8-1-2020

Novel evolutionary algorithm identifies interactions driving infestation of triatoma dimidiata, a chagas disease vector

John P. Hanley
University of Vermont

Donna M. Rizzo
University of Vermont

Lori Stevens
University of Vermont

Sara Helms Cahan
University of Vermont

Patricia L. Dorn
Loyola University New Orleans

See next page for additional authors

Recommended Citation

Hanley JP, Rizzo DM, Stevens L, Helms Cahan S, Dorn PL, Morrissey LA, Rodas AG, Orantes LC, Monroy C. Novel Evolutionary Algorithm Identifies Interactions Driving Infestation of Triatoma dimidiata, a Chagas Disease Vector. The American Journal of Tropical Medicine and Hygiene. 2020 Jun 8:tpmd180733.

Am. J. Trop. Med. Hyg., 103(2), 2020, pp. 735–744
doi:10.4269/ajtmh.18-0733

Novel Evolutionary Algorithm Identifies Interactions Driving Infestation of *Triatoma dimidiata*, a Chagas Disease Vector

John P. Hanley,[1]* Donna M. Rizzo,[1] Lori Stevens,[2] Sara Helms Cahan,[2] Patricia L. Dorn,[3] Leslie A. Morrissey,[4]† Antonieta Guadalupe Rodas,[5] Lucia C. Orantes,[4] and Carlota Monroy[5]

[1]*Department of Civil and Environmental Engineering, University of Vermont, Burlington, Vermont;* [2]*Department of Biology, University of Vermont, Burlington, Vermont;* [3]*Department of Biological Sciences, Loyola University New Orleans, New Orleans, Louisiana;* [4]*Rubenstein School of Environment and Natural Resources, University of Vermont, Burlington, Vermont;* [5]*Laboratorio de Entomología Aplicada y Parasitología, Escuela de Biología, Universidad de San Carlos de Guatemala, Ciudad de Guatemala, Guatemala*

Abstract. Chagas disease is a lethal, neglected tropical disease. Unfortunately, aggressive insecticide-spraying campaigns have not been able to eliminate domestic infestation of *Triatoma dimidiata*, the native vector in Guatemala. To target interventions toward houses most at risk of infestation, comprehensive socioeconomic and entomologic surveys were conducted in two towns in Jutiapa, Guatemala. Given the exhaustively large search space associated with combinations of risk factors, traditional statistics are limited in their ability to discover risk factor interactions. Two recently developed statistical evolutionary algorithms, specifically designed to accommodate risk factor interactions and heterogeneity, were applied to this large combinatorial search space and used in tandem to identify sets of risk factor combinations associated with infestation. The optimal model includes 10 risk factors in what is known as a third-order disjunctive normal form (i.e., infested households have chicken coops AND deteriorated bedroom walls OR an accumulation of objects AND dirt floors AND total number of occupants ≥ 5 AND years of electricity ≥ 5 OR poor hygienic condition ratings AND adobe walls AND deteriorated walls AND dogs). Houses with dirt floors and deteriorated walls have been reported previously as risk factors and align well with factors currently targeted by Ecohealth interventions to minimize infestation. However, the tandem evolutionary algorithms also identified two new socioeconomic risk factors (i.e., households having many occupants and years of electricity ≥ 5). Identifying key risk factors may help with the development of new Ecohealth interventions and/or reduce the survey time needed to identify houses most at risk.

INTRODUCTION

The WHO identifies Chagas disease as one of the most difficult neglected tropical diseases to control.[1] Village-scale infestation by the native Guatemalan vector, *Triatoma dimidiata*, can be transiently reduced by using pyrethroid insecticide; however, the vector often rebounds within months of application[2,3] and can achieve pre-application infestation levels within 3 years.[4] Consequently, different strategies are necessary for reducing infestation and lowering disease transmission risk.

In Guatemala, sustainable Ecohealth interventions have been successful in the long-term reduction of *T. dimidiata* infestation.[5–8] These interventions include plastering walls[8]; replacing dirt floors with locally sourced, cement-like materials[5,6]; educational awareness of the disease, including steps villagers can take to reduce their risk[7,8]; and the construction and distancing of chicken coops from the house.[7] Similar to insecticide application, Ecohealth interventions minimize infestation, with the benefit of doing so over longer time frames,[5,6] and often provide other social, economic, and health benefits at a lower overall cost.[7] For example, the installation of cement-like floors using locally sourced material[5,6] can have the dual benefit of preventing *T. dimidiata* infestation and lowering the incidence of hookworm,[9] another important neglected tropical disease in Guatemala.[10]

Limited financial and human resources often lead to interventions that focus on houses most at risk. However, understanding house infestation requires identifying the best combinations of risk factors most likely associated with household infestation. Of particular importance in this work is the ability to identify risk factor interactions, where the combinations of risk factors are associated with a particular outcome (i.e., infestation), but individually have no main effects. The statistical community uses the term "interaction" when referring to this type of multivariate model (i.e., where a single risk factor will not correlate with an outcome unless it is combined with at least one other risk factor).[11] Identifying risk factor interactions is an inherent challenge for complex diseases[12,13] such as Chagas disease.

Studies have suggested that the drivers of *T. dimidiata* infestation may be heterogeneous.[14,15] As a result, another challenge is identifying the combinations of risk factors in these types of real-world heterogeneous systems. For example, in a very broad sense, household infestation may occur when a household is able to provide a food source and shelter for the vector. Heterogeneity occurs because not every infested house offers the same combination of vector food sources and shelter; there are often risk factor combinations that are only associated with a subset of the infested houses. Thus, although multiple models may be necessary to fully predict the outcome, challenges arise because many parametric statistical methods are designed to identify only a single best model.

In addition, socioeconomic and entomologic surveys often limit questions to determine risk factors most associated with an outcome (e.g., infestation). Despite painstaking efforts to reduce the number of questions (whether it be to increase the sample size, reduce time and effort, overcome language barriers, etc.), the search space that results when all risk factors and their range of values are combined quickly becomes larger than is possible to exhaustively search with most statistical techniques. As a result, the aim was to find the most

* Address correspondence to John P. Hanley, Department of Civil and Environmental Engineering, University of Vermont, 180 Chipman Park, Middlebury, VT 05753. E-mail: jhanley@uvm.edu
† Deceased

parsimonious model (i.e., model that uses the fewest risk factors necessary to explain an outcome). Another shortcoming encountered when constructing traditional multivariate models is that ordinal- and continuous-valued risk factors often need to be reduced (e.g., binned a priori by domain experts) to reduce the computational challenges and increase statistical power. Data reduction often adds bias and obfuscates important relationships between risk factors and infestation.[14–17]

A number of studies have used traditional multivariate statistics to identify risk factors for *T. dimidiata* infestation to help prioritize interventions.[14–17] However, these multivariate methods relied on identifying single risk factors (i.e., main effects) a priori and then use this limited subset of the previously identified "significant" risk factors in a multivariate analysis. More advanced methods such as random forests and principal component analysis (PCA) have been used to model other vector-borne disease such as leishmaniasis and West Nile virus.[18–20] However, random forests are not designed to find parsimonious solutions because, by definition, they are combinations of decision trees, where each tree is a complex multivariate model. In addition, decision trees are not designed to find true risk factor interactions because the underlying methodology relies on main effects. When Hanley et al.[21] tested decision trees and random forests on a benchmark problem designed by the machine-learning community to contain both risk factor interactions and heterogeneity, no single decision tree or combination of trees was able to identify the most parsimonious solution (i.e., a solution that comprises only the risk factors, their associated values, and interactions that define the problem). In addition, individual decision trees had poor classification accuracy, whereas random forests were prone to overfitting.[21] Principal component analysis is another popular unsupervised learning technique often used to reduce the dimension of the original (continuous-valued) risk factors, whereby the original variables are transformed into the same number of orthogonal eigenvectors—each linear combinations of the original risk factors. The eigenvectors are ordered by their ability to explain total variance in the dataset. Whereas variance greater than zero is needed to differentiate between outcomes, PCA is not a supervised learning tool; thus, eigenvectors are not designed to be associated with any given outcome (or designed to classify outcomes). Therefore, selecting a subset of eigenvectors based on the amount of total variance explained may likely eliminate vectors important (highly associated) to a given outcome. In addition, the weights associated with select eigenvectors have been used for feature selection, but the weights are only associated with explaining the variance in the dataset for a particular eigenvector (i.e., how much of the total variance is explained by each of the principal components with respect to the total sum). Thus again, PCA is not designed to identify individual risk factors (especially nominal risk factors) most associated with a given outcome (e.g., infestation). Because *T. dimidiata* infestation at its core involves risk factor interactions (i.e., at a minimum, the vector requires a food source and shelter for survival), the development of tools for detecting risk factor interactions is important.

In this work, we assessed the utility of the conjunctive clause evolutionary algorithm (CCEA) and disjunctive normal form evolutionary algorithm (DNFEA), a new supervised, data-mining tool that has been shown to overcome these statistical challenges. The method is designed and was successfully tested on benchmark datasets to accommodate heterogeneity, interactions among risk factors, missing data, and multiple data types potentially associated with the risk of *T. dimidiata* infestation[21]; a brief comparison to logistic regression is presented in the Supplemental Material. We apply the tool to a real-world dataset containing socioeconomic and entomologic survey data designed to understand risk factors for Chagas vector house infestation in two rural communities in Jutiapa, Guatemala. We discuss how these multivariate models might be implemented by domain experts familiar with local stakeholder needs.

METHODS

We use two new evolutionary algorithms, the CCEA and a DNFEA, in tandem to identify risk factors most closely associated with house infestation by the Chagas vector *T. dimidiata*, in two rural Guatemalan villages. The CCEA first searches for combinations of risk factors (i.e., conjunctive clauses [CCs]) that are associated with infestation, whereas the DNFEA further refines the search for heterogeneous combinations of these CCs. The DNFEA takes statistically strong models of infestation that apply to subsets of the infested houses and searches for the best combinations of CCs to cover a larger portion of the search space: in this case, all infested houses.

***Triatoma dimidiata* infestation dataset.** We applied the CEAA and DNFEA frameworks to Chagas disease infestation risk factors using field surveys conducted in the two rural Guatemalan towns, El Carrizal and El Chaperno. Both study sites are located in the dry highlands of the department of Jutiapa, Guatemala, bordering El Salvador (Figure 1). From October 1, 2012 to October 3, 2012 in El Chaperno and February 4, 2013 to February 5, 2013 in El Carrizal, personnel from the University of San Carlos of Guatemala and the Guatemalan Ministry of Health Office, vector-borne disease division conducted socioeconomic and entomologic surveys of 182 and 129 houses, respectively. Informed consent was obtained from all adult participants and parents or legal guardians of minors with ethical clearance from the Ministry of Health in Guatemala, the University of San Carlos of Guatemala bioethics committee, and the Pan American Health Organization.

The household surveys contain 64 risk factors (Supplemental Table S1) that were either previously shown or are newly hypothesized to be associated with house infestation with *T. dimidiata*. These risk factors include vector shelter (e.g., cracks in bedroom walls), vector food sources (e.g., number of dogs), and socioeconomic conditions (e.g., source of household income). Previous studies on *T. dimidiata* focused on the risk factors associated with infestation and did not focus on risk factors associated with colonization nor dispersion.[14–17,22–24] Whereas colonization and dispersion are calculated in one study,[22] to the best of our knowledge, no previous study has calculated the risk factors associated with colonization of *T. dimidiata*. Also, given the challenges of finding live *T. dimidiata*,[25] we would expect a number of false negatives when calculating the domestic infestation and colonization according to the WHO equations for the entomological indicators of Chagas disease control.[26] Therefore, to limit the number

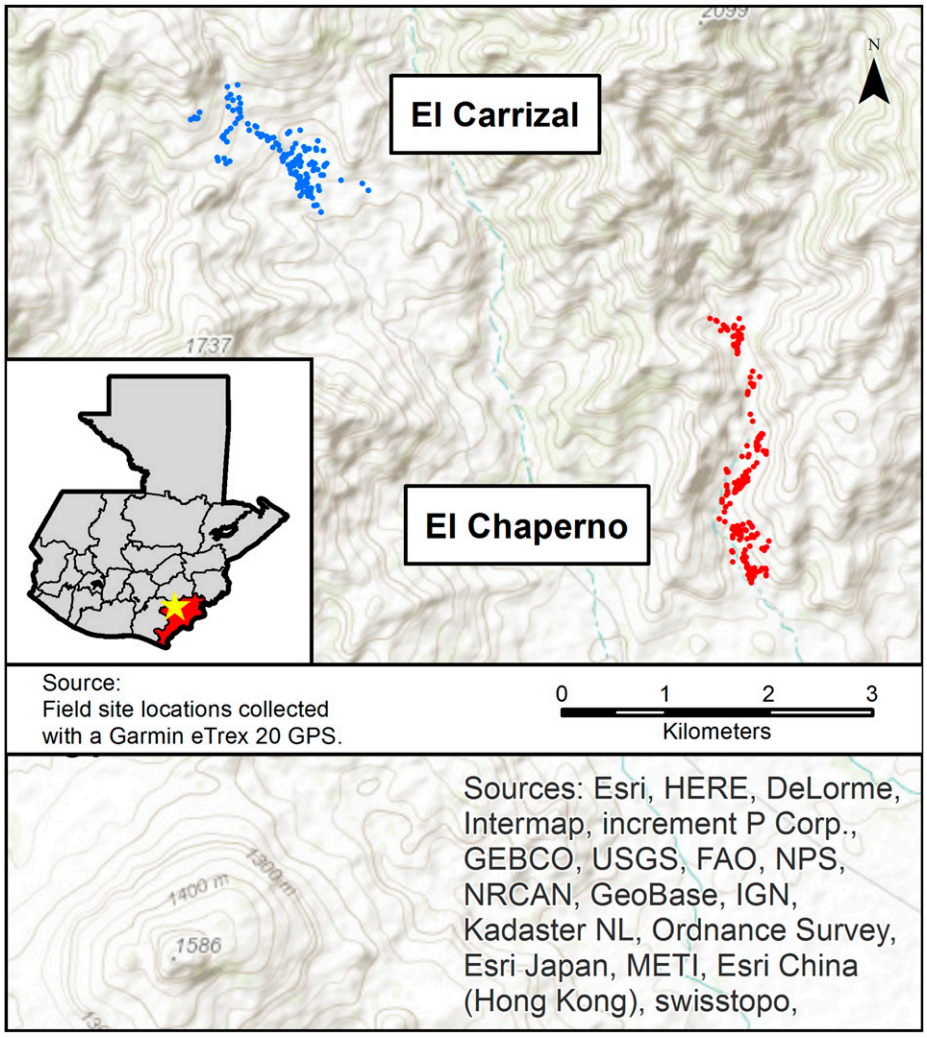

FIGURE 1. Topographic map of Guatemala study sites showing locations of houses in El Carrizal (blue dots) and El Chaperno (red dots). The inset highlights the department of Jutiapa, Guatemala (red), and the location of El Carrizal and El Chaperno (yellow star). Each house location was determined using a Garmin eTrex® 20 GPS (Garmin Ltd., Olathe, KS). The figure was created using ArcGIS® software by Esri (Esri, Redlands, CA). ArcGIS® and ArcMap™ are the intellectual properties of Esri and are used herein under license. Copyright © Esri. This figure appears in color at www.ajtmh.org.

of false negatives, we defined infestation as any sign that the house was infested with *T. dimidiata*, either at the time of survey or in the recent past, including live or dead *T. dimidiata*, eggs, exuviae, or characteristic fecal streaks. In this work, we combine the data from both towns, which have a minority of infested homes (combined 32.2%; see Table 1), to find more general associations of infestation in this area.

Conjunctive clause evolutionary algorithm and disjunctive normal form evolutionary algorithm. The CCEA is designed to identify multivariate risk factor interactions associated with large, complex datasets containing multiple data types (i.e., nominal, ordinal, and continuous) having large ranges of values, missing data, and imbalanced outcome classes.[21,27] In addition, the ranges of values evolved for each of the risk factors need not be monotonic. For instance, a continuous risk factor whose values most associated with infestation occupy the middle of a bell curve (or either extreme) may be discovered, unlike logistic regression whose models are based on exponentially increasing or decreasing values (for more detail, see Supplemental Comparison to Logistic Regression). The CCEA selects for the best CC of the form:

$$\text{CC}_k \text{ is defined as } F_i \in a_i \wedge F_j \in a_j \ldots, \quad (1)$$

where F_i represents a risk factor i whose value lies in the range a_i and the symbol $\wedge$ represents a conjunction (i.e., logical

TABLE 1
Summary data metrics for El Chaperno, El Carrizal, and the two towns combined

Dataset	Number of houses	% Infested houses	Number of ordinal, nominal, and binary risk factors	% Missing data	% Missing data per risk factor (median)	% Missing data per risk factor (min, max)
El Chaperno	182	26.9	12, 8, and 44	28.9	15.7	0.5, 86.8
El Carrizal	129	39.5	14, 8, and 42	22.3	3.9	0.8, 77.5
Combined	311	32.2	14, 8, and 42	26.1	10.3	1.2, 78.5

AND). The benefit of the CCEA is that it produces parsimonious models that are correlated with an outcome (e.g., infestation). One example of a parsimonious CC is houses with deteriorated walls AND dogs AND years of electricity ≥ 8 are more likely to be infested with *T. dimidiata* than houses that do not match each of these criteria. The number of risk factors in a CC constitutes the order of the CC. Thus, a CC with three risk factors joined by logical AND operators is referred to as a third-order CC, whereas an example of a second-order CC associated with an infested house is the household has dirt floors AND number of poultry ≥ 5.

The DNFEA tests for heterogeneity in datasets; it combines the CCs identified by the CCEA with logical OR statements to evolve sets of CCs using the DNF in the following structure:

$$\text{DNF}_k \text{ is defined as } \text{CC}_i \vee \text{CC}_j \ldots, \quad (2)$$

where each CC_i has the form of Equation 1 and the symbol $\vee$ represents a disjunction (i.e., logical OR).[19] The number of CCs constitutes the order of the DNF. Thus, a second-order DNF consists of two CCs joined by a logical OR operator. An example of a second-order DNF is an infested house that has deteriorated walls AND dogs AND years of electricity ≥ 8 (i.e., third-order CC_i) OR an infested house has dirt floors AND number of poultry ≥ 5 (i.e., second-order CC_j). The advantage of a logical OR operator is that it can account for heterogeneity in a given problem domain. For example, a simple CC (e.g., deteriorated walls AND dogs) that identified as highly correlated with infestation for one subset of infested houses might be just as likely as another clause (e.g., deteriorated walls AND poultry) for a different subset of infested houses because poultry have replaced the dog as a potential food source for *T. dimidiata*.

The fitness of each CC and DNF is evaluated using the hypergeometric probability mass function (PMF) and only the most fit are archived (i.e., saved). Unlike traditional statistics, the hypergeometric PMF is not a *P*-value and, thus, is not constrained by issues associated with what threshold is "significant."[28–30] However, much like the Akaike information criterion, the hypergeometric PMF in conjunction with the CC and DNF orders is used to compare the relative strengths of models for a given dataset.[31] The hypergeometric PMF considers the positive predictive value and infested house coverage. Positive predictive value is the number of true positives divided by the sum of true and false positives (also known as precision), and infested house coverage is the number of true positives divided by the sum of true positives and false negatives (also known as true positive rate, recall, sensitivity, probability of detection, and power). To prevent overfitting, the CCEA performs risk factor sensitivity on each CC to ensure each factor contributes to the overall fitness. For each risk factor in a CC, the sensitivity is calculated by taking the difference between the CC fitness and the fitness when that risk factor is removed. Thus, a risk factor's sensitivity may be viewed as the amount of fitness that each risk factor contributes to the CC. In this study, we only archived CCs in which all risk factors contribute at least $\log_{10}(0.05)$ to the fitness. To visualize the fitness landscape, both positive predictive value and infested house coverage are calculated.

The CCEA is first run to archive CCs most likely associated with *T. dimidiata* house infestation. After the first run, risk factors that were not archived were removed, and continuous risk factors were transformed into newly dichotomized risk factors based on the risk factor values archived during the first run. With this new set of risk factors, the CCEA was rerun. The CCs archived during the second CCEA run become the input data for the DNFEA. To contextualize the size of the search space of the CCEA, there are approximately 6×10^5, 2×10^8, and 3×10^{10} unique second-, third-, and fourth-order CCs in the infestation dataset. If one did not run the CCEA and DNFEA in tandem, then there would be $\sim 10^{20}$, $\sim 10^{30}$, and $\sim 10^{40}$ unique second-, third-, and fourth-order DNFs assuming that one limited the search space to fourth-order (or lower) CCs. The CCEA and DNFEA codes and example scripts can be found at Matlab Central[32]; further detail is provided in the Supplemental Methods and Hanley et al.[21]

RESULTS

The tandem CCEA and DNFEA archived more than 1,000 multivariate models associated with the infestation of *T. dimidiata*. These models contain a subset of the initial risk factors that our stakeholders believed to be associated with infestation. The first CCEA run evolved a set of 1,289 CCs that contained 48 of the original 64 input risk factors; 43 of these 48 were archived in second- or higher order CCs. In addition, nine of the continuous, ordinal, and discrete risk factors were dichotomized (Supplemental Figure S1, Table 2). Of the nine dichotomized risk factors, two (number of poultry and years of electricity for the household) were dichotomized into multiple risk factors covering different ranges (Table 2).

After rerunning the CCEA on the reduced dataset, 128 CCs were archived (~10% of the 1,289 from the first run) and used as input to the DNFEA to find the best disjunctive normal form (i.e., set of CCs); they contained 36 of the original 64 risk factors, with 32 of these risk factors embedded in second- or higher order CCs (Figure 2). The DNFEA results (green squares of Figure 2) contain 105 of the 128 CCs archived by the CCEA. The dashed contour lines represent the hypergeometric PMF model fitness. Overall, fitness increases from the lower left to the upper right corner of Figure 2. The blue circles represent the DNFEA output (2,571 archived second- to sixth-order disjunctive normal forms) (Figure 2, Supplemental Table S2).

Most of the archived DNFs are more fit than the most fit CC (Figure 2A). Thus, multiple models joined by logical OR statements are usually more fit than any single CC model. In general, as the order of the DNF increases, the gain in fitness diminishes, indicating there may be a fitness threshold for the number of models joined by logical OR statements. Although in theory all DNF models that lie along a given fitness contour may be considered "Pareto optimal" in terms of the hypergeometric PMF fitness, in practice, there is a desire to balance the trade-offs between multiple objectives. Selecting an optimal DNF is often constrained by the available resources. Ideally, the optimal DNF will have a 100% positive predictive value and 100% infested house coverage. In practice, there is a trade-off between the two. Therefore, as the percentage of infested house coverage increases, the positive predictive value decreases; thus, more resources are directed to houses that have false positives. However, a DNF with more risk factors may require more resources to reduce infestation.

TABLE 2

After the first conjunctive clause evolutionary algorithm run, nine of the original risk factor data types (continuous, ordinal, or discrete) were dichotomized into binary data or multiple binary (see number of poultry or years with electricity) data types

Risk factor	Data range	Dichotomized risk factor
Number of people	0, 15	≥ 5
Number of dogs	0, 13	≥ 1
Number of poultry	0, 35	≥ 5 and ≥ 6
Number of cats	0, 3	≥ 1
Number of pigs	0, 4	≥ 1
Number of beasts of burden	0, 4	≥ 1
Construction material location	Inside, against house, or outside	Inside, NOT inside
Chicken coop location	Inside, against house, or outside	Inside, NOT inside
Years with electricity	0.04, 20	≥ 3, ≥ 5, ≥ 6, ≥ 7, and ≥ 8

The third-order disjunctive normal form (red pentagram of Figure 2) identifies as an optimal solution because it is the most fit (fitness of 10^{-25}) third-order DNF as defined by our hypergeometric PMF fitness function. Higher order DNFs were not selected because they are more complex models (i.e., more risk factors) with only minimal gain in fitness, and the second-order DNFs are not selected because they either have much lower fitness and/or less infested house coverage. This third-order DNF has a hypergeometric PMF of 2.49×10^{-25}, an infested house coverage of 82%, a positive predictive value of 65%, and a classification accuracy of 80%. The three CCs that make up this DNF (orange hexagrams of Figure 2) are one

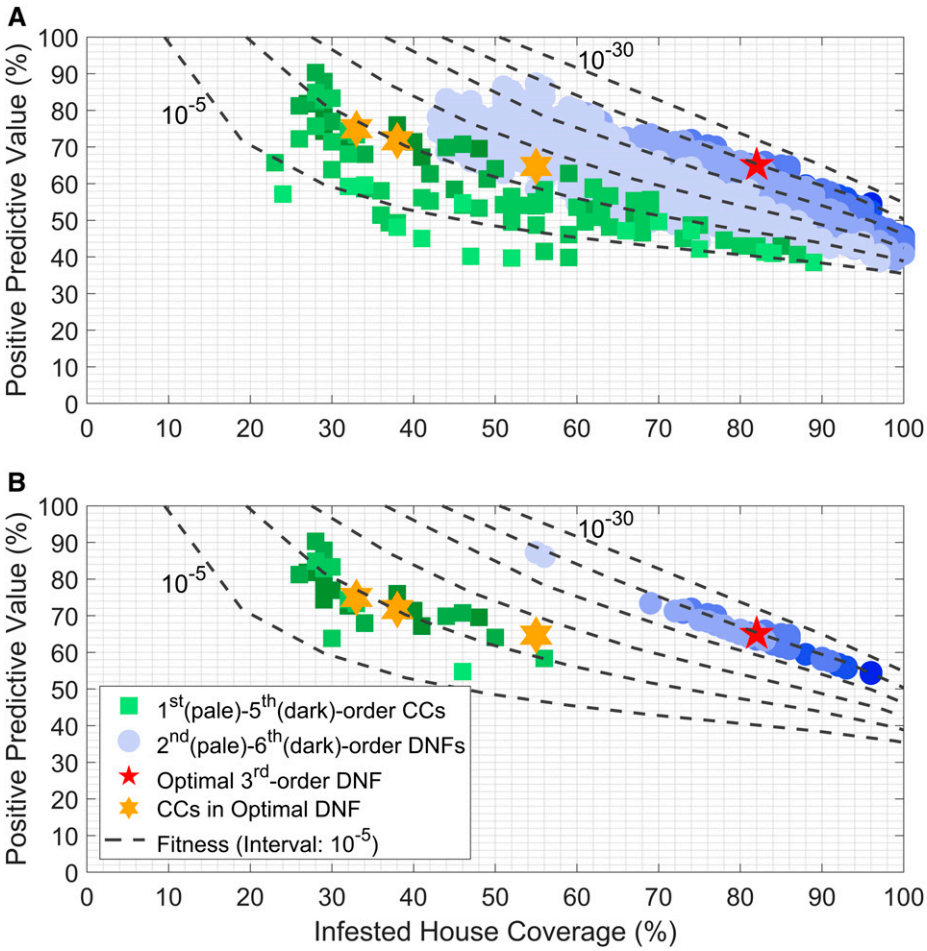

FIGURE 2. Archived results of the conjunctive clause (CC) evolutionary algorithm and disjunctive normal form evolutionary algorithm (DNFEA) output. The CCs are shown as green squares (where darker shades of green represent higher order CCs). The DNFEA output (archived disjunctive normal forms DNFs) are shown as blue circles (where darker shades of blue represent higher order DNFs). (**A**) All of the archived CCs and DNFs. (**B**) CCs present in the 100 most fit DNFs. The axes represent the positive predictive value and infested house coverage for towns of El Carrizal and El Chaperno. Dashed contour lines represent equally spaced fitness using the hypergeometric probability mass function. The optimal solution (third-order optimal disjunctive normal form) is identified by the red pentagram. The CCs embedded in this third-order disjunctive normal form are shown as orange hexagrams. This figure appears in color at www.ajtmh.org.

second-order CC and two fourth-order CCs, of Figure 3. The first CC, CC_1, has two risk factors (i.e., a chicken coop and deteriorated bedroom walls), a hypergeometric PMF of 3.31×10^{-10}, an infested house coverage of 33%, and a positive predictive value of 75%. The second conjunctive clause has a hypergeometric PMF of 3.28×10^{-12}, an infested house coverage of 38%, and a positive predictive value of 72%. The third conjunctive clause has a hypergeometric PMF of 1.05×10^{-10}, an infested house coverage of 55%, and a positive predictive value of 65%.

Logistic regression models for the three CCs that comprise the optimal disjunctive normal form (Figure 3) are presented in the Supplemental Section Comparison to Logistic Regression. Although a direct comparison between the CCEA models and logistic regression is not possible, the comparison was intended to help foster better understanding of how to interpret the CC results.

Another way to use the results of the evolutionary algorithms is to examine the individual risk factors that comprise the most-fit DNF models. As a result, we have organized the 25 risk factors into three groups: 1) vector food sources (i.e., potential blood meals for *T. dimidiata*), 2) vector shelter (daytime vector hiding places), and 3) socioeconomic conditions (Venn diagram of Figure 4). Based on this analysis, the main vector food sources associated with infestation are dogs, poultry, and humans. For dogs, only their presence or absence is associated with infestation, whereas for the other two vector food sources, infestation is most associated with larger numbers of poultry (≥ 5) and human household occupancy. With respect to vector shelter, the risk factors most associated with household infestation are having a poor hygiene rating and an accumulation of building materials, with the latter associated with a higher socioeconomic status. The socioeconomic risk factors most associated with infestation include the household income is something other than salary and the household has electricity for many years.

Further analysis of the 100 most fit DNF models shows they span second- to sixth-order DNFs, with the majority being fourth-order DNFs (Figure 2B). Of these most fit DNFs, only 29 distinct CCs are represented. The most common CC is fourth order and represented in 47 of the most fit 100 DNFs and is defined as having objects in the bedroom AND dogs AND number of poultry ≥ 5 AND years of electricity ≥ 8; the CC has a hypergeometric PMF of 6.64×10^{-9}, an infested house coverage of 29%, and a positive predictive value of 74%. Most importantly, CCs in the optimal DNF (Figure 3) are present in 16, 13, and 37 of the most fit 100 DNF models for CC_1, CC_2, and CC_3, respectively.

DISCUSSION

We applied a new, supervised, data mining tool for identifying socioeconomic and entomologic risk factors most associated with household infestation of *T. dimidiata*, the native vector of Chagas disease in Central America. The tandem CCEA and DNFEA provide a powerful and rigorous way to exhaustively explore the complex interactions among large numbers of risk factors (e.g., Supplemental Table S3). Although the combined algorithm produces multiple models that strongly correlate with infestation, we can identify a near optimal, third-order DNF model when the trade-offs among model simplicity, infested house coverage, and positive predictive value are balanced. The heterogeneity inherent in the models archived using this evolutionary framework provides support for previous studies,[14,15] which found no statistical support for a single "best" model of *T. dimidiata* infestation. We view the third-order DNF model as a tool to help domain experts identify households most at risk for *T. dimidiata* infestation (e.g., reducing the number of survey questions needed to identify houses most at risk) and promote best Ecohealth interventions.

The optimal DNF model (Figure 3) contains 10 risk factors associated with infestation. These include the following: household construction materials, such as 1) adobe walls and 2) dirt floors; 3) deteriorated household walls; 4) deteriorated bedroom walls; 5) the presence of chicken coops; 6) the presence of dogs; 7) objects accumulated in the bedroom; 8) household occupants ≥ 5; 9) a household hygiene rating of poor; and 10) years of electricity ≥ 5. Some of these risk factors (i.e., dirt floors, adobe walls, the deterioration of household walls, and the presence of chicken coops) have been previously identified as risk factors of *T. dimidiata* infestation.[14,15,22–24] Moreover, seven of these risk factors have been successfully targeted using Ecohealth interventions developed with local expertise, such as the replacement of dirt floors with cement-like material,[5,6] the replastering of deteriorated adobe walls,[8] and the construction of chicken coops with better materials and placement further from the house.[7] The overlap between these efforts and risk factors

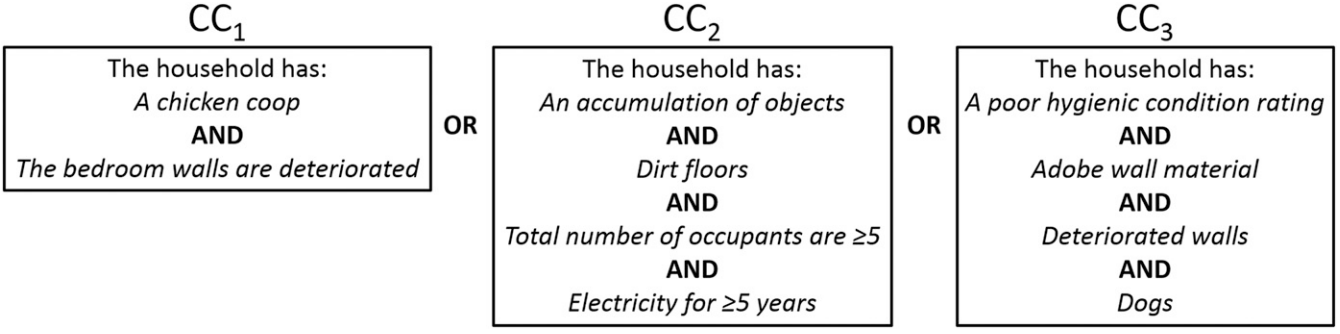

FIGURE 3. The three conjunctive clauses (CCs), joined by logical or statements, comprise the third-order DNF with the best fitness. Moving from left to right for the orange hexagrams in Figure 2, the first orange hexagram is CC_1, a second-order CC). The next two orange hexagrams are CC_2 and CC_3, both are fourth-order CCs.

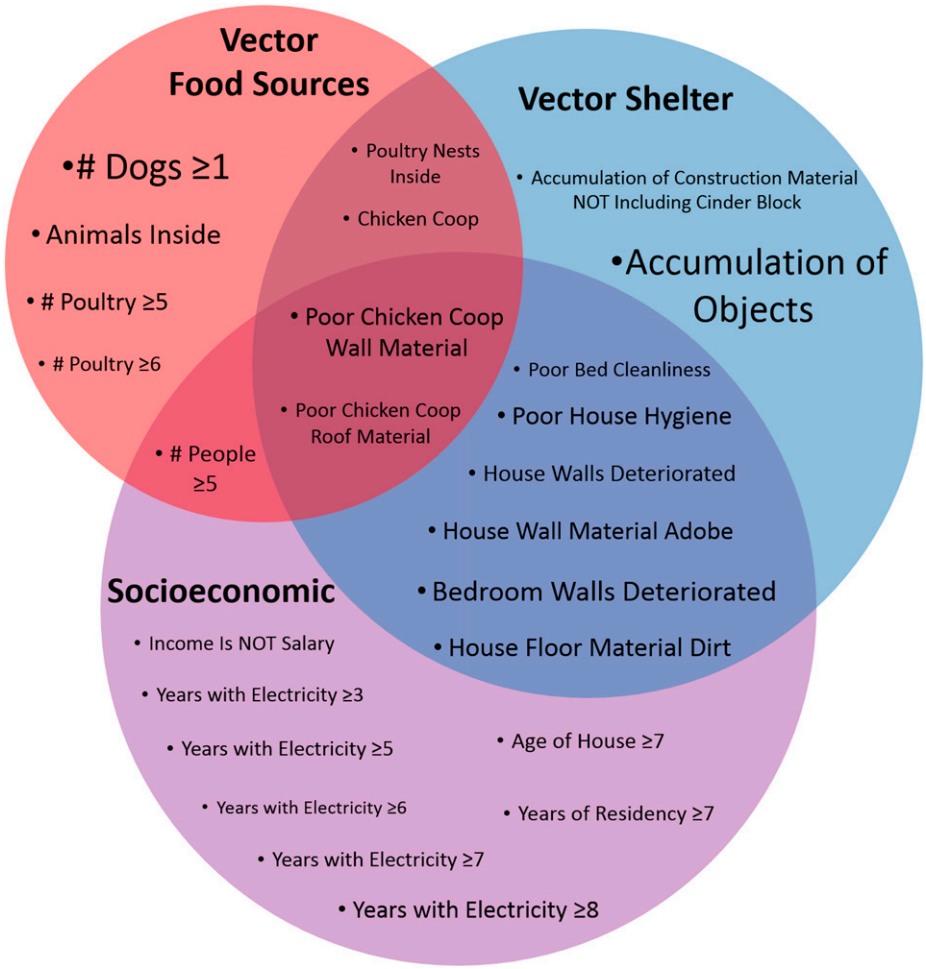

FIGURE 4. Venn diagram showing the 25 risk factors present in the 100 most fit DNFs defined by three groups, 1) vector food sources, 2) vector shelter, and 3) socioeconomic. The font size directly corresponds to the number of times the risk factor was archived in these DNFs. This figure appears in color at www.ajtmh.org.

identified by this new tandem evolutionary selection method is comforting from a model validation point of view.

In addition to determining the optimal DNF model, a macro-analysis approach was performed to identify all the risk factors present in the optimal DNF model as well as 99 nearly optimal models. In addition to the 10 risk factors associated with *T. dimidiata* infestation in the optimal DNF model (Figure 3), the macro-analysis provides a broader view of important risk factors associated with *T. dimidiata* infestation that are not represented in the optimal DNF model. It highlights another 15 risk factors that are present in nearly optimal models (Figure 4). The macro-analysis risk factors may be grouped as vector food sources, vector shelter, and/or socioeconomic risk factors associated with infestation. Some of these risk factors associated with *T. dimidiata* infestation, such as years of household electricity, are risk factors identified in the optimal DNF model; however, other factors (e.g., the source of household income not being salary) are not associated with the optimal DNF model.

Both the optimal DNF model and the macro-analysis indicate that the presence of dogs as the most important vector food source associated with the infestation of *T. dimidiata*. In fact, the presence of dogs as a risk factor associated with infestation is present in all of the top 100 most fit DNF models. Previous studies also show the presence of dogs to be a risk factor for *T. dimidiata* infestation.[15,22] Dogs are of particular concern because they are known reservoirs of Chagas disease in Central America, with a high infection prevalence (27.7–65.4%),[33–35] yet identifying a viable Ecohealth intervention could prove challenging. The construction of dog houses using materials and methods similar to those identified for chicken coops might be one solution. Alternatively, a proposed spay and neuter campaign might help reduce the risk of human infection,[36] albeit not necessarily domestic infestation unless all dogs are removed from the household.

The presence of poultry is another important vector food source associated with *T. dimidiata* infestation. Six of the 25 risk factors identified in the macro-analysis (Figure 4) involve poultry (e.g., households having poultry nests inside the house, poor materials for chicken coop walls, and number of poultry ≥ 5). For summary purposes, poor construction materials for chicken coops include any wall material that is not chicken wire and any roof material that is not corrugated metal. The Ecohealth interventions for chicken coops involved using both chicken wire and corrugated metal as the primary wall and roof materials, respectively.[7] Unlike adobe and thatched materials, these building materials provide little

shelter for *T. dimidiata* to hide during the day. In addition to improved construction materials, the chicken coops are relocated away from the house to help distance a *T. dimidiata* food source. There is support in the Venn diagram (Figure 4) for moving chicken coops away from the house because the presence of any animals in the house is an important risk factor. However, the algorithm did not find a relationship between chicken coop location and infestation. This may be due to a large number of chicken coops that are leaning against the house, making the location of the chicken coop irrelevant.

In addition to the vector food sources, vector shelters are present in the optimal DNF model and macro-analysis. Clutter (e.g., the accumulation of objects in the bedroom and a poor household hygiene rating) have been suggested as being correlated with infestation but never identified statistically.[14] A study in Costa Rica found correlation between *T. dimidiata* infestation and peri-domestic clutter both outside and under, but not within the house,[37] whereas a study in Guatemala posited but did not demonstrate a correlation between domestic infestation of *T. dimidiata* and objects accumulated inside houses and poor house hygiene.[8] As a result, they used an educational campaign to encourage cleaning and uncluttering houses as part of their Ecohealth interventions.[8] In this work, not only is an accumulation of objects present in the optimal DNF model of *T. dimidiata* infestation but it is also present in nearly all of the 100 most fit DNF models. In addition, other risk factors associated with clutter (e.g., construction material and poor house hygiene) are risk factors in the 100 most fit DNF models associated with infestation (Figure 4). Our findings add statistical support to the contributions of these studies and the resulting educational campaign to reduce infestation.

The macro-analysis presented here identifies a large number of socioeconomic risk factors, most of which are directly related to a steady source of shelter (e.g., deteriorated bedroom walls, dirt floors, and residency $\geq$ 7 years) for *T. dimidiata*.[25,38] The negative association between the accumulation of cinder blocks and infestation is interesting. Typically, the accumulation of objects and construction materials serves as a potential source of *T. dimidiata* shelter. However, cinder blocks are expensive, perhaps indicative of households with higher income, and are less likely to be left unused for any length of time.

The household income being identified as not salary based (i.e., the source of income is something other than salary such as day laborer or farmer) is the only purely economic risk factor identified to be associated with infestation. Determining the importance of an income that is not salary based in relation to other risk factors identified in the macro-analysis would require further study beyond the scope of this manuscript. However, given the health benefits that a guaranteed salary had on a rural community in a developed country[39] and the association of poverty with high levels of illness,[40] a future study on the impact of a guaranteed salary in a *T. dimidiata*–endemic community would be interesting.

A household having many years of electricity is a socioeconomic risk factor associated with infestation that was neither previously identified in the literature nor is it currently targeted by Ecohealth interventions. Unlike the number of household occupants, dichotomizing the number of years that a household has electricity into a single risk factor was difficult because there was no clear minimum number of years associated with infestation. However, given that the macro-analysis approach contains all five of the dichotomized "years of electricity" factors in the 100 most fit DNF models shows the robustness of the positive association between many years of electricity and *T. dimidiata* infestation. Despite the nocturnal nature of *T. dimidiata*, their affinity for light has been well documented in studies involving their capture using light traps.[41,42] In Mexico, studies also showed an association between public street lights and household *T. dimidiata* infestation yet no association with house light in the peridomicile.[43,44] In fact, we were not able to find any prior studies specifically identifying household light as a risk factor for *T. dimidiata* infestation. Field observations identify most of the household electrical devices, such as light bulbs, televisions, and cell phones, all of which emit light. In an indirect way, both the risk factor selection and macro-analysis of the 100 most fit DNFs (i.e., Venn diagram, Figure 4) in this study suggest the presence of light within the household to be a risk factor (i.e., there is high correlation between increased household infestation and the number of years a household has electricity). However, light attracts *T. dimidiata*, and there are a number of electrical devices that emit light in El Carrizal and El Chaperno; therefore, barriers blocking the entry of *T. dimidiata* into the household could help prevent infestation. In the Yucatan where *T. dimidiata* infestation is seasonal, window screens installed in houses in rural villages moderately reduced domestic infestation.[45] Although the use of window screens might help mitigate the risk of infestation due to household electricity in Jutiapa, it is costly and may not be effective, as not all houses have windows, and of those with windows, they are often left closed. In addition, a number of homes have gaps around the door frame and between the walls and the roof that provide additional points of entry.

Another challenging socioeconomic risk factor associated with infestation is a household having the total number of occupants $\geq$ 5. The latter is present in the optimal model and in 45 of the 100 most fit DNF models. Much like the risk factor of a house having many years of electricity, there is no obvious intervention because it is not currently known if households in these villages have large occupancy out of choice or necessity. With that being said, it may not be necessary (or possible) to develop interventions for every risk factor associated with reducing *T. dimidiata* infestation.

The evolutionary algorithms used in this analysis helped identify additional risk factors associated with infestation not previously identified in the literature and those that do not currently have viable interventions, specifically households having a larger number of occupants (i.e., $\geq$ 5), household income not being salary based, and households having electricity for an extended period of time. Targeting risk factors identified in CCs that do not have viable interventions would be an inefficient use of limited resources; efforts should be directed toward those factors with viable interventions, for instance, CC_2 in the optimal DNF (Figure 3). This clause contains two risk factors (the total number of occupants $\geq$ 5 and households having years of electricity $\geq$ 5) that do not have viable interventions. However, it also contains two risk factors (an accumulation of objects and dirt floors) with viable Ecohealth interventions. Based on this CC model, a decluttering campaign[8] and replacing dirt floors with cement-like floors[5,6] may reduce the risk of infestation. Alternatively, risk factors that do not have a viable intervention might be used to help

streamline household surveys that are designed to identify households most at risk.

Unlike traditional multivariate statistics, this novel methodology uses the hypergeometric PMF that like the Akaike information criterion is not subject to selecting a *P*-value.[28–30] The method is specifically designed to search for risk factor interactions and heterogeneity. The existence of heterogeneity shows that targeting the risk factors present in any one CC will only help prevent infestation of *T. dimidiata* for a subset of the households. Thus, local stakeholders will have to use their domain expertise when balancing positive predictive value, infested house coverage, and risk factors that have viable interventions in terms of resources and their applications. When stakeholder resources are limited, selecting "best" solutions often becomes an "art form" that should be performed in concert with domain experts[46] to ensure fewer complex models with fewer yet more manageable risk factors. Thus, in addition to identifying risk factors associated with infestation, this novel tandem evolutionary algorithm is able to provide a suite of solutions that leverage interactions among large numbers of risk factors associated with socioeconomic and entomologic survey data to assist domain experts familiar with local resources and stakeholder needs.

Received September 6, 2018. Accepted for publication April 29, 2020.

Published online June 8, 2020.

Note: Supplemental material, methods, figure, and tables appear at www.ajtmh.org.

Acknowledgments: We would like to thank the Laboratory for Applied Entomology and Parasitology (LENAP) at La Universidad de San Carlos Guatemala, especially Gabriela Rodas, Raquel Asuncion Lima, Elizabeth Solórzano, Bethany Richards, and Dulce Bustamante, for their help in collecting and managing the infestation and socioeconomic datasets. We also thank the Guatemalan Ministry of Health for their help with collecting the data. Finally, we would like to thank the people of El Chaperno and El Carrizal for welcoming us and collaborating with us in this study.

Financial support: This work was supported by the National Science Foundation (BCS-EEID-1216193) and in part by the National Science Foundation under VT EPSCoR (NSF-OIA-1556770 D.M.R.) and grant R03AI26268/1-2 from the National Institute of Allergy and Infectious Diseases (NIAID) at the National Institutes of Health (NIH) (L. S., S. H. C., P. L. D.).

Disclaimer: Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding organizations.

Disclosures: J. P. H. reports grants from the National Science Foundation during the conduct of the study. D. M. R. reports grants from the National Science Foundation during the conduct of the study. L. S. reports grants from the National Science Foundation and the National Institutes of Health during the conduct of the study. S. H. C. reports grants from the National Science Foundation and the National Institutes of Health during the conduct of the study. P. L. D. reports grants from the National Science Foundation and the National Institutes of Health during the conduct of the study. A. G. R. reports grants from the National Science Foundation during the conduct of the study. L. C. O. reports grants from the National Science Foundation during the conduct of the study. C. M. reports grants from the National Science Foundation during the conduct of the study.

Authors' addresses: John P. Hanley and Donna M. Rizzo, Department of Civil and Environmental Engineering, University of Vermont, Burlington, VT, E-mails: jhanley@uvm.edu and drizzo@cems.uvm.edu. Lori Stevens and Sara Helms Cahan, Department of Biology, University of Vermont, Burlington, VT, E-mails: lori.stevens@uvm.edu and scahan@uvm.edu. Patricia L. Dorn, Department of Biological Sciences, Loyola University New Orleans, New Orleans, LA, E-mail: dorn@loyno.edu. Leslie A. Morrissey and Lucia C. Orantes, Rubenstein School of Environment and Natural Resources, University of Vermont, Burlington, VT, E-mail: leslie.morrissey@uvm.edu and lucia.orantes@uvm.edu. Antonieta Guadalupe Rodas and Carlota Monroy, Laboratorio de Entomología Aplicada y Parasitología, Escuela de Biología, Universidad de San Carlos de Guatemala, Ciudad de Guatemala, Guatemala, E-mails: antonieta55@yahoo.com and mcarlotamonroy@gmail.com.

REFERENCES

1. World Health Organization, 2013. Savioli L, Daumerie D, eds. *Sustaining the Drive to Overcome the Global Impact of Neglected Tropical Diseases: Second WHO Report on Neglected Tropical Diseases*. Geneva, Switzerland: WHO.
2. Helms Cahan S, Orantes LC, Wallin K, Rizzo DM, Stevens L, Dorn PL, Rodas AG, Monroy C, 2019. Residual survival and local dispersal drive reinfestation by *Triatoma dimidiata* following insecticide application in Guatemala. *Infect Genet Evol 74:* 104000.
3. Dumonteil E, Ruiz-Piña H, Rodriguez-Félix E, Barrera-Pérez M, Ramirez-Sierra MJ, Rabinovich JE, Menu F, 2004. Reinfestation of houses by *Triatoma dimidiata* after intradomicile insecticide application in the Yucatán Peninsula, Mexico. *Mem Inst Oswaldo Cruz 99:* 253–256.
4. Hashimoto K, Cordon-Rosales C, Trampe R, Kawabata M, 2006. Impact of single and multiple residual sprayings of pyrethroid insecticides against *Triatoma dimidiata* (Reduviiade; Triatominae), the principal vector of Chagas disease in Jutiapa, Guatemala. *Am J Trop Med Hyg 75:* 226–230.
5. Lucero DE, Morrissey LA, Rizzo DM, Rodas A, Garnica R, Stevens L, Bustamante DM, Monroy MC, 2013. Ecohealth interventions limit triatomine reinfestation following insecticide spraying in La Brea, Guatemala. *Am J Trop Med Hyg 88:* 630–637.
6. Pellecer MJ, Dorn PL, Bustamante DM, Rodas A, Monroy MC, 2013. Vector blood meals are an early indicator of the effectiveness of the ecohealth approach in halting Chagas transmission in Guatemala. *Am J Trop Med Hyg 88:* 638–644.
7. Monroy C, Castro X, Bustamante DM, Pineda SS, Rodas A, Moguel B, Ayala V, Quiñonez J, 2012. An ecosystem approach for the prevention of Chagas disease in rural Guatemala. Charron DF, ed. *Ecohealth Research in Practice*. New York, NY: Springer.
8. Monroy C, Bustamante DM, Pineda S, Rodas A, Castro X, Ayala V, Quiñonez J, Moguel B, 2009. House improvements and community participation in the control of *Triatoma dimidiata* reinfestation in Jutiapa, Guatemala. *Cad Saude Publica 25 (Suppl 1):* S168–S178.
9. Hotez P, 2008. Hookworm and poverty. *Ann N Y Acad Sci 1136:* 38–44.
10. Murray CJL et al., 2012. Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1990–2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet 380:* 2197–2223.
11. Cox DR, 1984. Interaction. *Int Stat Rev 52:* 1–31.
12. Thornton-Wells TA, Moore JH, Haines JL, 2004. Genetics, statistics and human disease: analytical retooling for complexity. *Trends Genet 20:* 640–647.
13. Moore JH, 2003. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum Hered 56:* 73–82.
14. Bustamante Zamora DM, Hernandez MM, Torres N, Zuniga C, Sosa W, de Abrego V, Monroy Escobar MC, 2015. Information to act: household characteristics are predictors of domestic infestation with the Chagas vector *Triatoma dimidiata* in central America. *Am J Trop Med Hyg 93:* 97–107.
15. Bustamante DM, De Urioste-Stone SM, Juárez JG, Pennington PM, 2014. Ecological, social and biological risk factors for continued *Trypanosoma cruzi* transmission by *Triatoma dimidiata* in Guatemala. *PLoS One 9:* e104599.
16. Bustamante DM, Monroy C, Pineda S, Rodas A, Castro X, Ayala V, Quiñónes J, Moguel B, Trampe R, 2009. Risk factors for intradomiciliary infestation by the Chagas disease vector *Triatoma dimidiata* in Jutiapa, Guatemala. *Cad Saude Publica 25 (Suppl 1):* S83–S92.

17. Campbell-Lendrum D et al., 2007. House-level risk factors for triatomine infestation in Colombia. *Int J Epidemiol 36:* 866–872.
18. Chaves L, Calzada JE, Rigg C, Valderrama A, Gottdenker NL, Saldaña A, 2013. Leishmaniasis sand fly vector density reduction is less marked in destitute housing after insecticide thermal fogging. *Parasite Vector 6:* 1–13.
19. Chaves LF, Hamer GL, Walker ED, Brown WM, Ruiz MO, Kitron UD, 2011. Climatic variability and landscape heterogeneity impact urban mosquito diversity and vector abundance and infection. *Ecosphere 2:* 1–21.
20. Ruiz MO, Chaves LF, Hamer GL, Sun T, Brown WM, Walker ED, Haramis L, Goldberg TL, Kitron UD, 2010. Local impact of temperature and precipitation on West Nile virus infection in *Culex* species mosquitoes in northeast Illinois, USA. *Parasite Vectors 3:* 1–16.
21. Hanley JP, Rizzo DM, Buzas JS, Eppstein MJ, 2020. A tandem evolutionary algorithm for identifying causal rules from complex data. *Evol Comput 28:* 87–114.
22. Parra-Henao G, Cardona ÁS, Quirós-Gómez O, Angulo V, Alexander N, 2015. House-level risk factors for *Triatoma dimidiata* infestation in Colombia. *Am J Trop Med Hyg 92:* 193–200.
23. Weeks ENI, Cordón-Rosales C, Davies C, Gezan S, Yeo M, Cameron MM, 2013. Risk factors for domestic infestation by the Chagas disease vector, *Triatoma dimidiata* in Chiquimula, Guatemala. *Bull Entomol Res 103:* 634–643.
24. King RJ, Cordon-Rosales C, Cox J, Davies CR, Kitron UD, 2011. *Triatoma dimidiata* infestation in Chagas disease endemic regions of Guatemala: comparison of random and targeted cross-sectional surveys. *PLoS Negl Trop Dis 5:* e1035.
25. Monroy C, Mejia M, Rodas A, Rosales R, Horio M, Tabaru Y, 1998. Comparison of indoor searches with whole house demolition collections of the vectors of Chagas disease and their indoor distribution. *Med Entomol Zoolog 49:* 195–200.
26. World Health Organization, 2002. *Control of Chagas Disease: Second Report of the WHO Expert Committee*. Geneva, Switzerland: WHO, 1–120.
27. Hanley JP, Eppstein MJ, Buzas JS, Rizzo DM, 2016. *Evolving Probabilistically Significant Epistatic Classification Rules for Heterogeneous Big Datasets*. Proceedings of the 18th Annual Conference on Genetic and Evolutionary Computation, 445–452.
28. Wasserstein RL, Schirm AL, Lazar NA, 2019. Moving to a world beyond "$p < 0.05$". *Am Stat 73 (Suppl 1):* 1–19.
29. Wasserstein RL, Lazar NA, 2016. The ASA's statement on *p*- values: context, process, and purpose. *Am Stat 70:* 129–133.
30. Nuzzo R, 2014. Scientific method: statistical errors. *Nature 506:* 150–152.
31. Akaike H, 1974. A new look at the statistical model identification. *IEEE Trans Automat Contr 19:* 716–723.
32. Hanley JP, 2019. Available at: https://www.mathworks.com/matlabcentral/fileexchange/69950-ccea-and-dnfea. Accessed February 21, 2020.
33. Carrillo-Peraza J, Manrique-Saide P, Rodríguez-Buenfil J, Escobedo-Ortegón J, Rodríguez-Vivas R, Bolio-González M, Barrera-Pérez M, Reyes-Novelo E, Sauri-Arceo C, 2014. Estudio serológico de la tripanosomiasis Americana y factores asociados en perros de una comunidad rural de Yucatán, México. *Arch Med Vet 46:* 75–81.
34. Hernández JL, Rebollar-Téllez EA, Infante F, Morón A, Castillo A, 2010. Indicadores de infestación, colonización e infección de *Triatoma dimidiata* (Latreille) (Hemiptera: Reduviidae) en Campeche, México. *Neotrop Entomol 39:* 1024–1031.
35. Montenegro VM, Jimenez M, Dias JCP, Zeledon R, 2002. Chagas disease in dogs from endemic areas of Costa Rica. *Mem Inst Oswaldo Cruz 97:* 491–494.
36. Lima-Cordón RA, Stevens L, Solórzano Ortíz E, Rodas GA, Castellanos S, Rodas A, Abrego V, Zúniga Valeriano C, Monroy MC, 2018. Implementation science: epidemiology and feeding profiles of the Chagas vector *Triatoma dimidiata* prior to ecohealth intervention for three locations in central America. *PLoS Negl Trop Dis 12:* e0006952.
37. Zeledón R, Rojas JC, 2006. Environmental management for the control of *Triatoma dimidiata* (Latreille, 1811), (Hemiptera: Reduviidae) in Costa Rica: a pilot project. *Mem Inst Oswaldo Cruz 101:* 379–386.
38. Zeledón R, Zúñiga A, Swartzwelder JC, 1969. The camouflage of *Triatoma dimidiata* and the epidemiology of Chagas' disease in Costa Rica. *Bol Chil Parasitol 24:* 106–108.
39. Forget EL, 2011. The town with No poverty: the health effects of a Canadian guaranteed annual income field experiment. *Can Public Pol 37:* 283–305.
40. Marmot M, Friel S, Bell R, Houweling TAJ, Taylor S, 2008. Closing the gap in a generation: health equity through action on the social determinants of health. *Lancet 372:* 1661–1669.
41. Rebollar-Téllez EA, Reyes-Villanueva F, Escobedo-Ortegón J, Balam-Briceño P, May-Concha I, 2009. Abundance and nightly activity behavior of a sylvan population of *Triatoma dimidiata* (Hemiptera: Reduviidae: Triatominae) from the Yucatan, México. *J Vector Ecol 34:* 304–310.
42. Zeledón R, Ugalde JA, Paniagua LA, 2001. Entomological and ecological aspects of six sylvatic species of triatomines (Hemiptera, Reduviidae) from the collection of the National Biodiversity Institute of Costa Rica, central America. *Mem Inst Oswaldo Cruz 96:* 757–764.
43. Dumonteil E, Nouvellet P, Rosecrans K, Ramirez-Sierra MJ, Gamboa-León R, Cruz-Chan V, Rosado-Vallado M, Gourbière S, Gürtler RE, 2013. Eco-bio-social determinants for house infestation by non-domiciliated *Triatoma dimidiata* in the Yucatan Peninsula, Mexico. *PLoS Negl Trop Dis 7:* e2466.
44. Pacheco-Tucuch FS, Ramirez-Sierra MJ, Gourbière S, Dumonteil E, 2012. Public street lights increase house infestation by the Chagas disease vector *Triatoma dimidiata*. *PLoS One 7:* e36207.
45. Waleckx E et al., 2018. Non-randomized controlled trial of the long-term efficacy of an ecohealth intervention against Chagas disease in Yucatan, Mexico. *PLoS Negl Trop Dis 12:* e0006605.
46. Reklaitis GV, Ravindran A, Ragsdell KM, 1983. *Engineering Optimization: Methods and Applications*. New York, NY: Wiley.