CONCISE COMMUNICATION

Evaluation of geospatial methods to generate subnational HIV prevalence estimates for local level planning

The Subnational Estimates Working Group of the HIV Modelling Consortium*

See related paper on page 1481

Objective: There is evidence of substantial subnational variation in the HIV epidemic. However, robust spatial HIV data are often only available at high levels of geographic aggregation and not at the finer resolution needed for decision making. Therefore, spatial analysis methods that leverage available data to provide local estimates of HIV prevalence may be useful. Such methods exist but have not been formally compared when applied to HIV.

Design/methods: Six candidate methods – including those used by the Joint United Nations Programme on HIV/AIDS to generate maps and a Bayesian geostatistical approach applied to other diseases – were used to generate maps and subnational estimates of HIV prevalence across three countries using cluster level data from household surveys. Two approaches were used to assess the accuracy of predictions: internal validation, whereby a proportion of input data is held back (test dataset) to challenge predictions; and comparison with location-specific data from household surveys in earlier years.

Results: Each of the methods can generate usefully accurate predictions of prevalence at unsampled locations, with the magnitude of the error in predictions similar across approaches. However, the Bayesian geostatistical approach consistently gave marginally the strongest statistical performance across countries and validation procedures.

Conclusions: Available methods may be able to furnish estimates of HIV prevalence at finer spatial scales than the data currently allow. The subnational variation revealed can be integrated into planning to ensure responsiveness to the spatial features of the epidemic. The Bayesian geostatistical approach is a promising strategy for integrating HIV data to generate robust local estimates.

Department of Infectious Disease Epidemiology, Imperial College London, London, UK.

Correspondence to Sarah-Jane Anderson, Department of Infectious Disease Epidemiology, Imperial College London, St Mary's Campus, Norfolk Place, London W2 1PG, UK.

Tel: 02075943288; e-mail: sarah-jane.anderson@imperial.ac.uk

*The members of the working group are listed in the Acknowledgements section.

Introduction

Historically, country epidemics have been considered as fairly homogenous and been broadly classified as 'generalized' or 'concentrated' [1]. However, such an approach fails to capture the often substantial local level variation in the patterns of risk and transmission, key drivers of the epidemic, and availability of services observed [2]. Indeed, 13 of 33 countries in sub-Saharan Africa report at least five-fold differences in adult prevalence between provinces [3].

This spatial heterogeneity has profound implications for all aspects of monitoring the epidemic and planning the response. There have been a large number of examples of ways in which spatial data can be used to improve HIV planning, including identification of places or populations at highest risk [4], allocation of resources across locations [5], in understanding local level changes and monitoring the epidemic [6], in interpreting gaps in service provision [7], in understanding reasons for different biases in available surveillance data [8], and in tailoring services and local level targeting of intervention [9].

Although it is critical that planners have an understanding of local level variations in the intensity of the HIV epidemic, practical and financial constraints that restrict the size of surveys and surveillance systems commonly used to monitor generalized epidemics may inhibit this. Robust estimates of subnational HIV prevalence are typically available at the first administrative level (often termed 'provinces'), but not at more local levels which may be needed for programme planning (such as the district or county levels).

There are several candidate methods that could be used to generate local subnational estimates of HIV prevalence (Table 1). These include the method that the Joint United Nations Programme on HIV/AIDS (UNAIDS) has used to generate maps of high-burden countries [23] and a method which has been used to estimate malaria prevalence patterns [14,15]. However, the performance of these methods has not been evaluated or compared in the context of HIV epidemiology. Here, we conduct a formal evaluation and comparison of subnational HIV prevalence estimates generated by these candidate methods.

Methods

Six candidate methods were included in this study (labelled models 1−6 in Table 1). Key characteristics are described in Table 1.

Some mapping strategies use sampled HIV prevalence data from household surveys only and do not use ancillary spatial information (e.g. road network or vegetation coverage). PrevR (model 1) is such a method [10], and it has been recommended by UNAIDS on the basis of it being straightforward to implement and available for immediate use.

This approach is compared with methods that do leverage information on other features ('covariates') which can enhance predictions of HIV prevalence (data sources are described in Table S1, http://links.lww.com/QAD/A890). Such data are particularly useful if they are sampled at a greater geographical resolution than the HIV prevalence data. Models 2−6 fall into this category. Among these, there are important differences in the theoretical framework behind them, which influence their ability to deal with uncertainty and the computational load (Table 1).

Finally, some of these methods estimate a continuous 'surface' of HIV prevalence (models 1−3), whereas others instead aim to provide predictions for the aggregate level, namely the subnational units under consideration directly, for example, a district level prevalence estimate (models 4−6).

Two different validation procedures were used to assess the performance of the methods, depending on whether the method produces a 'surface' of prevalence or can give estimates at aggregate subnational level. Data from three countries (Tanzania, Kenya, Malawi) with generalized epidemics were used. These countries were chosen as they encompass variation in epidemic patterns and data availability.

Internal validation

The performance of those models that produce continuous HIV prevalence surfaces (models 1−3) was assessed using internal validation. A proportion of observed data is held back as a 'test' dataset, before the method is used. Then, the test data are used to challenge the model prediction at sample locations. This test dataset was either a single point, namely a demographic and health survey (DHS) cluster [leave out one cross validation (LOOCV), (models 1 and 3)] or a larger proportion of the data [partitioned data hold back (PDHD), (models 1 and 2)]. Which strategy was applied (i.e. LOOCV or PDHD) for a given spatial method was dependent on the computational intensity of the mapping approach.

The resulting root mean squared error (RMSE) (Eq. 1) between prediction and data was calculated. The RMSE estimates are in the same units as the surface (i.e. prevalence) and the lower the RMSE, the closer the model prediction is to the observed data.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(E_i - O_i)^2}{N}} \tag{1}$$

where E is the predicted values, O the observed values, N the total number of locations, and i each omitted location.

Table 1. Key characteristics of the candidate spatial methods.

Methods

Method name	Summary of method	References	Use of covariates	Estimation of uncertainty	Ease of use	Unit of analysis
Model 1: kernel density estimation with adaptive bandwidth (prevR)	The numbers of individuals tested and found positive at each DHS cluster location are plotted, and smoothed intensity surfaces are generated by computing average infection rates within a moving window (a 'kernel'). Creation of the smoothed surface relies on the specification of the kernel at each cluster; that are combined to make the intensity surface. The prevalence surface is the ratio of the intensity surface of the number who tested positive to the surface of the total number tested	Larmarange et al., 2011 [10]; Davies and Hazelton, 2010 [11]	Does not use information on covariates	Able to provide an indication of how much data were used in producing the estimate	Straightforward- well documented, open source	Pixel level-provides estimates for small areas to give a continuous surface
Model 2: model-based geostatistics	Bayesian, or 'model-based' geostatistical approaches are a special class of generalized linear mixed models. They extend classical geostatistical methods such as kriging to allow, among other features, formal incorporation of: sampling error in the observed data; relationships with covariates (and the uncertainty in the form of these relationships); uncertainty in the spatial autocorrelation structure of the outcome variable. The model can be fit using computational methods like MCMC	Diggle et al., 1998 [12]; Patil et al., 2011 [13]; Gething et al., 2011 [14]; Gething et al., 2012 [15]	Uses information on covariates	Conducted within a Bayesian framework, and is able to provide a rigorous assessment of the uncertainty of the estimated value at each location on the mapped surface	Complex but could be adapted for the end user	Pixel level-provides estimates for small areas to give a continuous surface
Model 3: kriging of covariates with logistic regression	This method involves two steps, the prediction of the spatial distribution of each covariate independently using kriging and combination of these layers using a nonspatial logistic regression model to generate HIV prevalence predictions	Cuadros, 2014 [16]; Klienschmidt et al., 2000 [17]	Uses information on covariates	Does not provide estimates of uncertainty	Moderately complex involves multiple steps that require expertise	Pixel level-provides estimates for small areas to give a continuous surface
Model 4: shared spatial component	The shared spatial component model was originally developed to map different diseases likely to have similar spatial distributions because of shared risk factors. The model has both shared components and components specific to each disease of interest. The shared spatial component model was applied here to incorporate information from both the DHS and ANC data sources. This model looks at discrete areas (administrative units) and is constructed within a Bayesian framework	Manda et al., 2009 [18]; Knorr-Held et al., 2001 [19]	Uses information on covariates	Able to provide estimates of uncertainty for the unit of analysis (administrative region)	Moderately complex involves multiple steps that require expertise	Provides estimates for the administrative unit

Table 1 (*continued*)

Methods

Method name	Summary of method	References	Use of covariates	Estimation of uncertainty	Ease of use	Unit of analysis
Model 5: spatio-statistical aggregate method: regression at an aggregated scale (administrative unit)	Following an exploration of regression kriging at the cluster level for Malawi, it was found that the spatial predictive power of the model was limited because of weak spatial structure. As a result an aggregate scale was explored instead. This method involves aggregation (through simple averaging) of the DHS clusters to administrative regions, followed by OLS multiple regression to examine the relationship between selected covariates and HIV prevalence. The administrative unit could be the target administrative unit, or it may be that the optimal unit for predictive power is lower	Moise et al., 2014 [20]; Moise et al., 2014 [21]	Uses information on covariates	Does not provide estimates of uncertainty	Moderately complex involves multiple steps that require expertise	Provides estimates for the administrative unit
Model 6: Bayesian geo-addative mixed model	This approach is similar to model 2 in that it is a Bayesian approach that is able to incorporate the sampling error, relationships with covariates, and uncertainty in the spatial autocorrelation structure. This model differs from model 2 in that it looks at the patterns of prevalence across discrete areas (administrative units) rather than continuous space as in model 2, and accounts for the spatial dependencies between these neighbouring areas. This model uses computational methods (MCMC) for inference and checking of the model. It is conducted using BayesX software.	Kandala et al., 2011 [22]	Uses information on covariates	Able to provide estimates of uncertainty for the unit of analysis (administrative unit)	Moderately complex involves multiple steps that require expertise	Provides estimates for the administrative unit

ANC, antenatal clinic; DHS, demographic and health survey; MCMC, Markov chain Monte Carlo; OLS, ordinary least squares.

External validation (cross-year comparison of estimates)

External validation was also conducted through comparing mapped predictions with data from an earlier survey year. This approach takes advantage of the location of clusters often being different in different survey years in a country, and assumes that the true spatial pattern of HIV prevalence is conserved over time. All three countries considered have surveys for more than one DHS round (Malawi: 2004 and 2010, Tanzania: 2007/2008 and 2012, Kenya: 2003 and 2008/2009) allowing for such an analysis. For methods that produce a continuous surface (models 1−3), prevalence surfaces were produced using only the data from the later survey. Then, a comparison was made between the observed values for each cluster in the earlier survey year and the predicted prevalence in the corresponding location in the year of the later survey.

The approach was repeated for all methods at the level of the first administrative unit for Malawi (district level), as not all methods produce continuous prevalence surfaces. For those methods which produced predicted surfaces (models 1−3), the average of the values of the surface within each district boundary was calculated, to allow for comparison with those methods which produce district level estimates directly (models 4−6). The accuracy of district level predictions in comparison to the data from the earlier survey year was summarized by the RMSE.

Results

Figure 1a presents the continuous prevalence surfaces for each country using models 1−3. Very substantial within-country variations in HIV prevalence are revealed by all methods. In particular, methods indicate a prevalence gradient from east to west in Kenya, south to north in Malawi, and a focus of high prevalence in southwest Tanzania.

There is substantial variation in the degree of local level variation suggested by the models. Methods that bring in additional data tend to produce estimates with a more complex spatial structure, which is related to road networks, among other factors, although uncertainty in the estimates about this is great (not shown). In contrast, the PrevR approach gives a smooth surface.

The internal validation procedure suggests that all methods can produce estimates of HIV prevalence at unsampled locations with a similar, and reasonable, level of accuracy (RMSE values: Fig. 1a). Among the methods, the Bayesian geostatistical approach (model 2), gives marginally the best RMSE values consistently across all the countries (Fig. 1: RMSE values displayed in the key for each panel).

The external validation for the continuous surfaces shows that the methods typically are successful in predicting prevalence in unsampled locations (Fig. 1b). The greatest difference in prediction error is between countries rather than between methods, and all methods have similar spatial pattern in the errors. Among models 1−3 (those that could do this test), model 2 gives the lowest RMSE across all countries. All methods give higher RMSE values than in the internal validation exercise described above as they are being used to predict the spatial distribution from a different year and the epidemic will have changed.

The external validation for the district level in Malawi (Fig. 1c), shows that whereas all the methods give the same broad spatial trends, differences in some districts are quite pronounced and overall errors are much greater. As prevalence may vary widely within an administrative region, particularly between urban and rural areas, an averaged value for each administrative region gives wider differences than the point by point comparison. Again, model 2 gives the lowest RMSE, and gives the most accurate prediction.

Discussion

Generating subnational estimates of HIV prevalence will be crucial to informing a locally tailored response to the HIV epidemic. This analysis has provided a number of key insights as to how countries can best utilise available spatial data.

First, the magnitude of the error and accuracy of predictions appears to depend most on the prevalence level in the country of interest and the characteristics of the survey sample, rather than the estimation method used. Because of this, we see a greater difference in the accuracy of predictions between countries rather than between methods. For this reason, some confidence in predictions from these spatial methodologies comes from their relative consistency in performance across all validation procedures described. The method used already by UNAIDS (model 1: PrevR) performs similarly to most other methods and so greater confidence can be afforded in the results. As a result of these analyses, we recommend that Bayesian geostatistical approach (model 2) be developed further, as the performance of this method was consistently the strongest. This method has been applied extensively to other infections [13], and has many desirable characteristics, in particular a formal accounting of uncertainty and the explicit leveraging of other geographic data.

Second, the methods appear to work reasonably well, and can capture the broad spatial trends in prevalence observed across countries. Arguably, in these high prevalence settings, these methods would usefully distinguish areas
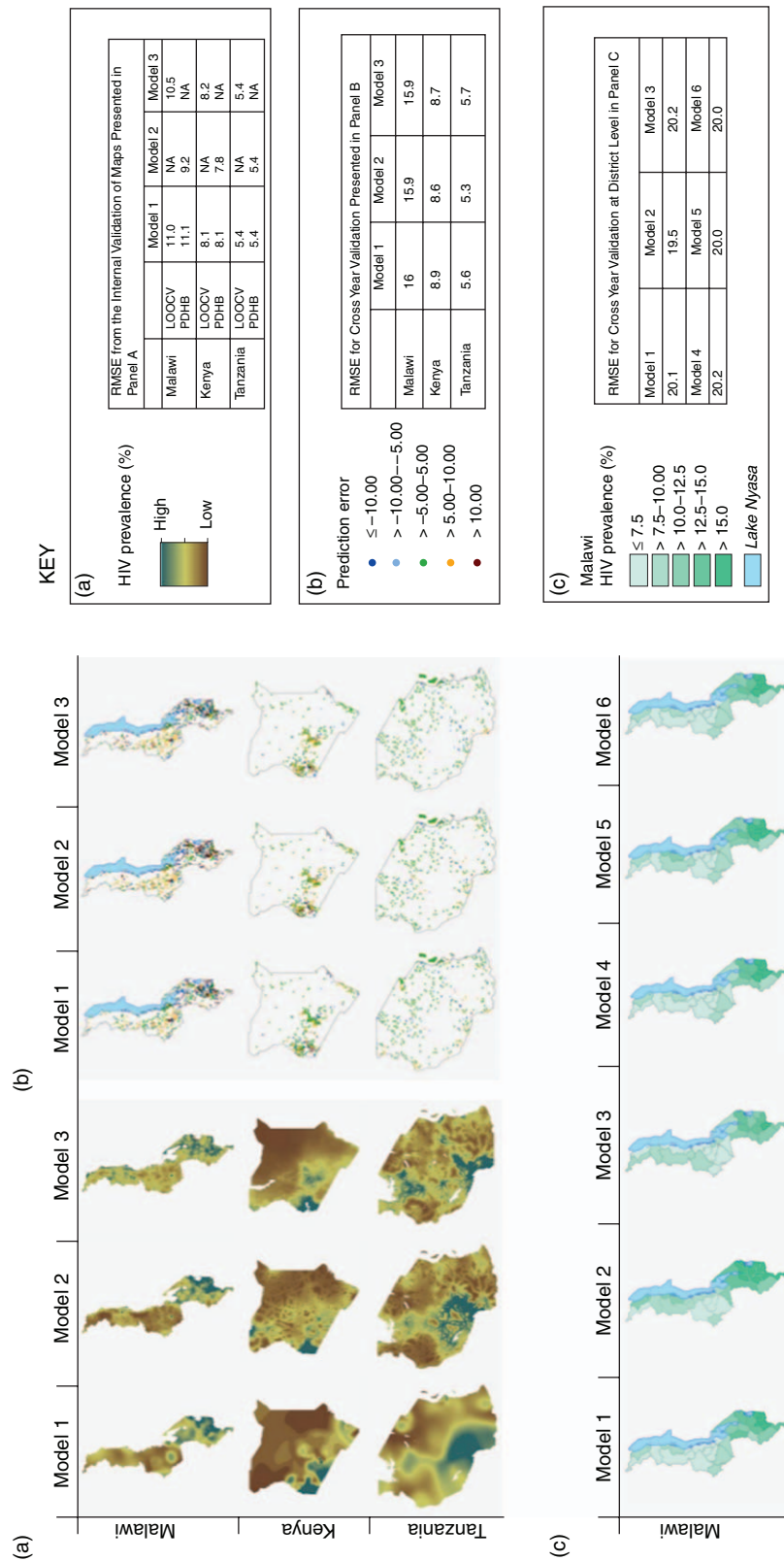
KEY

(a) HIV prevalence (%) — High / Low

RMSE from the Internal Validation of Maps Presented in Panel A

		Model 1	Model 2	Model 3
Malawi	LOOCV	11.0	NA	10.5
	PDHB	11.1	NA	NA
Kenya	LOOCV	8.1	9.2	8.2
	PDHB	8.1	7.8	NA
Tanzania	LOOCV	5.4	NA	5.4
	PDHB	5.4	5.4	NA

(b) Prediction error

- ≤ −10.00
- > −10.00−−5.00
- > −5.00−5.00
- > 5.00−10.00
- > 10.00

RMSE for Cross Year Validation Presented in Panel B

	Model 1	Model 2	Model 3
Malawi	16	15.9	15.9
Kenya	8.9	8.6	8.7
Tanzania	5.6	5.3	5.7

(c) Malawi HIV prevalence (%)

- ≤ 7.5
- > 7.5−10.00
- > 10.0−12.5
- > 12.5−15.0
- > 15.0
- Lake Nyasa

RMSE for Cross Year Validation at District Level in Panel C

Model 1	Model 2	Model 3
20.1	19.5	20.2
Model 4	Model 5	Model 6
20.2	20.0	20.0

Fig. 1. Results from the validation exercises. Panel (a) gives the HIV prevalence surfaces from models 1–3, and the corresponding Root Mean Squared Error (RMSE) values from the internal validation [either Leave Out One Cross Validation (LOOCV) or Partitioned Data Hold Back (PDHB)] are given in the tables in the Key. Panel (b) gives the error maps from the external validation procedure using the cross year comparison of spatial patterns and the corresponding RMSE values for models 1–3. Panel (c) gives the estimates of HIV prevalence for each district in Malawi and the corresponding RMSE value for all models (1–6). Note that the results presented in panel (a) and panel (b) are only for models 1–3 as they are only evaluating those methods which produce continuous prevalence surfaces.

of very high prevalence from those with very low prevalence.

The mapping methods described can be further developed, particularly through integration of different data sources alongside the DHS, in particular, antiretroviral therapy and prevention of mother-to-child transmission programme data, antenatal clinic surveillance, and, in the future, case-reporting data. Doing so would require building upon earlier work [8] to assess how different data sources may feed into prevalence mapping in a manner that reflects the different biases, underlying populations and spatial coverage of these data. Furthermore, although the tools outlined can fill critical gaps at this time, they do not mitigate the need for future additional local data collection and reporting to strengthen more localized responses to the HIV epidemic.

National epidemics cannot continue to be assessed as a whole when there is clear evidence of substantial subnational heterogeneity. Subnational indicators should be integrated into all national planning, monitoring, and evaluating processes performed routinely. Existing tools, such as Spectrum/Estimation and Projection Package modelling software, are already being adapted to explicitly examine the epidemic in subnational areas [24]. The increasing availability of georeferenced data and mapping tools provides us with the opportunity to be responsive to the subnational features of HIV epidemics to improve intervention planning.

References

1. UNAIDS/WHO Working Group on Global HIV/AIDS and STI Surveillance. Guidelines for second generation HIV surveillance. Geneva, Switzerland: 2000. Available: http://www.who.int/hiv/pub/surveillance/pub3/en/. [Accessed 27 October 2014]
2. Tanser F, de Oliveira T, Maheu-Giroux M, Bärnighausen T. **Concentrated HIV subepidemics in generalized epidemic settings.** *Curr Opin HIV AIDS* 2014; **9**:115–125.
3. Joint United Nations Programme on HIV/AIDS (UNAIDS). Location, location. Connecting people faster to HIV services. Geneva, Switzerland: 2013. Available: http://unaids.org.vn/images/New_2013/LocationLocation_EN_EMBARGOED_19%20Nov-Noon%20%283%29.pdf. [Accessed 23 October 2014]
4. Cuadros D, Awad S, Abu-Raddad L. **Mapping HIV clustering: a strategy for identifying populations at high risk of HIV infection in sub-Saharan Africa.** *Int J Health Geogr* 2013; **12**:28.
5. Anderson SJ, Cherutich P, Kilonzo N, Cremin I, Fecht D, Kimanga D, et al. **Maximising the effect of combination HIV prevention through prioritisation of the people and places in greatest need: a modelling study.** *Lancet* 2014; **384**:249–256.
6. Kayeyi N, Fylkesnes K, Michelo C, Makasa M, Sandoy I. **Decline in HIV prevalence among young women in Zambia: national-level estimates of trends mask geographical and socio-demographic differences.** *PLoS One* 2012; **7**:e33652. PubMed PMID: WOS:000304855200019.
7. Wilson DP, Blower SM. **Designing equitable antiretroviral allocation strategies in resource-constrained countries.** *PLoS Med* 2005; **2**:132–141.
8. Montana LS, Mishra V, Hong R. **Comparison of HIV prevalence estimates from antenatal care surveillance and population-based surveys in sub-Saharan Africa.** *Sex Transm Infect* 2008; **84**:I78–I84.
9. Weir S, J.T.B. **From people to places: overview of the PLACE method and lessons learned.** *MEASURE Eval Bull* 2002; **2002**:1–6.
10. Larmarange J, Vallo R, Yaro S, Msellati P, Meda N. **Methods for mapping regional trends of HIV prevalence from Demographic and Health Surveys (DHS).** *Cybergeo Europ J Geo* 2011; **558**:http://cybergeo.revues.org/24606; doi: 10.4000/cybergeo.24606.
11. Davies TM, Hazelton ML. **Adaptive kernel estimation of spatial relative risk.** *Stat Med* 2010; **29**:2423–2437.
12. Diggle PJ, Tawn JA, Moyeed RA. **Model-based geostatistics.** *J R Stat Soc Ser C Appl Stat* 1998; **47**:299–350.
13. Patil AP, Gething PW, Piel FB, Hay SI. **Bayesian geostatistics in health cartography: the perspective of malaria.** *Trends Parasitol* 2011; **27**:245–252.
14. Gething P, Patil A, Smith D, Guerra C, Elyazar I, Johnston G, et al. **A new world malaria map: plasmodium falciparum endemicity in 2010.** *Malar J* 2011; **10**:378.
15. Gething PW, Elyazar IRF, Moyes CL, Smith DL, Battle KE, Guerra CA, et al. **A long neglected world malaria map: plasmodium vivax endemicity in 2010.** *Plos Negl Trop Dis* 2012; **6**:e1814. doi: 10.1371/journal.pntd.0001814.
16. Cuadros D. **Incorporating spatial variability to generate sub-national estimates of HIV prevalence.** *Technical Consultation on Methods for Generating Sub-National Estimates of HIV Epidemiology To Support Country Program Planning and Evaluation.* 24–25 March 2014. Ole Sereni Hotel, Nairobi.
17. Kleinschmidt I, Bagayoko M, Clarke GP, Craig M, Le Sueur D. **A spatial statistical approach to malaria mapping.** *Int J Epidemiol* 2000; **29**:355–361.
18. Manda SO, Feltbower RG, Gilthorpe MS. **Investigating spatio-temporal similarities in the epidemiology of childhood leukaemia and diabetes.** *Eur J Epidemiol* 2009; **24**:743–752.
19. Knorr-Held L, Best NG. **A shared component model for detecting joint and selective clustering of two diseases.** *J R Stat Soc Ser A Stat Soc* 2001; **164**:73–85.
20. Moise IK, Zulu L, Shortridge A, Kalipeni E. **Sub-National HIV modeling: Malawi Example.** *Technical Consultation on Methods for Generating Sub-National Estimates of HIV Epidemiology To Support Country Program Planning and Evaluation.* 24–25 March 2014. Ole Sereni Hotel, Nairobi.
21. Moise IK, Zulu L, Shortridge A, Kalipeni E. **Evaluating DHS Cluster Data.** *Technical Consultation on Methods for Generating Sub-National Estimates of HIV Epidemiology To Support Country Program Planning and Evaluation.* 24–25 March 2014. Ole Sereni Hotel, Nairobi.
22. Kandala NB, Brodish P, Buckner B, Foster S, Madise N. **Millennium development goal 6 and HIV infection in Zambia: what can we learn from successive household surveys?** *AIDS* 2011; **25**:95–106.
23. Larmarange J, Bendaud V. **HIV estimates at second sub-national level from national population-based surveys and adjusted to UNAIDS estimates.** *AIDS* 2014; **28 (Suppl 4)**:S469–S476. doi: 10.1097/QAD.0000000000000480.
24. Mahy M, Nzima M, Ogungbemi MK, Ogbang DA, Morka MC, Stover J. **Redefining the HIV epidemic in Nigeria: from national to state level.** *AIDS* 2014; **28 (Suppl 4)**:S461–S467.