Dynamics of data availability in disease modeling: An example evaluating the trade-offs of ultra-fine-scale factors applied to human West Nile virus disease models in the Chicago area, USA

[*1]Uelmen, J.A., [2,3]Irwin, P., [1]Brown, W.M., [1,4]Karki, S., [1]Ruiz, M.O., [4]Li, B., and [1]Smith, R. L.

[1]Department of Pathobiology, College of Veterinary Medicine, University of Illinois at Urbana-Champaign, Urbana, Illinois, United States of America

[2]Northwest Mosquito Abatement District, Wheeling, Illinois, United States of America

[3]Department of Entomology, University of Wisconsin-Madison, Madison, WI, United States of America

[4]Department of Epidemiology and Public Health, Himalayan College of Agricultural Sciences and Technology, Kirtipur, Kathmandu, Nepal

[4]Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, Illinois, United States of America

[*]Corresponding Author

E-mail: uelmen@illinois.edu

Keywords: WNV, disease modeling, *Culex*, vector index, human illness,

Abstract

Background: Since 1999, West Nile virus (WNV) has moved rapidly across the United States, resulting in tens of thousands of human cases. Both the number of human cases and the minimum infection rate (MIR) in vector mosquitoes vary across time and space and are driven by numerous abiotic and biotic forces, ranging from differences in microclimates to socio-demographic factors. Because the interactions among these multiple factors affect the locally variable risk of WNV illness, it has been especially difficult to model human disease risk across varying spatial and temporal scales. Cook and DuPage Counties, comprising the city of Chicago and surrounding suburbs, experience some of the highest numbers of human neuroinvasive cases of WNV in the United States. Despite active mosquito control efforts, there is consistent annual WNV presence, resulting in more than 285 confirmed WNV human cases and 20 deaths from the years 2014-2018 in Cook County alone.

Methods: A previous Chicago-area WNV model identified the fifty-five most high and low risk locations in the Northwest Mosquito Abatement District (NWMAD), an enclave ¼ the size of the combined Cook and DuPage county area. In these locations, human WNV risk was stratified by model performance, as indicated by differences in studentized residuals. Within these areas, an additional two-years of field collections and data processing was added to a 12-year WNV dataset that includes human cases, MIR, vector abundance, and land-use, historical climate, and socio-economic and demographic variables, and was assessed by an ultra-fine-scale (1 km spatial x 1 week temporal resolution) multivariate logistic regression model.

Results: Multivariate statistical methods applied to the ultra-fine-scale model identified fewer explanatory variables while improving upon the fit of the previous model. Beyond MIR and

climatic factors, efforts to acquire additional covariates only slightly improved model predictive performance.

Conclusions: These results suggest human WNV illness in the Chicago area may be associated with fewer, but increasingly critical, key variables at finer scales. Given limited resources, these findings suggest large variations in model performance occur, depending on covariate availability, and provide guidance in variable selection for optimal WNV human illness modeling.

Introduction

West Nile virus (WNV; Family *Flaviviridae*), a mosquito-borne disease originating from the West Nile region of Uganda, first arrived to the United States (U.S., New York, NY) in 1999. Once arriving in New York, the virus took only three years to traverse the contiguous U.S., reaching California in 2002 (1). The virus has now become one of the most widespread arboviruses in the world, and is present in every continent except Antarctica (2). In the Midwestern U.S., mosquitoes of the *Culex* (*Cx.*) genus are the main vectors for transmitting WNV (3). *Culex* mosquitoes are capable of feeding on several hosts to satisfy one blood meal, increasing the opportunity for multiple infections across species (4). Although primarily ornithophilic, prior studies indicate that *Cx.* species may shift feeding preferences to humans later in the summer months (5,6).

From 1999-2018, there have been a total of 50,830 human cases resulting in 2,330 deaths across the US (7). At local scales, drivers of human disease, including WNV, vary in actual effect and magnitude from values reported in studies that more commonly assess disease dynamics at state, regional, or national scales (8). Previous studies have identified similar abiotic and biotic factors associated with human WNV illness, including prior weather conditions (weekly temperature and precipitation lags), mosquito infection and abundance, socio-demographic characteristics of the local population, and level of public awareness and education, but these were all at state or regional scales (9–17).

Karki et al. (2020) and Ruiz et al. (2010) are two of the few studies to evaluate weekly spatiotemporal factors and their associations with human WNV illness at a smaller scale (1-km hexagonal spatial units), in a highly urban 2-county area (Cook & DuPage counties, encompassing the greater Chicago, IL area). This region consistently experiences one of the

highest annual WNV incidences in the country (20). While an excellent overall model fit was achieved by using a large number of explanatory variables (n=40), the relative importance of covariates and the resulting strength of disease prediction across the study area varied widely. Understanding how and why these relationships change at specific spatiotemporal locales has been a major conceptual challenge when modeling human WNV illness, and is the central focus for this study.

The Northwest Mosquito Abatement District (NWMAD), occupying the northwest corner of Cook County, is one of Chicago's four abatement districts responsible for mosquito control, and has an excellent long-term mosquito abundance and testing data throughout its jurisdiction. Human and environmental factors are heterogenous throughout the NWMAD, presenting a strong gradient of human population density, household size and age, socio-economic values, and land-use and land-cover, providing a highly representative enclave of the greater Chicago region.

Specifically, the main objectives of this study were to: (i) evaluate and contrast key variables in this study to the larger Cook and DuPage model, (ii) assess the similarities and differences among locations that were predicted accurately by the larger model and those that were predicted poorly, and (iii) quantify the impact of newly acquired data on prediction of human WNV illness. The authors hypothesize that evaluating human WNV risk at an ultra-fine-scale (UFS) will improve overall model performance (as compared to broader scale models). Additionally, the authors hypothesize that by including several additional covariates that are specific to the UFS study area, the eco-epidemiologic relationships of human WNV transmission will be improved.

Methods

Ethics Statement

All data collected from the Illinois Department of Health (IDPH) were through a user agreement approved by the University of Illinois Institutional Review Board and the Illinois Department of Public Health Institutional Review Board. The human activity observation protocol was approved by the University of Illinois Institutional Review Board. Field collections and any use of generated data were approved by the University of Illinois Biosafety Committee.

Study area

This study was conducted within the NWMAD, a 605-km^2 area that comprises the northwest suburbs of Chicago (Cook County, IL, Fig 1). The NWMAD study area is an enclave of the Cook & DuPage counties model, the previous research site conducted by Karki et al. (2020). Within the NWMAD study area, fifty-five 1-km hexagonal units were specially selected. These fifty-five 1-km units denote the "ultra-fine-scale" (UFS) study area and contained a total of forty human WNV cases from 2005-2016 (Table S1). By focusing on the spatiotemporal dynamics of WNV transmission in humans in this UFS study area, research efforts have focused on additional data collection, more than doubling the total amount of covariates related to WNV in the Chicago region than the previous Karki et al. (2020) study. Through these additional collection efforts, this study aims to better control, assess, and ultimately, understand the relationships among key predictors of human WNV disease at very fine scales. All model data were summarized and processed within 1-km diameter hexagons, as a neutral configuration in both size and shape, free of any political boundaries. Using statistical selection processes (described below), fifty-five of the 1,019 hexagons within the NWMAD were selected as the observational units for this study.

126 **Fig 1. The UFS study area, contained within the Northwest Mosquito Abatement District**
127 **(NWMAD), in relation to Cook and DuPage Counties**. Overlaid are 1-km diameter hexagons,
128 the observational units used in this study. Northwest Mosquito Abatement District comprises
129 1,019 of the total 5,345 hexagons in all of Cook and DuPage Counties.

130 # Model covariates

131 The Cook and DuPage model evaluated forty covariates derived from a variety of abiotic and
132 biotic factors associated with human WNV transmission, including climate and weather records,
133 mosquito infection, environmental land use, and socio-demographic census data. For this study,
134 additional data processing and field collections resulted in forty-two additional independent
135 variables, each determined to be ecologically- or epidemiologically-related to human WNV
136 illness in our study areas of focus (Table 1). Each variable was independently calculated by
137 hexagon and averaged for each Centers for Disease Control and Prevention (CDC)
138 epidemiological week (18-38, Sunday-Saturday) of the years 2005 through 2016 (7). Previously
139 collected data used in this study are explained in detail in Karki et al (2020) and can also be
140 found in supplemental materials.

Table 1. Ecologically- and epidemiologically-related human WNV illness variables assessed.

Covariate Information				Cook/DuPage Model	Ultra-fine-scale Model
Designation		**Description**	**Notation**		
Environmental	Land Cover	Proportion of developed open space	dospct	X	X
		Proportion of developed low intensity	dlipct	X	X
		Proportion of developed medium intensity	dmipct	X	X
		Proportion of developed high intensity	dhipct	X	X
		Proportion of deciduous forests	dfpct	X	X
		Proportion of evergreen forests	efpct	X	X
		Proportion of mixed forests	mfpct	X	X
		Proportion of barren land	blpct	X	X
		Proportion of shrubs	shrubspct	X	X
		Proportion of grassland	glandpct	X	X
		Proportion of pasture	pasturepct	X	X
		Proportion of cultivated land	clpct	X	X
		Proportion of woody wetlands	wwpct	X	X
		Proportion of herbaceous wetlands	hwpct	X	X
		Proportion of total forest	ftotpct		X
		Proportion of total wetlands	wtotpct		X
		Proportion of open water	owpct	X	X
		Normalized Difference Vegetation Index	NDVI		X
Biological	Minimum Infection Rate (MIR)	MIR one week before	mirlag1	X	X
		MIR two weeks before	mirlag2	X	X
		MIR three weeks before	mirlag3	X	X
		MIR four weeks before	mirlag4	X	X
		Average MIR current week	MIRmean		X
		Difference in weekly average MIR from 12-year average	MIRdiff		X
	Mosquito Abundance	Vector Index current week	Vector Index		X
		Vector Index one week before	VIlag1		X
		Vector Index two weeks before	VIlag2		X
		Vector Index three weeks before	VIlag3		X
		Vector Index four weeks before	VIlag4		X
		Light and gravid trap collection mean current week	Trap_Mean		X
		Light and gravid trap collection mean one week before	Trap_Meanlag1		X
		Light and gravid trap collection mean two weeks before	Trap_Meanlag2		X
		Light and gravid trap collection mean three weeks before	Trap_Meanlag3		X
		Light and gravid trap collection mean four weeks before	Trap_Meanlag4		X
	Mosquito Biting Rates (HLC)	Mosquitoes per visit	mosquitoes per visit		X
		Culex spp. per visit	Cx per visit		X
Weather	Temperature	Average temperature current week	tempc		X
		Average temperature of one week before	templag1	X	X
		Average temperature of two weeks before	templag2	X	X
		Average temperature of three weeks before	templag3	X	X
		Average temperature of four weeks before	templag4	X	X
		Mean January temperature	Jantemp	X	X
	Precipitation	Average precipitation current week	preci		X
		Average precipitation of one week before	precilag1	X	X
		Average precipitation of two weeks before	precilag2	X	X
		Average precipitation of three weeks before	precilag3	X	X
		Average precipitation of four weeks before	precilag4	X	X
Anthropogenic	Socio-demographic	Percentage of White population	whitepct	X	X
		Percentage of African American population	blackpct	X	X
		Percentage of Asian population	asianpct	X	X
		Percentage of Hispanic population	hispanicpct	X	X
		Median household income	Income	X	X
		Percentage of housing constructed before WWII	hpctpreww	X	X
		Percentage of housing constructed post WWII (1945-1969)	hpctpostww	X	X
		Percentage of housing constructed from 1970-1989	hpct7089	X	X
		Percentage of housing constructed in 1990 or later	hpctpost90	X	X
	Land change & manipulation	Catch basin density	CB		X
		Total area of building structures	bldg_footprint_area_total		X
		Average area of building structures	bldg_footprint_area_avg		X
		Total perimeter of building structures	Building_Footprint_peri_total		X
		Average perimeter of building structures	Building_Footprint_peri_avg		X
		Total area of residential lot	Residential_lot_area_total		X
		Average area of residential lot	Residential_lot_area_avg		X
		Total perimeter of residential lot	Residential_lot_peri_total		X
		Average perimeter of residential lot	Residential_lot_peri_avg		X
		Ratio of total building area by total lot area	total_bldg_area/total_lot_area		X
		Ratio of average building area by average lot area	avg_bldg_area/avg_lot_area		X
		Ratio of total building perimeter by total lot area	total_bldg_peri/total_lot_area		X
		Ratio of average building perimeter by average lot area	avg_bldg_peri/avg_lot_area		X
		Number of buildings	buildings		X
		Building density per mi.^{2}	bldg_density		X
	Human population	Number of residents per building	persons_per_bldg		X
		Total human population	totpop	X	X
		Mean light pollution	lightpol		X
	Activity Observations	Senior Citizen Observations per visit	Senior_obs per visit		X
		Adults Observations per visist	Adults_obs per visit		X
		Children Observations per visit	Child_obs per visit		X
		Male Observations per visit	Male_obs per visit		X
		Female Observations per visit	Female_obs per visit		X
		Total Observations per visit	Total_obs per visit		X
Other		Year	yr	X	X
		Hexagon Designation	hexid	X	X
			Total Covariates Evaluated	**40**	**82**

141 List of covariates used previously in Cook & DuPage Counties WNV model and those newly acquired variables used in newly revised 55 hexagon UFS model.

Previously existing data

Human illness

Human WNV cases in Illinois were classified as either confirmed or probable, as reported to the IDPH by public health or licensed medical professionals (mandatory reporting of WNV cases is required in the state). Human cases were converted into binary form (presence/absence of illness) and weekly case rate, controlling for human population, for each hexagon.

Abiotic Predictors

Thirty meter resolution land cover from the 2011 United States Geological Survey (21) National Land Cover Database (NLCD) provided 30 m resolution classified raster data for the NWMAD. There were 15 unique land cover types, ranging from various forests and vegetation to built up urban space. Weekly mean temperatures and weekly precipitation totals, acquired from the PRISM Climate Group (22), were extracted for each hexagon in this study using ArcGIS 10.5.1(23).

Newly added data

Abiotic Predictors

Catch basin density: Due to the high preference for breeding in catch basins (e.g. sewers) by *Culex pipiens,* the density of catch basins ($\frac{\#\ catch\ basins}{unique\ hexagon}$) was calculated and assessed. The NWMAD provided point data for each catch basin within its jurisdiction. All point data were then aggregated to each hexagon using the spatial location join feature in ArcGIS. A combined total of 8,443 catch basins were recorded among all hexagons (min = 1, max = 543).

Building and residential structures: Previous WNV studies in the Chicago area found a link between the density, size, and age of housing and human cases (9). Through high-resolution (1 m) aerial imagery from ArcGIS and USDA (2018), every permanent structure (e.g. residence,

shed, garage, deck) was traced and converted to polygons in ArcGIS. The area and perimeter of each polygon was calculated and aggregated for each hexagon. Commercial and residential lots were provided by Cook County Data Catalog (2019), using 2016 tax appropriations. In total, there were a combined 22,892 lots with 24,468 buildings or permanent structures.

Light pollution: A recent study by Kernbach et al. (2019) has linked increases in light pollution to WNV in the environment. Because the NWMAD consists of a metropolitan area with an abundance of artificial light, pollution values were evaluated. Light pollution was provided by the New World Atlas of Artificial Night Sky Brightness (27,28). Light pollution was acquired from 2014 data of the VIIRS DNB sensor on the Suomi National Polar-orbiting Partnership satellite. Pixel resolution was 0.75 km; mean value for each 1-km hexagon was calculated in ArcGIS.

Biotic Predictors

Historical mosquito abundance: The NWMAD consistently collected and diligently maintained their mosquito trapping and identification data throughout the study period. Once deployed, traps were usually checked at least twice a week. Over the 2005-2016 study period, there were a total of 59 traps used in the NWMAD, resulting in a total of 48,406 female *Culex.* spp. from 22 light traps, and 1,110,024 from 37 gravid traps. Weekly mosquito collections by trap were geocoded and interpolated across all hexagons via IDW and extracted using the zonal statistics as table function for each hexagon in ArcGIS. The regular maintenance, collection, and identification, frequency of mosquitoes caught, and distribution of traps within the NWMAD provided strong evidence that mosquitoes collected were representative for the remainder of the study area. Additionally, standard error values as a result of IDW methods were very low, and thus, the assumption for spatial dependency is satisfied (S1 Fig). Mosquito abundance was calculated as

the weekly cumulative number of captured female *Culex* spp. from each respective gravid trap (GT) and light trap (LT). Since *Cx. pipiens* and *Cx. restuans* are very difficult to morphologically identify, and with prior studies establishing these as the major *Culex*. species present in this area, all collected specimens from the genus *Culex* were pooled.

Normalized Difference Vegetation Index (NDVI): Trees and shrubs are a major source of nectar and serve as resting places for mosquitoes, especially those that recently blood-fed (29). To evaluate the magnitude of all vegetation, NDVI was incorporated by hexagon, recorded as an average value at three timepoints of each year: CDC epidemiologic weeks 21 (3^{rd}-4^{th} week of May), 28 (2^{nd}-3^{rd} week of July), and 35 (4^{th} week of August-1^{st} week of September). These CDC epidemiologic weeks mark the center of each the three 8-week active WNV periods in the Midwest, represented as T1 = low WNV activity, T2 = high WNV activity, and T3 = moderate WNV activity. The best available Landsat 7 or 8 bands for each respective time period were acquired from EarthExplorer (30) and processed in ArcGIS.

Human activity observations: To provide the most complete measurement of human risk to potential mosquito vectors in nature, this study attempted to quantify human exposure during crepuscular time periods. Human activity observations were conducted in public spaces inside each hexagon, during the crepuscular hours between 6-9:30pm, the preferred feeding period for *Cx. pipiens/restuans*. Observations were conducted within each hexagon for a total of ten minutes per visit. Specifically, a researcher remained stationary for 2 minutes, walked 2 minutes, remained stationary in the new position for 2 minutes, walked back to origination point for 2 minutes, then remained stationary in the original position for 2 final minutes. Human exposure was determined as any period in time a person was outside of any building, vehicle, or enclosed

dwelling during the observation period. Observations were classified by apparent gender and age (child, adult, or senior citizen).

Human landing catch (HLC): In conjunction with human activity observations, the number of human-seeking mosquitoes that attempted to blood-feed were collected via human landing catch methods for a fifteen-minute period at each hexagon weekly. To mitigate actual biting events, the researcher would expose only one limb (arm or leg) at a given time. Any mosquito that landed was collected via mechanical aspirator and transferred to a 2 ml collection vial. All collected mosquitoes were transported to the NWMAD within 2 hours and stored at -80°C. All mosquito specimens were identified to species within three days. Any mosquitoes identified as *Culex* spp. were sent to the Fritz Lab at the University of Maryland for species confirmation by *Cx. pipiens* group-specific primers via PCR.

Vector Index: The vector index (VI) was calculated as an estimate of the relative number of WNV-infected mosquitoes. For this study, VI was calculated as the average number of pooled *Culex* spp. collected per trap-week multiplied by the proportion of mosquitoes infected with WNV. The following equation was modified from the CDC (2013):

$$\mathrm{VI} = \sum_{i=Culex\ \mathrm{spp.(pooled)}} \overline{N}_i \hat{P}_i,$$

where $\overline{N}_i$ = average density (number of mosquitoes per trap week) and $\hat{P}_i$ = estimated MIR (proportion of mosquito pools testing positive for WNV). Calculated weekly VI for each trap by week was then interpolated via IDW method for estimations across the NWMAD.

Nuisance Factor and Human WNV Added Risk: The combination of human activity observations, serving as a proxy for potential mosquito bloodmeals, and HLC data, serving as a proxy for potential rate of mosquito biting, formed two unique WNV disease indices: the Nuisance Factor and Human WNV Added Risk. Since the majority of mosquitoes collected were

non-*Culex*, a quantitative index, nuisance factor, was created to provide a risk spectrum of encountering nuisance mosquitoes in a given hexagon. The following equation defines the nuisance factor:

$$Nuisance\ Factor = \frac{\frac{Human\ Observations}{Hour} * \frac{Nuisance\ Mosquitoes}{Hour}}{100}$$

Nuisance factor values ranged from a low of 0 to a high of 32.3. To quantitatively estimate potential risk for exposure to disease within a given hexagon, the human WNV added risk factor was created. This index is defined by the following equation:

$$Human\ WNV\ Added\ Risk = \frac{\frac{Human\ Observations}{Hour} * \frac{Culex\ \text{spp.}}{\text{Hour}}}{100}$$

Human WNV added risk ranged from a low of 0 to a high of 1.44.

Statistical methods

Location selection

Of the total 1019 hexagons within the NWMAD, fifty-five (5.4%) were selected as the maximum number of sites that our research team could visit for fifteen minutes each, weekly. The subset of fifty-five hexagons were selected based on two criteria: (1) human population was > 0, and (2) the previous Cook and DuPage model either predicted human WNV extremely well or extremely poorly, as determined by the 2005-2016 average residual output. Furthermore, the residual output was stratified by those locations that had or had not experienced a human case during the 12-year period. These processes created a performance spectrum consisting of five categories of hexagons: negative residuals without a human case (NR0), low residuals without a case (LR0), low residuals with a case (LR1), positive residuals without a case (PR0), and positive residuals with a case (PR1) (Table S1). No hexagons with negative residuals in the Cook

and DuPage model had experienced a human case. The spatial arrangements of these hexagons provide adequate coverage of the NWMAD's jurisdiction (Fig 2).

Fig 2. Location of the 55-hexagon study area within the Northwest Abatement District. Hexagons are labeled by field season visited for mosquito collections and human activity observations (color outline) and by total human cases from 2005-2016 (gray scale shaded interior).

Model Selection

Two seasons of field collections and processing of new data provided the UFS model with an additional 42 covariates not made available in the previous Cook & DuPage model. The generation of linear and logistic regression models began with a two-step selection process for the initial covariate inclusion: (1) conduct a univariate analysis with each predictor (independent variable) to the WNV disease outcome (binary = logistic, case rate = linear, dependent variable). Candidate variables for multivariate analysis were selected using slightly more conservative p-value than Bursac et al. (2008), p-value ≤ 0.20 vs. ≤ 0.25). Models that create cut-off values of p-value ≤ 0.1 for purposeful univariate covariate selection can erroneously prevent important variables from entering final models (33,34); (2) the final model, a generalized linear model with a Poisson distribution and probit link function, was selected using forward selection method, selecting the final model based on the Bayesian information criterion (BIC). Non-significant covariates were removed from the final model as a product of the iterative selection process. Secondarily, a receiver operating characteristic (ROC) curve was used to visualize overall model performance and Area Under the Curve (AUC) was calculated. All predictors were evaluated for multicollinearity using the PROC REG procedure (SAS Institute Inc. Cary, NC, USA) (S2

Table). Regression analyses were analyzed using the Fit Model feature in JMP 14.2.0 (SAS Institute Inc. Cary, NC, USA). Binary WNV case outcome was analyzed with as a nominal logistic personality. The continuous WNV case rate outcome was analyzed as a standard least squares personality.

Model Comparisons

Human WNV illness in the NWMAD was assessed under four model environments, each expressing a defined set of specific parameters. The four model environments were:

1. MIR & Mosquito Abundance (contains no VI covariates),
2. Vector Index (contains no MIR or mosquito abundance covariates),
3. Best-Fit (best fit with all covariates in respective assessment), and
4. Global (all covariates made available in respective assessment)

As a comparison, the original Cook & DuPage model (Karki et al. 2020) was fit using only 40 covariates. Each of these four model environments were assessed using four different covariate sets:

1. All covariates (82 available covariates),
2. Excluding HLC and human observations covariates (74 available covariates),
3. Force-fitting HLC and human observations covariates (8 forced covariates, 82 available covariates), and
4. Only the covariates made available to the Cook & DuPage 2019 model (control model, 40 available covariates).

Under each model environment and covariate set, the outcome of human WNV illness was analyzed using:

1. Logistic regression (presence/absence human WNV illness) and

2. Linear regression (WNV case rate) methods.

In total, there were 36 models assessed (Fig 3); models are named using the convention $E_xC_yO_z$, where x is the model environment number (0-4, with number 0 assigned to the control environment), C is the covariate set number (1-4), and O is the outcome number (1-2). For both logistic and linear regression, each of the four model environments was fit using each of the four covariate sets. In addition, the control models using only the covariates from the final Cook & DuPage model applied to the UFS region were fit with and without force fitting HLC and human observation covariates.

Fig 3. Flow diagram displaying how models were characterized, assembled, and compared in this study. Global models failed to converge and were excluded from the final results. The control model (optimal Cook & DuPage Counties (2019) model) was only used as a comparison for covariates made available only to that original model. Of the original 36 models initially assessed, 8 were removed, resulting in 28 final models assessed in this study.

Half of the models were assessed under logistic and linear outcomes, respectively, and based on the *# of Significant Covariates* (quantity of variables included in final model with $p<0.05$) and *Degrees of Freedom* (the number of values in the final model that are free to vary). Overall model performance was determined by BIC. While BIC and Aikake's Information Criterion (AIC) are both maximum likelihood estimators, BIC was chosen to determine model strength due to its stronger penalty term for covariate inclusion (35).

Covariate Performance

Similarly to the model performance index, to evaluate the performance for all covariates across 18 logistic and 18 linear models, each of the 82 covariates were standardized by creating the following index:

$$\bar{p}_{\text{Covariate}} = \frac{Significance\ Level}{Data\ Availability}$$

where: *Significance Level* = significance level of covariate in each of the 36 final models ($p<0.001 = 4$, $p<0.01 = 3$, $p<0.05 = 2$, included in the final model = 1), and *Data Availability* = resources required to acquire a respective covariate (level 1 = data widely available, no processing needed, level 2 = data available, requires minimal to moderate processing/analyses, level 3 = data available, requires extensive processing/analyses, level 4 = data not available, needs to be collected, processed, and analyzed, S3 Table). The final net prediction:availability tradeoff used to create the Data Availability variable are categorical and based on the authors' personal experiences with data used in this study.

Results

Model Comparison

The highest performing WNV human risk models were $E_3C_4O_3$ (Cook & DuPage Best Fit, df = 8, BIC = -227444) and $E_2C_4O_1$ (Cook & DuPage + VI, df = 14, BIC = 576.2), for linear and logistic regressions, respectively (S4 & S5 Tables).

The top five models that predicted human WNV cases strongest were represented by the control (E_0, n=2), best-fit (E_3, n=2) and vector index (E_2, n=1) environments (Fig 4B, Table 2). These models' corresponding covariate sets were represented by variables only available to the original Cook & DuPage models (C_4, n=4), and force-fitting HLC covariates (C_3, n=1) environments.

Fig 4. Overall performance for each predictor and final model used in this study. Each of the 70 covariates used in the study, listed in alphabetic order by data availability/work load to acquire score (1-4), were evaluated by mean performance (A). Highest performing covariates are noted by enlarged label text and darker blue bar color. The overall performance for each linear and logistic model (n=14 for both) was evaluated by BIC value (B). Means for each outcome ($\bar{x}_{\text{covariate}}$ = 0.48; $\bar{x}_{\text{linear}}$ = -193406; $\bar{x}_{\text{logistic}}$ = 670.9) are designated by vertical dashed lines. Details of scoring for each covariate and model are provided in 3 & S3 Tables.

348 **Table 2. Overall assessment of performance for each model environment, covariate, and outcome combination.**

Model[a]	Cumulative Significance Total	# of Significant Covariates	DF (lower is better)	BIC value	BIC value (lower is better) rank	Regression Type
E0C3O2	6	5	21	-227300	3	Linear
E0C4O2	6	5	15	-227354	2	
E1C1O2	7	4	13	-182037	11	
E1C2O2	7	4	13	-182037	12	
E1C3O2	7	4	19	-181982	14	
E1C4O2	7	5	17	-182001	13	
E2C1O2	8	4	8	-185362	7	
E2C2O2	10	4	7	-185373	6	
E2C3O2	7	3	12	-185322	10	
E2C4O2	10	5	9	-185347	8	
E3C1O2	6	2	4	-185389	5	
E3C2O2	8	3	4	-185395	4	
E3C3O2	6	2	9	-185344	9	
E3C4O2	10	6	8	-227444	1	
E4C1O2						
E4C2O2			Global Models Excluded			
E4C3O2						
E4C4O2						
E0C3O1	4	4	21	683.40	9	Logistic
E0C4O1	3	3	15	632.30	3	
E1C1O1	4	3	30	742.50	12	
E1C2O1	5	4	34	768.70	14	
E1C3O1	9	7	32	757.70	13	
E1C4O1	7	6	21	653.30	6	
E2C1O1	4	3	26	692.70	10	
E2C2O1	7	4	23	661.10	7	
E2C3O1	4	3	26	696.60	11	
E2C4O1	5	5	14	576.20	1	
E3C1O1	5	3	19	634.60	4	
E3C2O1	8	5	21	640.40	5	
E3C3O1	5	3	23	672.70	8	
E3C4O1	8	6	12	580.80	2	
E4C1O1						
E4C2O1			Global Models Excluded			
E4C3O1						
E4C4O1						

Detailed assessment of each model evaluated in this study. Overall model strength was determined by BIC value (by linear and logistic regression types), with the following characteristics denoted as follows: Cumulative Significance Total, sum of each variable score, denoted as: $p<0.001 = 4$, $p<0.01 = 3$, $p<0.05 = 2$, included in model = 1; # of Significant Covariates = summation of included covariates with p-value<0.05; DF = degrees of freedom denoted in model; BIC value = overall model rank (best model = 1, worst model = 14)/14 for each logistic and linear model group, respectively.

[a]All global models were excluded from analysis as they were all overfit and statistically biased

Covariate Performance

Of the 82 available covariates, 70 (85.4%) were included at least once among a given model, excluding the overfit global models (individual predictor summaries located in S6 Table). Of the 41 covariates (58.6%) that were greater than the mean covariate performance, seven were highly efficient (determined by natural break in the distribution), providing a crude estimation as most valuable variables for human WNV estimation (Fig 4A). These covariates are provided here in descending order of most importance: tempc (temperature (°C), $\bar{p}$ = 1.15), preci (precipitation (mm), $\bar{p}$ = 1.14), Yr (year, $\bar{p}$ = 1.0), templag3 (temperature lagged by 3 weeks, $\bar{p}$ = 0.92), blpct (barren land (%), $\bar{p}$ = 0.92), precilag1 (precipitation lagged by 1 week, $\bar{p}$ = 0.90), and VIlag4 (vector index 4 weeks prior, $\bar{p}$ = 0.88). All eight HLC and human observation covariates were included in at least one final model, but none performed highly ($\bar{p}_{\text{each HLC Covariate}}$ = 0.25). Estimates and calculations for individual covariates are available in S3 Table.

The eight HLC and human observation covariates provided significant differences ($P \leq 0.05$) in observations and mosquito collections by hexagon type (Figs 5A & 5B). The indices, nuisance mosquito exposure and human WNV added risk, significantly differed by hexagon type (Fig 5C). Hexagons designated as PR1 (positive residual (underpredicted actual cases) with a prior human WNV case) were found to have the most human observations and collected mosquitoes (from both *Culex* and non-*Culex* spp.) per visit. This combination of factors provides hexagons among the PR1 type as the most "risky" in regard to human WNV added risk and increased nuisance mosquito exposure (Fig 5).

Fig 5. Relationship of hexagon by type. Hexagon type (LR = low residual, PR = positive residual, NR = large, negative residual; 0 = no human case, 1 = human case) are detailed by

372 human observations per visit (A), mosquitoes collected per visit (B), and a product of the two
373 former variables, nuisance factor and WNV added risk (C). Letters above each box and whisker
374 plot designate significantly different groups by hexagon type, as calculated by Tukey's HSD.

375 Discussion

376 In addition to model comparisons, this study evaluated the performance of the newly acquired
377 VI in comparison to the more commonly used MIR in combination with mosquito abundance.
378 Overall, when fit to the UFS study area, adding mosquito abundance and associated 4-week lags
379 improved this model. When evaluating WNV prediction as a linear outcome, the best-fit model
380 using only covariates available to the original Cook & DuPage model was the highest performing
381 in WNV predictability. However, when evaluating WNV prediction as a binary outcome, VI and
382 its associated 4-week lags replaced MIR as the best predictor of human WNV. While no model
383 emphasizing MIR and abundance was selected as one of the best predictive models, at least one
384 of these variables (and their associated lags) were represented in 4 of the 5 best models (control
385 and best-fit, n=2 for each model). On the contrary, VI, as an emphasized model environment,
386 was selected as the best performing logistic model. Both MIR and VI are critical components in
387 predicting WNV. Under ideal settings, the VI is the preferred method for estimating the risk of
388 mosquito infection, as opposed to MIR. However, deciding between the two biological indicators
389 will be largely dependent upon the data availability for each model of interest. Our study
390 suggests that if resources are limited, net model value leans in favor of using MIR.
391 The addition of 42 new covariates required a significant allocation of resources but provided
392 minimal benefits towards reducing variance in human WNV prediction. Fortunately, this study
393 suggests that excellent disease prediction models can be achieved with conventional covariates
394 that are publicly available, requiring little to no processing and/or analyses (data availability

scores ≤ 2, Fig 4B). However, any covariate used should be adjusted and properly designed for the highest spatial and/or temporal resolution possible, which may require additional efforts to accomplish.

Extensive review of literature indicated no other studies have evaluated covariate strength given limited resources, particularly in the context of making decisions to acquire data. Therefore, the categorizations of covariates by resource allocation (values ranging from 1 (low) to 4 (high)) are based on the experiences of the authors during this study. These values are subjective and may vary across institution or research group, but they may be used as a general estimation in model selection and decision-making. For example, variables related to building and lot size (avg bldg. area: avg lot area, bldg. footprint area avg, bldg. footprint area total, bldg. footprint peri avg., bldg. footprint peri total, and total bldg. area: total lot area) were all ranked a value of 4 because of extensive data processing and review. The authors downloaded high resolution, cloud-free satellite images that were used as a basemap for digital tracing of every building structure (houses, businesses, sheds, detached garages, storage units, etc.) and lots (residential and commercial). This resulted in >47,000 structures and lots digitally traced manually. On the other hand, weather variables (e.g. preci, tempc) were ranked a value of 1 because very little resources were devoted to have the data in a "ready" state. The source of these data, PRISM Climate Group, allows for monthly summaries to be downloaded and extracted with one quick geostatistic process.

This study also aimed to address a key missing index that few studies have evaluated: the relationship of human activity, mosquito exposure, and WNV disease risk. While the related variables did not greatly impact overall model strength, they did provide key insight into a potential key in WNV ecology – the areas that were previously underpredicted with recorded

418 human WNV (hex type: PR1) were consistently found to have the most human activity at
419 crepuscular times, the most mosquitoes overall, and the most *Culex* mosquitoes. However, our
420 results appear to contradict the findings of Read et al. (1994), who discovered that as reports of
421 biting nuisance mosquitoes increased beyond 2 per minute, outdoor human activity rapidly
422 declined. Our results indicate that as mosquito collections increased, human observations also
423 increased (Fig 5). Not only is this a potentially dangerous combination that can foster
424 environments ideal to mosquito-human spillover, previous modeling efforts failed to capture
425 these cases. Future directions will target these highly susceptible locations and aim to capture
426 any additional unaccounted variance.
427 Like all disease modeling efforts, there are always reporting biases that directly affect true case
428 prevalence. Unfortunately, many vector-borne diseases are largely underreported (37–40), as
429 human cases are vastly overlooked or misdiagnosed, largely due to low severity in disease
430 manifestation in a majority of cases (41,42). This creates difficulties in predicting when and
431 where VBD incidence will arise. Specifically regarding WNV, it is estimated that about 80% of
432 human infections are unreported, as clinical signs are minor or asymptomatic (43,44). The
433 remaining 20% of humans develop West Nile fever, and among this group, about 1% will
434 develop severe and sometimes fatal neuroinvasive disease. In the Chicago area, models in both
435 the UFS and Cook & DuPage locations have very high human WNV prediction capabilities.
436 Despite having among the highest total number of human WNV cases in the U.S. (20), this
437 region has more observational units denoted as non-cases than cases. That has resulted in models
438 with excellent accuracy in predicting where there are no human cases, thus inflating the true
439 accuracy of our models. Nonetheless, while our models are able to reliably predict where human

cases are present, the magnitude of effect can be missed (e.g. “hot spots” with greater than 1 case may not be represented).

Disease modelers need to be cognizant of saturating their efforts, both statistically and biologically. Statistically, additional and meaningful covariates will usually improve model fit parameters. However, the inclusion of too many variables can result in overfitting, resulting in models failing to converge (45–47). It is possible that no matter the amount of effort to improve model fit, there is an element of variability attributed with infected humans not seeking medical attention and thus, reducing true disease prevalence (48).

Overall, when compared to the Cook & DuPage model, the best UFS models required fewer predictors and produced a stronger overall fit using most, if not all, the same covariates made available to both model types. Spending the resources (time, money, human-power, processing, analyses, logistic, etc.) to acquire additional covariates may not necessarily be worth the impact on improving human WNV modeling predictions. Rather, fine-tuning the traditional covariates (climatic, weather, and MIR, for example), to the highest spatiotemporal resolution possible may be the most efficient use of resources to minimize variance in VBD prediction models.

Conclusions

1. The factors and their overall effect on the prediction of human WNV cases differs across scale. Although improved, in comparison to the control Cook & DuPage model applied to the same study region, the "best fit" UFS model AUC = 0.89, suggesting newly unaccounted variances are present.
2. Both vector index and MIR contribute to high performing human WNV prediction models under UFS study areas. In direct comparison, VI is favorable to MIR. However, given limited resources in acquiring and processing additional data, MIR is more efficient for predicting human WNV illness.
3. The effort and resources required to acquire additional covariates, most of which are not publicly available, demonstrate a slight improvement in model prediction and appear less important in reducing variance.
4. In addition to the conventional WNV covariates, namely weather and infection rates, land-use and land-cover and SES/demographic information are widely available with little to no processing or analyses required, and provide the breadth to develop excellent prediction models. However, any covariate utilized must be structured at the finest spatial and/or temporal resolution possible.
5. Human exposure to mosquito biting rates provided minimal benefits to model prediction. More importantly however, these two covariates provided potentially key insight to the susceptibility of humans in locations where WNV is prevalent. Additionally, where WNV is less of a concern, these results provide insight into nuisance mosquito exposure that may lead to improvements in targeted control efforts.

Declarations

Acknowledgements

The authors would like to thank Roger Nasci for providing expert insight in vector biology and modeling efforts, Dan Bartlett for providing geospatial data and detailed field information, the Megan Fritz lab for providing confirmatory *Culex* identification, and Chris Stone and Andrew Mackay from the Illinois Medical Entomology lab for providing guidance and expertise with human landing catch methodology. Dr. Marilyn O'Hara Ruiz passed away before the submission of the final version of this manuscript. The corresponding author, Dr. Johnny Uelmen, accepts responsibility for the integrity and validity of the data collected and analyzed.

Availability of Data and Materials

The dataset supporting the conclusions of this article is available in the University of Illinois repository, https://doi.org/10.13012/B2IDB-5901636_V1.

Authors' contributions

JAU conceived the presented idea, collected field samples and provided data analysis and processing. PI provided research assistance and expertise in mosquito collection and biology. SK provided expertise and additional datasets for analysis. WMB provided analytical assistance and provided data sources. BL provided statistical oversight and expertise. MOR provided the initial product idea, planning, and supervision. RLS provided oversight of all aspects throughout the study. All authors discussed the results and contributed to the final manuscript.

Funding

This publication was supported by Cooperative Agreement #U01 CK000505, funded by the Centers for Disease Control and Prevention. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Centers of Disease Control and Prevention or the Department of Health and Human Services.

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

References

1. Sejvar JJ. West Nile virus: An historical overview. Ochsner J. 2003;5(3):6–10.

2. Kramer LD, Styer LM, Ebel GD. A Global Perspective on the Epidemiology of West Nile Virus. Annu Rev Entomol. 2008;53(1):61–81.

3. Goddard LB, Roth AE, Reisen WK, Scott TW. Vector competence of California mosquitoes for West Nile virus. Emerg Infect Dis. 2002;8(12):1385–91.

4. Hamer GL, Kitron UD, Goldberg TL, Brawn JD, Loss SR, Ruiz MO, et al. Host selection by *Culex pipiens* mosquitoes and west nile virus amplification. Am J Trop Med Hyg. 2009;80(2):268–78.

5. Russell C, Hunter FF. *Culex pipiens* (Culicidae) is attracted to humans in southern Ontario, but will it serve as a bridge vector of West Nile virus? Can Entomol. 2012;

6. Kilpatrick AM, Kramer LD, Jones MJ, Marra PP, Daszak P. West Nile virus epidemics in North America are driven by shifts in mosquito feeding behavior. PLoS Biol. 2006;4(4):606–10.

7. CDC. National Notifiable Diseases Surveillance Systems (NNDSS): MMWR Week Fact Sheet. 2019.

8. Cohen JM, Civitello DJ, Brace AJ, Feichtinger EM, Ortega CN, Richardson JC, et al. Spatial scale modulates the strength of ecological processes driving disease distributions. Proc Natl Acad Sci U S A. 2016;113(24):E3359–64.

9. Ruiz MO, Tedesco C, McTighe TJ, Austin C, Uriel K. Environmental and social

525 determinants of human risk during a West Nile virus outbreak in the greater Chicago area,
526 2002. Int J Health Geogr. 2004;3(May).

527 10. Winters AM, Eisen RJ, Delorey MJ, Fischer M, Nasci RS, Zielinski-Gutierrez E, et al.
528 Spatial risk assessments based on vector-borne disease epidemiologic data: Importance of
529 scale for West Nile virus disease in Colorado. Am J Trop Med Hyg. 2010;

530 11. Kilpatrick AM, Pape WJ. Predicting human west nile virus infections with mosquito
531 surveillance data. Am J Epidemiol. 2013;178(5):829–35.

532 12. Manore CA, Davis JK, Christofferson RC, Wesson DM, Hyman JM, Mores CN. Towards
533 an Early Warning System for Forecasting Human West Nile Virus Incidence. PLoS Curr.
534 2014;1–21.

535 13. Roiz D, Ruiz S, Soriguer R, Figuerola J. Climatic effects on mosquito abundance in
536 Mediterranean wetlands. Parasites and Vectors. 2014;7(1):1–13.

537 14. Rosà R, Marini G, Bolzoni L, Neteler M, Metz M, Delucchi L, et al. Early warning of
538 West Nile virus mosquito vector: Climate and land use models successfully explain
539 phenology and abundance of *Culex pipiens* mosquitoes in north-western Italy. Parasites
540 and Vectors. 2014;7(1):1–12.

541 15. Wimberly MC, Lamsal A, Giacomo P, Chuang TW. Regional variation of climatic
542 influences on West Nile virus outbreaks in the United States. Am J Trop Med Hyg.
543 2014;91(4):677–84.

544 16. Hahn MB, Monaghan AJ, Hayden MH, Eisen RJ, Delorey MJ, Lindsey NP, et al.
545 Meteorological conditions associated with increased incidence of west nile virus disease

in the United States, 2004-2012. Am J Trop Med Hyg. 2015;92(5):1013–22.

17. Giordano B V., Kaur S, Hunter FF. West Nile virus in Ontario, Canada: A twelve-year analysis of human case prevalence, mosquito surveillance, and climate data. PLoS One. 2017;12(8):1–15.

18. Karki S, Brown WM, Uelmen J, O'Hara Ruiz M, Smith RL. The drivers of West Nile virus human illness in the Chicago, Illinois, USA area: Fine scale dynamic effects of weather, mosquito infection, social, and biological conditions. PLoS One. 2020;

19. Ruiz MO, Chaves LF, Hamer GL, Sun T, Brown WM, Walker ED, et al. Local impact of temperature and precipitation on West Nile virus infection in Culex species mosquitoes in northeast Illinois, USA. Parasites and Vectors. 2010;

20. CDC. National arboviral surveillance system (ArboNET), Arboviral Diseases Branch, Centers for Disease Control and Prevention. Fort Collins, CO; 2020.

21. USGS. National Geospatial Data Asset (NGDA) Land Use Land Cover. 2011.

22. Oregon State University. PRISM Climate Group [Internet]. [cited 2019 Nov 12]. Available from: http://prism.oregonstate.edu

23. Environmental Systems Research Insititute. ArcGIS Desktop. Redlands, CA; 2011.

24. USDA. National Agriculture Imagery Program (NAIP) Imagery [Internet]. 2018 [cited 2018 Dec 11]. Available from: https://catalog.data.gov/dataset/national-geospatial-data-asset-ngda-naip-imagery

25. County C. Cook County Government Open Data. Cook County Data Catalog. 2019.

566 26. Kernbach ME, Newhouse DJ, Miller JM, Hall RJ, Gibbons J, Oberstaller J, et al. Light
567 pollution increases West Nile virus competence of a ubiquitous passerine reservoir species.
568 Proc R Soc B Biol Sci. 2019;286(1907).

569 27. Falchi F, Cinzano P, Duriscoe D, Kyba CCM, Elvidge CD, Baugh K, et al. The new world
570 atlas of artificial night sky brightness. Sci Adv. 2016;

571 28. Falchi, F., Cinzano, P., Duriscoe, D., Kyba, C. C. M., Elvidge, C. D., Baugh, K., Portnov,
572 B., Rybnikova, N. A., Furgoni R. Supplement to the New World Atlas of Artificial Night
573 Sky Brightness. GFZ Data Serv. 2016;

574 29. Barredo E, DeGennaro M. Not Just from Blood: Mosquito Nutrient Acquisition from
575 Nectar Sources. Trends in Parasitology. 2020.

576 30. USGS. EarthExplorer [Internet]. 2019 [cited 2019 Dec 17]. Available from:
577 https://earthexplorer.usgs.gov/

578 31. CDC. West Nile Virus in the United States: Guidelines for Surveillance, Prevention, and
579 Control. Appendix 2: Calculation and Application of a Vector Index (VI) Reflecting the
580 Number of WN Virus Infected Mosquitoes in a Population. 4th Revision. 2013.

581 32. Bursac Z, Gauss CH, Williams DK, Hosmer DW. Purposeful selection of variables in
582 logistic regression. Source Code Biol Med. 2008;3:1–8.

583 33. Bendel RB, Afifi AA. Comparison of stopping rules in forward “stepwise” regression. J
584 Am Stat Assoc. 1977;72(357):46–53.

585 34. Greenland S, Mickey RM. The impact of confounder selection criteria on effect estimation.
586 Am J Epidemiol. 1989;130(5):1066.

587 35. Schwarz G. Estimating the Dimension of a Model. Ann Stat. 1978;6:461–4.

588 36. Read NR, Rooker JR, Gathman JP. Public perception of mosquito annoyance measured by
589 a survey and simultaneous mosquito sampling. J Am Mosq Control Assoc.
590 1994;10(1):79–87.

591 37. Bowden SE, Magori K, Drake JM. Regional differences in the association between land
592 cover and West Nile virus disease incidence in humans in the United States. Am J Trop
593 Med Hyg. 2011;84(2):234–8.

594 38. Nelson CA, Saha S, Kugeler KJ, Delorey MJ, Shankar MB, Hinckley AF, et al. Incidence
595 of clinician-diagnosed lyme disease, United States, 2005–2010. Emerg Infect Dis.
596 2015;21(9):1625–31.

597 39. Waterman SH, Margolis HS, Sejvar JJ. Surveillance for dengue and dengue-associated
598 neurologic syndromes in the United States. Am J Trop Med Hyg. 2015;92(5):996–8.

599 40. CDC. WNV Symptoms, Diagnosis, & Treatment [Internet]. 2018. Available from:
600 https://www.cdc.gov/westnile/symptoms/index.html

601 41. CDC. National Notifiable Disease Surveillance System (NNDSS): Arboviral Diseases,
602 Neuroinvasive and Non-neuroinvasive 2015 Case Definition. 2015.

603 42. Rosenberg R, Lindsey NP, Fischer M, Gregory CJ, Hinckley AF, Mead PS, et al. Vital
604 signs: Trends in reported vectorborne disease cases — United States and Territories,
605 2004-2016. Morb Mortal Wkly Rep. 2018;67(17):496–501.

606 43. Centers for Disease Control and Prevention (CDC). Morbidity and Mortality Weekly
607 Report Surveillance for Human West Nile Virus Disease — United States , 1999 – 2008.

MMWR Wkly Rep. 2010;59(31):1999–2008.

44. Petersen LR, Brault AC, Nasci RS. West Nile virus: Review of the literature. JAMA - J Am Med Assoc. 2013;310(3):308–15.

45. Babyak MA. What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. Psychosom Med. 2004;66(3):411–21.

46. Lever J, Krzywinski M, Altman N. Points of Significance: Model selection and overfitting. Nat Methods. 2016;13(9):703–4.

47. Hawkins DM. The Problem of Overfitting. J Chem Inf Comput Sci. 2004;44(1):1–12.

48. Petersen LR, Carson PJ, Biggerstaff BJ, Custer B, Borchardt SM, Busch MP. Estimated cumulative incidence of West Nile virus infection in US adults, 1999-2010. Epidemiol Infect. 2013;141(3):591–5.

Supporting Information

S1 Fig. Measurement of standard error associated with interpolated *Culex* species abundance by light (A) and gravid (B) traps (averaged for all traps), and mean MIR (C) for each of the 55 hexagons, from 2005-2016. The average weekly mosquito abundance multiplied by average weekly mean MIR created a third infection parameter, the vector index.

S2 Table. Correlation matrices for each of the 82 covariates assessed in this study. Tables are grouped by anthropogenic (A), biological (B), environmental (C), weather (D), or other (E), as indicated in Table 1.

S3 Table. Detail of scoring and categorization of covariates used across all models assessed in this study, organized by data availability/work load to acquire score. Scoring was denoted as follows: Cumulative Significance, $p<0.001 = 4$, $p<0.01 = 3$, $p <0.05 =2$, included in model = 1; Data Availability/Work Load to Acquire, Data Unavailable/Fieldwork required = 4, Data available, but requires many resources to use = 3, Data available, but requires moderate resources to acquire = 2, Data readily available and requires little to no resources to use = 1; Covariate Value = Quotient of previous two columns.

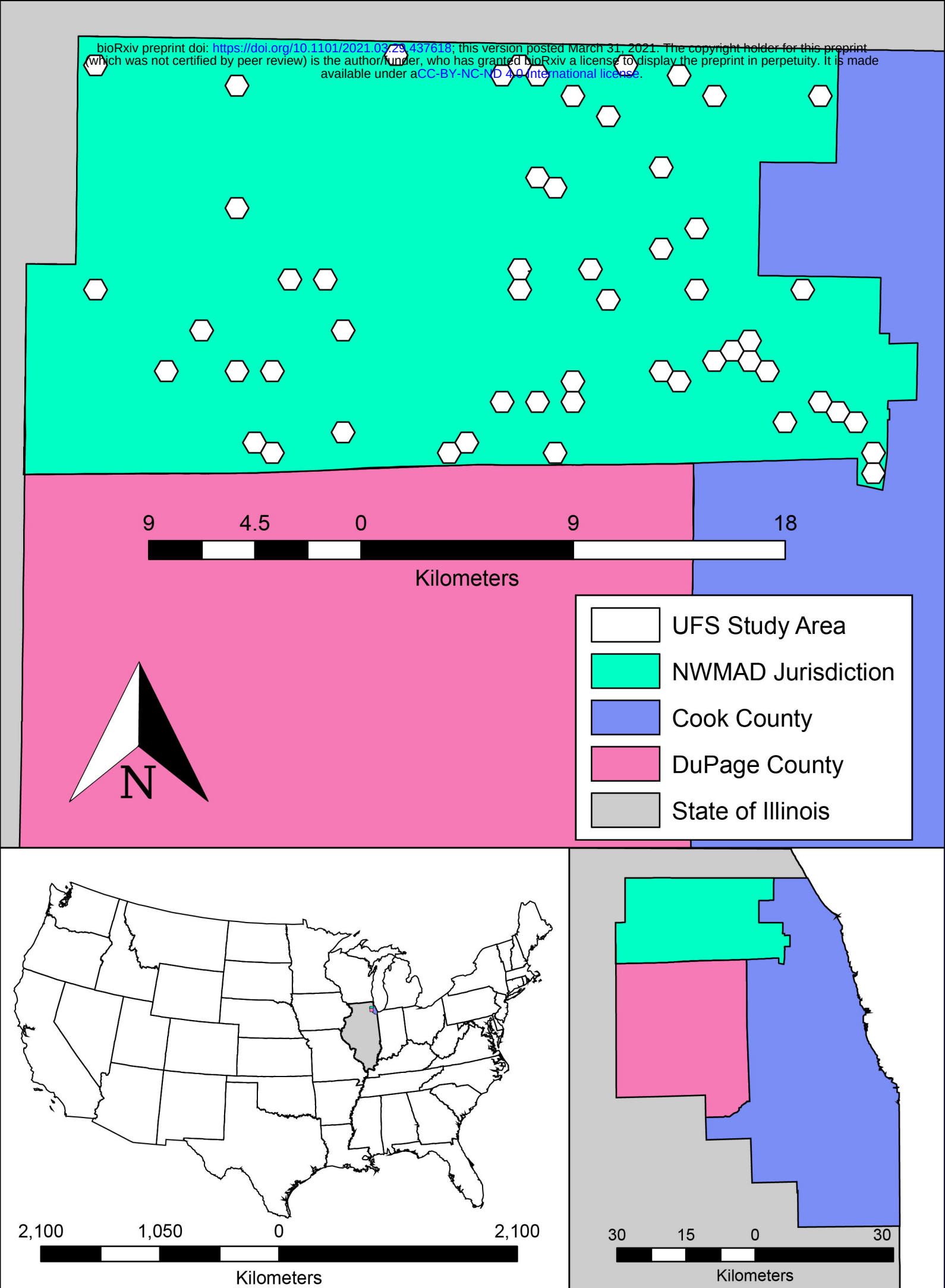
S4 Table. Model fit comparisons of the UFS hexagons, applying (A) newly acquired data (excluding HLC and human observations, covariate set 2), or (B) only the covariates made available to the previously published Cook & DuPage model (covariate set 4). Each model outcome was assessed using logistic (presence/absence WNV human illness case) and generalized linear (WNV case rates, controlling for human population) methods. Asterisks indicate level of statistical significance (* = $p \leq 0.05$, ** = $p \leq 0.001$, *** = $p \leq 0.0001$

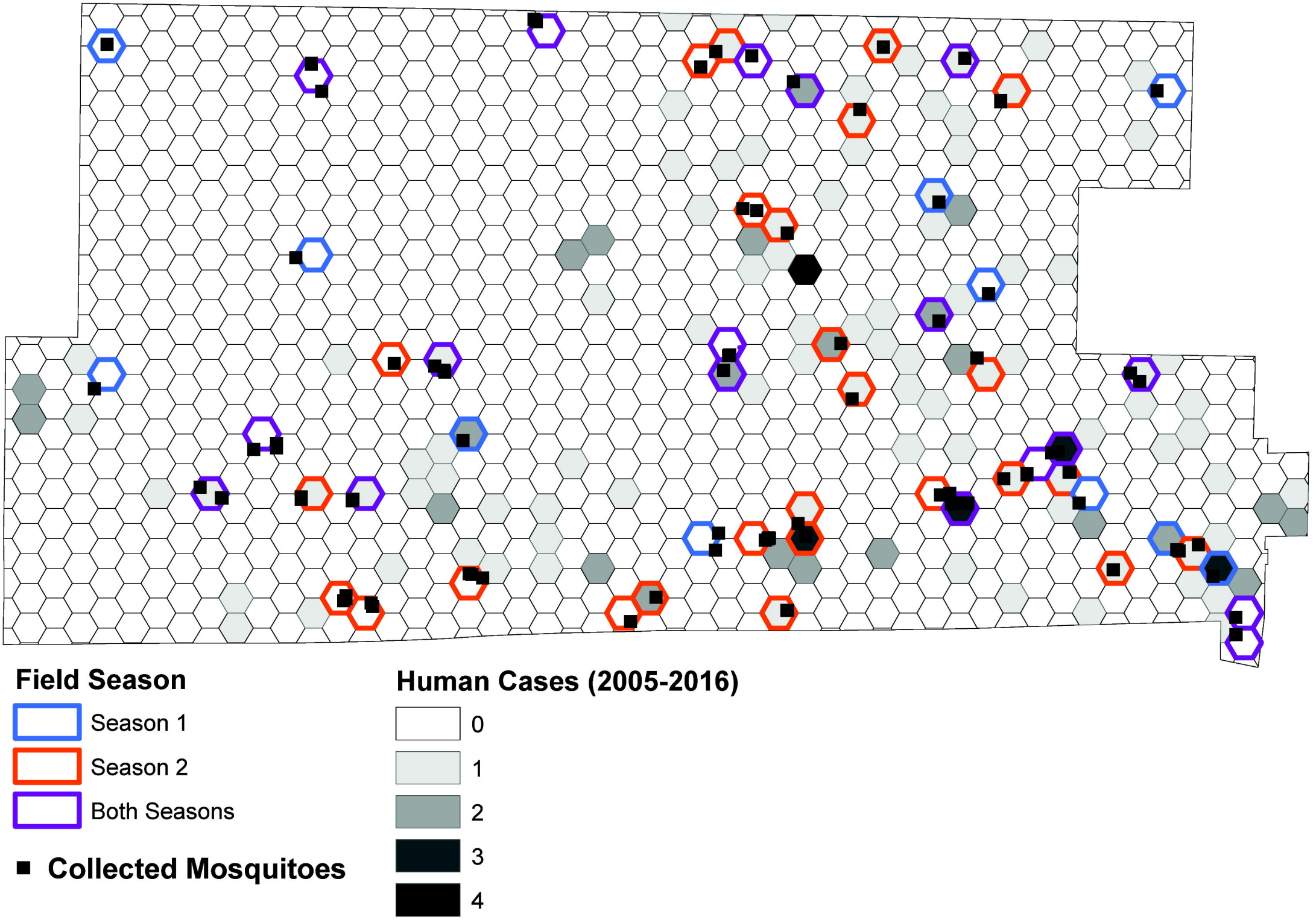
641 [a]Logistic regression outcome = human WNV presence/absence per hexagon, per week; GLM outcome = WNV
642 human case rate (per hexagon, per week).
643 [b]ROC applies to only logistic regression
644 [c]As the final selected model in the Original Cook & DuPage paper (2019), this model environment was assessed
645 only for the comparison to the Cook & DuPage models for this study and not applied to the UFS model. The original
646 model covariates, eftpct and ehwpct, have 0 observations among the selected 55 hexagons and were removed.

647 **S5 Table. Model fit comparisons of the UFS hexagons, using best-fit models with additional**
648 **human landing catch and human activity observations to incorporate added human risk.**
649 Human risk covariates were added to the UFS model by (**A**) best-fit integration (covariate set 1)
650 and (**B**) force-fitting (covariate set 3). Asterisks indicate level of statistical significance (* = $p \leq$
651 0.05, ** = $p \leq 0.001$, *** = $p \leq 0.0001$

652 [a]Logistic regression outcome = human WNV presence/absence per hexagon, per week; GLM outcome = WNV
653 human case rate (per hexagon, per week).
654 [b]ROC applies to only logistic regression
655 [c]As the final selected model in the Original Cook & DuPage paper (2019), this model environment was assessed
656 only for the comparison to the Cook & DuPage models for this study and not applied to the UFS model. The original
657 model covariates, eftpct and ehwpct, have 0 observations among the selected 55 hexagons and were removed.

658 **S6 Table. Mean and standard error values for each predictor evaluated in this study.**
659 Values represent averages for all hexagons (pooled) over the entire study period (2005-2016).




Field Season
Season 1
Season 2
Both Seasons
Collected Mosquitoes
Human Cases (2005-2016)
0
1
2
3
4

Model Environment

(E_x) X 3.5

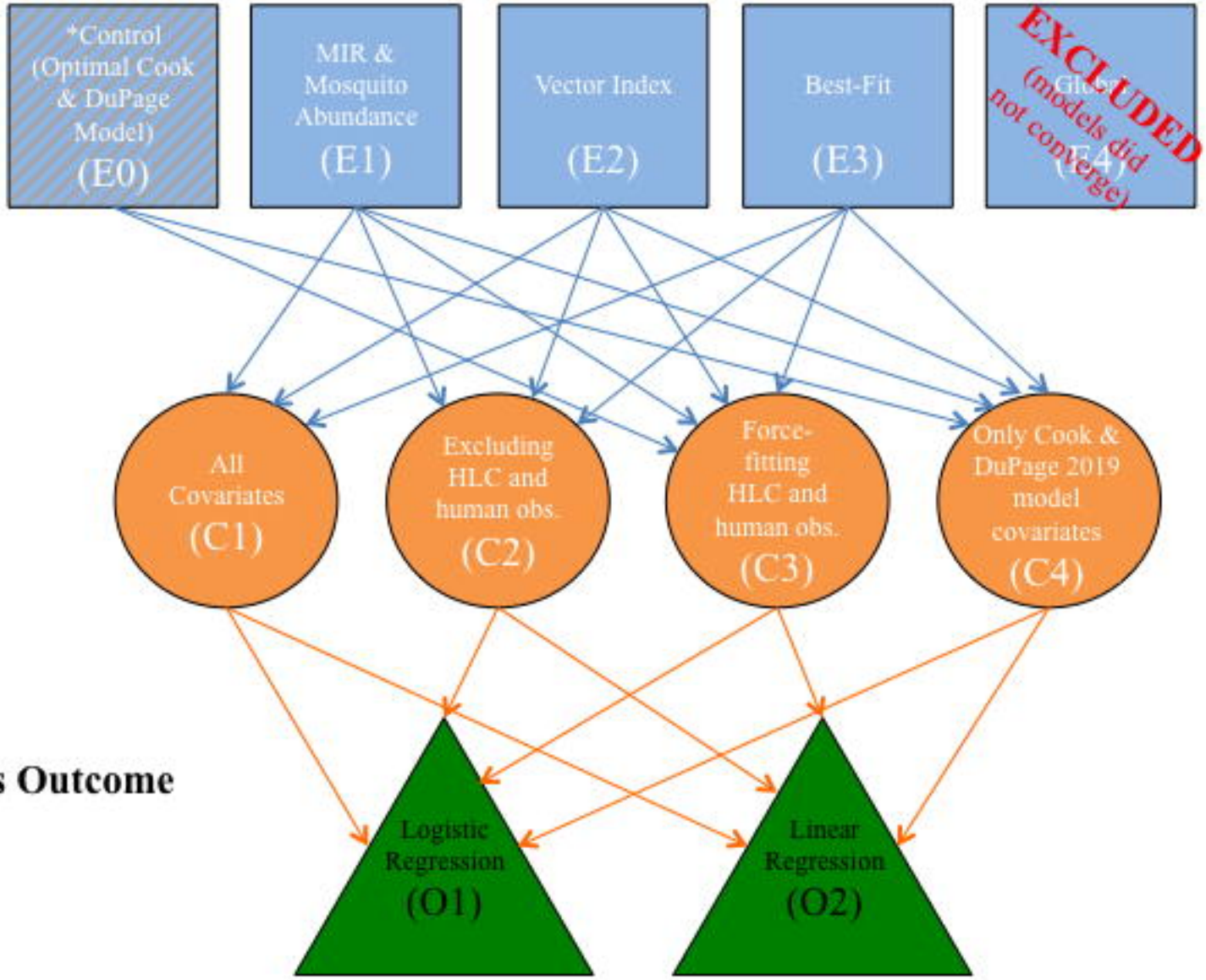
*Control used only for comparison to Cook & DuPage model covariate sets

Covariate Set

(C_y) X 4

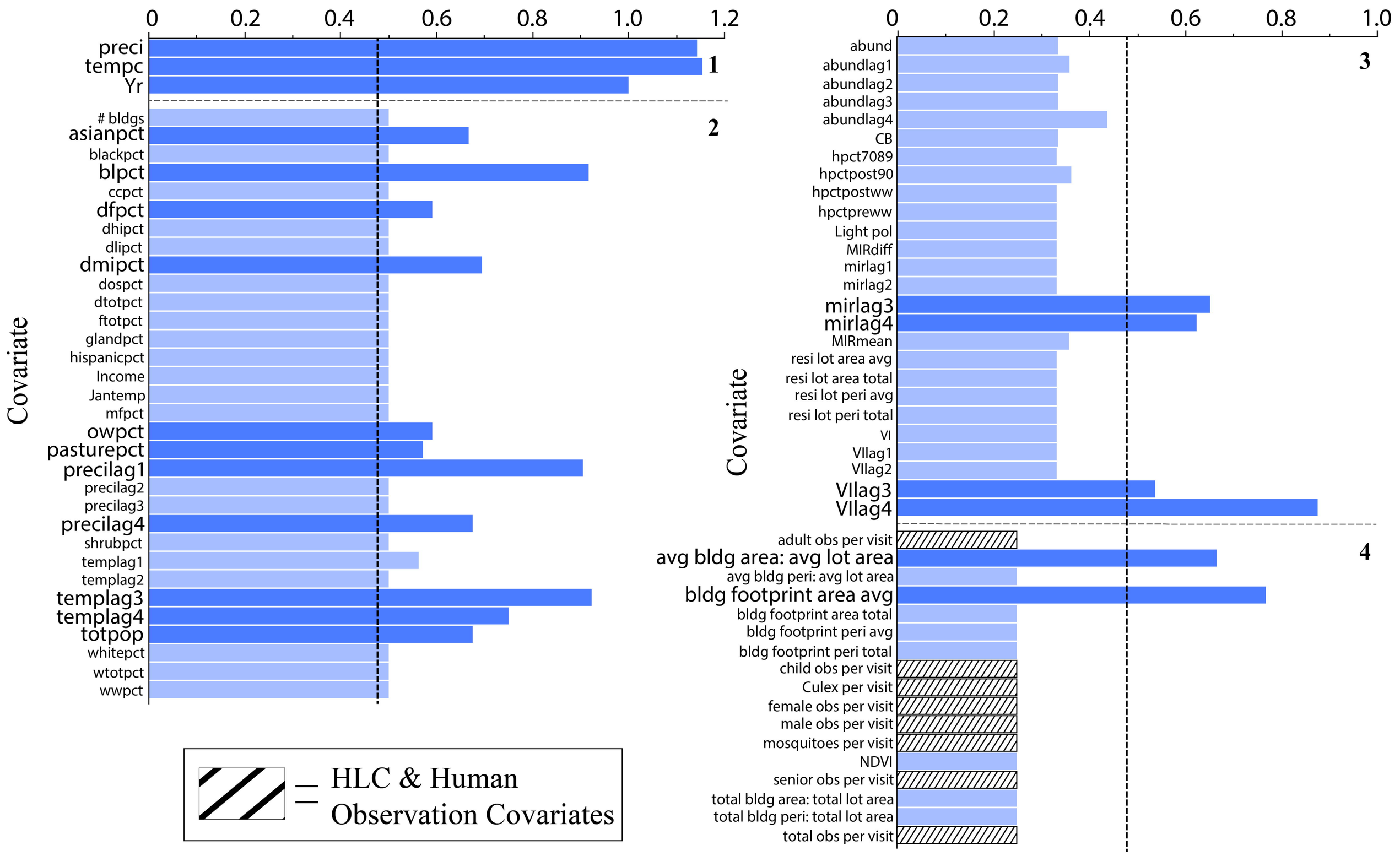
Human WNV Illness Outcome

(O_z) X 2

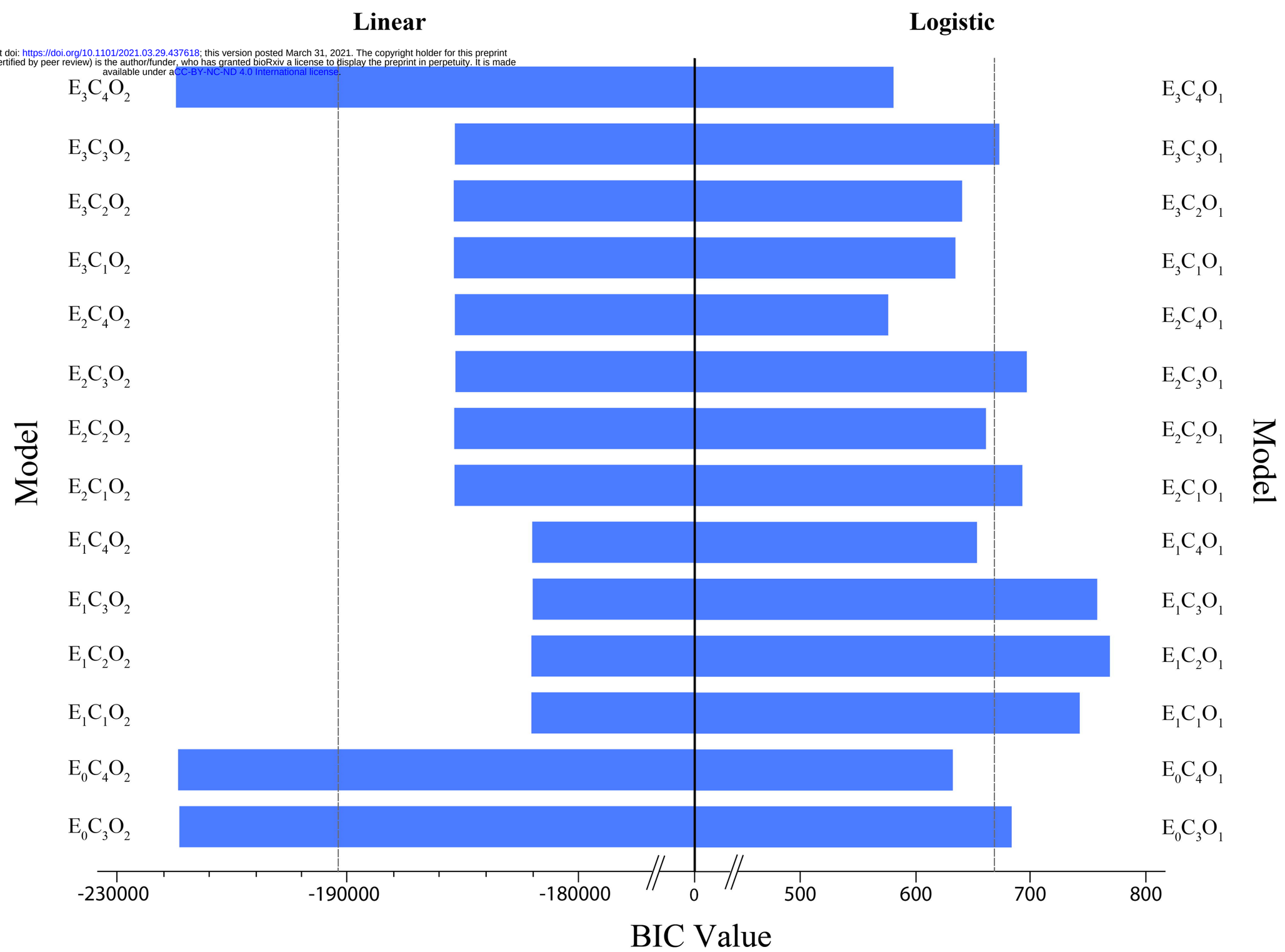


$3.5(E_x) \times 4(C_y) \times 2(O_z) = 28$ total models

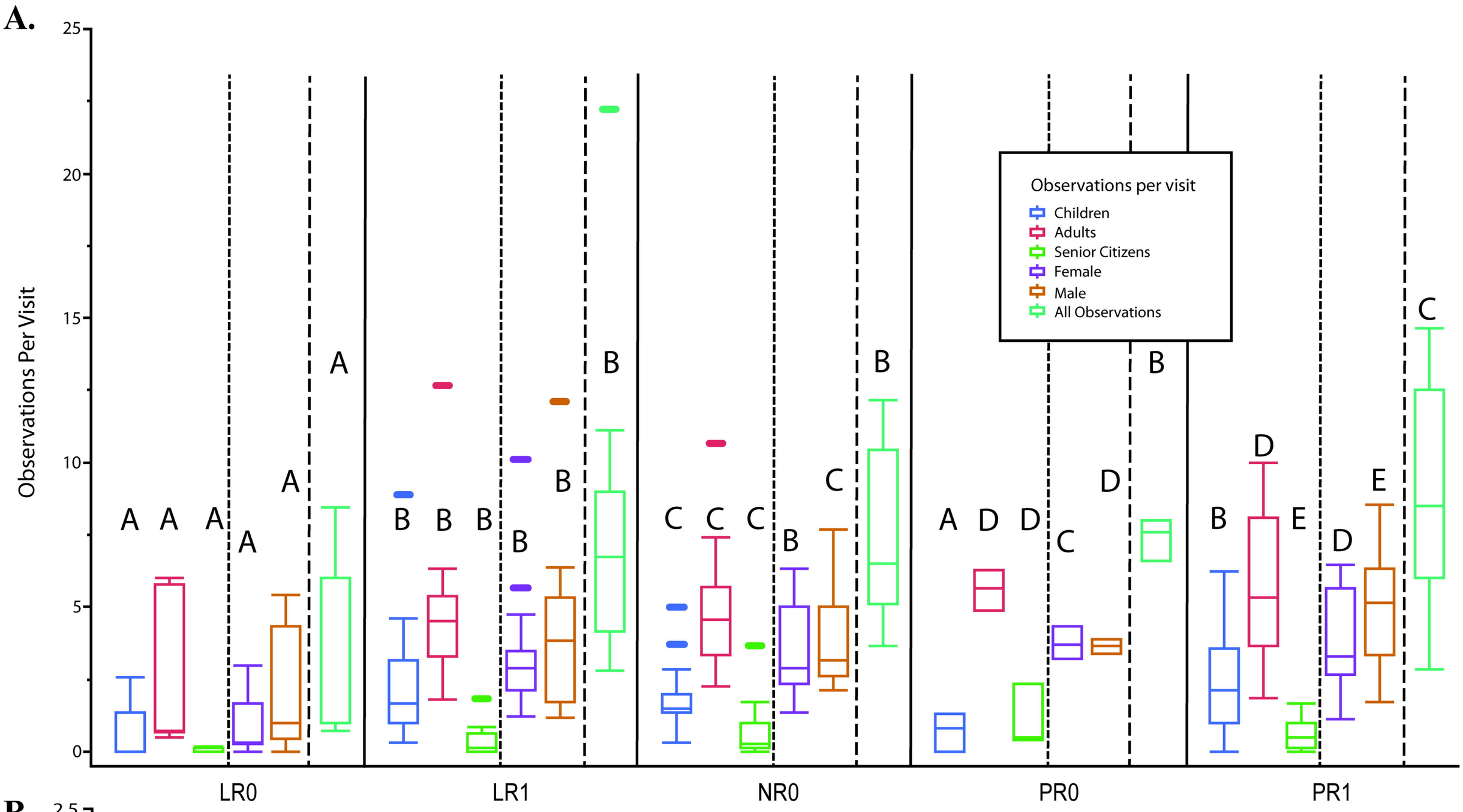
A.



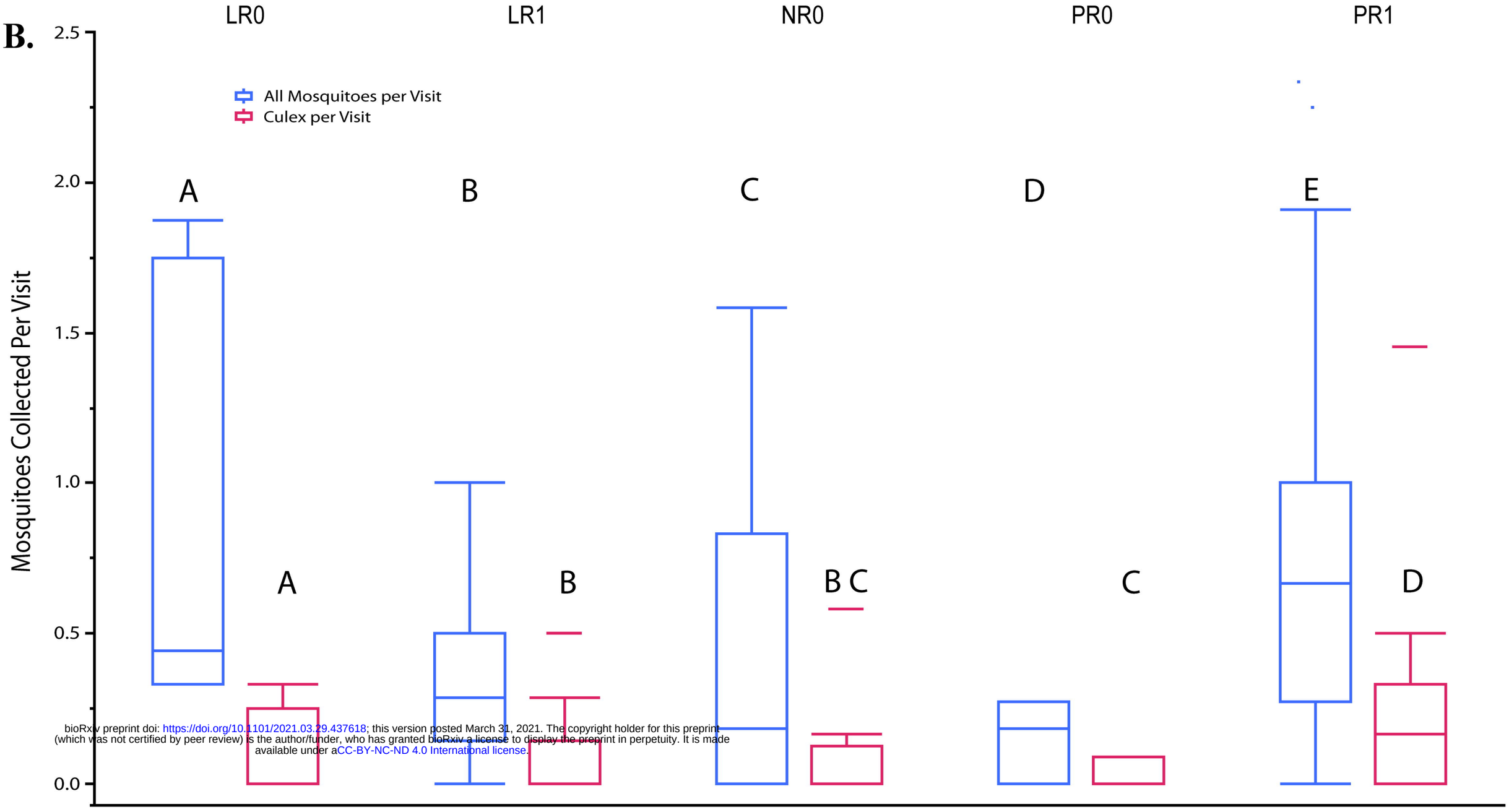
B.



A.



B.



C.