Abstract:

Neglected Tropical Diseases (NTDs) are a group of twenty diseases identified by the World Health Organisation (WHO) as disproportionately affecting people living in poverty. Those most affected by NTDs often live in remote, rural locations, mainly in tropical regions. It is estimated that 1 billion people are affected by the burden of these diseases, and 1.6 billion require treatment. The WHO has laid out a roadmap towards the control, prevention, elimination and eradication of NTDs. This roadmap details global goals to reach by 2030 which include, among others, a 90% reduction in those requiring treatment and for at least 100 countries to eliminate at least one NTD. In order to meet these goals, the roadmap states that “New approaches and mapping tools are necessary to obtain a granular view of disease epidemiology”. Model-based geostatistics is a branch of spatial statistics which has been increasingly used to support the achievement of this goal through multidisciplinary effort of the statistical and epidemiological research communities. In this article we provide an overview of how advanced statistical methodology has been used to support disease control programmes from low middle income countries in the elimination of NTDs, with a focus on soil transmitted helminths. The example described is drawn from our involvement as a World Health Organization Collaborating Centre with Ministries of Health in Africa. In particular, we highlight the hurdles statisticians encounter in promoting the adoption of advanced methodologies, including the communication barriers that must be overcome to ensure their effective utilisation.

Article

Title:

A Spatial Solution: Using Geostatistics in the fight against Neglected Tropical Diseases

Authors:

Jana Purkiss[1*], Freya N. Clark[1+], Emanuele Giorgi[1]

1.Centre for Health Informatics, Computation and Statistics (CHICAS), Lancaster University Medical School.

* Funded by NIHR.

+ Funded by Evidence Action.

Acronyms: (sidebar)
NTD - Neglected Tropical Diseases
WHO - World Health Organisation
LMIC - Low-middle-income countries

STH - Soil Transmitted Helminths
MDA - Mass Drug Administration
SAC - School-aged children (5–14 years)

Definitions: (sidebar)
Geostatistics - A branch of statistics that analyses spatial patterns in data.
Prevalence - The proportion of people who are infected with a disease in an area.
Covariate - A factor that can influence the outcome of the relationship we are studying. For example, soil quality would be a covariate when studying the relationship between sunlight and plant growth.
Spatial correlation - The tendency for things that are close together in space to be more alike or similar to each other than things that are farther apart.
Residuals - The variation in an outcome which we have not been able to account for using covariates in a model.
Log-Odds - A transformation which can be applied to prevalence data so that it follows a normal distribution and linear relationships with covariates can be assessed.

Geostatistics: (as a box coloured so that it matches the highlighted text)
Geostatistics refers to a branch of statistics which can be used to model spatially continuous processes using data collected from a finite set of survey locations. Geostatistics was first used in the South African mining industry in the 1950s to predict the distribution of ores and has since been used across many fields. The main principle of geostatistics lies in acknowledging the first law of geography in statistical models; that is that phenomena which are geographically closer to one another are more similar than those which are further apart. Its application can be used to model environment-driven diseases and is particularly important in the field of Neglected Tropical Diseases (NTDs) due to the low-resource settings in which many NTDs are found, and therefore the limited number of surveys which are conducted. The basic structure of a geostatistical model is as follows:
Variation in outcome of interest = Covariate effects + Spatially correlated residuals + Non-spatial residuals.
In summary, this model is saying that the variation in the outcome of interest can be explained by known factors (covariate effects), plus an additional patterns related to spatial location (spatially correlated residuals), and finally, some random noise that isn't explained by either the covariates or spatial patterns (non-spatial residuals).

NTDs - What are they?

On average 1 in 8 people globally are infected with a neglected tropical disease (NTD) yet many have never heard of them. NTDs are a group of 20 diseases identified by the World Health Organisation (WHO) as disproportionately affecting those living in poverty. Those most affected by NTDs often live in remote, rural locations, mainly in tropical regions. NTDs often have low mortality, but impact lives by causing developmental issues, disability, stigmatisation, social

exclusion and discrimination. These diseases are not a new phenomenon, and in fact there is evidence of what we now call NTDs being referenced in historical texts such as the Bible. Today, it is estimated that 1 billion people are affected by the burden of these diseases, and 1.6 billion require treatment (https://bit.ly/3xMLtBw).

The WHO has laid out a roadmap towards the control, prevention, elimination and eradication of NTDs. This roadmap details global goals to reach by 2030 which include, among others, a 90% reduction in those requiring treatment and for at least 100 countries to eliminate at least one NTD. In order to meet these goals, the roadmap states that "New approaches and mapping tools are necessary to obtain a granular view of disease epidemiology" [1]. Model-based geostatistics is a branch of spatial statistics which has been increasingly used to support the achievement of this goal through multidisciplinary effort of the statistical and epidemiological research communities. In this article we provide an overview of how advanced statistical methodology has been used to support disease control programmes from low middle income countries (LMICs) in the elimination of NTDs, with a focus on Soil Transmitted Helminths (STH).

NTDs - What can be done?

The burden of many NTDs, including STH, can be easily relieved with cheap, easy-to-administer treatments, however, ensuring that those most in need have access to these treatments presents a significant challenge. It thus becomes crucial to identify those who are infected in order to prioritise treatment for the most vulnerable. Given the remoteness of many disease-stricken populations and limited resources, this can prove challenging. Surveys conducted to monitor the burden of NTDs are typically limited to a finite set of locations. Thus, optimising the use of this information becomes crucial for informing policy decisions by government and health organisations. Geostatistical methods have been playing an essential role to address this issues in two different ways: 1) to infer prevalence of NTDs at locations where data has not been collected by exploiting spatial correlation; 2) to develop survey designs that, by making use of data collected in the past, are more efficient than standard sample surveys.

Case Study: STH in Kenya:

STH infections are caused by parasitic worms. These worms are transmitted via eggs present in human faeces, which contaminate soil; for this reason, STH affects those with poor access to sanitation facilities. For those with a high intensity of infection (many worms in their intestines), this disease can cause a range of symptoms including diarrhoea, abdominal pain, malnutrition, and impaired growth and development. The worms feed on human tissues, including blood, leading to a loss of iron and protein which can cause anaemia, especially in adolescent girls and women of reproductive age. Fortunately, repeated mass drug administration (MDA) with relatively cheap medication can be used to control morbidity.

We present an example investigating the prevalence of STH in Kenya [2]. In 2017, 153 randomly selected schools in 12 counties in south Kenya were visited, and a sample of school-aged children (SAC) were tested for STH. The prevalence of STH in the sampled schools can

be seen in Figure 1. We will create a simple model using only covariates (Model 1) and compare it to a second, improved model utilising the geostatistical method (Model 2).

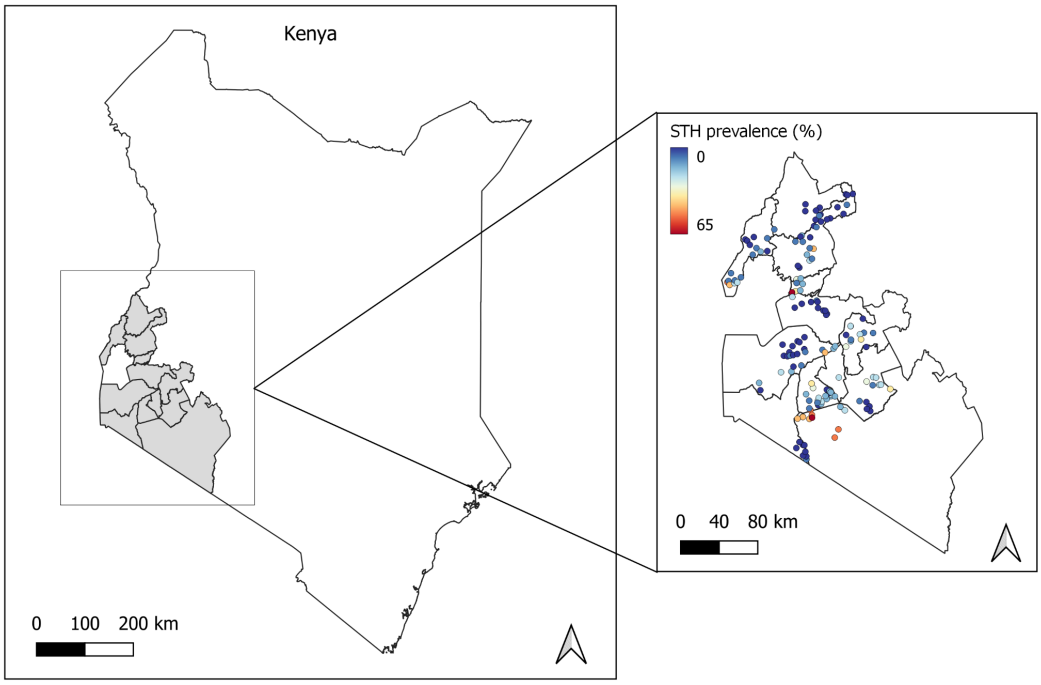

Figure 1: Maps showing the sampled schools in Kenya, coloured by % prevalence (proportion of children who tested positive for STH in each school).

Modelling STH - Using characteristics to explain prevalence:

Scientific research has delved into the environmental conditions conducive to the proliferation of STH, delineated by factors such as soil pH and land surface temperature. In addition to the environment, socio-economic characteristics also play a significant role in influencing exposure to STH. For instance, factors like school attendance rate are associated with higher prevalence of STH. It's worth noting these relationships are not causative; a low school attendance rate does not *cause* high STH prevalence but areas with low school attendance rates are likely to be poorer areas with poorer sanitation, and hence have a higher STH prevalence. Regression methods, of which model-based geostatistical methods are an extension, allow us to combine the effects of climactic, environmental, social, and demographic characteristics to predict the prevalence of STH in that area. When we use variables, other than the outcome, in this way, we will refer to them as the covariates of the model.

STH in Kenya - Modelling using Covariates:

In our Kenya example, the enhanced vegetation index (a measure of vegetation greenness), land surface temperature for day and night, and soil acidity were selected in the modelling process as important covariates. We retrieved a satellite image for each covariate which we used to determine the vegetation greenness, soil acidity and land surface temperature across

Kenya at pixel level. We will refer to these covariates as EVI, LST day, LST night and soil pH, respectively. EVI, LST day and LST night were retrieved from the MODIS data portal (https://bit.ly/3w7J83u). Soil pH was obtained from the ISRIC Soil Data Hub (https://bit.ly/3UGLWOJ).

Standard regression models that ignore spatial correlation provide a simpler naive approach to use those covariates for mapping STH prevalence. This model, which we refer to as Model 1, can be expressed as

Model 1: $Log - Odd(STH\ Prevalence) = \beta_0 + \beta_1 \times EVI + \beta_2 \times LST\ day +$

$$\beta_3 \times LST\ night + \beta_4 \times soil\ pH$$

Model 1 is a standard logistic regression model from which we can predict the prevalence at other locations which were not sampled by extracting the values of each covariate and combining them according to the equation above. In Model 1, β_0 refers to the intercept value (the log-odds of the STH prevalence we would expect if the EVI, LST day, LST night and soil pH all had value 0). We define β_1 through β_4 to be the amount we would expect the log-odds of the STH prevalence to increase (or decrease) when the value of its corresponding covariate increases by 1.

The results show that in regions with a greater EVI (indicating more vegetation), we observed a higher prevalence of STH. This can be attributed to vegetation offering shelter, which prevents eggs from being washed away and enables soil moisture retention [3]. Furthermore, our findings revealed a negative association between soil pH and STH prevalence, or in other words areas with more alkaline soil exhibited a lower proportion of positive tests, as the eggs struggle to survive in highly alkaline conditions. Similarly, a negative association was found between LST and STH prevalence, as eggs fail to hatch in temperatures exceeding 40°C [4].

Modelling STH - Improving the predictions by modelling spatial correlation

Whilst including covariates is useful in modelling, we wish to extend our model to improve the estimates of prevalence in unsampled locations by accounting for the variation in prevalence that our covariates are unable to account for and which causes spatial correlation. Modelling of spatial correlation is a topic that has been addressed in different epidemiological problems and the methods proposed are often tailored to the specific data-formats. In our context, our sampling units correspond to locations that may represent a village, a household or, as in the STH example, a school. Model-based geostatistics aims to exploit the information collected at this finite set of locations to draw inferences on spatially continuous surfaces of disease prevalence. For further information about how geostatistics can be used in global public health we recommend the book “Model-Based Geostatistics for Global Public Health: Methods and Applications” [5]

STH in Kenya - Implementing a geostatistical model:

Based on the established association that exists between water, sanitation and hygiene affects with the likelihood that someone is infected with STH, we can assume that those sharing water sources and toilet facilities are likely to have a similar risk of infection. However, these variables may not always be available, thus making the use of geostatistical models essential. In more formal statistical terms, we extend Model 1 by introducing an additional spatial term to the model, technically referred to as the spatial Gaussian process. We will refer to the model which includes covariates and the spatial Gaussian process as Model 2.

Model 2: $Log - Odds(STH\ Prevalence) = \beta_0 + \beta_1 \times EVI + \beta_2 \times LST\ day +$

$$\beta_3 \times LST\ night + \beta_4 \times soil\ pH + Spatial\ gaussian\ process$$

The important feature of this additional element of the model is that the values it takes at two different locations are correlated based upon how far apart the two locations are.

STH in Kenya - Comparing Model 1 and Model 2:

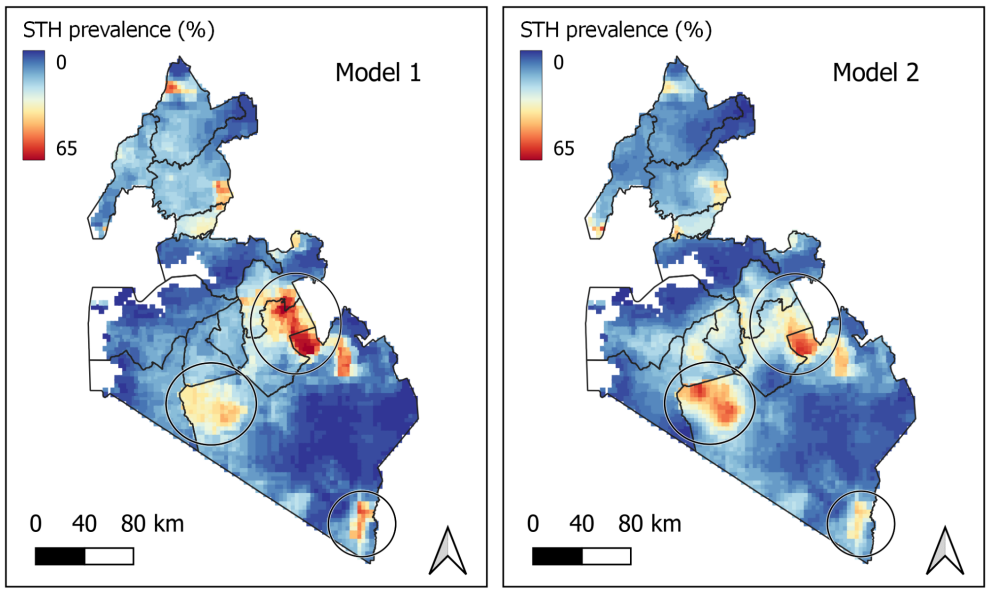

Figure 2: Maps showing the predicted STH prevalence using Model 1 (left) and Model 2 (right).

Figure 2 shows the predictions made from Model 1 and Model 2 across the area of interest in Kenya. Whilst there are similarities between the two maps, notable differences in the prevalence predictions can also be noticed.

STH in Kenya - Informing policy:

In addition to the point predictions of prevalence, we also can compare Model 1 and Model 2 by considering the probability, at a given location, of exceeding 2% prevalence, which we refer to as exceedance probability. The reason for this is that the WHO defines the STH prevalence thresholds for treatment as follows: suspend MDA if STH prevalence is less than 2%; conduct

MDA every 2 years if STH prevalence is between 2% and 10%; annually if between 10% and 20%; twice yearly if between 20% and 50%; and thrice yearly if greater than 50%. Figure 3 shows the probabilities of exceeding the 2% prevalence threshold using Model 1 and Model 2.

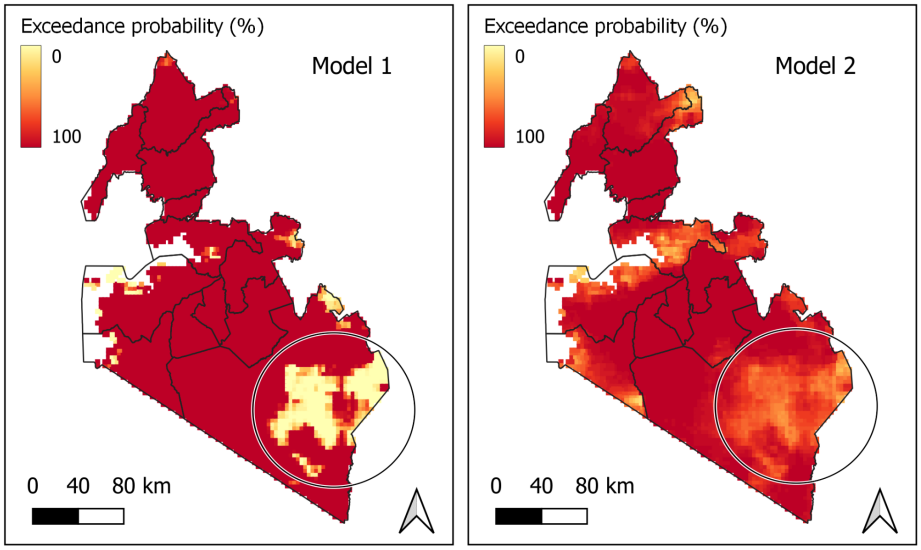

Figure 3: Maps showing the exceedance probability of 10% prevalence using Model 1 (left) and Model 2 (right).

Model 1 delineates two distinct areas: one marked in red, indicating a high likelihood of exceeding the 2% prevalence threshold, and another in yellow, corresponding to a very low likelihood (circled area). However, accounting for residual spatial correlation, induced by unmeasured covariates, alters the result substantially. Notably, the region initially deemed unlikely to exceed 2% now, under Model 2, exhibits exceedance probabilities closer to 50%. In essence, disregarding spatial correlation would lead to overly optimistic assessments regarding non-exceedance in the circled area depicted in Figure 3, whereas Model 2 advocates for a more cautious approach to its delineation.

For logistical reasons, decisions about treatment are often made over areal units. In the case of Kenya, these correspond to counties. To help inform these decisions, geostatistical models can also be used to generate predictive inferences for the county-level average prevalence as shown in Fig. 4. We can see there are 4 counties in the study area where MDA should be carried out every 2 years, 7 counties in which MDA should be carried out annually, and 1 in which MDA should be administered twice yearly.

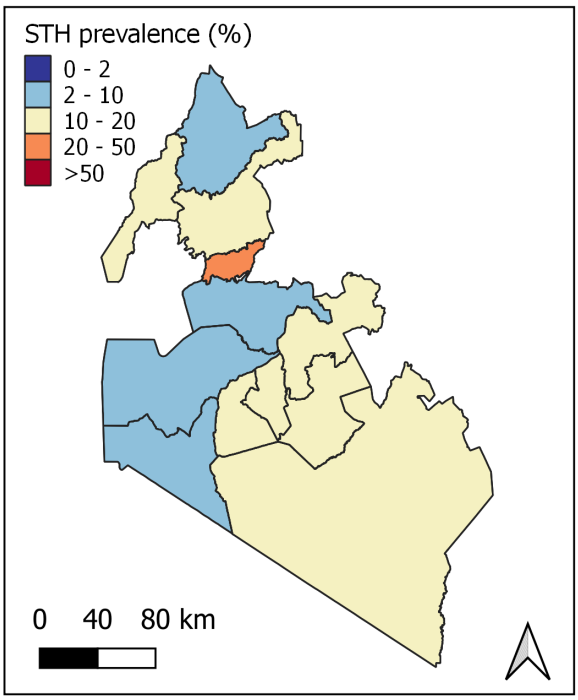

Figure 4: Maps showing the STH prevalence threshold of each county using Model 2.

NTDs - The current and future use of geostatistics:

As we have shown, the use of geostatistical methods enables us to better understand how prevalence of diseases varies over space. This knowledge can be used to inform decisions about where treatment should be carried out, and can also be used to assess how successful previous control programmes were. In addition to the one we have shown, there exist several other applications of geostatistics both in modelling and in survey design. Geostatstical methods can be used to model data combined from many NTD surveys collected across both space and time, and to understand the relationships between the different NTDs. Using geostatistical methods in survey designs entails identifying the locations which should be sampled in future surveys in order to gain the most information possible. Current research is also investigating the integration of geostatistical models with mathematical dynamic modelling to forecast potential changes in prevalence based on various control scenarios and understand how the impact of control interventions may vary over space.

NTDs - What are the obstacles?

Two of the main hurdles in widening the adoption of geostatistics for tackling both old and new public health threats, especially in LMICs, are the skills and infrastructure that are required. As a result, simpler but statistically less efficient methods are often preferred over geostatistical models. Training of scientists from various backgrounds thus becomes a crucial activity in which experienced statisticians should invest more to support countries in the fight against public health threats, including NTDs.

You - The next geostatistician?

This article has centred on the application of geostatistics in epidemiology, particularly in the context of NTDs. However, the utility of these methods extends across various disciplines such as ecology, geology, forestry, soil science, logistics, and meteorology. In light of the versatility of geostatistics and its rigorous approach to deal with uncertainty, exploring how these techniques could be applied within your own field may offer valuable insights for addressing real-world challenges more effectively.

Acknowledgements: We would like to thank Evidence Action's Deworm the World Initiative and the Kenya Medical Research Institute (KEMRI) for obtaining and providing the Kenya dataset.

References:

[1] "Ending the neglect to attain the Sustainable Development Goals: a road map for neglected tropical diseases 2021–2030". Geneva: World Health Organization; 2020. Licence: CC BY-NC-SA 3.0 IGO.

[2] Okoyo, Collins, Mark Minnery, Idah Orowe, Chrispin Owaga, Suzy J. Campbell, Christin Wambugu, Nereah Olick, et al. "Model-Based Geostatistical Design and Analysis of Prevalence for Soil-Transmitted Helminths in Kenya: Results from Ten-Years of the Kenya National School-Based Deworming Programme." *Heliyon* 9, no. 10 (October 2023). https://doi.org/10.1016/j.heliyon.2023.e20695.

[3] Maikai, B.V., J.U. Umoh, O.J. Ajanusi, and I. Ajogi. "Public Health Implications of Soil Contaminated with Helminth Eggs in the Metropolis of Kaduna, Nigeria." *Journal of Helminthology* 82, no. 2 (June 2008): 113–18. https://doi.org/10.1017/s0022149x07874220.

[4] Blum, Alexander J., and Peter J. Hotez. "Global 'Worming': Climate Change and Its Projected General Impact on Human Helminth Infections." *PLOS Neglected Tropical Diseases* 12, no. 7 (July 19, 2018). https://doi.org/10.1371/journal.pntd.0006370.

[5] Diggle, P.J., & Giorgi, E. (2019). Model-based Geostatistics for Global Public Health: Methods and Applications (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781315188492

Contributor Guidelines

- Articles must be interesting, engaging and easy to read.
- Readers should finish your article knowing more about statistics, or the application of statistics, than they did before.
- Technical terms and mathematics should be kept to a minimum, and explained clearly where used (we recommend you do this in a box or sidebar, using real-life analogies wherever possible).
- The target reader is someone with an interest in data, who knows some of the basics but is by no means an expert.
- Articles submitted to *Significance* will be vetted by the editor and an editorial board of statistical experts. If your article is accepted they will provide comments or questions to help you improve it and get it publication-ready.
- Avoid the formal tone and structure of an academic paper, and draw inspiration from intelligent, upmarket, mainstream magazines and websites. The Conversation, and any mainstream science magazine, would be good places to start.
- Remember to tell a story. It's not enough simply to describe a process.
- A strong 'hook' at the outset is invaluable for grabbing a reader's attention, and a real-life anecdote that 'humanises' the subject and sets the context of what follows often works very well.
- Your conclusion is extremely important. What do you want your readers to remember and think about once they've finished reading?
- You are encouraged to include charts, graphs, tables and figures**.** Please refer to the RSS data visualisation guide (https://royal-statistical-society.github.io/datavisguide/). Ensure all supporting figures are presented simply and neatly, are labelled correctly and clearly, and that accompanying captions are written to support the reader's understanding of the visual material. Charts and graphs should be supplied as Adobe Illustrator-compatible EPS files to allow our designers to update text and colour elements to fit house style. Editable PDFs are also suitable.

If you wish to use charts or graphs that are not your own work, please ensure that they are correctly sourced and referenced, and that you have permission to republish them from the original author or copyright owner. A letter or email confirming this permission is required.

You do not need to source and include stock photos - we will do that (although personal photos relating to the article are welcome)

- All we need is a Word doc and hi-res versions of your figures and tables. Please do not 'design' the article, but do indicate clearly to us where a text box begins and ends.
- End references are optional but should be limited to five. Use the Chicago referencing style, not Harvard - flagging each reference with a number in superscript, then ordering them as endnotes. If a reference is not an academic journal (ie if it is to a web page), please add a Bitly link in brackets after it, rather than including as a reference.
- Explain your data, quantity and quality of evidence, assumptions, methods, and models, and the limitations of your findings.
- Where estimates are made, be sure to quantify their accuracy, reliability, reproducibility, and validity.
- Keep mathematical details, such as symbols, notations and equations, to a minimum.
- Specific terminology should be used carefully and flagged as such, or explained as necessary; novel or unusual meanings must be explained either within the text or in sidebars. The editor and reviewers may point out such issues and help devise suitable wordings.
- The *Significance* Editorial Board requires that authors include within their articles any links and/or references to the sources of data, computer code and/or software and software packages on which their analyses are based. We understand that some of these sources may not be publicly available, whether for legal, ethical or commercial reasons. However, readers should still be told where the data come from, even if they are not able to access the data directly.
- Work count:
    - For our **Notebook** section: 500-1,000 words
    - For our **Features** section: 2,000-3,000 words
    - For our **StatsComm, Profiles and Perspectives** sections: 500-2,500 words
    - For **website** articles: 500-1,500 words
- Authors are required to declare, at the point of submission of their article, any financial or other interests they may have or hold, any conflicts (personal or professional), or

any affiliations that are relevant to the content of their submission. Examples include, but are not limited to:

- Paid employment by companies or organisations, whether full or part time
- Voluntary positions with companies or organisations, including committee memberships and appointments to advisory, management or oversight boards
- Directorships or shareholdings
- Research grants and funding
- Involvements in prior disputes, whether academic, civil or legal
- Authors with no conflicts of interest or affiliations to declare beyond their academic appointment or main employment should state that they have no conflicts of interest to declare. Declarations of interest may be used by the editor and editorial board to inform editorial decisions, and declarations made will be published alongside accepted articles.