive $R^2$ | 0.818 | 0.771 | 0.617 | 0.719 | 0.776 | 0.649 | 0.781 | 0.728 | 0.402 |
| Overlap | - | - | - | 0.531 | 0.289 | 0.545 | 0.474 | 0.355 | 0.510 |

Ind. = Indigenous.

*Supp. Table 3: Number of municipalities where the MAD Coefficient of Variation of the predictions of the model averaging is above 2 for each of the 3 approaches and for predictions in 1980 and 2010.*

| | Approach 1 | | | | Approach 2 | | | | Approach 3 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1980 | | 2010 | | 1980 | | 2010 | | 1980 | | 2010 | |
| | n | % | n | % | n | % | n | % | n | % | n | % |
| Urban | 163 | 15.31 | 243 | 2282 | 359 | 33.71 | 354 | 33.24 | 286 | 26.85 | 299 | 28.08 |
| Rural | 285 | 26.76 | 284 | 26.67 | 271 | 25.45 | 275 | 25.82 | 266 | 24.98 | 265 | 24.88 |
| Indigenous | 341 | 3202 | 334 | 31.36 | 337 | 31.64 | 348 | 32.68 | 239 | 22.44 | 236 | 22.16 |
| | | | | | | | | | | | | |
| Total | 789 | 24.69 | 861 | 26.95 | 967 | 30.27 | 977 | 30.58 | 791 | 27.76 | 800 | 25.04 |

*Supp. Table 4: Median values of the MAD Coefficient of Variation (MAD-CV) of the predictions of the model averaging in areas where serosurveys have been conducted (in catchment area) and where no data were available (out of catchment area) and number of municipalities where the MAD-CV is greater than 5*

| | Median MAD-CV values (range) | | Number of municipalities with MAD-CV> 5 | | |
|---|---|---|---|---|---|
| | in catchment area | out catchment area | urban | rural | all |

| | | | | | |
|---|---|---|---|---|---|
| A1 | 1.29 (0.43-4.06) | 1.49 (0.23-11.98) | 2 | 14 | 39 |
| A2 | 1.28 (0.44-4.56) | 1.49 (0.16-12.12) | 9 | 13 | 81 |
| A3 | 1.33 (0.44-2.76) | 1.49 (0.24-11.00) | 6 | 11 | 17 |

*Supp. Table 5: Predicted FoI averaged across all Colombian municipalities and among municipalities where serosurveys have been conducted and used in the analyses in 1980, 1990 and 2010, the percentage of decrease between 1980 and 2010 (trend)*

| | All Municipalities | | | | Municipalities in catchment area | | | |
|---|---|---|---|---|---|---|---|---|
| | 1980 | 1990 | 2010 | trend | 1980 | 1990 | 2010 | trend |
| | mean (sd) | mean (sd) | mean (sd) | % | mean (sd) | mean (sd) | mean (sd) | % |
| Urban | $2.2 \times 10^{-3}$ | $2.1 \times 10^{-3}$ | $1.7 \times 10^{-3}$ | -23* | $2.2 \times 10^{-3}$ | $2.1 \times 10^{-3}$ | $1.6 \times 10^{-3}$ | -25* |
| | $(1.1 \times 10^{-3})$ | $(1.1 \times 10^{-3})$ | $(9.9 \times 10^{-4})$ | | $(9.6 \times 10^{-4})$ | $(9.1 \times 10^{-4})$ | $(8.7 \times 10^{-4})$ | |
| Rural | $1.7 \times 10^{-3}$ | $1.7 \times 10^{-3}$ | $1.7 \times 10^{-3}$ | -0.07 | $1.7 \times 10^{-3}$ | $1.7 \times 10^{-3}$ | $1.7 \times 10^{-3}$ | -0.10 |
| | $(1.0 \times 10^{-3})$ | $(1.0 \times 10^{-3})$ | $(1.0 \times 10^{-3})$ | | $(6.3 \times 10^{-4})$ | $(6.3 \times 10^{-4})$ | $(6.3 \times 10^{-4})$ | |
| Indigenous | $2.0 \times 10^{-2}$ | $2.0 \times 10^{-2}$ | $1.8 \times 10^{-2}$ | -7* | $2.3 \times 10^{-2}$ | $2.3 \times 10^{-2}$ | $2.1 \times 10^{-2}$ | -9* |
| | $(4.5 \times 10^{-3})$ | $(4.5 \times 10^{-3})$ | $(4.4 \times 10^{-3})$ | | $(2.9 \times 10^{-3})$ | $(2.9 \times 10^{-3})$ | $(3.0 \times 10^{-3})$ | |

*Statistically significant at a 5% significance level according to Student's *t* test comparing FoI values between 1980 and 2010

NB. The average FoI estimates are significantly higher before 2010 than after 2010 in all settings (Supp. Table 6) meaning that the year when the serosurvey was organised impacted all the settings with the greater impact in the rural settings.

*Supp. Table 6: Comparison of the observed FoI of Chagas Disease for serosurveys organised before and after 2005 in urban, rural, indigenous and mixed settings, Colombia, 1998-2014.*

| | 2010 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Before | | | | | After | | | | | t-test |
| | n | min | med | mean | max | n | min | med | mean | max | p_value |
| Urban | 368 | 0.0005 | 0.0029 | 0.0031 | 0.0089 | 421 | 0.0003 | 0.0012 | 0.0017 | 0.0055 | << 0.001 |
| Rural | 433 | 0.0005 | 0.0035 | 0.0053 | 0.0166 | 566 | 0.0002 | 0.0008 | 0.0013 | 0.0064 | << 0.001 |
| Indigenous | 230 | 0.0196 | 0.0279 | 0.0342 | 0.0772 | 18 | 0.0196 | 0.0201 | 00203 | 0.0215 | << 0.001 |
| Mixed | 99 | 0.0019 | 0.0061 | 0.0051 | 0.0075 | 218 | 0.0001 | 0.0006 | 0.0025 | 0.0081 | << 0.001 |

*Supp. Table 7: predictors used in Random Forest predictive model of the Force-of-Infection of Chagas disease at the municipal level in Colombia.*

| Name | Description | time span | time unit | space unit | Source |
|---|---|---|---|---|---|
| Serosurvey characteristics: | | | | | |
| Year when the serosurvey was conducted | | | | | (21) |
| Setting type | urban/rural/indigenous of mixed | | | | (21) |
| Spatiotemporal Coordinates: | | | | | |
| Year | year of concern | | | | (21) |
| Environmental predictors: | | | | | |
| Year of certification | Year when the municipalities have been certified free from intradomiciliated vector. Municipalities without elimination year have been set to 1900 | 2009-2019 | total period | Municipal | (175) |
| Isothermality Bio3 | | 1979-2013 | total period | 1km | (176) |
| Minimum temperature of the coldest month Bio6 | | 1979-2013 | total period | 1km | (176) |
| Seasonality of precipitation Bio15 | | 1979-2013 | total period | 1km | (176) |
| NDVI | Vegetation indicator. Years before 1981 received values of 1981 and year after 2013 received values of 2013 | 1981-2015 | year | municipal | (177) |
| Elevation | STRM DEM | | | 1km | (178) |
| Demographic predictors: | | | | | |

| | | | | | |
|---|---|---|---|---|---|
| Population size | Estimated total population by municipality | 1985-2020 | year | municipality | (88) |
| Proportion of urban population | Proportion of the municipal population defined as urban | 1985-2020 | year | municipality | (88) |
| Unfinished floor | Median percentage of households in the geographic unit that have a dirt/unfinished floor | 1973, 1985, 1993 and 2005 | year | municipality | (133) |

The Random Forest predictive model showed good performance within each bootstrapping loop with R2 calculated on the training and validation set being around 0.92. When looking at all of the bootstrapped models and the entire posterior distribution of the response variable (FoI), the performance still is good with a general R2 of 0.627 while the one calculated only on the cross-validation sets is 0.625.

*Supp. Table 8: Performances of the Random Forest models*

| | urban | rural |
|---|---|---|
| cross validation R2 | 0.64 | 0.71 |
| proportion of overlap | 0.59 | 0.59 |
| performance indicator | 0.61 | 0.65 |

*Supp. Table 9: Comparison of the number of cases at the departmental level obtain through the modelling pipeline or the surveillance system in 2019*

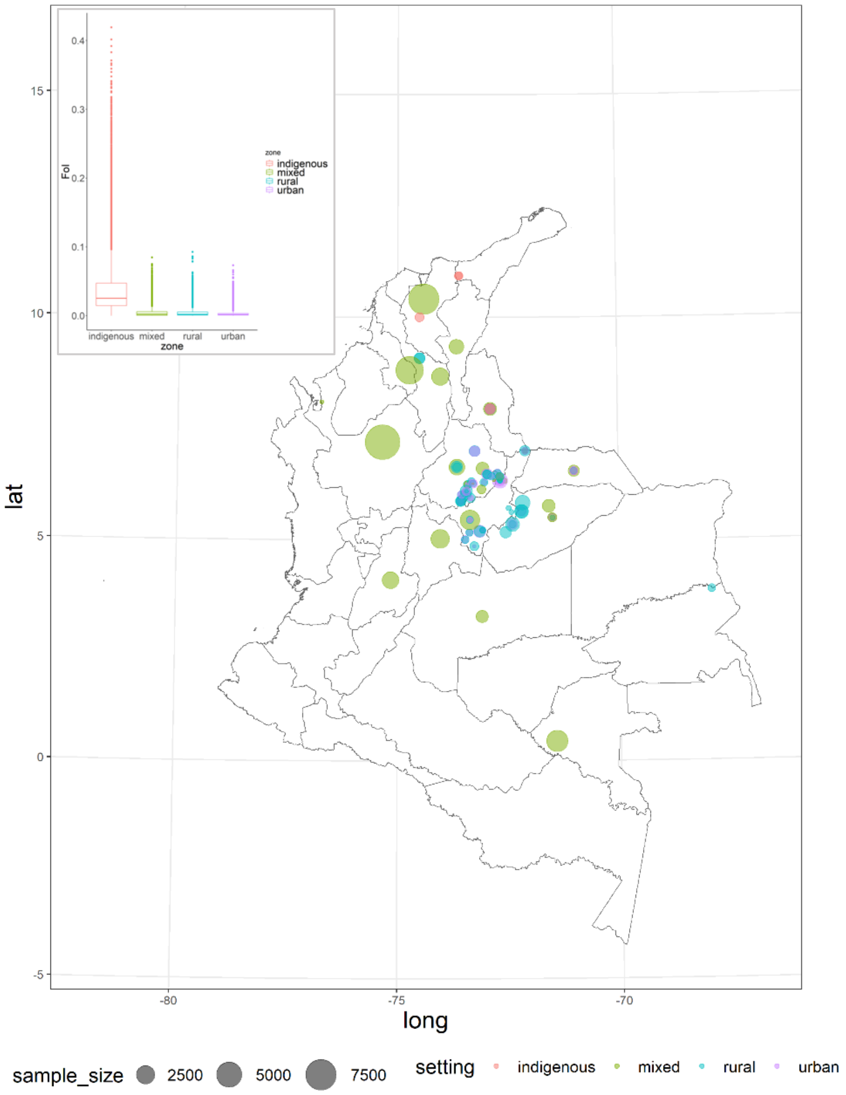
| | Median estimates from model pipeline | | | | | Confirmed cases from the surveillance system | |
|---|---|---|---|---|---|---|---|
| Department | Total | Acute | Asymptomitic | Chronic mild | Chronic severe | Acute | Chronic |
| Antioquia | 54,488 | 2,399 | 38,210 | 11,433 | 2,381 | 4 | 0 |
| Atlántico | 15,381 | 753 | 10,862 | 3,135 | 675 | 4 | 0 |
| Bogotá, D.C. | 27,503 | 1,255 | 19,639 | 6,240 | 1,353 | 0 | 0 |
| Bolívar | 20,812 | 962 | 14,508 | 4,222 | 895 | 0 | 0 |
| Boyacá | 18,579 | 752 | 13,208 | 3,852 | 847 | 0 | 21 |
| Caldas | 9,786 | 397 | 6,804 | 2,087 | 453 | 0 | 0 |
| Caquetá | 4,861 | 229 | 3,431 | 993 | 194 | 0 | 0 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Cauca | 22,707 | 973 | 15,865 | 4,746 | 973 | 0 | 0 |
| Cesar | 11,699 | 563 | 8,265 | 2,406 | 510 | 3 | 22 |
| Córdoba | 21,548 | 949 | 15,222 | 4,487 | 964 | 0 | 0 |
| Cundinamarca | 39,343 | 1,705 | 27,543 | 8,097 | 1,831 | 0 | 0 |
| Chocó | 4,659 | 227 | 3,223 | 939 | 198 | 2 | 0 |
| Huila | 11,989 | 530 | 8,396 | 2,446 | 512 | 0 | 0 |
| La Guajira | 19,354 | 1,068 | 13,586 | 3,908 | 809 | 0 | 0 |
| Magdalena | 15,337 | 719 | 10,747 | 3,117 | 674 | 0 | 0 |
| Meta | 13,862 | 621 | 9,898 | 2,793 | 529 | 1 | 0 |
| Nariño | 20,377 | 862 | 14,378 | 4,231 | 964 | 0 | 0 |
| Norte de Santander | 25,326 | 1,156 | 17,880 | 5,054 | 1,027 | 0 | 0 |
| Quindio | 3,456 | 143 | 2,431 | 740 | 150 | 0 | 0 |
| Risaralda | 6,293 | 268 | 4,356 | 1,328 | 293 | 0 | 0 |
| Santander | 27,245 | 1,177 | 18,935 | 5,840 | 1,267 | 0 | 33 |
| Sucre | 10,244 | 449 | 7,232 | 2,110 | 461 | 16 | 0 |
| Tolima | 14,827 | 587 | 10,448 | 3,081 | 652 | 0 | 0 |
| Valle del Cauca | 25,545 | 1,112 | 17,619 | 5,307 | 1,077 | 0 | 0 |
| Arauca | 3,707 | 170 | 2,593 | 737 | 156 | 0 | 91 |
| Casanare | 6,152 | 281 | 4,348 | 1,259 | 272 | 28 | 0 |
| Putumayo | 3,664 | 165 | 2,589 | 757 | 166 | 0 | 0 |
| Archipiélago de San Andrés | 864 | 37 | 601 | 185 | 35 | 0 | 0 |
| Amazonas | 584 | 31 | 414 | 113 | 22 | 0 | 0 |
| Guainía | 319 | 18 | 231 | 59 | 11 | 0 | 0 |
| Guaviare | 1,005 | 46 | 717 | 191 | 37 | 0 | 0 |
| Vaupés | 327 | 20 | 237 | 59 | 11 | 0 | 0 |
| Vichada | 1,598 | 91 | 1,122 | 310 | 65 | 0 | 0 |

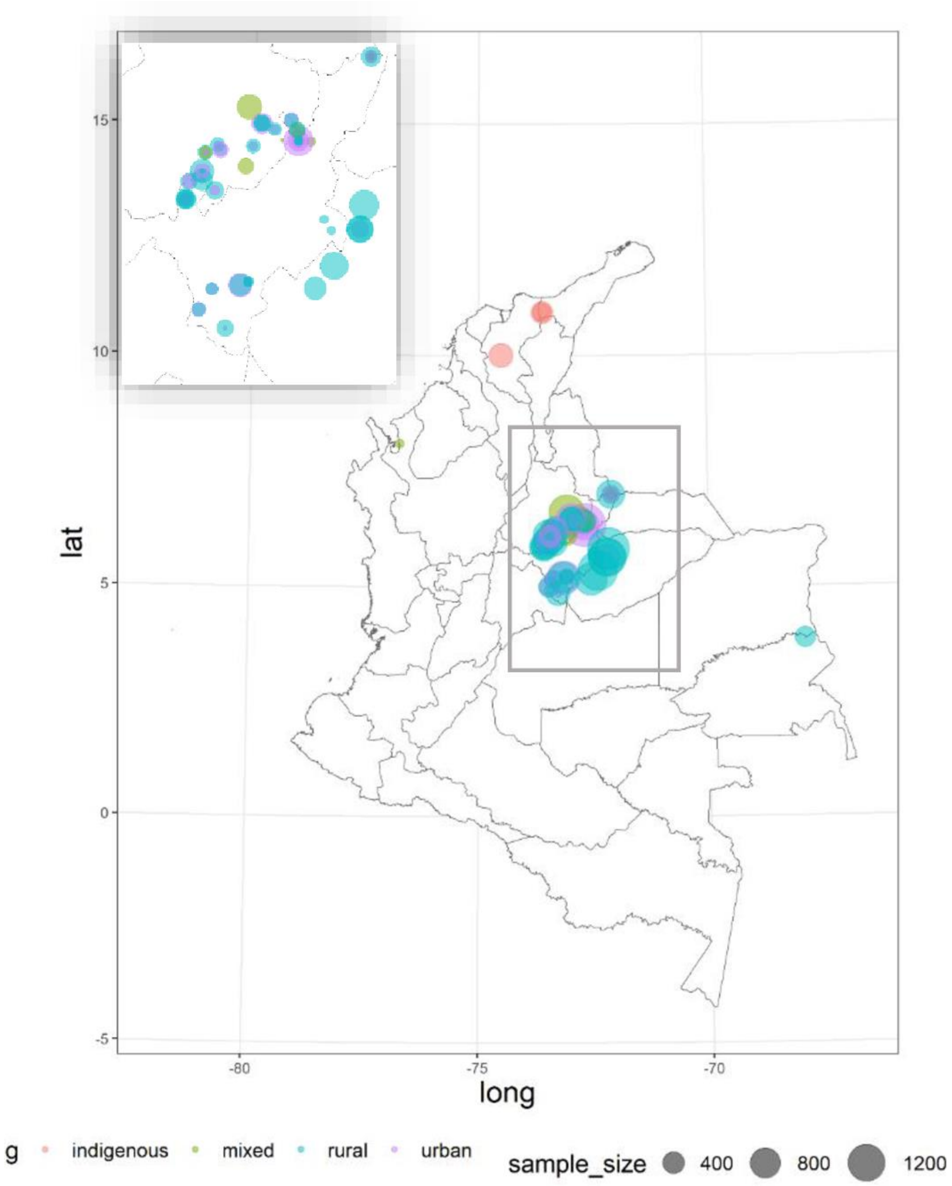
## Supplementary Figures

In Colombia, 109 serosurveys were conducted after 1980 and they only provide information on the catchment area, either at Municipality or Departmental levels. The presence of domiciliated vectors, which transmit most of the infection, can vary greatly from one area to another. Even at the village level, vector infestation strongly depends on the materials used to build houses, as well as on knowledge of the risk and vector control activities. Therefore, in these analyses, we only used the serosurveys with information on the location at municipality level. Being able to use all the data available as well as using a smaller geographical scale would provide a more coherent model. This is not possible within the linear framework but some more advanced methods, such as machine learning, could handle this issue.

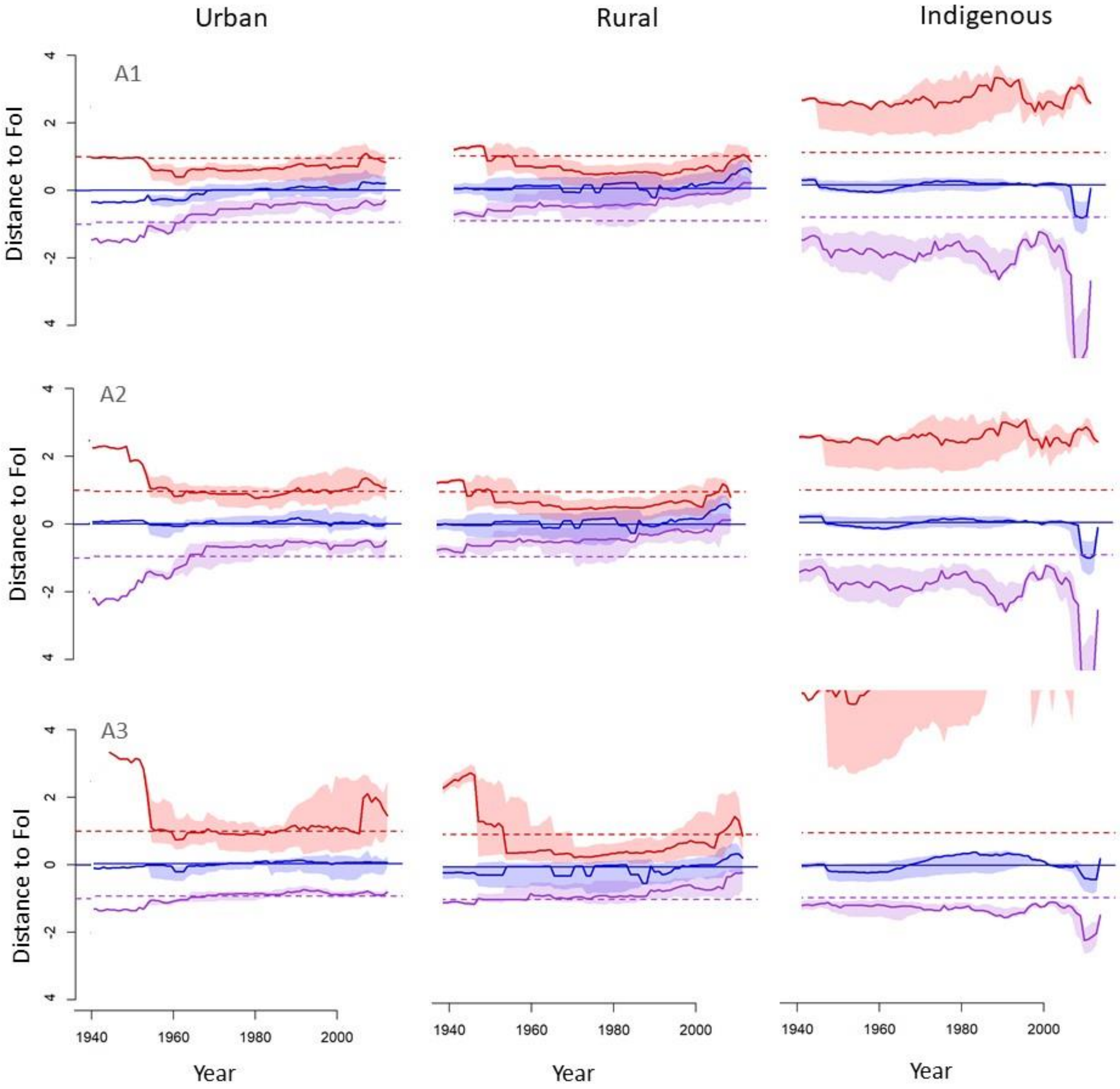
From the 76 serosurveys used in the analyses, 27 were conducted in urban settings, 36 in rural settings, 5 in indigenous settings and 8 were mixed (including urban, rural, and unknown settings).



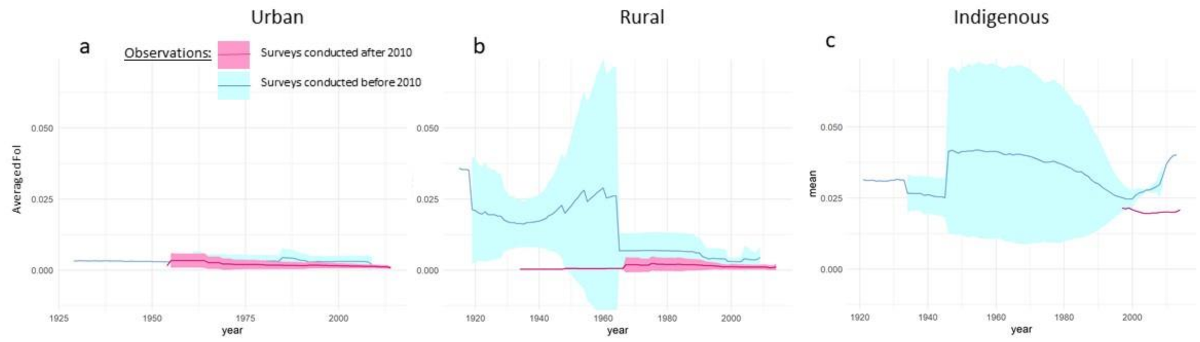
*Supp. Figure 1: Locations and sample sizes of Chagas disease serosurveys conducted in Colombia with information on the location at the departmental level, from 1998 to 2014. The grey boundaries delimitate the Departments.*



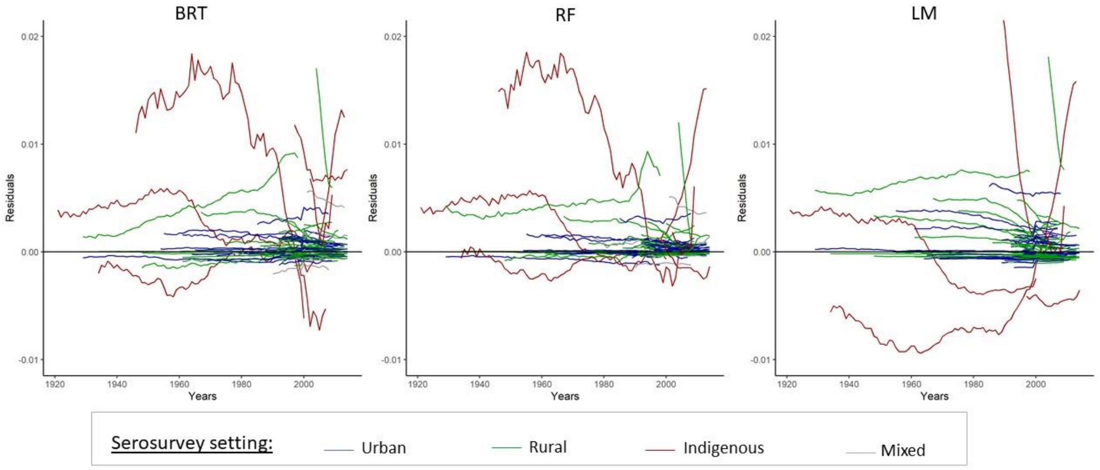
*Supp. Figure 2: Locations and sample sizes of Chagas disease serosurveys conducted in Colombia with information on the location at the municipality level, from 1998 to 2014.* The grey boundaries delimitate the Departments.



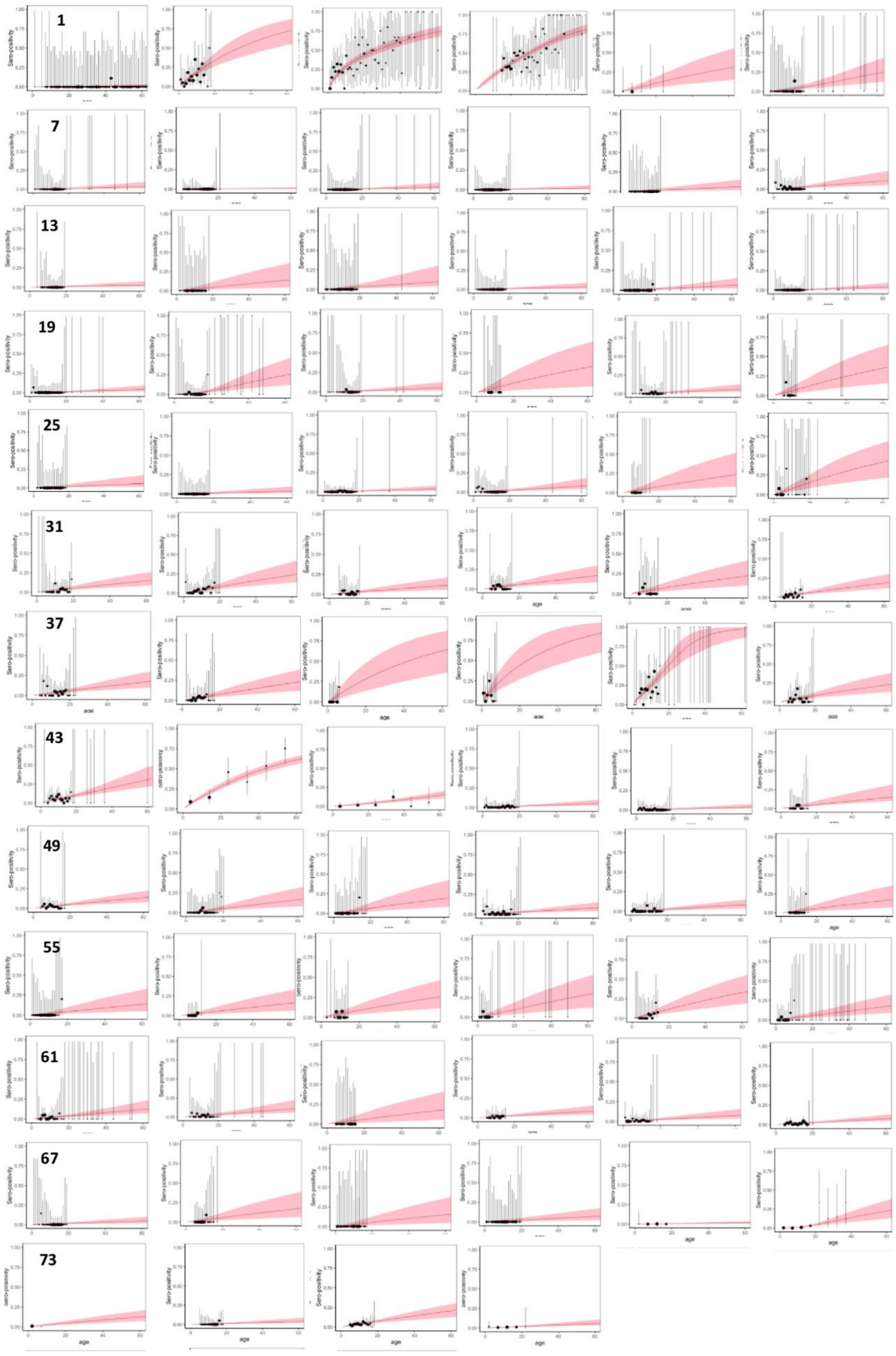
*Supp. Figure 3: Goodness of fit of the model averaging of the 3 modelling approaches for each setting (urban, rural and indigenous). The lines and envelopes are the distance between observations and predictions' median (blue), and 95%CI (upper bound in red and lower bound in purple); Approach 1: models fitted with median FoI estimates and selected based on Predictive $R^2$; Approach 2: models fitted with median FoI estimates and selected based on Predictive $R^2$ and overlap; Approach 3: models fitted with the full posterior distribution of FoI estimates and selected based on the Predictive $R^2$ and overlap.*



*Supp. Figure 4 :Averaged observed FoI values by year for serosurveys conducted before 2010 in blue (n=39) and after or in 2010 in pink (n=40) for a) urban, b) rural and c) indigenous settings.*



*Supp. Figure 5: Comparison of residual serial correlation for Boosted Regression Trees (BRT), Random Forest (RF) and Linear Model (LM) models built on the log scale based on the FullPostFoI approach. Each line corresponds to one serosurvey.*



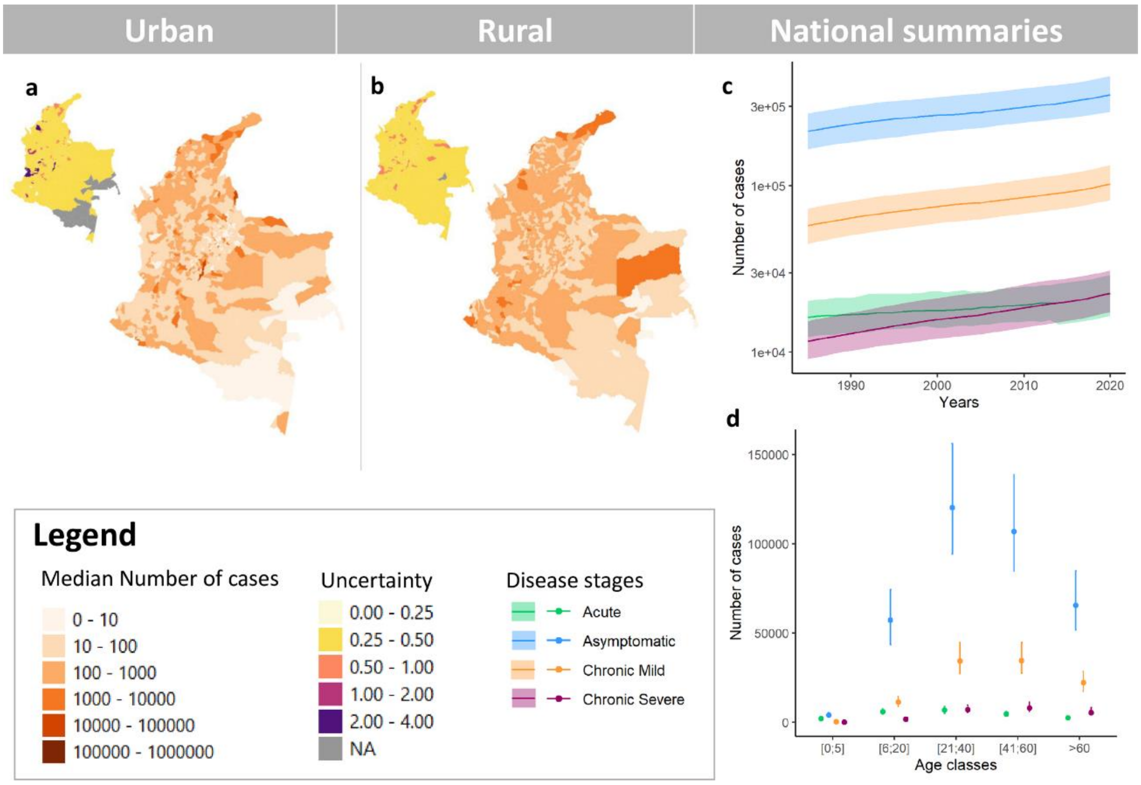
*Supp. Figure 6: Observed and predicted prevalence using the catalytic model. The figure numbering corresponds to the code in the table A.*

Table A: Characteristics of the serosurveys included in the analyses

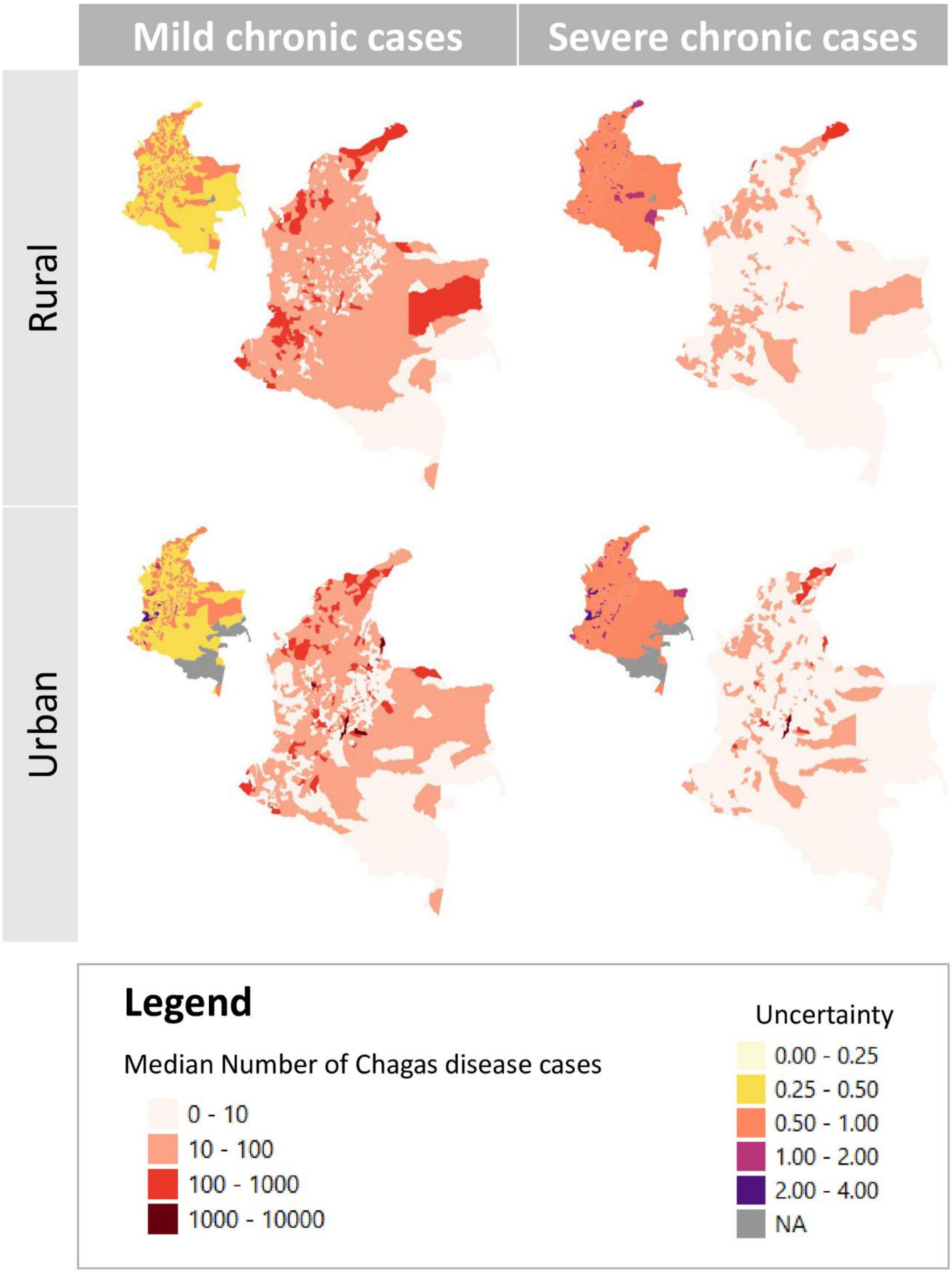
| Code | Year | Setting type | Department | Municipality | Sample size | Diagnostic test | Survey ID |
|---|---|---|---|---|---|---|---|
| 1 | 2013 | rural | Guainía | Puerto Inírida | 363 | ELISA . | COL-035-82 |
| 2 | 2014 | indigenous | Magdalena | Santa Marta (Dist. Esp.) | 287 | ELISA . | COL-035-92 |
| 3 | 2007 | indigenous | Magdalena | Santa Marta (Dist. Esp.) | 418 | ELISA . | COL-035-96 |
| 4 | 2000 | indigenous | Magdalena | Santa Marta (Dist. Esp.) | 493 | ELISA IFI | COL-035-97 |
| 5 | 2010 | mixed | Antioquia | Turbo | 66 | ELISA . | COL-035-81 |
| 6 | 2013 | urban | Santander | Coromoro | 146 | ELISA . | COL-035-54 |
| 7 | 2013 | rural | Santander | Coromoro | 354 | ELISA . | COL-035-53 |
| 8 | 2014 | mixed | Santander | Curití | 1152 | ELISA . | COL-035-55 |
| 9 | 2014 | mixed | Santander | Encino | 454 | ELISA . | COL-035-56 |
| 10 | 2014 | rural | Santander | Gámbita | 573 | ELISA . | COL-035-57 |
| 11 | 2014 | urban | Santander | Gámbita | 162 | ELISA . | COL-035-58 |
| 12 | 2013 | rural | Santander | Guadalupe | 390 | ELISA . | COL-035-60 |
| 13 | 2013 | mixed | Santander | Guadalupe | 291 | ELISA . | COL-035-59 |
| 14 | 2013 | urban | Santander | Guadalupe | 80 | ELISA . | COL-035-61 |
| 15 | 2014 | urban | Santander | Guapotá | 103 | ELISA . | COL-035-63 |
| 16 | 2014 | rural | Santander | Guapotá | 351 | ELISA . | COL-035-62 |
| 17 | 2013 | urban | Santander | Mogotes | 330 | ELISA . | COL-035-65 |
| 18 | 2014 | urban | Santander | Mogotes | 775 | ELISA . | COL-035-67 |
| 19 | 2013 | rural | Santander | Mogotes | 523 | ELISA . | COL-035-64 |
| 20 | 2014 | rural | Santander | Mogotes | 428 | ELISA . | COL-035-66 |
| 21 | 2013 | rural | Santander | Oiba | 263 | ELISA . | COL-035-69 |
| 22 | 2013 | mixed | Santander | Oiba | 30 | ELISA . | COL-035-68 |
| 23 | 2013 | urban | Santander | Oiba | 457 | ELISA . | COL-035-70 |
| 24 | 2014 | mixed | Santander | Onzaga | 40 | ELISA . | COL-035-71 |
| 25 | 2014 | urban | Santander | San Joaquín | 173 | ELISA . | COL-035-73 |
| 26 | 2014 | rural | Santander | San Joaquín | 305 | ELISA . | COL-035-72 |
| 27 | 2013 | rural | Santander | Suaita | 1014 | ELISA . | COL-035-74 |
| 28 | 2013 | urban | Santander | Suaita | 419 | ELISA . | COL-035-90 |
| 29 | 2014 | rural | Santander | Suaita | 61 | ELISA . | COL-035-75 |
| 30 | 2008 | urban | Boyacá | Tipacoque | 81 | ELISA . | COL-035-37 |
| 31 | 2008 | mixed | Boyacá | Tipacoque | 377 | ELISA . | COL-035-35 |
| 32 | 2008 | rural | Boyacá | Tipacoque | 467 | ELISA . | COL-035-36 |
| 33 | 2007 | urban | Boyacá | Chinavita | 277 | ELISA . | COL-035-15 |
| 34 | 2007 | rural | Boyacá | Chinavita | 274 | ELISA . | COL-035-14 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 35 | 2007 | urban | Boyacá | Chitaraque | 92 | ELISA . | COL-035-17 |
| 36 | 2007 | rural | Boyacá | Chitaraque | 803 | ELISA . | COL-035-16 |
| 37 | 2008 | urban | Boyacá | Covarachia | 392 | ELISA . | COL-035-30 |
| 38 | 2008 | rural | Boyacá | Covarachia | 323 | ELISA . | COL-035-29 |
| 39 | 2009 | rural | Boyacá | Cubara | 79 | ELISA . | COL-035-39 |
| 40 | 2009 | indigenous | Boyacá | Cubara | 85 | ELISA . | COL-035-38 |
| 41 | 2007 | indigenous | Boyacá | Cubara | 212 | ELISA . | COL-035-18 |
| 42 | 2007 | urban | Boyacá | Cubara | 285 | ELISA . | COL-035-20 |
| 43 | 2007 | rural | Boyacá | Cubara | 689 | ELISA . | COL-035-19 |
| 44 | 1998 | urban | Boyacá | Guateque | 356 | ELISA . | COL-001-02 |
| 45 | 1998 | rural | Boyacá | Guateque | 333 | ELISA . | COL-001-01 |
| 46 | 2009 | urban | Boyacá | Miraflores | 1035 | ELISA . | COL-035-41 |
| 47 | 2009 | rural | Boyacá | Miraflores | 853 | ELISA . | COL-035-40 |
| 48 | 2007 | urban | Boyacá | Moniquira | 205 | ELISA . | COL-035-22 |
| 49 | 2007 | rural | Boyacá | Moniquira | 590 | ELISA . | COL-035-21 |
| 50 | 2011 | urban | Boyacá | Moniquira | 311 | ELISA . | COL-035-50 |
| 51 | 2010 | urban | Boyacá | Moniquira | 134 | ELISA . | COL-035-46 |
| 52 | 2011 | rural | Boyacá | Moniquira | 760 | ELISA . | COL-035-49 |
| 53 | 2010 | rural | Boyacá | Moniquira | 561 | ELISA . | COL-035-45 |
| 54 | 2007 | rural | Boyacá | Paya | 124 | ELISA . | COL-035-23 |
| 55 | 2010 | mixed | Boyacá | Boavita | 154 | ELISA . | COL-035-44 |
| 56 | 2007 | rural | Boyacá | Pisba | 140 | ELISA . | COL-035-24 |
| 57 | 2007 | urban | Boyacá | San Eduardo | 84 | ELISA . | COL-035-26 |
| 58 | 2009 | urban | Boyacá | San Eduardo | 87 | ELISA . | COL-035-43 |
| 59 | 2007 | rural | Boyacá | San Eduardo | 166 | ELISA . | COL-035-25 |
| 60 | 2009 | rural | Boyacá | San Eduardo | 200 | ELISA . | COL-035-42 |
| 61 | 2008 | rural | Boyacá | San jose de Pare | 484 | ELISA . | COL-035-31 |
| 62 | 2008 | urban | Boyacá | San jose de Pare | 364 | ELISA . | COL-035-32 |
| 63 | 2007 | urban | Boyacá | Santa maria | 50 | ELISA . | COL-035-28 |
| 64 | 2007 | rural | Boyacá | Santa maria | 508 | ELISA . | COL-035-27 |
| 65 | 2010 | urban | Boyacá | Soata | 741 | ELISA . | COL-035-48 |
| 66 | 2008 | urban | Boyacá | Soata | 1680 | ELISA . | COL-035-34 |
| 67 | 2011 | urban | Boyacá | Soata | 311 | ELISA . | COL-035-52 |
| 68 | 2008 | rural | Boyacá | Soata | 121 | ELISA . | COL-035-33 |
| 69 | 2010 | rural | Boyacá | Soata | 76 | ELISA . | COL-035-47 |
| 70 | 2011 | rural | Boyacá | Soata | 158 | ELISA . | COL-035-51 |
| 71 | 2013 | rural | Casanare | Aguazul | 926 | ELISA IFI | COL-035-79 |

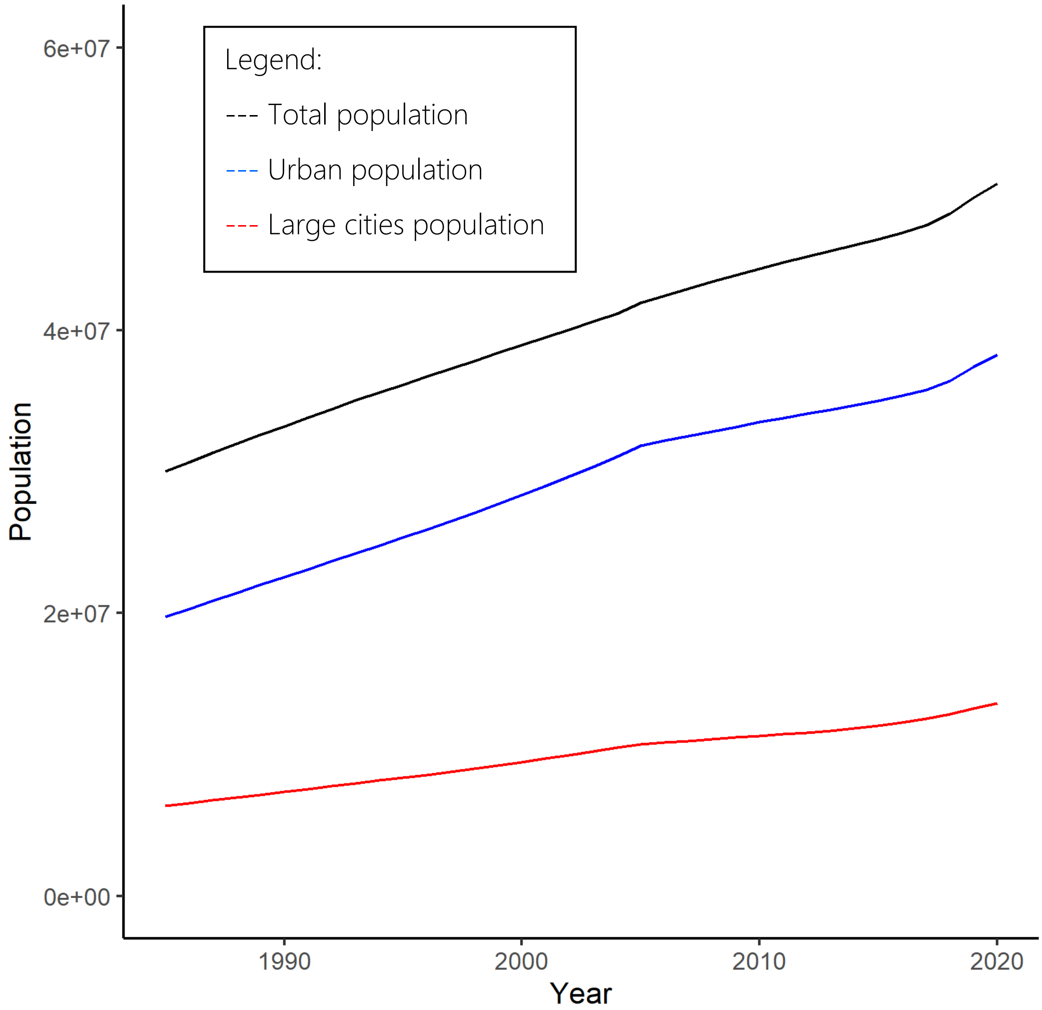
| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 72 | 2011 | rural | Casanare | Támara | 1597 | ELISA IFI | COL-035-77 |
| 73 | 2011 | rural | Casanare | Yopal | 1400 | ELISA IFI | COL-035-78 |
| 74 | 2009 | urban | Casanare | Nunchía | 528 | ELISA IFI | COL-005-02 |
| 75 | 2009 | rural | Casanare | Nunchía | 1338 | ELISA IFI | COL-005-01 |
| 76 | 2013 | rural | Casanare | Nunchía | 1215 | ELISA IFI | COL-035-80 |



*Supp. Figure 7: Spatial, temporal and age class distribution of the number of Chagas disease cases (per year and per individual), in Colombia. a and b: Municipal number of cases in 2020 in urban and rural areas (main maps); the associated uncertainty (small map insets) present the median divided by the interquartile. c: Yearly national number of cases (median, solid line and interquartile, ribbon) from 1985 to 2020; each color corresponds to a different disease stage. d: National number of cases by age class (median, point and interquartile, error bar) in 2020; each color corresponds to a different disease stage.*



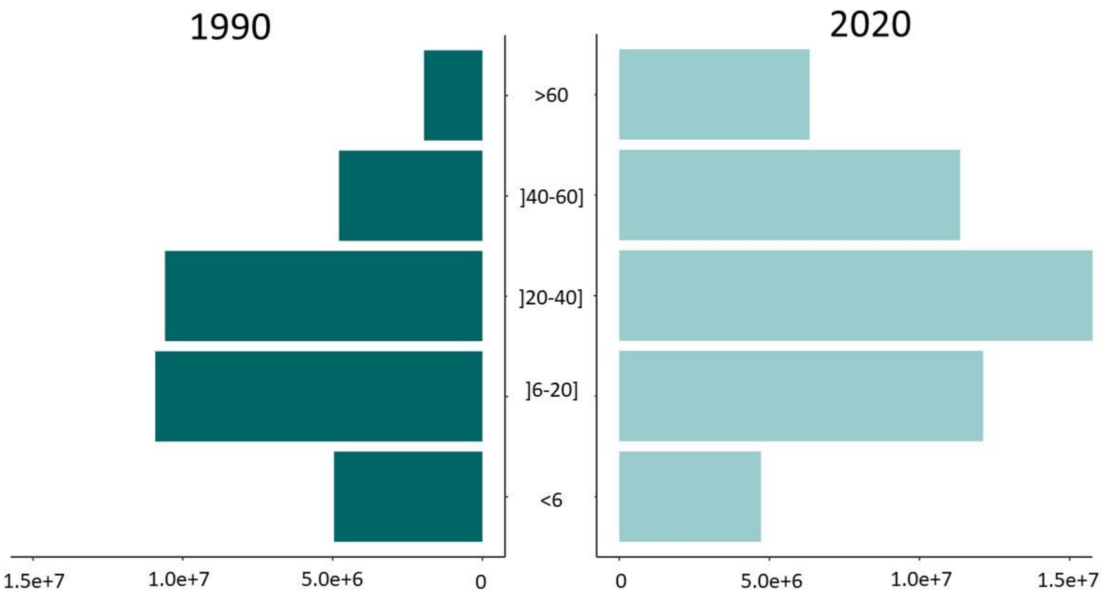
*Supp. Figure 8: Spatial distribution of the number of Chagas disease chronic cases (per year and per individual), in Colombia. Municipal number of cases in 2020 for chronic mild and chronic severe cases (main maps); the associated uncertainty (small map insets) presents the median divided by the interquartile.*



*Supp. Figure 9: Population size by year between 1985 and 2020 at the national level in Colombia. Urban population in blue, large cities population in red (i.e., cities with above 100,000 inhabitants in 1985) and total population in black*

Population increased by about 39% between 1990 and 2020.

In 2020, urban population represented 76% of the total population while the population in the large cities (i.e., cities with above 100,000 inhabitants in 1985) represented 27% of the total population.



*Supp. Figure 10: Population size by age classes in 1990 and 2020 at the national level in Colombia.*

# Supplementary method

*Supp. Method 1: Models used to estimate the Force-of-Infection: Extracted from Z. M. Cucunubá, "Modelling the epidemiology and healthcare burden of Chagas disease in Colombia," Imperial College of London. (2017)*

There are 1122 municipalities in Colombia, but the study area only covered 34 of them.

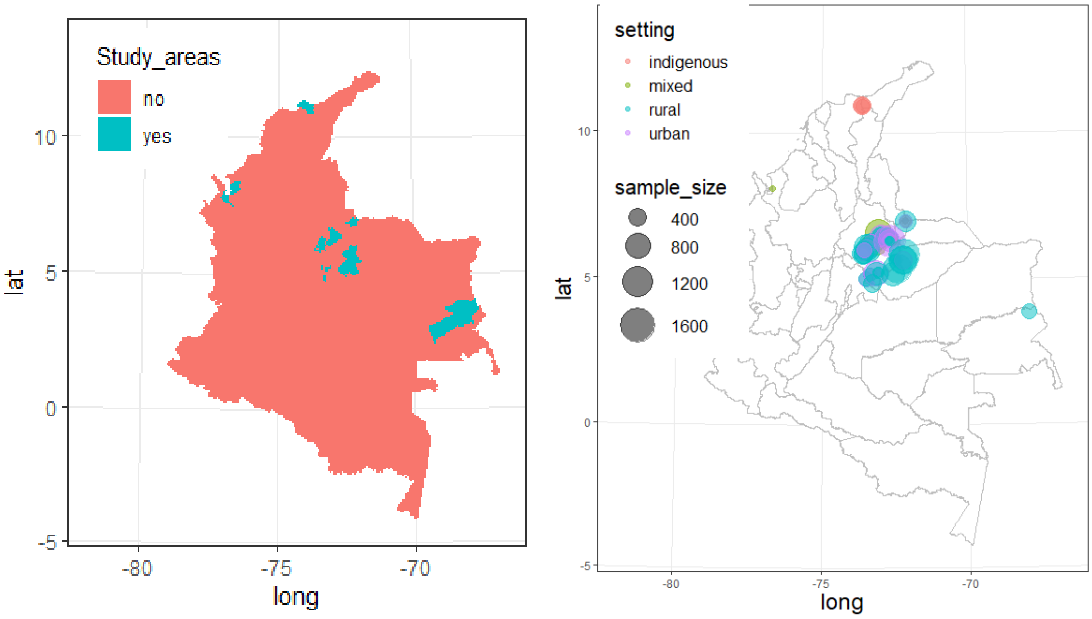


Figure a: Spatial distribution of Chagas disease serosurveys conducted in Colombia at ADM2 level, 1980–2014. Left panel, municipalities where at least one serosurvey has been conducted in green and municipality where no survey has been conducted in orange; right panel, locations, setting type and sample sizes of the serosurveys.

## Models used to estimate the Force-of-Infection: Extracted from Z. M. Cucunubá, "Modelling the epidemiology and healthcare burden of Chagas disease in Colombia," Imperial College of London. (2017)

"Descriptive prevalence results are reported as percentages and accompanied by 95% binomial (exact) confidence intervals (95% CI). For the force-of-infection models, we consider that if the rate of infection acquisition—here the rate of seroconversion—is constant over time, infection (sero)prevalence will increase monotonically with age as cumulative exposure increases. Formally, $P_a = 1 - exp(-\lambda_a)$ , with $P_a$ the age-specific seroprevalence and $\lambda$ the force-of-infection (the per susceptible incidence or FoI) as originally described by Muench, 1959 [9, 10]. More generally, the FoI may fluctuate over time t, modifying the seroprevalence age profiles. For a survey completed at time τ , $P_{a,\tau} = 1 - exp\left(-\int_{t=\tau-a}^{t=\tau} \lambda_t dt\right)$. Therefore, a serosurvey completed at time τ , and including ages from {$a_{min}$ , $a_{max}$} , is informative on exposure (and FoI) between τ - $a_{max}$ and τ . Other modelling assumptions included: a) no age-dependency in transmission [76], b) no seroreversion [76], and c) no specific migration due to Chagas infection status [11, 12]."

Based on the above, we used the posterior distribution of the FoI fitted with time-varying FoI ($\lambda_i$) following:"

$$P_{a,\tau} = 1 - exp\left(-\sum_{i=\tau-a=1}^{i=\tau} \lambda_i\right)$$

Models were fitted on a Bayesian framework using Stan's No-U-Turn Sampler (179) with four Markov chains and 20,000 iterations on each and with 50% of these iterations discarded as "warm-up".

Prior for constant FOI model

The prior for the FOI estimate in the constant model is based on a uniform distribution

$$FOI \sim uniform(0,2)$$

Prior for time-varying FOI model

The prior for FOI estimate in the time-varying follows a $\boldsymbol{normal}$ distribution informed by the FOI value from the previous decade:

$$FOI \sim normal\ (FOI_{previous\ year}, \sigma)$$
$$\sigma \sim cauchy\ (0, 1)$$
$$FOI_{first\ year} \sim normal\ (0, 1)$$

And, the probability of a positive case at age $\boldsymbol{a}$ follows a binomial distribution

$$\sim binomial(N_a,\ P_a)$$

Where $N$ is the total sample, and $P$ the observed prevalence at age $\boldsymbol{a}$

Convergence and Posterior Predictive Checks

Convergence was assessed by the use of $\mathbf{R}^{\wedge}$ statistic, which measures the "within chain" variability (W) and compares to the "between chains" variability (B):

$$\mathrm{R}^{\wedge} = \sqrt{\frac{\mathrm{W} + \frac{1}{n}(\mathrm{B} - \mathrm{W})\,\mathrm{W}}{W}}$$

This method assumes that the chains have been simulated in parallel, each with different starting points, which are over dispersed with respect to the target distribution.

If this metric is large, this suggests that either estimate of the variance can be further decreased by more simulations (180). It is expected that in convergence, B → W =⇒ R^ → 1.

A value of $\mathrm{R}^{\wedge} < 1.1$ was considered enough to achieve convergence. We show the convergence plots for the two instances where a time-varying FI model fit the data the best. We also did posterior predictive checks and examined the residuals

### *Supp. Method 2: Machine Learning tuning*

#### i. Resampling strategies

Spatial only, temporal only and spatiotemporal resampling strategies were tested and compared to a standard random resampling strategy. Also, the resampling strategy that was developed for the linear framework (which consisted of selecting one value (i.e. year) for each serosurvey at each iteration) was tested (Table a). Values of between 5 and 50 folds (number of subsets the data are divided into) were tested and eventually set at 10 as no substantial changes were observed on the performance indicators.

*Table a: Resampling strategies tested*

| Model Name | Parameters used | Task type | Resampling method |
|---|---|---|---|
| Random | default parameters | default | Random 10 folds |
| Temporal | default parameters | temporal | LTO: Leave Time Out 10 folds (122) |
| Spatial | default parameters | spatial | LLO: Leave Location Out 10 folds (122) |

| | | | |
|---|---|---|---|
| Spatiotemporal | default parameters | Spatiotemporal | LTLO: Leave Time and Location Out 10 folds (122) |
| Custom | default parameters | default | Custom stratified bootstrapping that only selects one FoI value for each of the serosurveys. Account for the temporal correlation inherited from a catalytic model used to obtain FoI estimates |

The Boosted Regression Tree (BRT) and Random Forest (RF) methods seemed unaffected by the resampling strategies, with only a marginal improvement in predictions (+/- 1%) (Table b).

However, the custom resampling strategy showed substantially worse performance and signs of overfitting, with an extremely low Resample $R^2$ when using both Machine Learning (ML) methods. This strategy had been developed for the Linear Model (LM) framework, but it performed poorly with the ML frameworks. ML seemed to suffer from a small number of observations selected at each iteration (n=76). However, ML was able to handle temporal correlation directly, by using specially designed spatiotemporal resampling methods. Thus, spatiotemporal resampling methods available within the ML framework can be a straightforward and efficient alternative to the stratified bootstrapping method used previously (103). The number of folds used in the resampling method did not impact the results and was set at 10.

*Table b: Median performance of Boosted Regression Trees (BRT) and Random Forest (RF) models for different resampling methods using median FoI (MedFoI) to fit the models and default parameters*

| Model Name | $R^2$ | | Resample $R^2$ | |
|---|---|---|---|---|
| | BRT | RF | BRT | RF |
| Random | 0.83 | 0.98 | 0.82 | 0.97 |
| Temporal | 0.84 | 0.98 | 0.83 | 0.98 |
| Spatial | 0.84 | 0.98 | 0.83 | 0.98 |
| Spatiotemporal | 0.84 | 0.98 | 0.83 | 0.98 |
| Custom | 0.50 | 0.66 | -3.11 | -1.27 |

BRT: Boosted Regression Trees; RF: Random Forest methods; $R^2$ is calculated on the entire dataset while the Resample $R^2$ is calculated on the test set at each resampling iteration.

### ii. Hyperparameters tuning

ML hyperparameters define the structure and complexity of the models and can be optimised to better fit the requirements of the dataset. Hyperparameters are tuned before the model is trained and tested.

Here, to minimize the computational time, the tuning of the hyperparameters was performed on the Median FoI only (MedFoI approach) and the list of hyperparameters tuned is presented in Table c.

*Table c: Hyperparameters tested for the Boosted Regression Trees (BRT) and Random Forest (RF) as described in the mlr3 framework (programming environment built on R to simplify the use of Machine Learning methods (126))*

| mlr3 name | generic name | function | default value | values tested |
|---|---|---|---|---|
| *BRT:* | | | | |
| interaction.depth | Depth of trees | Number of splits in each tree; also called the tree complexity., e.g. allows interactions between factors if set to 2. | 1 | 1,2 |
| n.minobsinnode | Minimum node size | Minimal number of observations in terminal nodes | 10 | 5-100 |
| n.trees | Number of trees | The total number of trees that will be included in the model | 100 | 1000-7000 |
| shrinkage | learning rate | Defines the pace at which the algorithm moves on the error surface | 0.001 | 0.001-0.01 |
| train.fraction | Training fraction | Defines the proportion of the dataset that is used for fitting; the remaining proportion is used to evaluate the performance of the model | 1 | 0.90-0.50 |
| *RF:* | | | | |
| ntree | Number of trees | Number of trees built | 500 | 5-1000 |
| nodesize | Minimum node size | Minimal number of observations in nodes | 5 | 1-100 |

BRT: Boosted Regression Trees; RF: Random Forest methods

The tuning process was applied in two steps. Firstly, a random search of 100 evaluations was implemented with a large set of values tested for each hyperparameter (Table c). Second, the search window was narrowed around the best values found at the first step and a grid search was applied with 100 further evaluations.

### iii. Results of tuning

Five and two hyperparameters were tuned for, respectively, BRT and RF models. Table d presents the hyperparameter values tested at each step of the process.

*Table d: Tuning process for ML frameworks' hyperparameters*

| R name | abbreviation | default value | values tested at step 1 | values tested at step 2 |
|---|---|---|---|---|
| *BRT*: | | | | |
| interaction.depth | dep | 1 | 1,2 | 2 |
| n.minobsinnode | n.ob | 10 | 5-100 | 5--25 |
| n.trees | n.tr | 100 | 1000-7000 | 3000-5000 |
| shrinkage | shr | 0.001 | 0.001-0.01 | 0.004-0.009 |
| train.fraction | t.fr | 1 | 0.90-0.50 | 0.75-0.90 |
| *RF:* | | | | |
| ntree | | 500 | 5-2000 | 5-100 |
| nodesize | | 5 | 1-100 | 1-30 |

The hyperparameters obtained from the tuning of the models (n.ob=5, n.tr=4556, shr=0.01, dep=2 and t.fr=0.82%) were used to obtain the final results.

For RF, the cross validation $R^2$ (CV $R^2$), calculated on the cross validation set, gradually improved but reached a plateau after 25 trees and 5 nodes size; therefore, these values were used to fit the model. Higher numbers of trees and nodes size led to substantially larger computational cost but with marginal impact on model performance.

*Supp. Method 3: Comparing observations and predictions across serosurveys (extracted from (103))*

> "For each serosurvey, we compared, across years, the median and 95%CI (Confidence Interval) of the predicted FoI against the median and 95%CrI (Credible Interval) of the originally estimated FoI (21) (i.e. the dependent variable or 'observed' FoI).

For each quantile of interest $q_x$ (i.e., median, 2.5%, and 97.5% percentiles, denoted $q_m$, $q_l$ and $q_u$ respectively), we computed a distance between the 'observed' and predicted quantile ($\delta_{q_x}$). This distance was standardised by the interval between the observed median and observed upper (or lower) 95% CrI,

$$\begin{cases} \delta_{q_x} = \frac{q_x(\hat{y}) - q_x(y)}{q_x(y) - q_l(y)} & if\ q_x(\hat{y}) < q_x(y) \\ \delta_{q_x} = \frac{q_x(\hat{y}) - q_x(y)}{q_u(y) - q_x(y)} & if\ q_x(\hat{y}) > q_x(y) \end{cases} \quad \text{(Eq. 3)}$$

When the predicted and 'observed' medians are equal, we expect $\delta_{q_m} = 0$. If the predicted median was equal to the upper (or lower) 95%CrI of the 'observed' FoI values, then we would have $\delta_{q_m} = 1$ ($\delta_{q_m} = -1$).

If the predicted and 'observed' upper (or lower) 95% CI/CrI were equal, then we would expect $\delta_{q_u} = 1$ ($\delta_{q_u} = -1$). A value $\delta_{q_u} = 2$ would indicate that the interval between the median and upper CI in the prediction is twice as wide as the interval between the median and upper CrI in the observations.

The change in the denominator reflects the non-symmetrical nature of the 95%CI.

As it is rescaled, this measure of bias allows an assessment of the predictive ability of our modelling approaches across serosurveys. For each year, we estimated the median and interquartile range in the bias. This was also done by setting.

### *Supp. Method 4: Predictive model for the FoI at the municipal level*

1. Serosurvey Characteristics
   1.1. Setting type

Serosurveys were classified as urban, rural, indigenous, or mixed (urban and rural mixed population) depending on the setting where the serosurveys have been conducted. The settings' definitions are matching Colombian's government definition of urbanicity. Indigenous setting refers to people living in traditional villages.

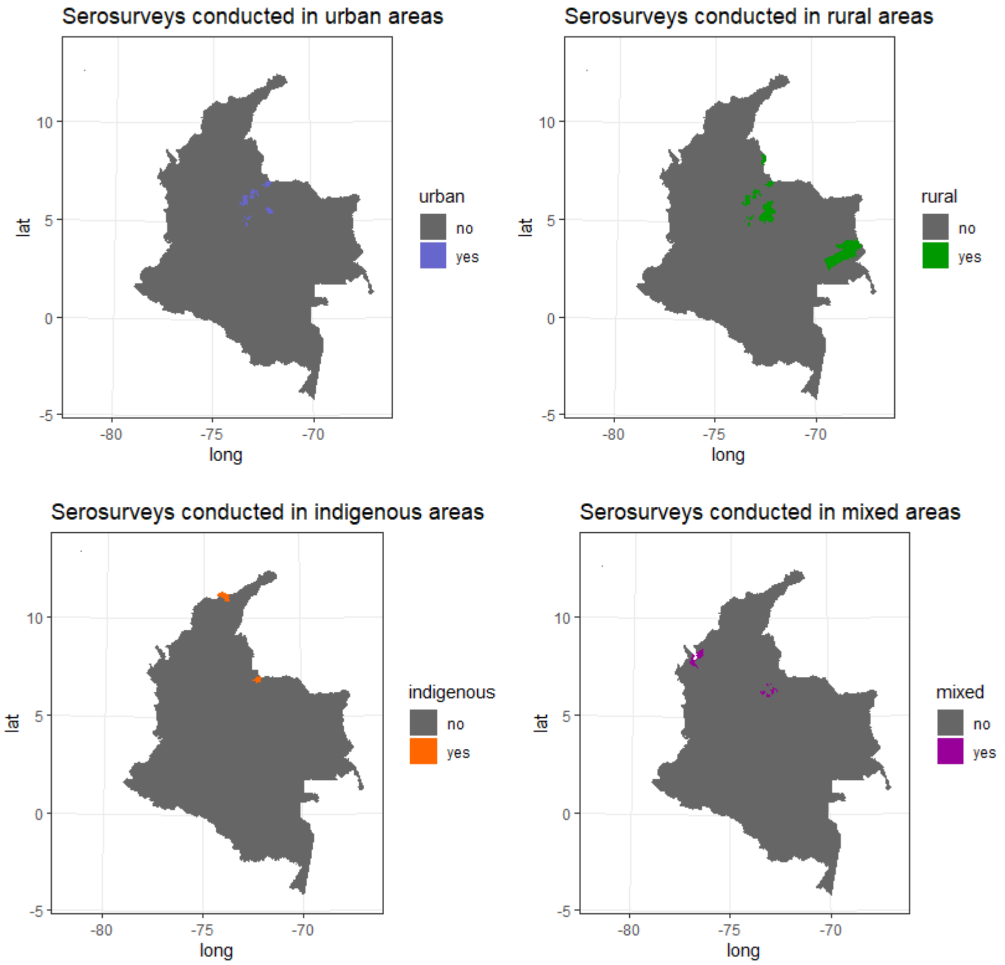

Serosurveys conducted in urban areas
urban
no
yes
Serosurveys conducted in rural areas
rural
no
yes
Serosurveys conducted in indigenous areas
indigenous
no
yes
Serosurveys conducted in mixed areas
mixed
no
yes
lat
long
10
5
0
-5
-80
-75
-70

*Figure 4.1: Spatial distribution of Chagas disease serosurveys conducted in Colombia at ADM2 level, 1980–2014. Upper left panel, municipalities where at least one serosurvey have been conducted in urban settings; upper right panel, idem for rural settings; bottom left panel, idem for indigenous setting and, bottom right idem for mixed setting (including urban and rural populations)*

## 3.3. Year when the serosurvey was conducted

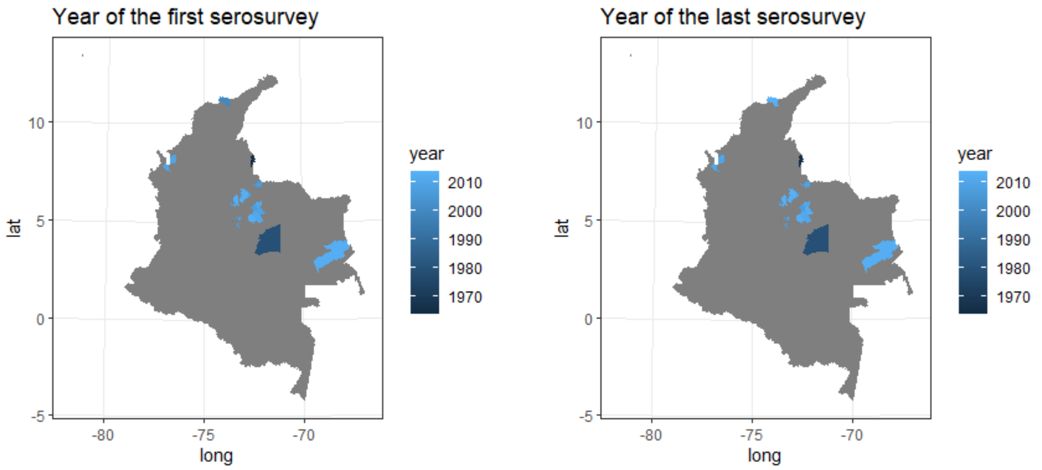


*Figure 4.2: Spatial distribution of the year when the serosurveys were conducted. Left panel, year of the first serosurvey conducted in the municipality; right panel, year when the last serosurvey have been conducted in the municipality.*

2. Environmental predictors
   2.1. Bio_03

Description: Median Isothermality (quantifies how large the day-to-night temperatures oscillate relative to the summer- to-winter (annual) oscillations)

Source: CHELSA (Climatologies at high resolution for the earth's land surface areas) data

Notes: Data processing is described here: https://www.nature.com/articles/sdata2017122. The Raw data is a raster with high spatial resolution, the median in each municipality has been extracted to create the dataset at the municipality level.

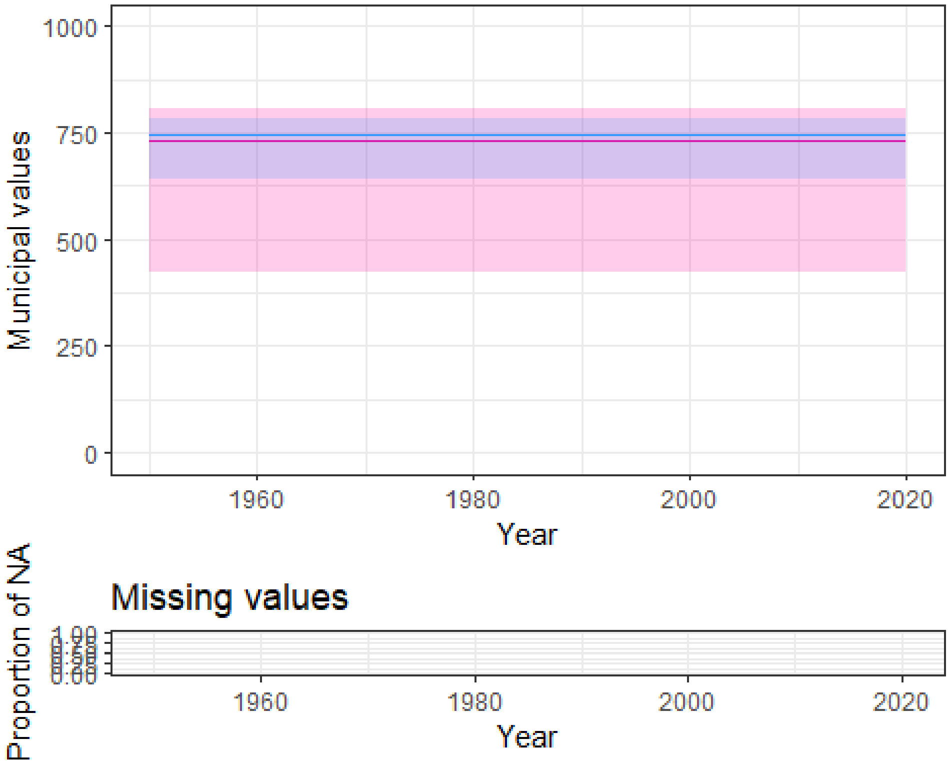


*Figure 4.3: Temporal distribution of the median isothermality in the study area (blue) and in entire Colombia (red) between 1950 and 2020. Top panel, municipal median (line) and range (ribbon); bottom panel, the proportion of missing values.*

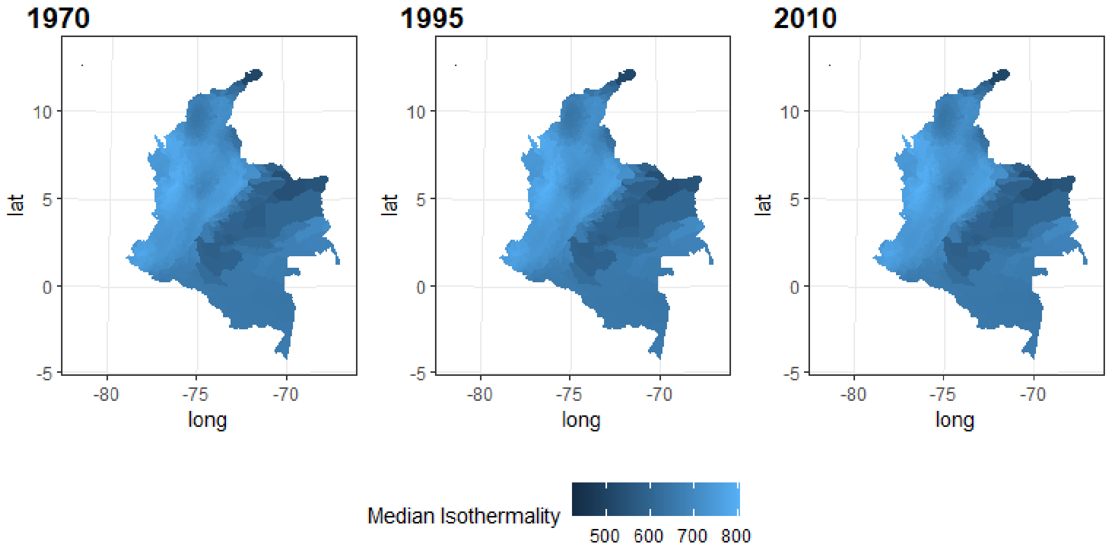


*Figure 4.4: Spatial distribution of the median isothermality at the municipal level for 1970, 1995 and 2010. Missing values appear in grey.*

### 2.2. Bio_06

Description: Median minimum temperature of the coldest month.

Source: CHELSA (Climatologies at high resolution for the earth's land surface areas) data

Notes: Data processing is described here: https://www.nature.com/articles/sdata2017122. The Raw data is a raster with high spatial resolution, the median in each municipality has been extracted to create the dataset at the municipality level.

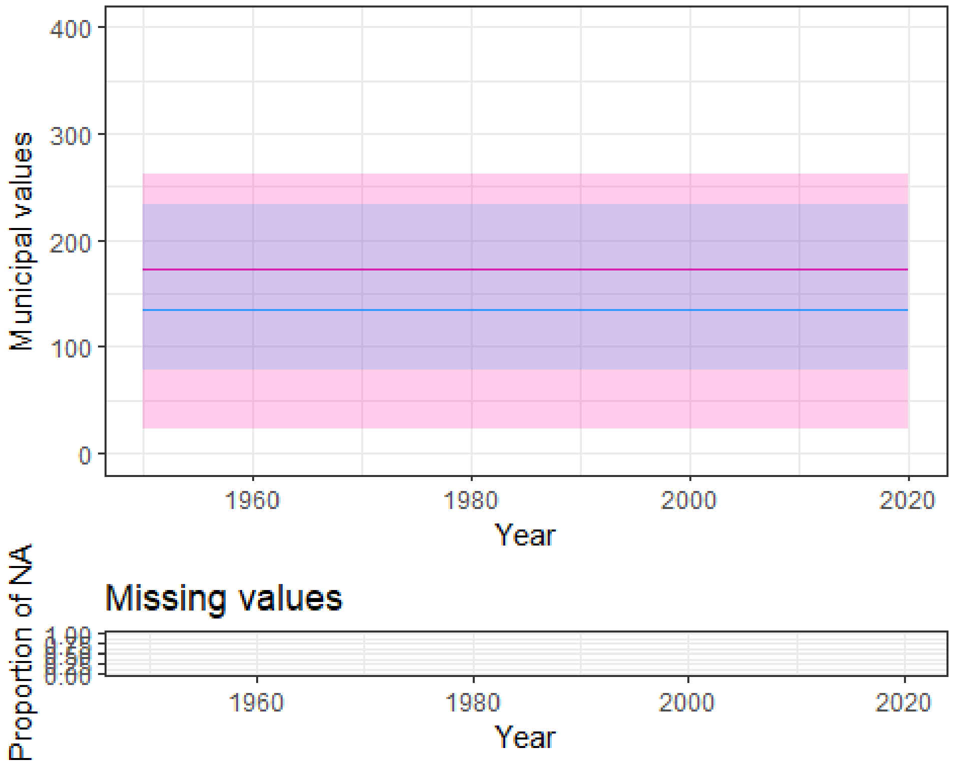


*Figure 4.5: Temporal distribution of the median minimum temperature of the coldest month in the study area (blue) and in the entire Colombia (red) between 1950 and 2020. Top panel, municipal median (line) and range (ribbon); bottom panel, the proportion of missing values.*

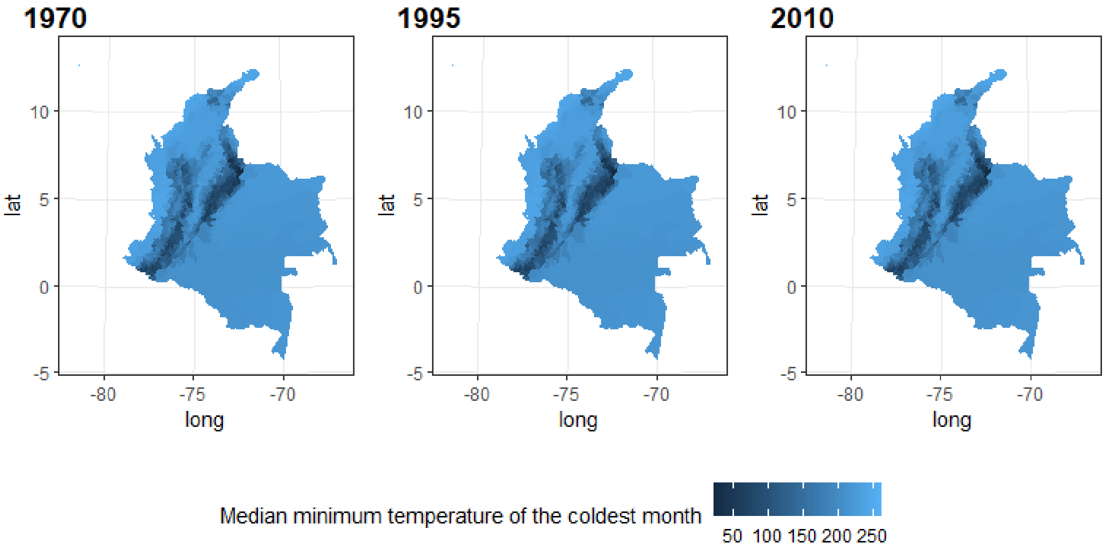


*Figure 4.6: Spatial distribution of the median minimum temperature of the coldest month at the municipal level for 1970, 1995 and 2010. Missing values appear in grey.*

### 2.3. Bio_15

Description: Median seasonality of precipitation

Source: CHELSA (Climatologies at high resolution for the earth's land surface areas) data

Notes: Data processing is described here: https://www.nature.com/articles/sdata2017122. The Raw data is a raster with high spatial resolution, the median in each municipality has been extracted to create the dataset at the municipality level.

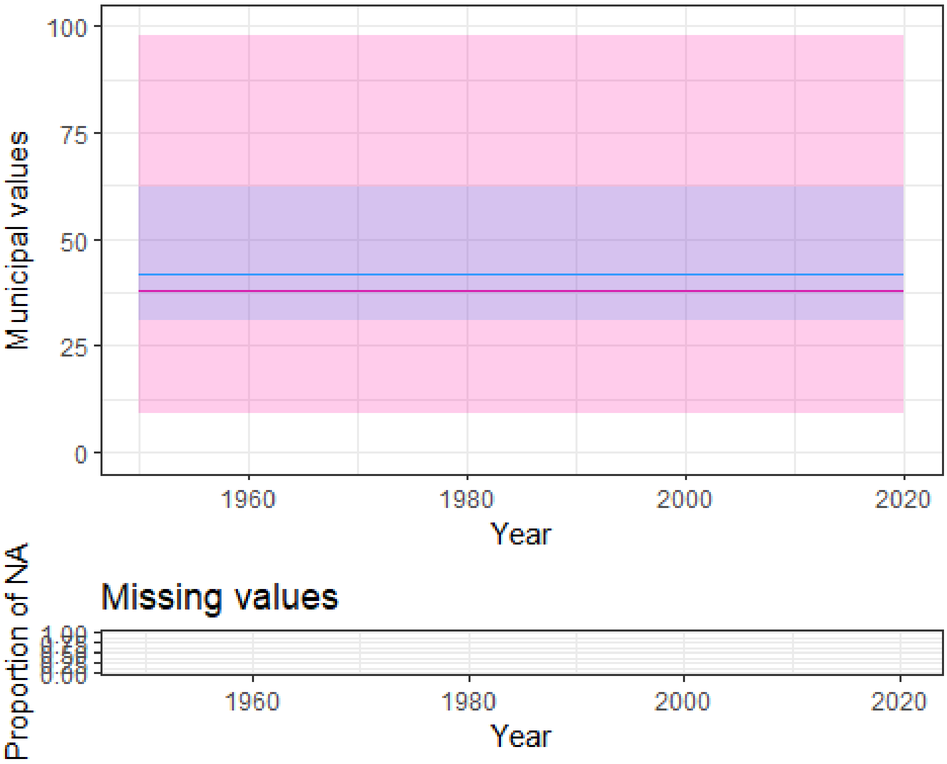


*Figure 4.7: Temporal distribution of the median seasonality of precipitation in the study area (blue) and in entire Colombia (red) between 1950 and 2020. Top panel, municipal median (line) and range (ribbon); bottom panel, the proportion of missing values.*

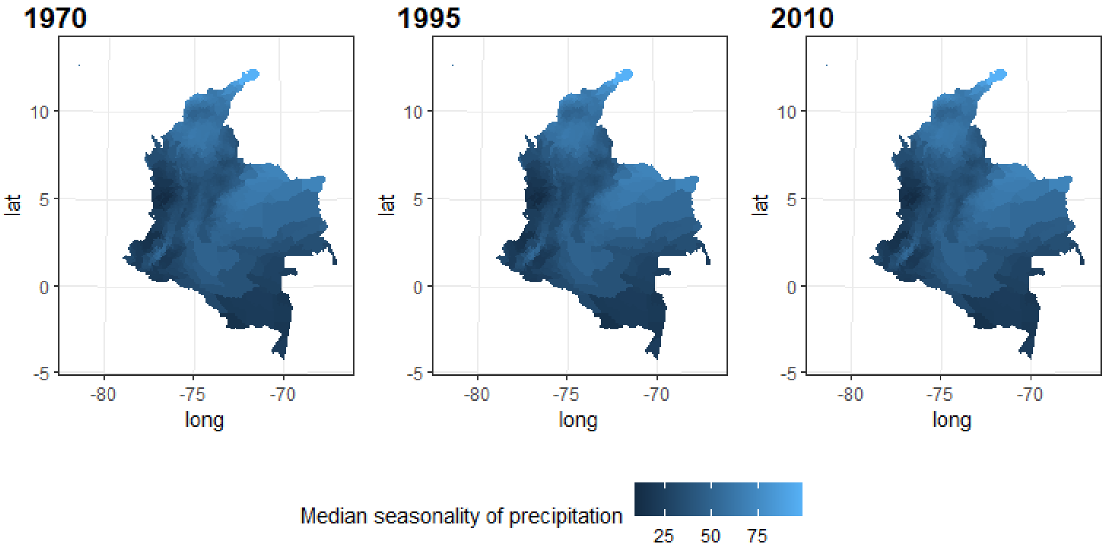


*Figure 4.8: Spatial distribution of the median seasonality of precipitation at the municipal level for 1970, 1995 and 2010. Missing values appear in grey.*

3. Averaged NDVI

Description: Averaged Normalized Difference Vegetation Index - NDVI (LTDR v5 - AVHRR) at municipality level

Source: AidData GeoQuery (Goodman, S., BenYishay, A., Lv, Z., & Runfola, D. (2019). GeoQuery: Integrating HPC systems and public web-based geospatial data tools. Computers & Geosciences, 122, 103-112.).

Notes: The original data have been aggregated at municipality level for 1981-2015 years by AidData GeoQuery (Goodman, S., BenYishay, A., Lv, Z., & Runfola, D. (2019). GeoQuery: Integrating HPC systems and public web-based geospatial data tools. Computers & Geosciences, 122, 103-112.). Original Remote sensing data used by AidData: Yearly value for Normalized Difference Vegetation Index (NDVI). Created using the NASA Long Term Data Record (v5) AVHRR data. In the analyses, the missing values before 1980 were replaced by the values for 1980 and the missing values after 2015 were replaced by the values of 2015.

Aggregation processes: Created by aggregating daily data to monthly by taking the maximum value, then averaging the monthly data to get yearly values. All negative NDVI values were truncated to 0 and saturated pixels were adjusted to the max of the normal NDVI range

(10000). Original source: Pedelty JA, Devadiga S, Masuoka E et al. (2007) Generating a Long-term Land Data Record from the AVHRR and MODIS Instruments. Proceedings of IGARRS 2007, pp. 1021–1025. Institute of Electrical and Electronics Engineers, NY, USA.(http://ltdr.nascom.nasa.gov/ltdr/ltdr.html)

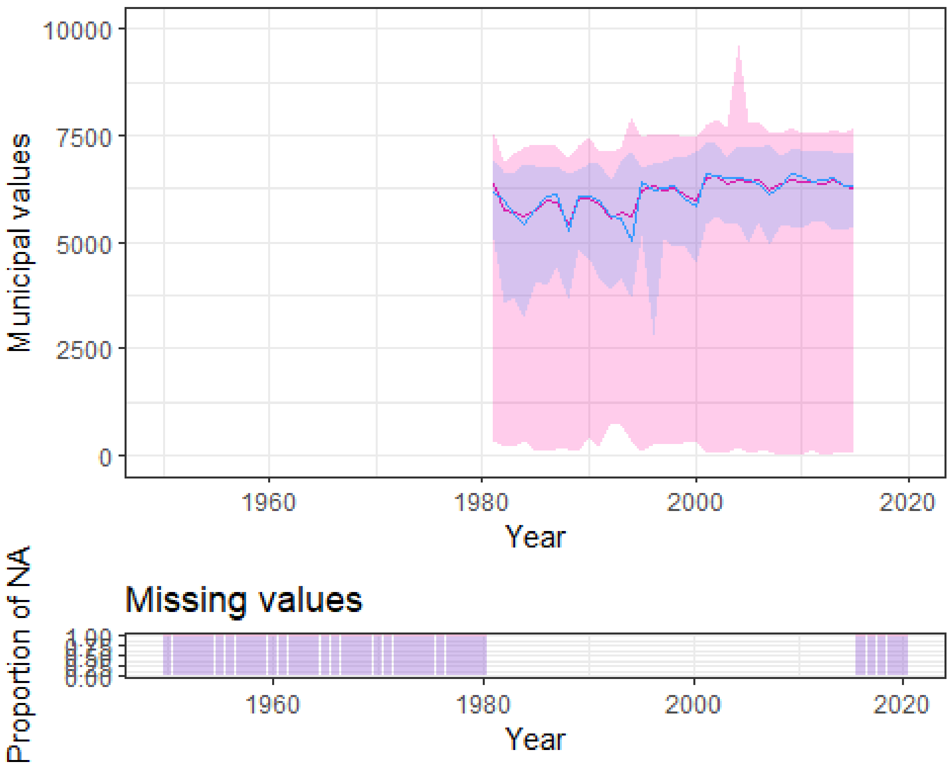


*Figure 4.9: Temporal distribution of the average NDVI in the study area (blue) and in the entire Colombia (red) between 1950 and 2020. Top panel, municipal median (line) and range (ribbon); bottom panel, the proportion of missing values.*

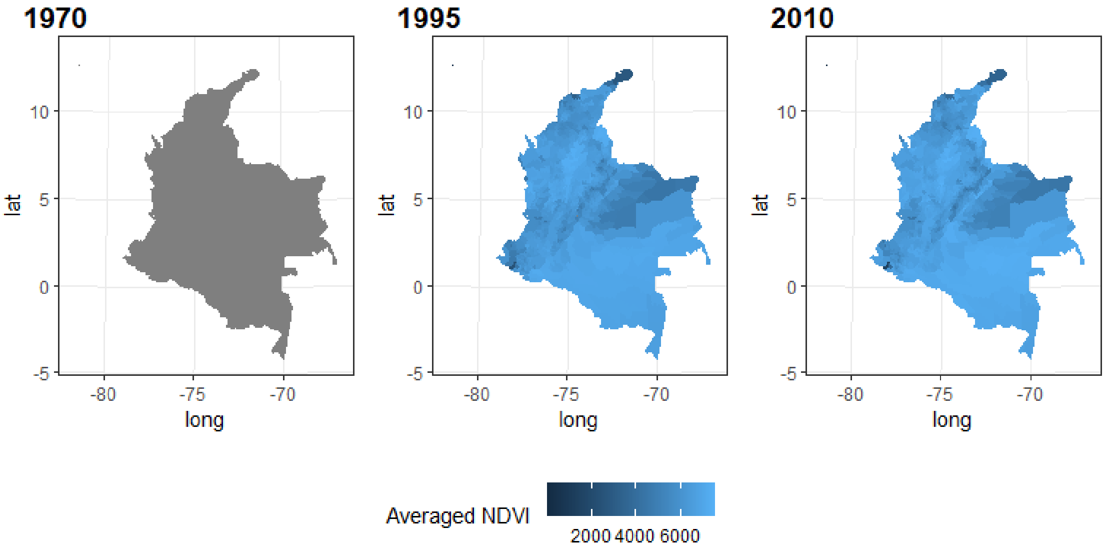


*Figure 4.10: Spatial distribution of the average NDVI at the municipal level for 1970, 1995 and 2010. Missing values appear in grey.*

3. Elevation

Description: Median Elevation

Source: http://www.earthenv.org/topography (Amatulli, G., Domisch, S., Tuanmu, M.-N., Parmentier, B., Ranipeta, A., Malczyk, J., and Jetz, W. (2018) A suite of global, cross-scale topographic variables for environmental and biodiversity modeling. Scientific Data volume 5, Article number: 180040. DOI: doi:10.1038/sdata.2018.40.)

Notes: The Raw data is a raster with high spatial resolution, the median in each municipality has been extracted to create the dataset at the municipality level.

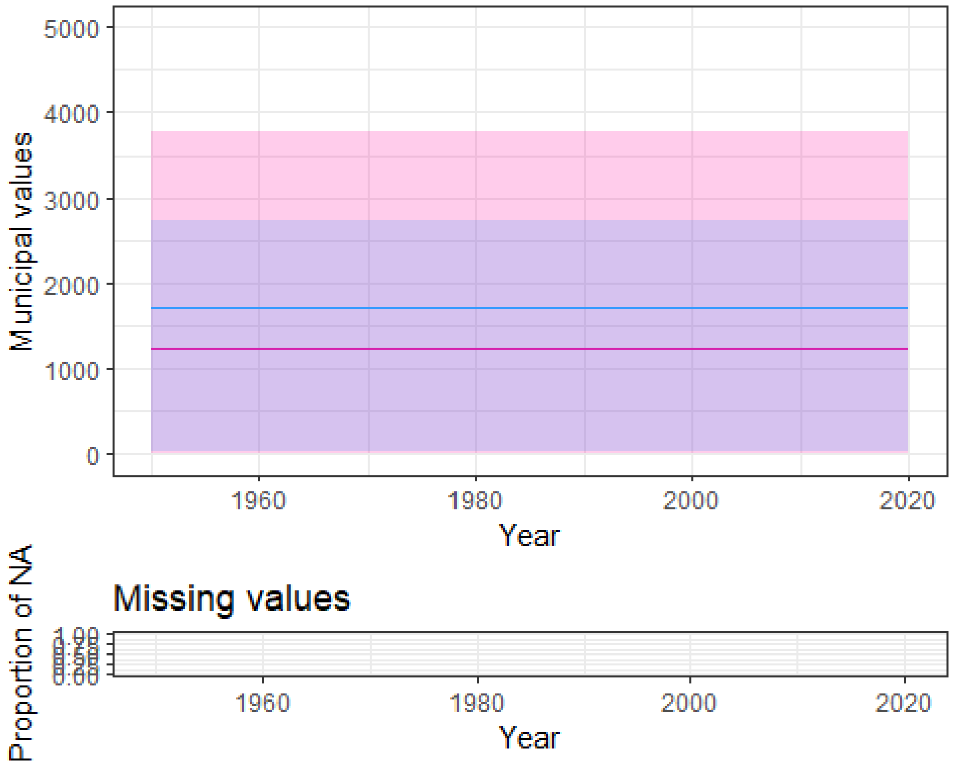


*Figure 4.11: Temporal distribution of the median elevation in the study area (blue) and in the entire Colombia (red) between 1950 and 2020. Top panel, municipal median (line) and range (ribbon); bottom panel, the proportion of missing values.*

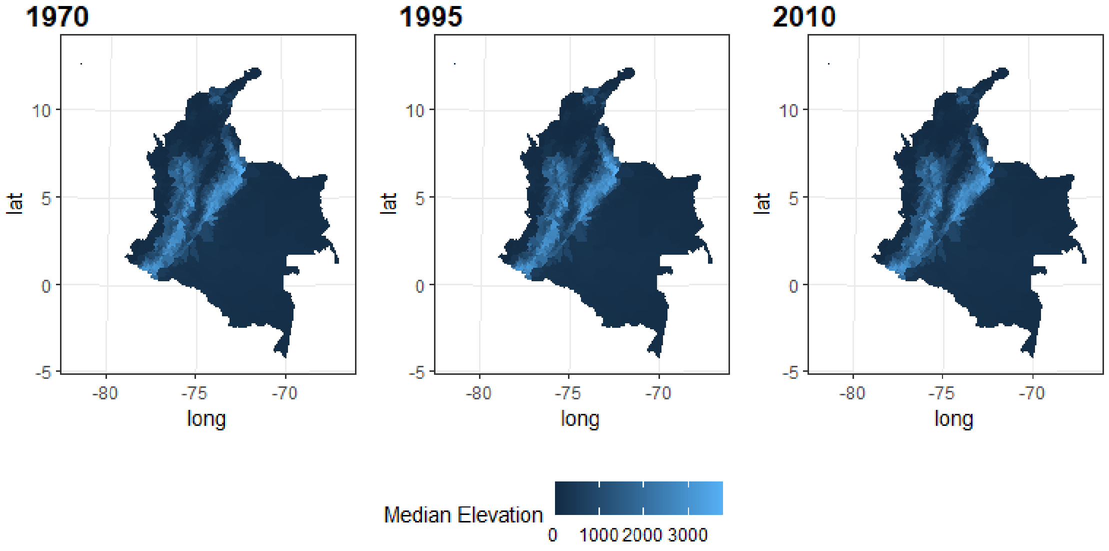


*Figure 4.12: Spatial distribution of the median elevation at the municipal level for 1970, 1995 and 2010. Missing values appear in grey.*

3. Certification year

Description: Year when the municipality has been certified free of domiciliated vector

Source: "Material para la homologación de la validación de municipios endémicos con interrupción de la transmisión vectorial domiciliaria de T.cruzi".

Notes: In the analyses, municipalities that have never been certified have received the value 1900 to avoid having missing values.

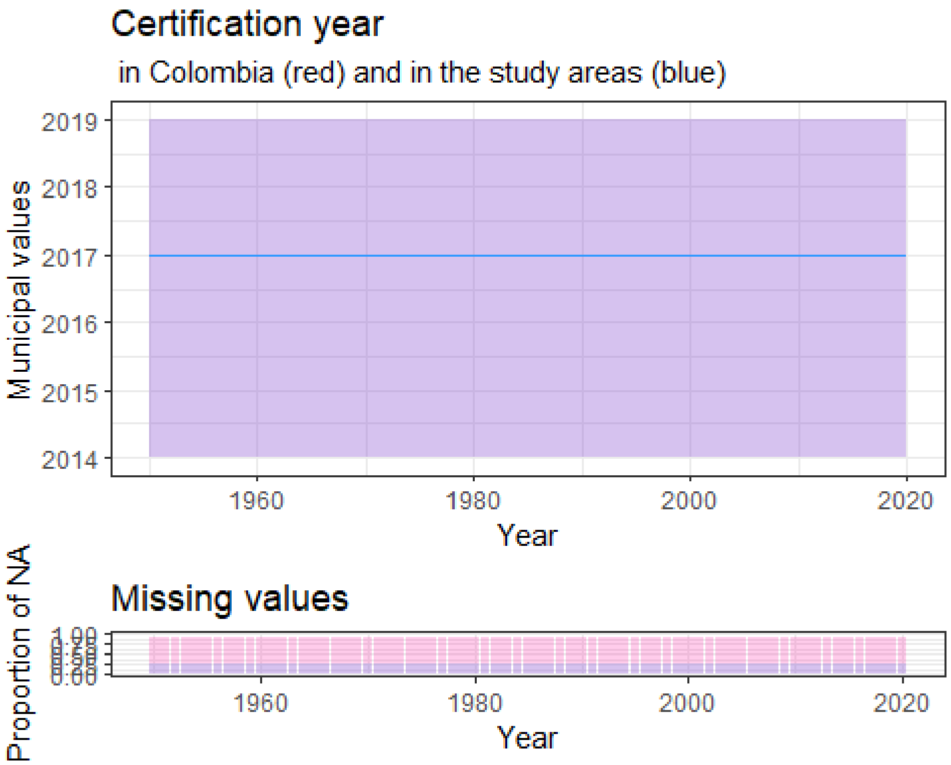


*Figure 4.13: Temporal distribution of the certification year in the study area (blue) and in the entire Colombia (red) between 1950 and 2020. Top panel, municipal median (line) and range (ribbon); bottom panel, the proportion of missing values.*

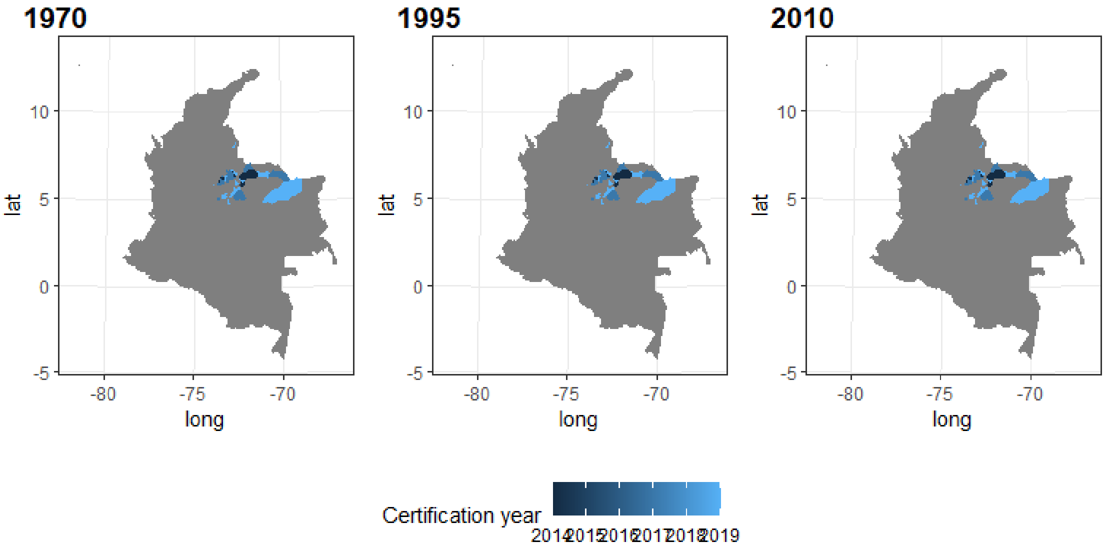


*Figure 4.14: Spatial distribution of the certification year at the municipal level for 1970, 1995 and 2010. Missing values appear in grey.*

3. Demographic predictors

3.3. Proportion of households with an unfinished floor

Description: Median percentage of households in the geographic unit that have a dirt/unfinished floor

Source: IPUMS International.

Notes: This variable is derived from Census data. The Raw data is a raster with high spatial resolution, the median in each municipality has been extracted to create the dataset at the municipality level. In Colombia, censuses were organised in 1973, 1985, 1993 and 2005. Thus, years from 1950 to 1979 received data from 1973; years from 1980 to 1989 received data from 1985; years from 1990 to 1999 received data from 1993 and years from 2000 to 2020 received data from 2005.

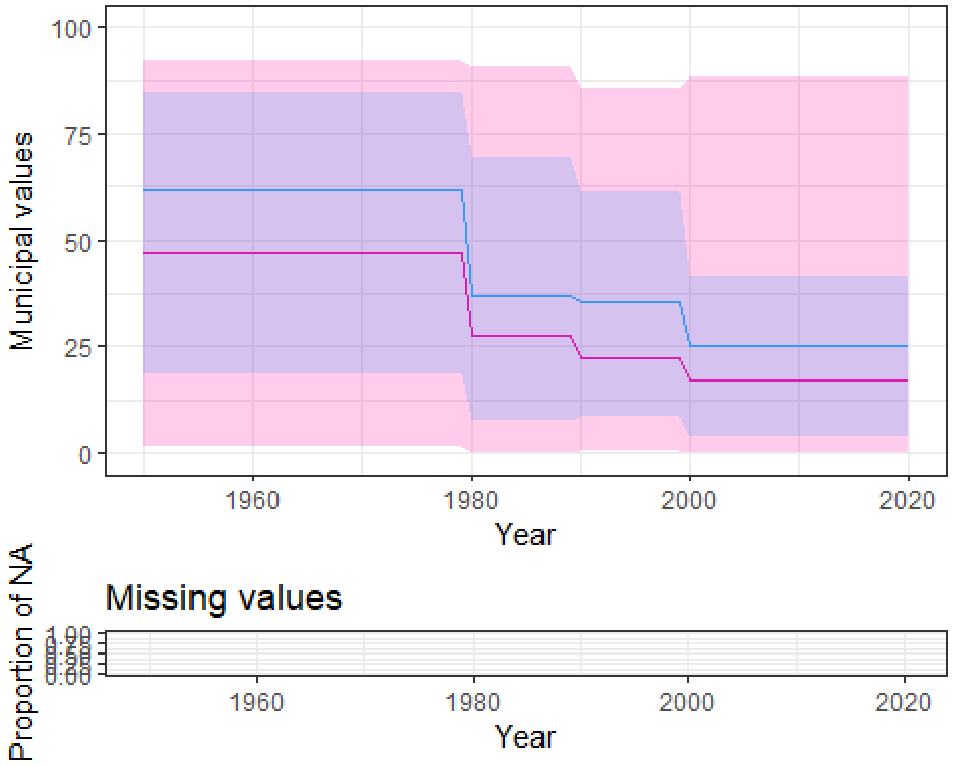


*Figure 4.15: Temporal distribution of the median proportion of households with an unfinished floor in the study area (blue) and the entire Colombia (red) between 1950 and 2020. Top panel, municipal median (line) and range (ribbon); bottom panel, the proportion of missing values.*

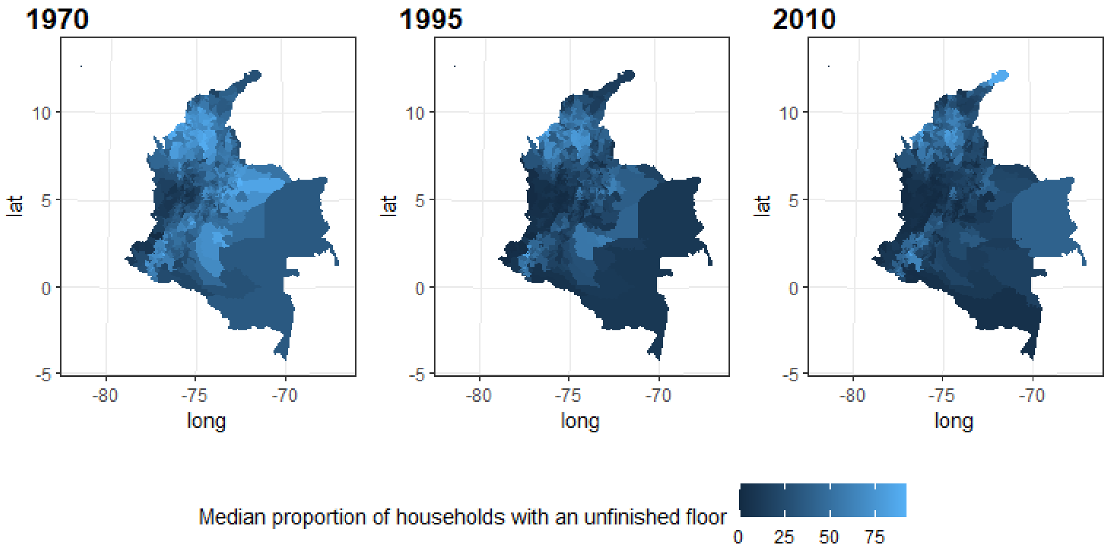


*Figure 4.16: Spatial distribution of the median proportion of households with an unfinished floor at the municipal level for 1970, 1995 and 2010. Missing values appear in grey.*

### 3.3. Population size

Description: Estimated size of the population.

Source: DANE

Notes: projections realised for the period 1985-2020. In the analyses, missing values before 1980 were replaced by the values in 1980.

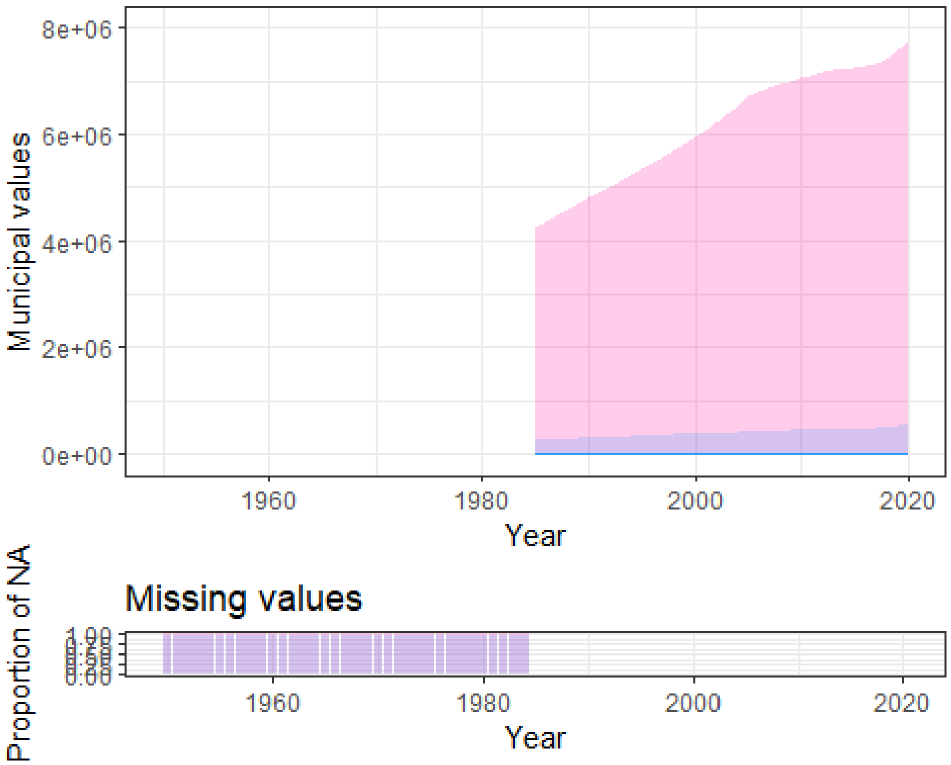


*Figure 4.17: Temporal distribution of the population size in the study area (blue) and in the entire Colombia (red) between 1950 and 2020. Top panel, municipal median (line) and range (ribbon); bottom panel, the proportion of missing values.*

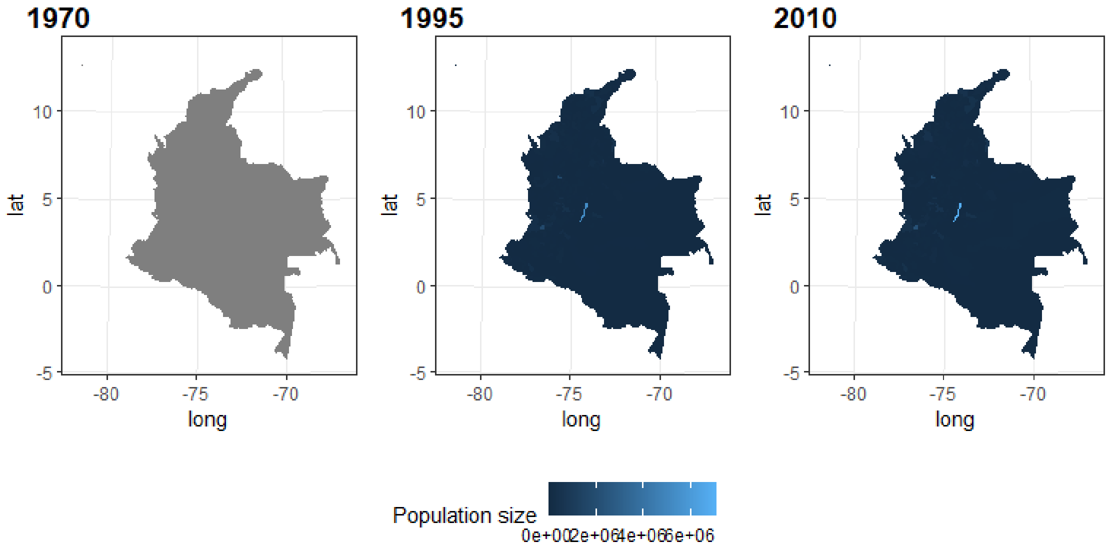


*Figure 4.19: Spatial distribution of the population size at the municipal level for 1970, 1995 and 2010. Missing values appear in grey.*

### 3.3. Proportion of the population living in urban settings

Description: Proportion of the population living in urban settings

Source: DANE

Notes: projections realised for the period 1985-2020. In the analyses, missing values before 1980 were replaced by the values in 1980.

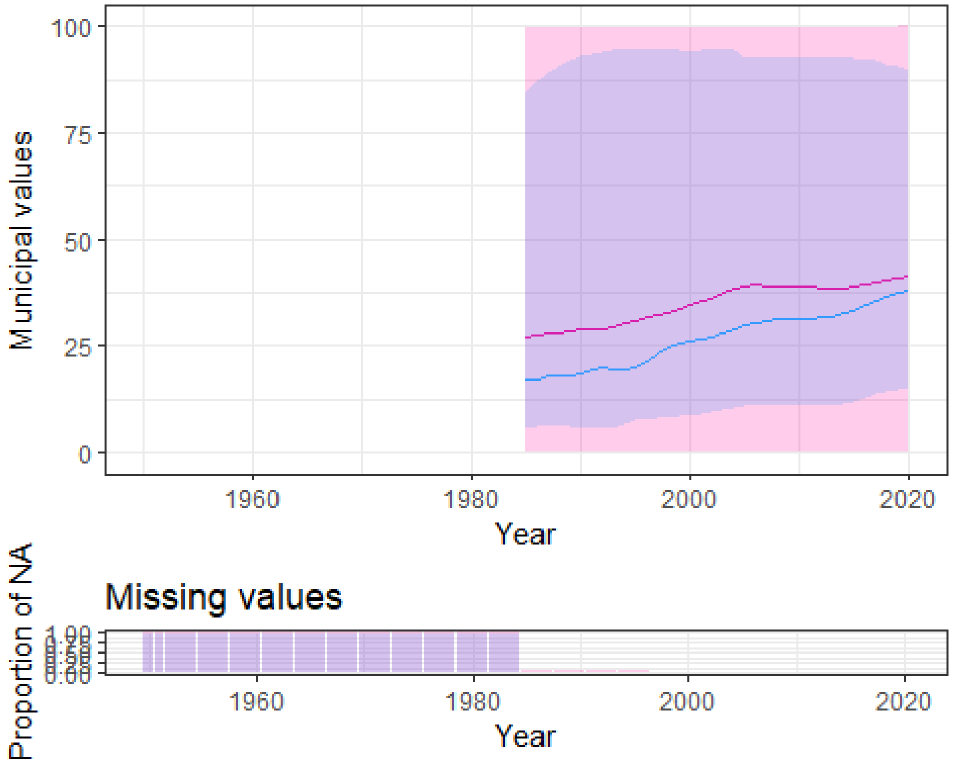


*Figure 4.19: Temporal distribution of the proportion of the population living in urban settings in the study area (blue) and the entire Colombia (red) between 1950 and 2020. Top panel, municipal median (line) and range (ribbon); bottom panel, the proportion of missing values.*

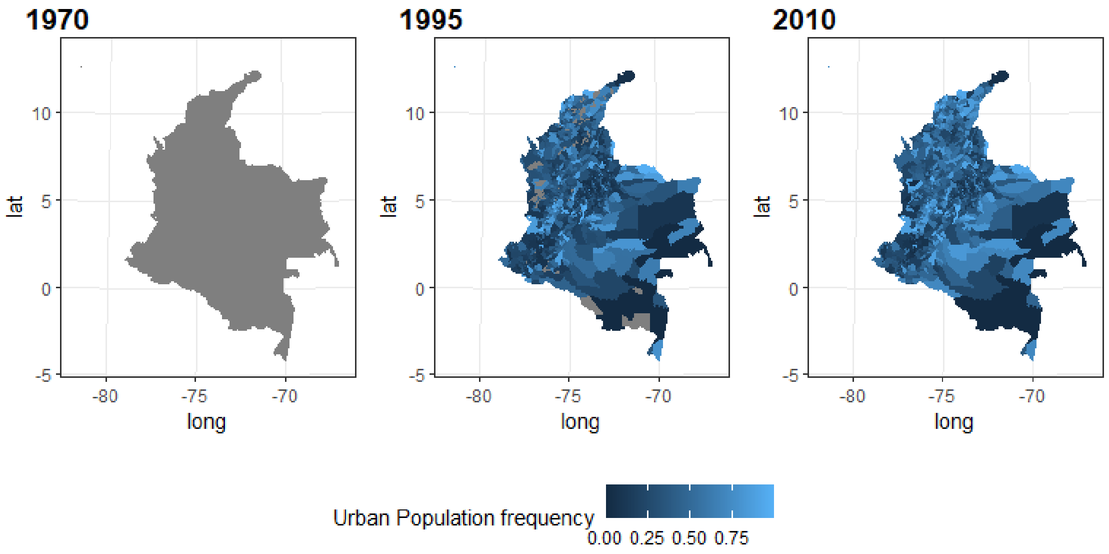


*Figure 4.20: Spatial distribution of the proportion of the population living in urban settings at the municipal level for 1970, 1995 and 2010. Missing values appear in grey.*

### *Supp. Method 5: Models used to estimate the FoI across Colombia*

In order to integrate the uncertainty on the response variable, generation and assessment of the model predictions relied on the full posterior distribution of the FoI estimated with the catalytic model. The performance indicator used is based on the central tendency, the $R^2$, and the amount of the "observed" and predicted distributions that are overlapping. The overlap has been calculated using the overlap function in the overlapping R-package (123).

Using the mlr3 framework in R Studio (126,134), a nested resampling strategy has been applied to tune model hyperparameters and realise a spatial resampling. The two hyperparameters tuned were the number of trees with values tested between 5 and 50 and the final node size with values tested between 2 and 10. The number of trees is a parameter that limits the number of trees that will be built by the model and thus has to be limited to avoid overfitting. The final node size also impacts overfitting as it defines the minimal number of observations that can be let in a final node, i.e. that will not be split again. If the number of trees is large and the final node size is small, the model can overfit the data.

Another way to limit overfitting is using a resampling strategy. Here, a 10 folds spatial resampling has been used, meaning the study area have been divided into 10 with each of them having the same number of "observations". At each resampling iteration, one fold is excluded from fitting and used to calculate the Resample $R^2$. A Training $R^2$ can be calculated on the training set and having these two $R^2$ of equivalent value show a limited overfitting and limited spatial correlation.

On top of the nested resampling process, cross-validation has been realised by leaving half of the dataset aside. $R^2$ calculated on the cross-validation set are used to assess overfitting in the model.

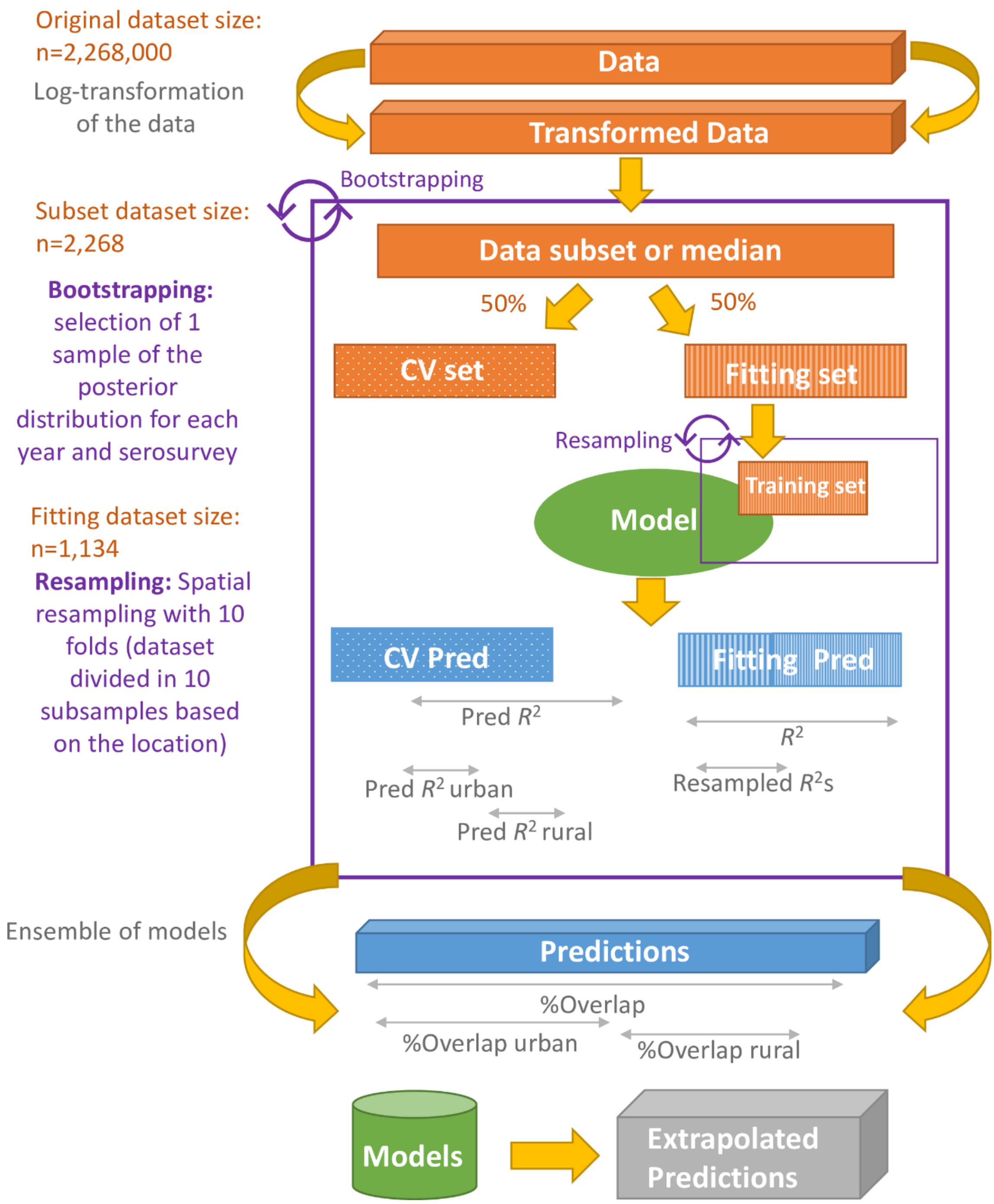


*Figure 5.1: Description of the modelling workflow for the Random Forest (RF) model. CV denotes cross-validation; Pred R2 urban and Pred R2 rural denote urban- and rural-specific predictive R2 values that were estimated based on the urban/rural data from the CV set; %Overlap indicates the proportion of the 'observed' and predicted distributions that overlap, assessed over all settings and for urban and rural settings separately.*

*Supp. Method 6: Models used to estimate the Burden of disease: Extracted from Z. M. Cucunubá, "Modelling the epidemiology and healthcare burden of Chagas disease in Colombia," Imperial College of London. (2017)*

"<u>Compartmental Disease Burden Model</u>

Based on the characterisation of the population susceptible and infected by T. cruzi [...], a progression model for Chagas disease was developed. The model uses information and parameters estimated to compare exposure to T. cruzi infection vs. other aetiologies of heart disease to estimate the specific burden due to Chagas disease. The general structure of the model is presented in Figure 5.1 and the parameters in Table 5.3.

Table 5.2. Disease stages included in the burden of Chagas disease model

| States | Description |
|---|---|
| *Sa* | Susceptible (asymptomatic) stage with no heart disease |
| *Sm* | Susceptible stage with mild/moderate heart disease |
| *Ss* | Susceptible stage with severe heart disease |
| *Am* | Acute mild *T. cruzi* infection |
| *As* | Acute severe *T. cruzi* infection |
| *Ci* | Indeterminate chronic *T. cruzi* infection |
| *Cm* | Mild/moderate chronic chagasic cardiomyopathy |
| *Cs* | Severe chronic chagasic cardiomyopathy |

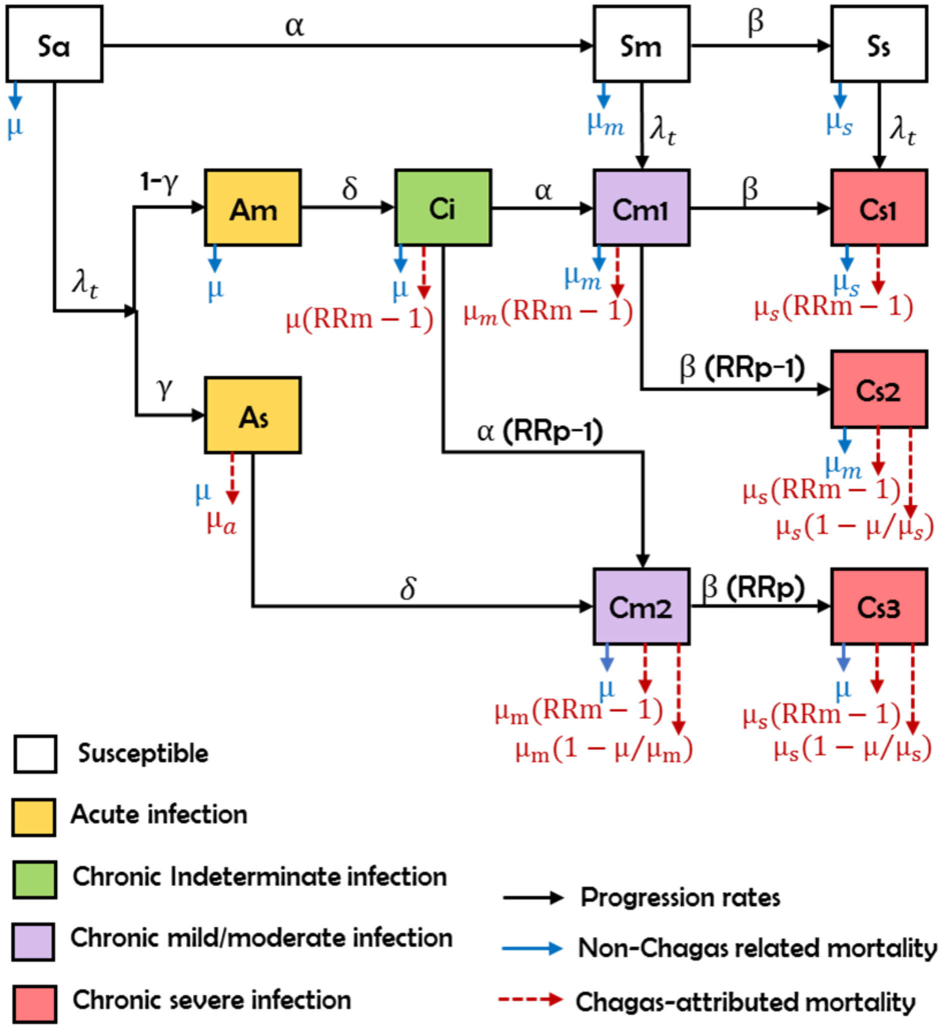


*Figure 5.1. Progression flowchart for the Chagas burden of disease model.*

$\lambda_t$ is the per susceptible rate of infection acquisition (the per susceptible incidence or force-of-infection [..]); $\gamma$ is the proportion of those infected who develop severe disease ($1-\gamma$ is the proportion with mild infection); $\delta$ is the rate of progression from acute mild T. cruzi infection to the indeterminate chronic infection state (same rate applies from acute severe T. cruzi infection to the stage of mild/moderate chronic chagasic cardiomyopathy); $\mu$ is the (background) death rate for susceptible asymptomatic individuals; the corresponding death rates for susceptibles with mild and severe (non-chagasic) CVD are $\mu_m$ and $\mu_s$ respectively (with $\mu_s > \mu_m > \mu$); $RRm$ is the relative risk of mortality and $RRp$ is the relative risk of disease progression […]. Table 5.3 presents the parameter notation, definitions, units and sources.

Table 5.3. State variables and parameters of the burden of Chagas disease model

| | Definition | Median | Lower | Upper | Units | Source |
|---|---|---|---|---|---|---|
| $\lambda_t$ | Force-of-Infection | | | | year$^{-1}$ | (8) |
| $\mu$ | Background per capita death rate (country specific) | 0.018 | 0.017 | 0.019 | year$^{-1}$ | (181) |
| $CFR_{As}$ [1] | Case fatality ratio in severe acute phase | 0.07 | 0.03 | 0.138 | proportion | (182) |
| $\mu_a$ | Mortality rate of severe acute phase | 0.93 | 0.36 | 2.06 | year$^{-1}$ | Estimated* |
| $\mu_m$ | Mortality rate of mild/ moderate chronic CVD | 0.06 | 0.04 | 0.09 | year$^{-1}$ | (8) |
| $\mu_s$ | Mortality rate of severe chronic CVD | 0.28 | 0.19 | 0.36 | year$^{-1}$ | (8) |
| $RRm$ | Relative risk of mortality | 1.74 | 1.49 | 2.03 | ratio | (8) |
| $\mu_m\left(1-\frac{\mu}{\mu_m}\right)$ | Excess mortality (relative to background) due to moderate CVD | | | | | |
| $\mu_s\left(1-\frac{\mu}{\mu_s}\right)$ | Excess mortality (relative to background) due to severe CVD | | | | | |
| $\mu_s\left(1-\frac{\mu_m}{\mu_s}\right)$ | Excess mortality (relative to moderate) due to severe CVD | | | | | |
| $\gamma$ | Proportion of severe acute cases | 0.01 | 0.005 | 0.02 | proportion | (182) |
| $\varepsilon = 1/\delta$ | Duration of acute phase | 4 | 2 | 8 | weeks | (182) |
| $\delta$ | Progression rate from acute (mild or severe) phase | 13 | 6.5 | 26 | year$^{-1}$ | [14] |
| $\alpha$ | Progression rate from indeterminate to determinate cardiomyopathy | 0.008 | 0.001 | 0.01 | year$^{-1}$ | (8) |
| $\beta$ | Progression rate from mild to severe cardiomyopathy | 0.01 | 0.001 | 0.04 | year$^{-1}$ | (8) |
| $RRp$ | Relative risk of disease progression due to Chagas | 4.39 | 2.63 | 7.33 | ratio | (8) |

CVD: cardiovascular disease; NA: Not applicable. * $\mu_a = \left[CFR_{As}(\delta+\mu)-\mu\right]/(1-CFR_{As})$

[1] While mortality rate, $\mu_a$, is unknown, we found estimate of the case fatality ratio for the acute severe stage of the disease, i.e. $CFR_{AS}$. By definition $CFR_{AS}$ is the proportion of individual that dies, at a rate $\mu_a + \mu$, without progressing, at rate $\delta$, to the mild stage. So we have: $CFR_{AS} = \dfrac{\mu+\mu_a}{\mu+\mu_a+\delta}$, therefore $\mu_a = \dfrac{CFR_{AS}(\mu+\delta)-\mu}{1-CFR_{AS}}$

| Stage | Equations | |
|---|---|---|
| Susceptible | $\dfrac{\partial Sa}{\partial t}+\dfrac{\partial Sa}{\partial a}=-(\lambda_t+\mu+\alpha)\,Sa(t,a)$ | (5.1) |
| | $\dfrac{\partial Sm}{\partial t}+\dfrac{\partial Sm}{da}=\alpha\,Sa(t,a)-(\lambda_t+\mu_m+\beta)\,Sm(t,a)$ | (5.2) |
| | $\dfrac{\partial Ss}{\partial t}+\dfrac{\partial Ss}{\partial a}=\beta\,Sm(t,a)-(\lambda_t+\mu_s)\,Ss(t,a)$ | (5.3) |
| Acute | $\dfrac{\partial Am}{\partial t}+\dfrac{\partial Am}{\partial a}=[(1-\gamma)\lambda_t]\,Sa(t,a)-(\delta+\mu)\,Am(t,a)$ | (5.4) |
| | $\dfrac{\partial As}{\partial t}+\dfrac{\partial As}{\partial a}=\gamma\,\lambda_t\,Sa(t,a)-(\delta+\mu+\mu_a)\,As(t,a)$ | (5.5) |
| *Chronic indeterminate* | $\dfrac{\partial Ci}{\partial t}+\dfrac{\partial Ci}{\partial a}=\delta\,As(t,a)-(\mu RRm+\alpha RRp)\,Ci(t,a)$ | (5.6) |
| *Chronic mild* | $\dfrac{\partial Cm1}{\partial t}+\dfrac{\partial Cm1}{\partial a}=\alpha Ci(t,a)+\lambda_t\,Sm(t,a)-(\mu_m RRm+\beta RRp)Cm1(t,a)$ | (5.7) |
| | $\dfrac{\partial Cm2}{\partial t}+\dfrac{\partial Cm2}{\partial a}=\delta As(t,a)+[\alpha(RRp-1)]Ci(t,a)$ <br> $-(\mu_m RRm+\beta RRp)Cm2(t,a)$ | (5.8) |
| *Chronic severe* | $\dfrac{\partial Cs1}{\partial t}+\dfrac{\partial Cs1}{\partial a}=\beta\,Cm1(t,a)+\lambda_t Ss(t,a)-(\mu_s RRm)Cs1(t,a)$ | (5.9) |
| | $\dfrac{\partial Cs2}{\partial t}+\dfrac{\partial Cs2}{\partial a}=\beta(RRp-1)Cm1(t,a)-(\mu_s RRm)\,Cs2(t,a)$ | (5.10) |
| | $\dfrac{\partial Cs3}{\partial t}+\dfrac{\partial Cs3}{\partial a}=\beta\,RRp\,Cm2(t,a)-(\mu_s RRm)\,Cs3(t,a)$ | (5.11) |
| Deaths caused by CD | $Da(t)_{direct}=\mu_a\int_a As(t,a)\,da$ | (5.12) |
| | $Di(t)_{direct}=\mu(RRm-1)\int_a Ci(t,a)\,da$ | (5.13) |
| | $Dm(t)_{direct}=\mu_m(RRm-1)\left\{\int_a Cm1(t,a)\,da+\int_a Cm2(t,a)\,da\right\}$ | (5.14) |
| | $Ds(t)_{direct}=\mu_s(RRm-1)\left\{\int_a Cs1(t,a)\,da+\int_a Cs2(t,a)\,da+\int_a Cs3(t,a)\,da\right\}$ | (5.15) |
| Deaths not caused by CD | $Dm(t)_{indirect}=\mu_m\left(1-\dfrac{\mu}{\mu_m}\right)\int_a Cm2(t,a)\,da$ | (5.16) |

| | | |
|---|---|---|
| | $Ds(t)_{indirect} = \mu_s\left(1-\frac{\mu_m}{\mu_s}\right)\int_a Cs2(t,a)\,da + \mu_s\left(1-\frac{\mu}{\mu_s}\right)\int_a Cs2(t,a)\,da$ | (5.17 ) |
| Deaths due to other chronic cardiovascular diseases | $Dh(t) = \mu_m\left\{\int_a Sm(t,a)\,da + \int_a Cm1(t,a)\,da\right\} + \mu_s\left\{\int_a Ss(t,a)\,da + \int_a Cs1(t,a)\,da\right\}$ | (5.18 ) |

″