OPEN
https://doi.org/10.1038/s41586-019-1200-9

Mapping HIV prevalence in sub-Saharan Africa between 2000 and 2017

Laura Dwyer-Lindgren[1], Michael A. Cork[1], Amber Sligar[1], Krista M. Steuben[1], Kate F. Wilson[1], Naomi R. Provost[1], Benjamin K. Mayala[2], John D. VanderHeide[1], Michael L. Collison[1], Jason B. Hall[1], Molly H. Biehl[1], Austin Carter[1], Tahvi Frank[1], Dirk Douwes-Schultz[1], Roy Burstein[1], Daniel C. Casey[1], Aniruddha Deshpande[1], Lucas Earl[1], Charbel El Bcheraoui[1], Tamer H. Farag[1], Nathaniel J. Henry[1], Damaris Kinyoki[1], Laurie B. Marczak[1], Molly R. Nixon[1], Aaron Osgood-Zimmerman[1], David Pigott[1], Robert C. Reiner Jr[1], Jennifer M. Ross[1,3,4], Lauren E. Schaeffer[1], David L. Smith[1], Nicole Davis Weaver[1], Kirsten E. Wiens[1], Jeffrey W. Eaton[1,5], Jessica E. Justman[6,7], Alex Opio[8], Benn Sartorius[9], Frank Tanser[10,11,12,13], Njeri Wabiri[14], Peter Piot[15], Christopher J. L. Murray[1] & Simon I. Hay[1]*

HIV/AIDS is a leading cause of disease burden in sub-Saharan Africa. Existing evidence has demonstrated that there is substantial local variation in the prevalence of HIV; however, subnational variation has not been investigated at a high spatial resolution across the continent. Here we explore within-country variation at a 5 × 5-km resolution in sub-Saharan Africa by estimating the prevalence of HIV among adults (aged 15–49 years) and the corresponding number of people living with HIV from 2000 to 2017. Our analysis reveals substantial within-country variation in the prevalence of HIV throughout sub-Saharan Africa and local differences in both the direction and rate of change in HIV prevalence between 2000 and 2017, highlighting the degree to which important local differences are masked when examining trends at the country level. These fine-scale estimates of HIV prevalence across space and time provide an important tool for precisely targeting the interventions that are necessary to bringing HIV infections under control in sub-Saharan Africa.

HIV/AIDS is a leading cause of morbidity and mortality in sub-Saharan Africa[1,2]. In the nearly four decades since HIV was first recognized, scientific breakthroughs have transformed the once invariably fatal illness to one that can be successfully managed with lifelong anti-retroviral therapy (ART)[3]. Despite the rapid increase in the use of ART since the mid-2000s and the resulting decline in mortality, 34% of people in east and southern Africa and 60% of people in west and central Africa who are living with HIV are not currently receiving any treatment[4] and HIV/AIDS remains the most common cause of death in sub-Saharan Africa[2]. The burden of the global HIV epidemic is disproportionately concentrated in sub-Saharan Africa, where—in 2017—75% of deaths and 65% of new infections occurred and where 71% of people living with HIV resided[1,2].

The global community has repeatedly called for the end of the HIV epidemic. Millennium Development Goal 6 (Combat HIV/AIDS, malaria, and other diseases) included the target: "To halt by 2015 and have started to reverse the spread of HIV/AIDS"[5]. More recently, Sustainable Development Goal 3 (Ensure healthy lives and promote well-being for all at all ages)[6] explicitly calls for the end of the epidemic by 2030. The Joint United Nations Programme on HIV/AIDS (UNAIDS) fast-track strategy has set diagnosis and treatment targets[7] for 2020 and 2030, with the goal of markedly reducing both new infections and deaths by 2030. Despite these goals, a recent review of the state of HIV concluded that the world is not on track to end the HIV epidemic[8]. Moreover, global spending on HIV in sub-Saharan Africa peaked in 2013 and has since declined[9], potentially compromising existing efforts to combat HIV.

Renewed commitment and new tools are required to get the world on track to bring HIV infection under control, in sub-Saharan Africa and globally. Local data on the current prevalence of HIV are such a tool, providing a means to target resources and interventions more efficiently.

Precision public health and HIV

Country-level estimates of HIV prevalence, produced by both the Global Burden of Disease (GBD) study[1] and UNAIDS[4], highlight extensive differences in HIV prevalence between countries within sub-Saharan Africa. Further differences in HIV prevalence within national borders have long been recognized[10] and recent evidence suggests that there is substantial within-country variation. Both GBD[1] and UNAIDS[4] estimate the prevalence of HIV at the first-level administrative subdivisions in select countries and a growing number of studies have examined subnational trends in the prevalence of HIV in a variety of locations and at various levels of granularity[11–19] (Supplementary Table 1); these studies consistently find extensive within-country geographical variation in HIV prevalence.

Subnational variation in HIV prevalence has important implications for efforts to bring HIV infection under control, related to the treatment of people living with HIV as well as other prevention efforts that are aimed at directly reducing the number of new infections. Local estimates of HIV prevalence—particularly the number of people living with HIV—are useful for estimating the location-specific need for ART and other HIV-related services, and complement routinely collected clinical data that in some locations provide estimates of the number of

[1]Institute for Health Metrics and Evaluation, University of Washington, Seattle, WA, USA. [2]DHS program, ICF International, Rockville, MD, USA. [3]Department of Global Health, University of Washington, Seattle, WA, USA. [4]Department of Medicine, University of Washington, Seattle, WA, USA. [5]Department of Infectious Disease Epidemiology, Imperial College London, London, UK. [6]ICAP, Mailman School of Public Health, Columbia University, New York, NY, USA. [7]Vagelos College of Physicians and Surgeons, Columbia University, New York, NY, USA. [8]Medireal Investment Uganda, Entebbe, Uganda. [9]Public Health Medicine, School of Nursing and Public Health, College of Health Sciences, University of KwaZulu-Natal, Durban, South Africa. [10]School of Nursing and Public Health, University of KwaZulu-Natal, Durban, South Africa. [11]Africa Health Research Institute, KwaZulu-Natal, South Africa. [12]Centre for the AIDS Programme of Research in South Africa (CAPRISA), University of KwaZulu-Natal, Durban, South Africa. [13]Research Department of Infection & Population Health, University College London, London, UK. [14]HIV/AIDS, STIs & TB Research Programme, Human Sciences Research Council, Pretoria, South Africa. [15]London School of Hygiene & Tropical Medicine, London, UK. *e-mail: sihay@uw.edu

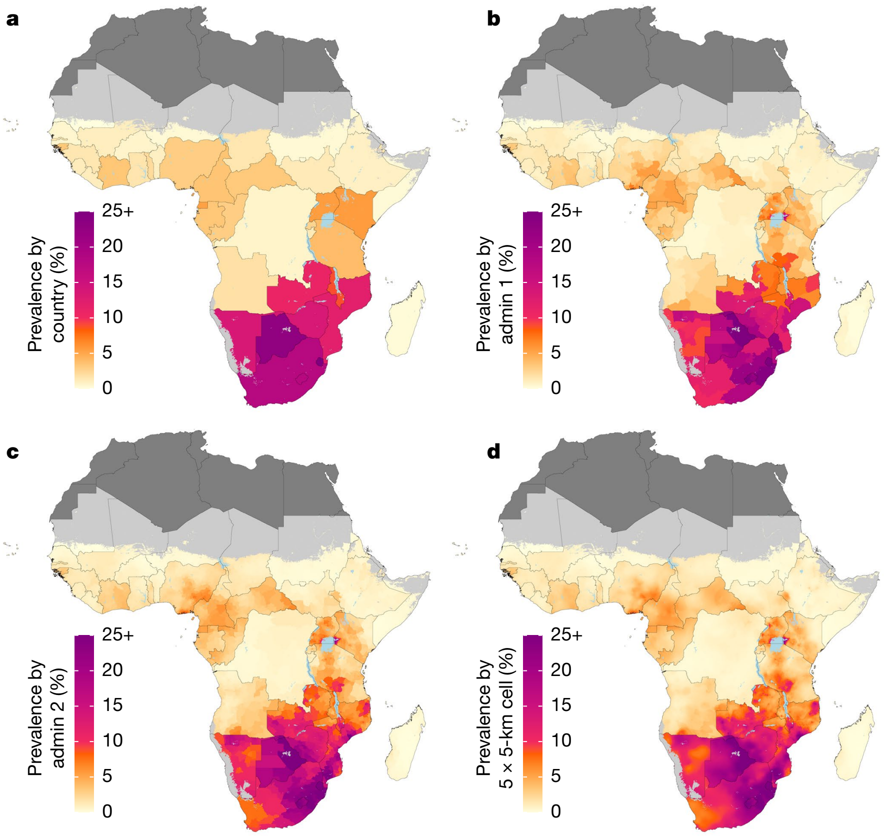

a
Prevalence by country (%)
25+
20
15
10
5
0
b
Prevalence by admin 1 (%)
25+
20
15
10
5
0
c
Prevalence by admin 2 (%)
25+
20
15
10
5
0
d
Prevalence by 5 × 5-km cell (%)
25+
20
15
10
5
0

Fig. 1 | Prevalence of HIV in adults aged 15–49 in 2017. a–**d**, Prevalence of HIV among adults aged 15–49 in 2017 at the country level (**a**), first administrative subdivision level (admin 1; **b**), second administrative subdivision level (admin 2; **c**) and 5 × 5-km grid-cell level (**d**). Maps reflect administrative boundaries, land cover, lakes and population; grid cells with fewer than 10 people per 1 × 1 km, and classified as barren or sparsely vegetated, are coloured light grey[25,26,37–40]. Countries in dark grey were not included in the analysis.

diagnosed individuals living with HIV. In terms of prevention, areas in which HIV prevalence is high and ART coverage is low are likely to have a high incidence of HIV[20,21]. In the absence of local information on HIV incidence, knowledge of the variation in HIV prevalence can be used to better target prevention efforts to those areas with the greatest need. Recognizing the importance of subnational heterogeneity in the HIV epidemic, UNAIDS and funding agencies—including the US President's Emergency Plan for AIDS Relief (PEPFAR) and the Global Fund to Fight AIDS, Tuberculosis and Malaria—have called for incorporating local data into strategies for addressing the HIV epidemic[22–24].

Although previous studies have examined subnational variation in HIV prevalence in select countries[11–19] (Supplementary Table 1), there is—to our knowledge—no comprehensive and comparable set of subnational HIV prevalence estimates for all of sub-Saharan Africa. Moreover, for most countries, existing estimates are for a single year and use a single data source. Here we present comprehensive space–time estimates of HIV prevalence among adults aged 15–49 years who reside in each area on a 5 × 5-km grid across 47 countries in sub-Saharan Africa, annually from 2000 to 2017. For this analysis, we constructed a geolocated database of HIV prevalence data from 134 surveys in 41 countries and 9,794 site-years of sentinel surveillance of antenatal care clinics at 1,858 unique sites in 46 countries (Extended Data Figs. 1–3). We adapted existing Bayesian spatiotemporal methods to analyse these data and produce gridded estimates of HIV prevalence, calibrated to national estimates from the GBD[1]. We additionally combined grid-cell-level estimates of HIV prevalence with grid-cell-level estimates of the population[25,26] aged 15–49 years to estimate the number of people living with HIV. Finally, for HIV prevalence, we calculated population-weighted averages of the grid-cell-level estimates to generate estimates for first-level administrative subdivisions (for example, provinces or regions) and second-level administrative subdivisions (for example, districts or departments) in each country. All estimates are publicly available from the Global Health Data Exchange (http://ghdx.healthdata.org/ihme-data/africa-hiv-prevalence-geospatial-estimates-2000-2017) and through a user-friendly data visualization tool (https://vizhub.healthdata.org/lbd/hiv).

Widespread differences in HIV prevalence

HIV prevalence varied substantially at the grid-cell level as well as among first and second administrative subdivisions throughout sub-Saharan Africa (Fig. 1, Extended Data Fig. 4 and Supplementary Figs. 1–4). This variation was apparent within countries with a relatively high overall HIV prevalence; for example, in Botswana (national prevalence, 22.8% (95% uncertainty interval, 19.8–26.1%)) prevalence among districts ranged from 15.1% (11.5–19.8%) in Ghanzi district to 27.7% (22.3–33.8%) in North-East district in 2017. This variation was also apparent in countries with a more moderate national HIV prevalence; for example, in Tanzania (national prevalence, 3.9% (3.6–4.3%)), prevalence among regions ranged from 0.4% (0.2–0.6%) in Kusini Pemba region to 9.1% (7.1–11.3%) in Njombe region in 2017. In countries in which levels of HIV prevalence are lower overall, the absolute differences among subnational units were necessarily smaller. However, in many instances, relative differences among subnational units remained large—for example, in the Democratic Republic of the Congo, in which national prevalence was 0.7% (0.6–0.9%), prevalence among second-level administrative subdivisions ranged from 0.3% (0.2–0.5%) in Lukaya district to 1.4% (0.8–2.3%) in the city Likasi in 2017. Most countries (36 out of 47) had a more than twofold difference in prevalence between the second-level administrative subdivisions with the lowest and highest estimated prevalence in 2017, and the largest difference was more than fivefold in 14 out of 47 countries.

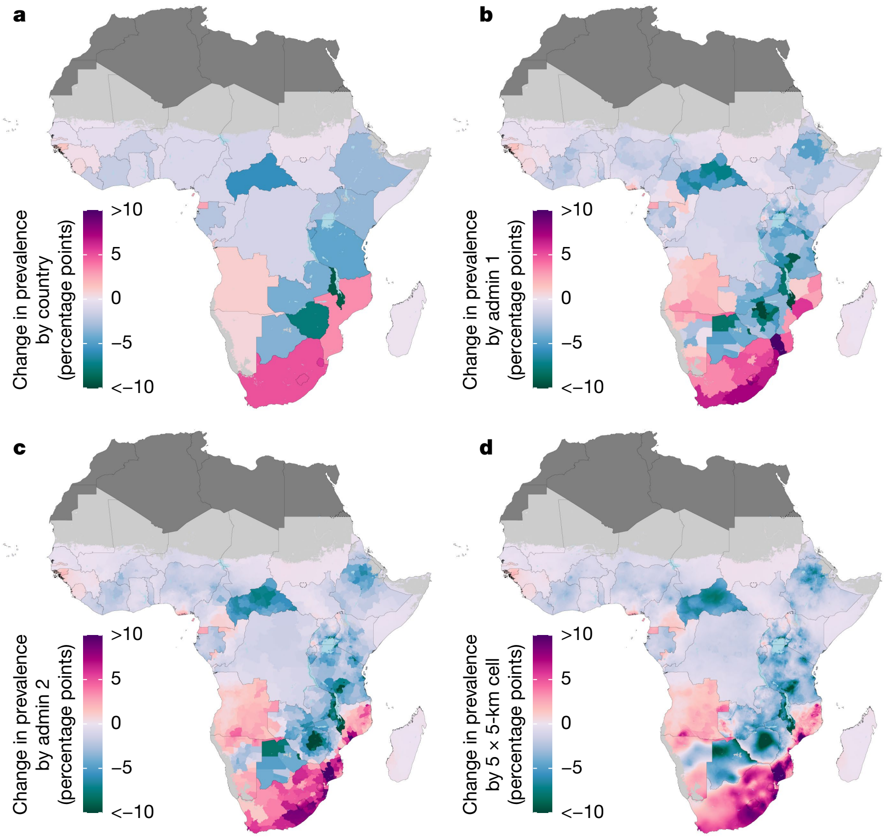

a
Change in prevalence by country (percentage points)
>10
5
0
−5
<−10
b
Change in prevalence by admin 1 (percentage points)
>10
5
0
−5
<−10
c
Change in prevalence by admin 2 (percentage points)
>10
5
0
−5
<−10
d
Change in prevalence by 5 × 5-km cell (percentage points)
>10
5
0
−5
<−10

Fig. 2 | Change in HIV prevalence in adults aged 15–49 from 2000 to 2017. a–**d**, Absolute change in HIV prevalence among adults aged 15–49 between 2000 and 2017 at the country level (**a**), first administrative subdivision level (**b**), second administrative subdivision level (**c**) and 5 × 5-km grid-cell level (**d**). Maps reflect administrative boundaries, land cover, lakes and population; grid cells with fewer than 10 people per 1 × 1 km, and classified as barren or sparsely vegetated, are coloured light grey[25,26,37–40]. Countries in dark grey were not included in the analysis.

At the country level (Fig. 1a), there was a clear divide between countries in southern sub-Saharan Africa (Botswana, Lesotho, Mozambique, Namibia, South Africa, Swaziland, Zambia and Zimbabwe), where estimated HIV prevalence exceeded 10% in 2017 and the rest of the continent, where prevalence was generally much lower. At subnational levels, however, there are areas outside of southern sub-Saharan Africa that nonetheless had a very high prevalence of HIV, including second-level administrative subdivisions in Kenya, Malawi, Uganda and Tanzania, where the estimated prevalence of HIV exceeded 10% in 2017 (Fig. 1c). Overall, the highest estimated prevalence observed in 2017 at the country level was 27.2% (23.6–31.1%) in Swaziland, compared to 28.3% (24.2–32.7%) in Lubombo province (Swaziland) at the first administrative level and 30.1% (25.2–35.4%) in Tikhuba constituency (Swaziland) at the second administrative level.

Local temporal changes in HIV prevalence

Between 2000 and 2017, estimated HIV prevalence at the country level increased in 15 out of 47 countries (Fig. 2a). At subnational levels, we estimated an increase in HIV prevalence in 22.9% of first-level administrative subdivisions (located in 24 countries) and in 25.0% of second-level administrative subdivisions (located in 28 countries) across sub-Saharan Africa (Fig. 2b, c; the posterior probability of an increase is shown in Supplementary Fig. 5). Although there was local heterogeneity, broad regional trends were apparent; the largest increases were found primarily in areas in coastal countries in southern sub-Saharan Africa and the largest decreases found primarily in a band stretching from Botswana to Kenya and in Central African Republic. Although in some places the direction and rate of change differed substantially on opposite sides of international borders (for example, between Botswana and South Africa), transnational patterns were also apparent—for example, the region that covered eastern South Africa and southern Mozambique.

There were substantial differences in both the direction and rate of change in HIV prevalence within many countries: 16 (34%) countries had areas in which the estimated HIV prevalence increased and areas in which the estimated HIV prevalence decreased among first-level administrative subdivisions (Fig. 2b). At the second administrative level this was true in 20 (42.6%) countries, and at the grid-cell level this was true in 28 (59.6%) countries (Fig. 2c, d). In some of these countries, the differences were substantial. For example, HIV prevalence declined by 5.8 percentage points (0.2–11.4 percentage points) in Manica district in Mozambique, whereas prevalence increased by 17.2 percentage points (9.3–26.1 percentage points) in Guija district. Similarly, prevalence declined by 14.3 percentage points (10.3–18.2 percentage points) in Chegutu district in Zimbabwe, whereas it increased by 0.6 percentage points (−4.1 to 5.0 percentage points) in Beitbridge district.

Changes in HIV prevalence from 2000 to 2017 in any given location were not generally linear or necessarily consistently in the same direction. Estimates of changes in prevalence over shorter periods within the overall 2000–2017 timeframe of this analysis highlight the variation within this period (Supplementary Figs. 6–8).

Local trends in the number of people living with HIV

Figure 3 shows the estimated number of people living with HIV by 5 × 5-km grid cell. As expected, given variation in population density and HIV prevalence, the number of people living with HIV per grid cell was highly variable and skewed: in 2017, we estimate that less than one person lives with HIV in 52.1% (50.6–53.4%) of grid cells, less than 10 people live with HIV in 83.8% (83.3–84.3%) of grid cells, less than 100 people living with HIV in 97.4% (97.2–97.5%) of grid cells and less than 1,000 people live with HIV in 99.8% (99.78–99.81%) of grid cells. Grid cells with large numbers of people living with HIV tend to have large populations in general. Although many of the grid cells that

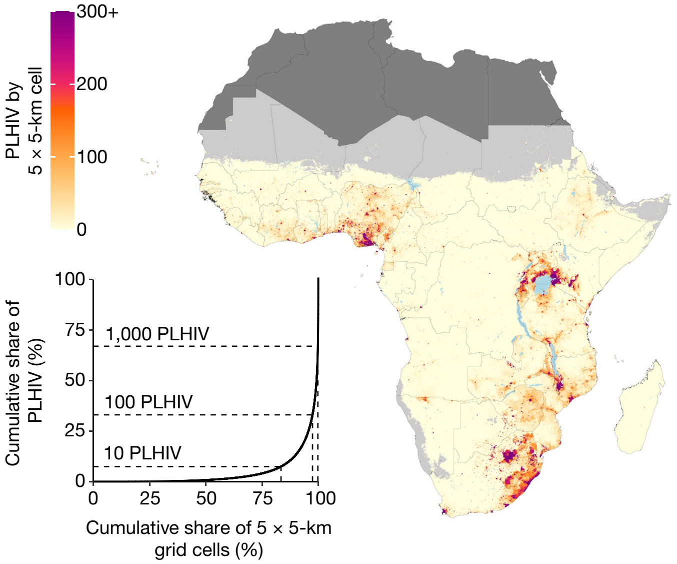


Fig. 3 | Number of people living with HIV for adults aged 15–49 in 2017. Number of people living with HIV (PLHIV) aged 15–49 in 2017 per 5 × 5-km grid cell (map) and Lorenz curve depicting the cumulative share of people living with HIV compared to the cumulative share of 5 × 5-km grid cells (inset). Maps reflect administrative boundaries, land cover, lakes and population; grid cells with fewer than 10 people per 1 × 1 km, and classified as barren or sparsely vegetated, are coloured light grey[25,26,37–40]. Countries coloured dark grey were not included in the analysis. In the inset, dotted lines indicate the cumulative share of people living with HIV and cumulative share of 5 × 5-km grid cells represented by grid cells with fewer than 10, 100 and 1,000 people living with HIV each.

have the largest number of people living with HIV are also grid cells with very high prevalence (which are located primarily in southern and south-eastern sub-Saharan Africa), there are also grid cells with more moderate HIV prevalence but large numbers of people living with HIV; these are located primarily in western Africa.

A large proportion of people who are living with HIV are concentrated in a small number of grid cells with high spatial concentrations of people who are living with HIV. Approximately one-third (34.3% (33.0–35.7%)) of people living with HIV in sub-Saharan Africa live in the 0.2% of grid cells in which it is estimated that there are more than 1,000 people living with HIV. A similarly large proportion of people living with HIV is distributed throughout the larger number of grid cells that have more moderate spatial concentrations of people living with HIV: 32.0% (30.6–33.4%) of people with HIV live in grid cells in which there are estimated to be fewer than 100 people with HIV, and 7.2% (6.7–7.7%) of people with HIV reside in grid cells in which there are estimated to be fewer than 10 people with HIV. The total number of people living with HIV aged 15–49 years in sub-Saharan Africa increased by 3.0 (1.8–4.4) million between 2000 and 2017, from 17.0 million (16.3–17.8) to 20.1 million (19.0–21.2). This increase was due to a corresponding increase in population, as prevalence in sub-Saharan Africa as a whole declined over this same period, from 5.5% (5.2–5.7%) in 2000 to 4.0% (3.8–4.2%) in 2017. The increase in people living with HIV was larger in locations with high spatial concentrations of people with HIV compared to those with fewer people living with HIV: in 2017, the total number of people with HIV in grid cells in which there are estimated to be fewer than 100 people with HIV was nearly identical (6.4 million (6.2–6.6)) to the number in 2000 (6.5 million (6.3–6.6)). However, the number of people living with HIV in grid cells in which there are estimated to be more than 1,000 people increased by 37.5%, from 5.0 million (4.7–5.3) to 6.9 million (6.3–7.5).

Discussion

This study provides a comprehensive quantification of subnational trends in HIV prevalence and the number of people living with HIV in sub-Saharan Africa. These estimates highlight substantial differences between and within countries in levels and trends in HIV prevalence and the spatial concentration of people living with HIV. For discussion of the advantages of this analysis compared to earlier analyses, important limitations of the present analysis and potential future directions, see Supplementary Discussion.

Subnational estimates of HIV prevalence can be used to more efficiently target resources and interventions. The WHO (World Health Organization) recommends ART for all people living with HIV[27], and the UNAIDS fast-track strategy emphasizes the importance of treatment and diagnosis[7]. Estimates of the prevalence of HIV and the number of people living with HIV at local levels provide important information about the number of people who are potentially in need of diagnosis and treatment services. Additionally, in the absence of local information on HIV incidence, information about HIV prevalence can be used to target primary prevention strategies: modelling studies that compare geographically targeted to non-geographically targeted prevention strategies have found that geographically targeted strategies are more efficient in preventing new HIV infections under the same budgetary constraints[11,28]. Moreover, previous research has highlighted the potential role of geographical 'hot spots' as a source of HIV transmission both locally and further afield, which suggests that targeted prevention strategies may reduce the incidence of HIV not only in targeted areas but also more broadly[29,30].

Our analysis highlights several challenges to bringing HIV infection under control in Africa. Growing population size coupled with continued high incidence[1,4] of new HIV infections and increased life expectancy among people living with HIV[31–34] has led to an increase in the number of people living with HIV in sub-Saharan Africa since 2000. Despite this increase, spending on HIV in sub-Saharan Africa has declined in recent years, largely as a result of a reduction in development assistance for health[9]. Our estimates also highlight the diversity of the HIV epidemic: although a large number of people living with HIV are concentrated in a few select areas (Fig. 3), a similarly large number are living in areas with a relatively low spatial concentration of people living with HIV. The most effective treatment and prevention strategies probably differ between areas in which many people live with HIV and those with a smaller number of people living with HIV, and economies of scale may be harder to realize in the latter case. Nonetheless, it is essential to ensure that people living with HIV have access to appropriate health services regardless of their location.

The results of this analysis describe a multifaceted picture of patterns of changing HIV prevalence across sub-Saharan Africa, with many areas experiencing increases over the same period in which other areas experienced declines. Changes in HIV prevalence are the outcome of a complex interaction between incidence, mortality and migration patterns. Globally, the large-scale expansion of ART coverage has reduced mortality among people living with HIV, offsetting declines in incidence and resulting in an overall increase in HIV prevalence since 2000[1,4,35]. At the region and country levels, trends in mortality and incidence have varied, which has resulted in differing trends in the prevalence of HIV[1,4,35]. Exploration of this dynamic at a subnational level is warranted, although it is complicated by the relative lack of directly observed empirical data on HIV incidence and mortality in sub-Saharan Africa[36]. Nonetheless, existing evidence indicates that subnational increases in prevalence should not be interpreted as inherently alarming without additional consideration of incidence and mortality trends.

Despite progress in recent decades, HIV continues to impose a substantial health burden on countries in sub-Saharan Africa. The estimates from this analysis highlight the degree to which the effect of this epidemic varies, even within countries. These local data provide a new tool for policymakers, programme implementers and researchers to use to assess local needs, efficiently target interventions and ultimately work towards bringing HIV infection under control in Africa.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-019-1200-9.

Received: 13 September 2018; Accepted: 10 April 2019;
Published online 15 May 2019.

1. GBD 2017 Disease and Injury Incidence and Prevalence Collaborators. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet* **392**, 1789–1858 (2018).
2. GBD 2017 Causes of Death Collaborators. Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet* **392**, 1736–1788 (2018).
3. Teeraananchai, S., Kerr, S. J., Amin, J., Ruxrungtham, K. & Law, M. G. Life expectancy of HIV-positive people after starting combination antiretroviral therapy: a meta-analysis. *HIV Med.* **18**, 256–266 (2017).
4. Joint United Nations Programme on HIV/AIDS. *AIDSinfo*. http://aidsinfo.unaids.org/ (UNAIDS, 2018).
5. United Nations Development Programme. *The Millennium Development Goals Report 2015*. http://www.undp.org/content/undp/en/home/librarypage/mdg/the-millennium-development-goals-report-2015.html. (United Nations, 2015).
6. United Nations. *Transforming our World: The 2030 Agenda for Sustainable Development*. https://sustainabledevelopment.un.org/post2015/transformingourworld/publication (2015).
7. Joint United Nations Programme on HIV/AIDS. *Fast-Track—Ending the AIDS Epidemic by 2030*. http://www.unaids.org/en/resources/documents/2014/JC2686_WAD2014report (UNAIDS, 2014).
8. Bekker, L.-G. et al. Advancing global health and strengthening the HIV response in the era of the Sustainable Development Goals: the International AIDS Society—*Lancet* Commission. *Lancet* **392**, 312–358 (2018).
9. Global Burden of Disease Health Financing Collaborator Network. Spending on health and HIV/AIDS: domestic health spending and development assistance in 188 countries, 1995–2015. *Lancet* **391**, 1799–1829 (2018).
10. Piot, P. et al. The global epidemiology of HIV infection: continuity, heterogeneity, and change. *J. Acquir. Immune Defic. Syndr.* **3**, 403–412 (1990).
11. Anderson, S.-J. et al. Maximising the effect of combination HIV prevention through prioritisation of the people and places in greatest need: a modelling study. *Lancet* **384**, 249–256 (2014).
12. Kleinschmidt, I., Pettifor, A., Morris, N., MacPhail, C. & Rees, H. Geographic distribution of human immunodeficiency virus in South Africa. *Am. J. Trop. Med. Hyg.* **77**, 1163–1169 (2007).
13. Kandala, N.-B., Campbell, E. K., Rakgoasi, S. D., Madi-Segwagwe, B. C. & Fako, T. T. The geography of HIV/AIDS prevalence rates in Botswana. *HIV AIDS* **4**, 95–102 (2012).
14. Larmarange, J. & Bendaud, V. HIV estimates at second subnational level from national population-based surveys. *AIDS* **28**, S469–S476 (2014).
15. Okano, J. T. & Blower, S. Sex-specific maps of HIV epidemics in sub-Saharan Africa. *Lancet Infect. Dis.* **16**, 1320–1322 (2016).
16. Carrel, M. et al. Changing spatial patterns and increasing rurality of HIV prevalence in the Democratic Republic of the Congo between 2007 and 2013. *Health Place* **39**, 79–85 (2016).
17. Coburn, B. J., Okano, J. T. & Blower, S. Using geospatial mapping to design HIV elimination strategies for sub-Saharan Africa. *Sci. Transl. Med.* **9**, eaag0019 (2017).
18. Cuadros, D. F. et al. Mapping the spatial variability of HIV infection in sub-Saharan Africa: effective information for localized HIV prevention and control. *Sci. Rep.* **7**, 9093 (2017).
19. Meyer-Rath, G. et al. Targeting the right interventions to the right people and places: the role of geospatial analysis in HIV program planning. *AIDS* **32**, 957–963 (2018).
20. Bärnighausen, T. et al. High HIV incidence in a community with high HIV prevalence in rural South Africa: findings from a prospective population-based study. *AIDS* **22**, 139–144 (2008).
21. Tanser, F. et al. Effect of population viral load on prospective HIV incidence in a hyperendemic rural African community. *Sci. Transl. Med.* **9**, eaam8012 (2017).
22. Joint United Nations Programme on HIV/AIDS. *On the Fast-Track to end AIDS by 2030: Focus on Location and Population*. http://www.unaids.org/en/resources/documents/2015/FocusLocationPopulation (UNAIDS, 2015).
23. Office of the Global AIDS Coordinator. *PEPFAR 3.0. Controlling the Epidemic: Delivering on the Promise of an AIDS-free Generation*. https://www.pepfar.gov/documents/organization/234744.pdf (US Department of State, 2014).
24. The Global Fund to Fight AIDS, Tuberculosis and Malaria. *The Global Fund Strategy 2017–2022: Investing to End Epidemics*. https://www.theglobalfund.org/media/2531/core_globalfundstrategy2017-2022_strategy_en.pdf (2017).
25. WorldPop. *WorldPop Dataset*. http://www.worldpop.org.uk/data/get_data/ (accessed 7 July 2017).
26. Tatem, A. J. WorldPop, open data for spatial demography. *Sci. Data* **4**, 170004 (2017).
27. World Health Organization. *Guideline on When to Start Antiretroviral Therapy and on Pre-exposure Prophylaxis for HIV*. http://www.ncbi.nlm.nih.gov/books/NBK327115/ (WHO, Geneva, 2015).
28. McGillen, J. B., Anderson, S.-J., Dybul, M. R. & Hallett, T. B. Optimum resource allocation to reduce HIV incidence across sub-Saharan Africa: a mathematical modelling study. *Lancet HIV* **3**, e441–e448 (2016).
29. Cuadros, D. F., Graf, T., de Oliveira, T., Bärnighausen, T. & Tanser, F. Assessing the role of geographical HIV hot-spots in the spread of the epidemic. In *Proc. Conference on Retroviruses and Opportunistic Infections* http://www.croiconference.org/sessions/assessing-role-geographical-hiv-hot-spots-spread-epidemic (2018).
30. Tanser, F., Bärnighausen, T., Dobra, A. & Sartorius, B. Identifying ‘corridors of HIV transmission’ in a severely affected rural South African population: a case for a shift toward targeted prevention strategies. *Int. J. Epidemiol.* **47**, 537–549 (2018).
31. Reniers, G. et al. Mortality trends in the era of antiretroviral therapy: evidence from the Network for Analysing Longitudinal Population based HIV/AIDS data on Africa (ALPHA). *AIDS* **28**, S533–S542 (2014).
32. Johnson, L. F. et al. Estimating the impact of antiretroviral treatment on adult mortality trends in South Africa: a mathematical modelling study. *PLoS Med.* **14**, e1002468 (2017).
33. Zaidi, J., Grapsa, E., Tanser, F., Newell, M.-L. & Bärnighausen, T. Dramatic increases in HIV prevalence after scale-up of antiretroviral treatment. *AIDS* **27**, 2301–2305 (2013).
34. Granich, R. et al. Trends in AIDS deaths, new infections and ART coverage in the top 30 countries with the highest AIDS mortality burden; 1990–2013. *PLoS ONE* **10**, e0131353 (2015).
35. GBD 2015 HIV Collaborators. Estimates of global, regional, and national incidence, prevalence, and mortality of HIV, 1980–2015: the Global Burden of Disease Study 2015. *Lancet HIV* **3**, e361–e387 (2016).
36. Ghys, P. D., Williams, B. G., Over, M., Hallett, T. B. & Godfrey-Faussett, P. Epidemiological metrics and benchmarks for a transition in the HIV epidemic. *PLoS Med.* **15**, e1002678 (2018).
37. GeoNetwork. *Global Administrative Unit Layers (GAUL)*. http://www.fao.org/geonetwork/srv/en/metadata.show?id=%2012691 (2015).
38. Land Processes Distributed Active Archive Center. *Combined MODIS 5.1 dataset. MCD12Q1 | LP DAAC: NASA Land Data Products and Services* (accessed 1 June 2017).
39. Lehner, B. & Döll, P. Development and validation of a global database of lakes, reservoirs and wetlands. *J. Hydrol.* **296**, 1–22 (2004).
40. World Wildlife Fund. *Global Lakes and Wetlands Database Level 3*. https://www.worldwildlife.org/pages/global-lakes-and-wetlands-database (World Wildlife Fund, 2004).

Acknowledgements This work was primarily supported by grant OPP1132415 from the Bill & Melinda Gates Foundation. J.W.E. was supported by grant R03-AI125001 from the National Institutes of Health.

Reviewer information *Nature* thanks Laith Jamal Abu Raddad, Emanuele Giorgi, Andrew B. Lawson and Brian Rice for their contribution to the peer review of this work.

Author contributions S.I.H. and L.D.-L. conceived and planned the study. L.D.-L., A.S., K.M.S., K.F.W., N.R.P., B.K.M., M.L.C., M.H.B., A.C., T.F., D.D.-S., J.W.E., A.O., B.S., F.T. and N.W. identified and obtained data for analysis. K.M.S., K.F.W., N.R.P., M.L.C. and J.B.H. extracted, processed and geopositioned the data. L.D.-L., M.A.C., B.K.M. and J.D.V. carried out the statistical analyses with assistance and input from R.B., D.C.C., A.D., N.J.H., D.K., A.O.-Z., D.P., R.C.R., J.M.R. and K.E.W. L.D.-L., M.A.C., A.S., K.M.S., K.F.W., N.R.P., B.K.M., J.D.V., M.L.C., J.B.H., M.H.B., A.C., T.F., D.D.-S., R.B., D.C.C., A.D., L.E., C.E.B., T.H.F., N.J.H., D.K., L.B.M., M.R.N., A.O.-Z., D.P., R.C.R., J.M.R., L.E.S., D.L.S., N.D.W., K.E.W., J.W.E., J.E.J., A.O., B.S., F.T., N.W., P.P., C.J.L.M. and S.I.H. provided intellectual input into aspects of this study. L.D.-L., M.A.C., K.M.S., K.F.W., N.R.P., J.D.V. and L.E. prepared figures and tables. L.D.-L. wrote the first draft of the manuscript with assistance from M.A.C., A.S., K.M.S., K.F.W., N.R.P. and J.D.V., and B.K.M., R.B., D.C.C., A.D., L.E., T.H.F., N.J.H., D.K., L.B.M., A.O.-Z., D.P., R.C.R., J.M.R., L.E.S., D.L.S., N.D.W., K.E.W., J.W.E., J.E.J., A.O., B.S., F.T., N.W., P.P., C.J.L.M. and S.I.H. contributed to subsequent revisions.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-019-1200-9.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1200-9.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to S.I.H.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Overview. Our study follows the Guidelines for Accurate and Transparent Health Estimates Reporting (GATHER). This analysis provides estimates of HIV prevalence among adults aged 15–49 on a 5 × 5-km grid in 47 countries in sub-Saharan Africa, with annual resolution, from 2000 to 2017. The period of 2000–2017 and the age group of 15–49 years were selected to optimize the contemporaneousness of the estimates and to maximize data availability—there were relatively few large-scale seroprevalence surveys conducted before 2000, and most seroprevalence surveys focus on adults, in which 15–49 years was the most commonly reported age range. The methodology used here is similar to that used for previous analyses of mortality in children under 5 years of age[41], child growth failure[42] and education[43] in Africa. We used a 5 × 5-km grid for consistency with these previous analyses; to align with the resolution available for pre-existing covariates incorporated in this analysis; and for flexibility in aggregating these estimates to other levels of interest (for example, first- and second-order administrative subdivisions). Extended Data Figure 5 provides an overview of the analytic process. Each step is described below and additional details are available in the Supplementary Information, including a discussion of the limitations of this approach.

HIV data. We compiled a dataset of 29,103 data points from 134 seroprevalence surveys in 41 countries and 9,794 data points from sentinel surveillance of antenatal care clinics (ANC data) in 46 countries. Data from seroprevalence surveys were originally in one of three forms: survey microdata (that is, individual-level survey responses), survey reports or published literature (Supplementary Table 2). For surveys with available microdata, we extracted variables related to age, HIV blood test result, location and survey weights. After subsetting the data to ages 15–49 years and excluding rows with missing information on any of these variables, we collapsed the data by calculating the weighted HIV prevalence at the finest spatial resolution available. Ideally, this was at the level of the GPS coordinates that represent the location of a survey cluster, but in instances for which GPS data were not available, the smallest areal unit (termed a polygon) possible was used instead, typically representing an administrative subdivision. For surveys for which microdata were unavailable but for which estimates with some subnational resolution were provided in a report or published literature, we extracted these estimates along with information about the sample size and location. Where possible, these data were matched to a specific set of GPS coordinates, and otherwise were matched to a polygon, which most-often represented an administrative subdivision. In some instances, estimates extracted from reports or published literature were for age groups other than 15–49 years (34 sources representing 1.76% of the total effective sample size; Supplementary Table 3). In these instances, we used a cross-walking model—that is, an approach for linking disparate data sources (in this case data sources reporting for different age groups)—that leveraged existing microdata and linear regression to translate the prevalence in the reported age range to the standard 15–49 age range (Supplementary Information, section 2.3).

ANC data were primarily derived from national HIV estimate files developed by national teams and compiled and shared via UNAIDS[44], and supplemented with data derived from sentinel surveillance country reports (Supplementary Table 4). In both instances, we extracted information on HIV prevalence and sample size by site and year. Sites were geolocated to specific GPS coordinates where possible and otherwise to a polygon that represents an administrative subdivision.

In instances in which data were matched to a polygon rather than specific GPS coordinates, we resampled these data to mimic point data. Specifically, for each observation, we randomly sampled 10,000 candidate locations within the associated polygon with a probability proportional to the population and then used *k*-means clustering to generate a reduced set of locations based on the centroid of each *k*-means cluster. Each of these resulting pseudo-points was assigned the HIV prevalence observed for the polygon as a whole, and the sample size was set to the observed sample size for the polygon as a whole multiplied by the fraction of candidate locations that belonged to that *k*-means cluster. Weighting by sample size, 78.0% of all data (including 61.1% of survey data and 83.5% of ANC data) were associated with GPS coordinates, and the remaining data were associated with polygons and were analysed using this approach.

Covariates. This analysis included five pre-existing covariates: (1) travel time to the nearest settlement of more than 50,000 inhabitants; (2) total population; (3) night-time lights; (4) urbanicity; and (5) malaria incidence (Supplementary Table 5). In addition, eight covariates were constructed explicitly for this analysis owing to their known association with HIV prevalence and data availability: (1) prevalence of male circumcision (all forms); (2) prevalence of self-reported STI symptoms; (3) prevalence of marriage or living with a partner as married; (4) prevalence of one's current partner living elsewhere; (5) prevalence of condom use at last sexual encounter; (6) prevalence of reporting ever having had intercourse among young adults; and (7) and (8) prevalence of multiple partners in the past year for men and for women (Extended Data Fig. 6). These eight covariates were constructed based on survey data collected and analysed analogously to the HIV data (described above), and using geostatistical models similar to those described in the next section (Supplementary Table 6 and Supplementary Figs. 9–16). In addition, calendar year was used as a covariate.

Statistical model. *Covariate stacking.* An ensemble covariate modelling approach was implemented to capture possible nonlinear effects and complex interactions among these covariates[45]. For each modelling region (Extended Data Fig. 7), three sub-models were fitted to the HIV survey data with the covariates as explanatory predictors: generalized additive models, boosted regression trees and lasso regression. Each sub-model was fitted using fivefold cross-validation to avoid overfitting, and the out-of-sample predictions from across the five folds were compiled into a single set of predictions that were used to fit the geostatistical model described below. In addition, each sub-model was also fitted to the full dataset to generate a complete set of in-sample predictions that were subsequently used when generating predictions from the geostatistical model (Supplementary Figs. 17–19).

Geostatistical model. We modelled HIV prevalence using a spatially and temporally explicit generalized linear mixed effects model:

$$Y_{i,t} \sim \text{binomial}(p_{i,t}, N_{i,t})$$

$$\text{logit}(p_{i,t}) = \beta_0 + \boldsymbol{\beta}_1 \mathbf{X}_{i,t} + \gamma_{c[i]} + Z_{i,t} + \epsilon_{i,t} + (\beta_2 + U_i) I_{\text{ANC}}$$

$$\gamma_{c[i]} \sim \text{normal}(0, \sigma^2_{\text{country}})$$

$$Z_{i,t} \sim \text{GP}(0, \Sigma_{\text{space}} \otimes \Sigma_{\text{time}})$$

$$\epsilon_{i,t} \sim \text{normal}(0, \sigma^2_{\text{nugget}})$$

$$U_i \sim \text{GP}(0, \Sigma_{\text{space}})$$

in which $\sim$ denotes 'distributed as'. We modelled the number of HIV-positive individuals ($Y_{i,t}$) among a sample ($N_{i,t}$) in location *i* and year *t* as a binomial variable. This model specified logit-transformed HIV prevalence ($p_{i,t}$) as a linear combination of a regional intercept (β_0), covariate effects ($\boldsymbol{\beta}_1 \mathbf{X}_{i,t}$), country random effects ($\gamma_{c[i]}$), spatially and temporally correlated random effects ($Z_{i,t}$) and an uncorrelated error term or nugget effect ($\epsilon_{i,t}$). HIV prevalence as measured by sentinel surveillance of antenatal care clinics is known to be biased as a measure of HIV prevalence in the general adult population, because it only covers pregnant women who attend ANC, compared to all adult men and women[46,47]. In instances in which data in our model were derived from ANC sentinel surveillance ($I_{\text{ANC}} = 1$), our model allowed for this bias using a fixed term (β_2) that captured the overall mean bias and a spatially varying term (U_i) that captured local differences in the extent of this bias. In this model, the spatially and temporally correlated random effect ($Z_{i,t}$) was modelled as a Gaussian process with mean 0 and a covariance matrix given by the Kronecker product of a spatial Matérn covariance function (Σ_{space}) and a temporal first-order autoregressive covariance function (Σ_{time}). U_i was modelled as a Gaussian process with mean 0 and spatial Matérn covariance (Σ_{space}). Sensitivity analyses were carried out to assess sensitivity to hyper-prior specification and are described in detail in the Supplementary Information, section 4.2.

This model was fitted in R-INLA[48] using the stochastic partial differential equation[49] approach to approximate the continuous spatial and spatio-temporal Gaussian random fields (U_i and $Z_{i,t}$, respectively). Owing to computational constraints, and to allow for regional differences in the relationship between the covariates and HIV prevalence, as well as differences in the temporal and spatial autocorrelation in HIV prevalence, separate models were fitted for each of the four regions (Extended Data Fig. 7). From each fitted model, we generated 1,000 draws from the approximated joint posterior distribution of all model parameters and used these to construct 1,000 draws of $p_{i,t}$, setting I_{ANC} to 0. Fivefold cross-validation was used to assess model performance and to compare among a number of alternative models that use covariates, ANC data and polygon data in a variety of ways (Supplementary Figs. 20–25 and Supplementary Information, section 4.3).

Post-estimation. To take advantage of the more structured modelling approach and additional national-level data used by GBD 2017, we performed post hoc calibration of our estimates to the corresponding national-level GBD estimates[1]. For each country and year in our analysis, we defined a raking factor equal to the ratio of the GBD estimate for this country and year to the population-weighted posterior mean HIV prevalence in all grid cells within this country and year (Supplementary Fig. 26). These raking factors were then used to scale each draw of HIV prevalence for each grid cell within that GBD geography and year. Point estimates for each grid cell were calculated as the mean of the scaled draws, and 95% uncertainty intervals were calculated as the 2.5th and 97.5th percentiles of the scaled draws. Grid cells that crossed international borders within modelling regions were fractionally allocated to multiple countries in proportion to the covered area during this process.

In addition to estimates of HIV prevalence on a 5 × 5-km grid, we constructed estimates of HIV prevalence for first- and second-level administrative subdivisions by calculating population-weighted averages of prevalence for all grid cells within a given area. This process was carried out for each of the 1,000 posterior draws (after calibration to GBD) with final point estimates derived from the mean of these draws and uncertainty intervals from the 2.5th and 97.5th percentiles. Additionally, estimates of the number of people living with HIV for each grid cell were derived by multiplying estimated prevalence in each grid cell by the corresponding population estimate from WorldPop[25,26], which was also calibrated to match GBD 2017[50] (Supplementary Information, section 4.4). As with calibration, grid cells that crossed borders were fractionally allocated to multiple areas when calculating aggregated prevalence estimates and estimates of people living with HIV.

Although the model makes predictions for all locations that are covered by available covariates, all final model outputs for which the land cover was classified as barren or sparsely vegetated on the basis of MODIS satellite data and for which the total population density was less than 10 individuals per 1 × 1 km in 2015 were masked for improved clarity when communicating with data specialists and policymakers.

Limitations. This analysis is subject to several limitations (further discussed in the Supplementary Information, section 5.2). Most importantly, the accuracy of our estimates is dependent on the quantity and quality of the underlying data. We have constructed a large database of geolocated HIV prevalence data for the purposes of this analysis. Nonetheless, important gaps in data coverage, both spatial and temporal, remain (Extended Data Figs. 1–3). Data quality is also likely to be variable and may be problematic for some data sources or locations. For HIV seroprevalence surveys, potential non-response bias is a particular concern[51] and the quality of the underlying data that are used to generate the covariate surfaces may also be suboptimal in some situations—for example, if cultural context influences the interpretation of a survey question or the response to potentially sensitive questions regarding sexual behaviour[52]. The information on locations that is associated with the data used in this analysis is also subject to some error and uncertainty. For example, in most surveys, GPS coordinates are randomly displaced (typically by 2–5 km) to protect the confidentiality of respondents [53] and some data sources have relatively non-specific location information (for example, districts or provinces instead of GPS coordinates). Primarily as a consequence of gaps in data coverage as well as the relative sparsity and small sample sizes in existing data sources disaggregated at small subnational levels, our estimates at the grid cell level—and to a lesser extent at the second and first administrative level—are associated with considerable uncertainty (Extended Data Fig. 4 and Supplementary Figs. 1–4). In the future, additional data collection, increased access to existing datasets (including detailed location information) and new strategies for using non-traditional data sources such as routine healthcare facility data[54] will be needed to improve the precision of these estimates at all levels.

The modelling strategy incorporates a number of assumptions, which—if incorrect—may lead to error. Additionally, the model fitting and prediction strategy used an integrated nested Laplace approximation to the posterior distribution, as implemented in R-INLA[48], as well as further approximations to generate predictions; these approximations may also introduce error. Although it is difficult to assess the effect of these assumptions and approximations, our validation analyses showed that our final model had minimal bias and a good coverage of the 95% prediction intervals, which provides some reassurance that the approximation method used—as well as other potential sources of error—did not result in appreciable bias or poorly described uncertainty in our reported estimates.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The findings of this study are supported by data that are available in public online repositories, data that are publicly available upon request from the data provider and data that are not publicly available owing to restrictions by the data provider and that were used under a license for the current study (including select data sources in Burkina Faso, Burundi, Chad, Eritrea, Nigeria, Sierra Leone, Uganda and Zambia, as indicated in Supplementary Tables 2, 6). A detailed description of data sources can be found in Supplementary Tables 2, 4–6. More information about each data source is available on the Global Health Data Exchange (http://ghdx.healthdata.org/), including information about the data provider and links to where the data can be accessed or requested (where available).

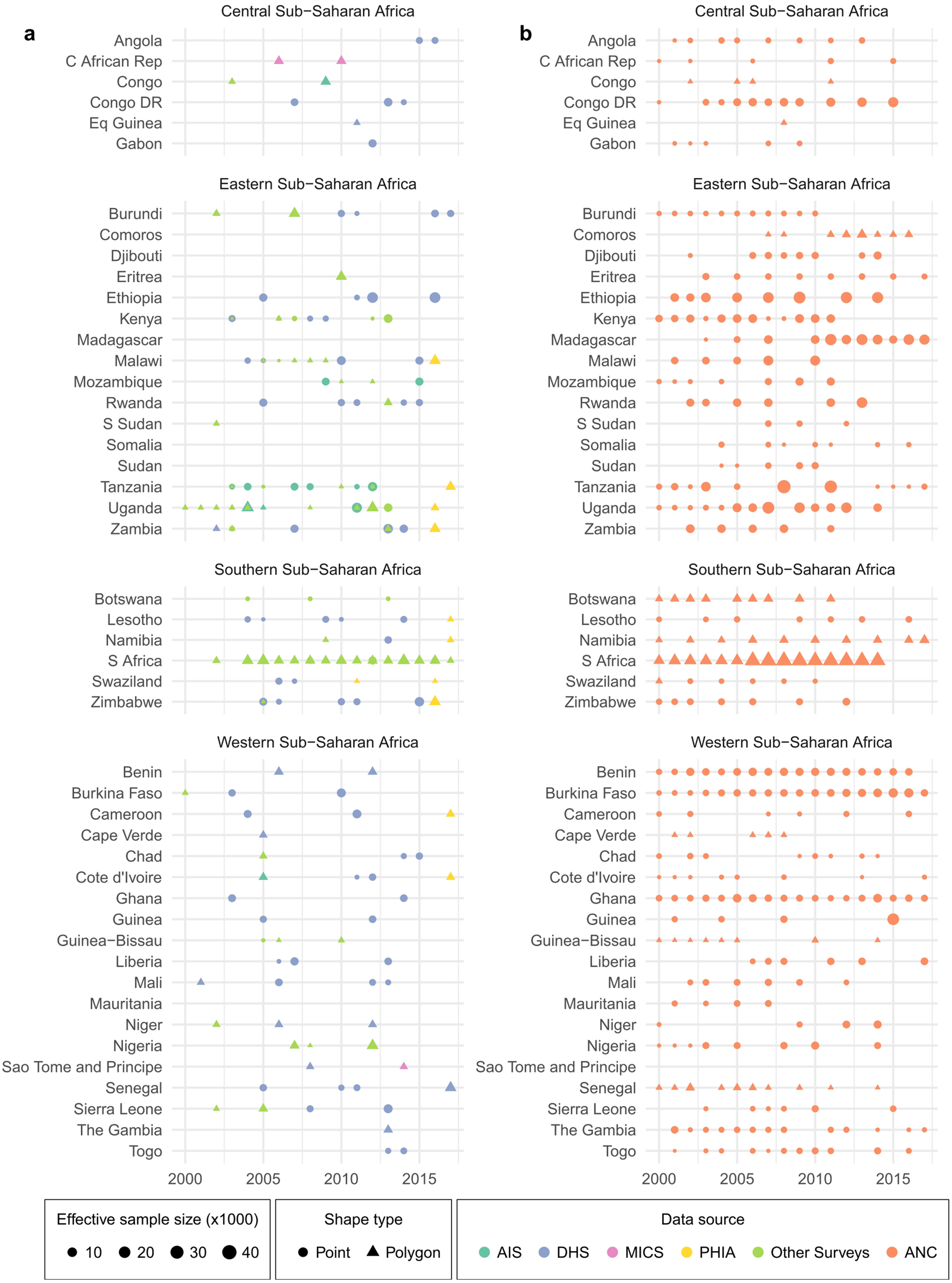
Administrative boundaries were retrieved from the Global Administrative Unit Layers dataset, implemented by FAO within the CountrySTAT and Agricultural Market Information System projects[37]. Land cover data were retrieved from the online Data Pool, courtesy of the NASA EOSDIS Land Processes Distributed Active Archive Center, USGS/Earth Resources Observation and Science Center[38]. Lakes were retrieved from the Global Lakes and Wetlands Database, courtesy of the World Wildlife Fund and the Center for Environmental Systems Research[39,40]. Populations were retrieved from WorldPop[25,26].

All estimates produced as part of this analysis are publicly available from the Global Health Data Exchange (http://ghdx.healthdata.org/ihme-data/africa-hiv-prevalence-geospatial-estimates-2000-2017) and via a user-friendly data visualization tool (https://vizhub.healthdata.org/lbd/hiv).

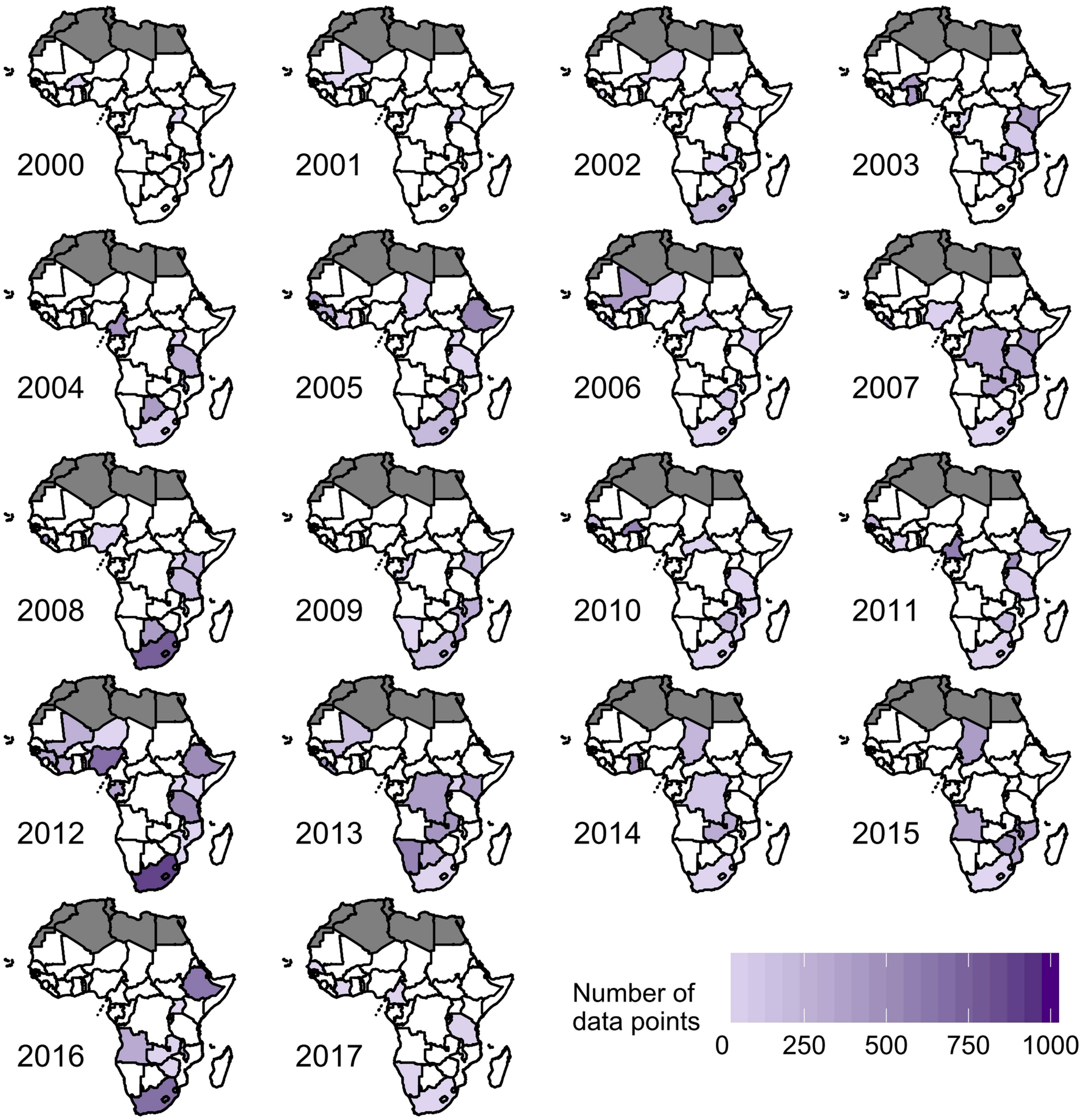
Code availability

All code used for these analyses is publicly available at https://github.com/ihmeuw/lbd/tree/hiv-africa-2019.

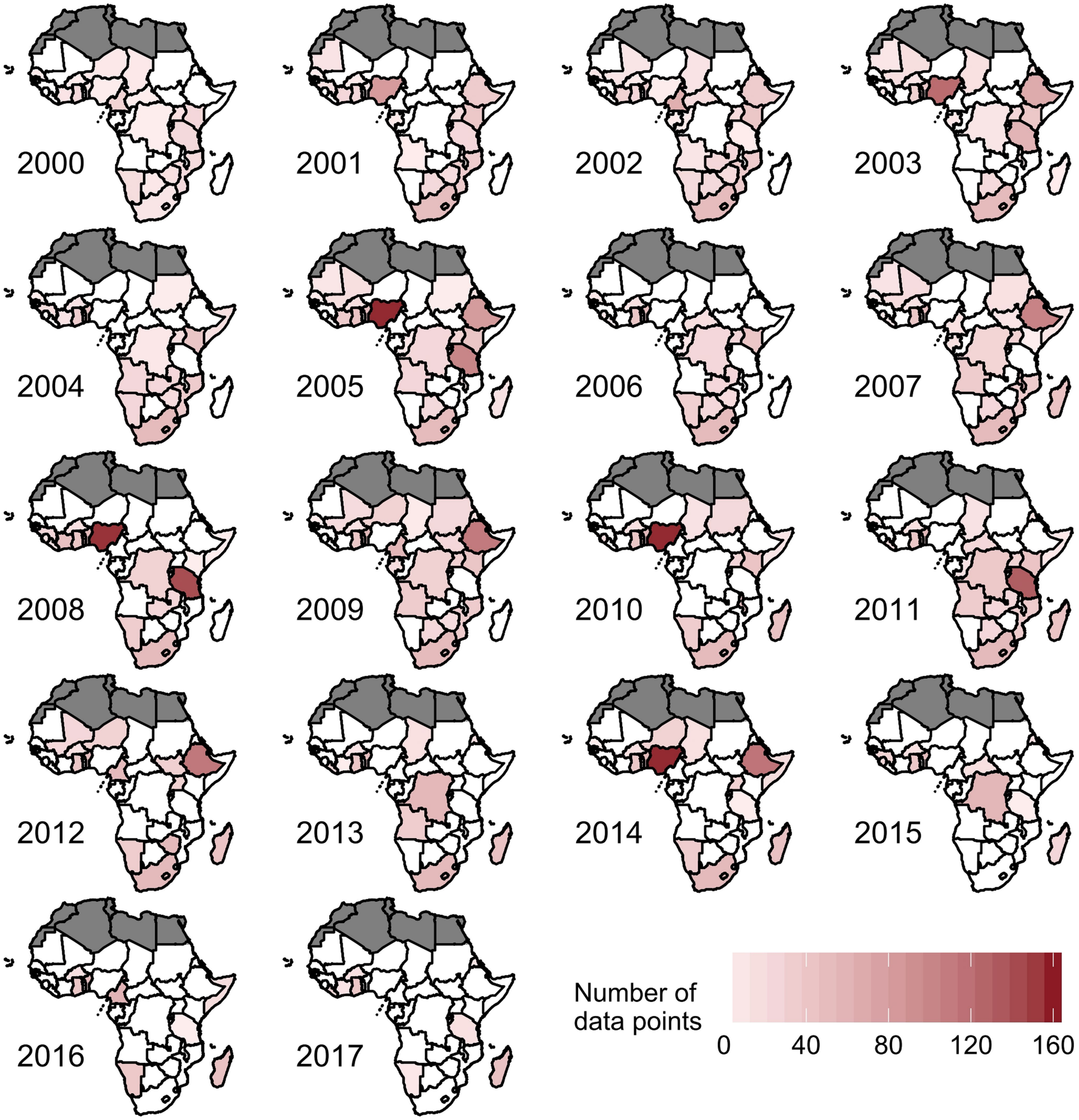
41. Golding, N. et al. Mapping under-5 and neonatal mortality in Africa, 2000–15: a baseline analysis for the Sustainable Development Goals. *Lancet* **390**, 2171–2182 (2017).
42. Osgood-Zimmerman, A. et al. Mapping child growth failure in Africa between 2000 and 2015. *Nature* **555**, 41–47 (2018).
43. Graetz, N. et al. Mapping local variation in educational attainment across Africa. *Nature* **555**, 48–53 (2018).
44. Joint United Nations Programme on HIV/AIDS. *National HIV Estimates File*. http://www.unaids.org/en/dataanalysis/datatools/spectrum-epp (UNAIDS, 2017).
45. Bhatt, S. et al. Improved prediction accuracy for disease risk mapping using Gaussian process stacked generalization. *J. R. Soc. Interface* **14**, https://doi.org/10.1098/rsif.2017.0520 (2017).
46. Gouws, E., Mishra, V. & Fowler, T. B. Comparison of adult HIV prevalence from national population-based surveys and antenatal clinic surveillance in countries with generalised epidemics: implications for calibrating surveillance data. *Sex. Transm. Infect.* **84**, i17–i23 (2008).
47. Marsh, K., Mahy, M., Salomon, J. A. & Hogan, D. R. Assessing and adjusting for differences between HIV prevalence estimates derived from national population-based surveys and antenatal care surveillance, with applications for Spectrum 2013. *AIDS* **28**, S497–S505 (2014).
48. Rue, H., Martino, S. & Chopin, N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc.* **71**, 319–392 (2009).
49. Lindgren, F., Rue, H. & Lindström, J. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *J. R. Stat. Soc.* **73**, 423–498 (2011).
50. GBD 2017 Population and Fertility Collaborators. Population and fertility by age and sex for 195 countries and territories, 1950–2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet* **392**, 1995–2051 (2018).
51. Mishra, V., Hong, R., Khan, S., Gu, Y. & Liu, L. *Evaluating HIV Estimates from National Population-Based Surveys for Bias Resulting from Non-Response. DHS Analytical Studies No. 12* http://dhsprogram.com/publications/publication-as12-analytical-studies.cfm (2008).
52. Curtis, S. L. & Sutherland, E. G. Measuring sexual behaviour in the era of HIV/AIDS: the experience of Demographic and Health Surveys and similar enquiries. *Sex. Transm. Infect.* **80**, ii22–ii27 (2004).
53. Burgert, C. R., Colston, J., Roy, T. & Zachary, B. *Geographic Displacement Procedure and Georeferenced Data Release Policy for the Demographic and Health Surveys*. https://dhsprogram.com/publications/publication-SAR7-Spatial-Analysis-Reports.cfm (Calverton, 2013).
54. Cuadros, D. F. et al. Capturing the spatial variability of HIV epidemics in South Africa and Tanzania using routine healthcare facility data. *Int. J. Health Geogr.* **17**, 27 (2018).



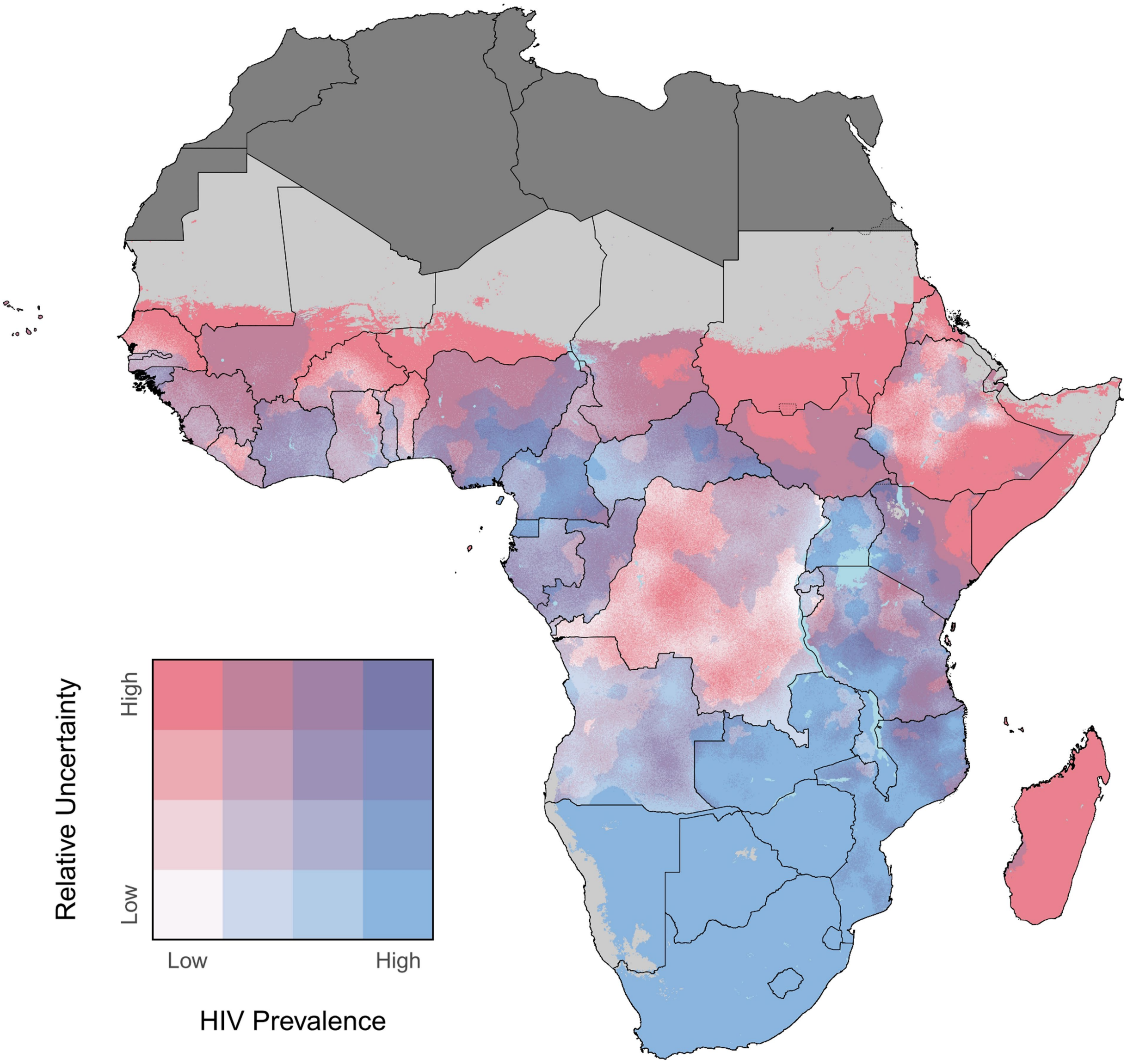
Extended Data Fig. 1 | HIV prevalence data by region and country. **a**, **b**, HIV seroprevalence survey data (**a**) and ANC sentinel surveillance data (**b**) used in this analysis, by region and country. Colour indicates the data source. AIS, AIDS Indicator Survey; DHS, Demographic and Health Survey; MICS, Multiple Indicator Cluster Survey; PHIA, Population-based HIV Impact Assessment Survey. Shape type indicates whether a data source has point (GPS) or polygon location information. Size indicates the relative effective sample size for each source. A full list of data sources with additional details about data type (such as survey microdata and survey reports) and geographical details are provided in Supplementary Tables 2, 4.



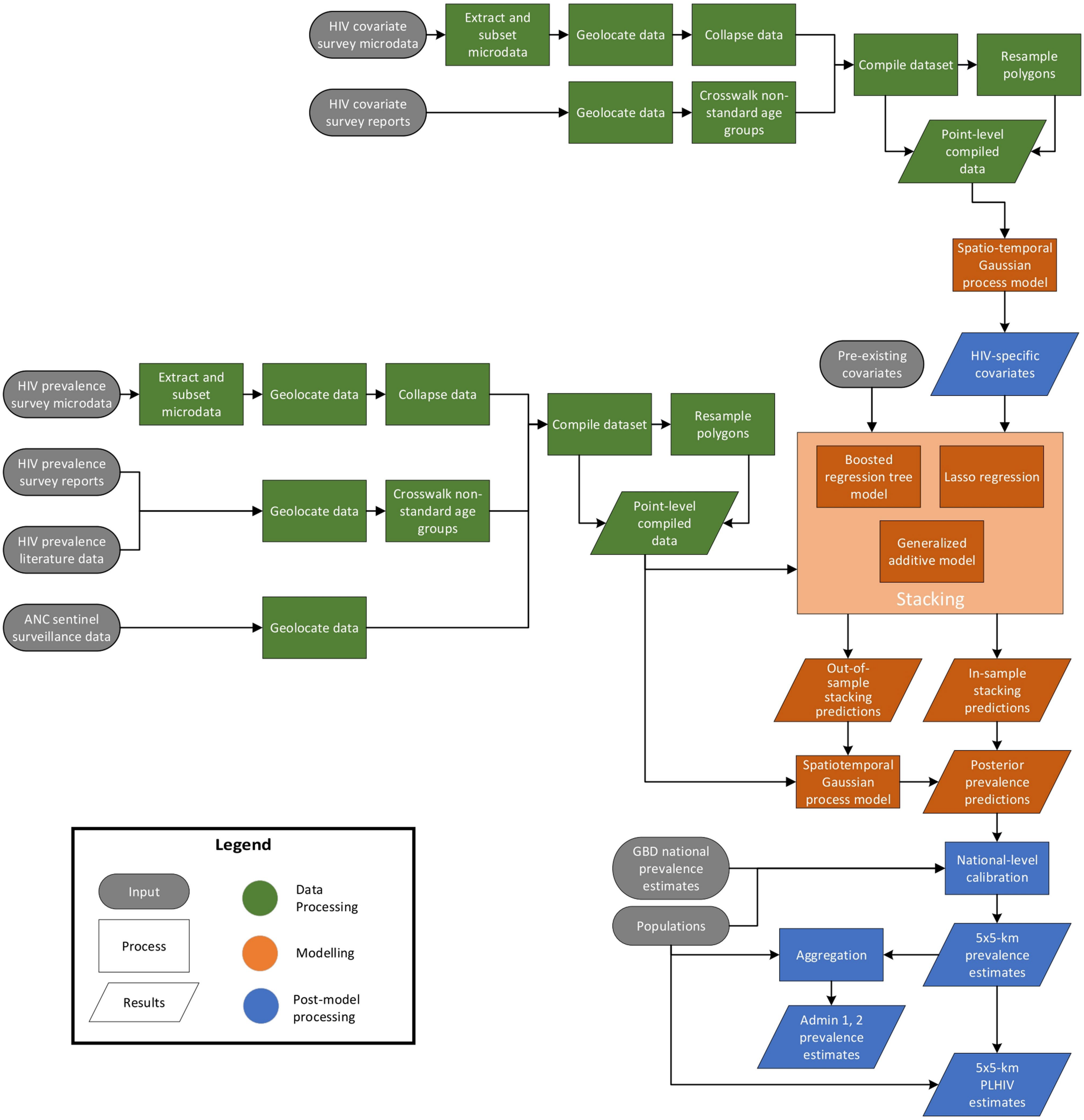
Extended Data Fig. 2 | HIV seroprevalence survey data coverage by year. A data point is defined as a cluster or polygon used in the analysis for the given year. There are a total of 29,103 data points for the HIV seroprevalence surveys from 2000 to 2017. Countries in white have no available survey data in the given year. Countries in dark grey were not included in the analysis.



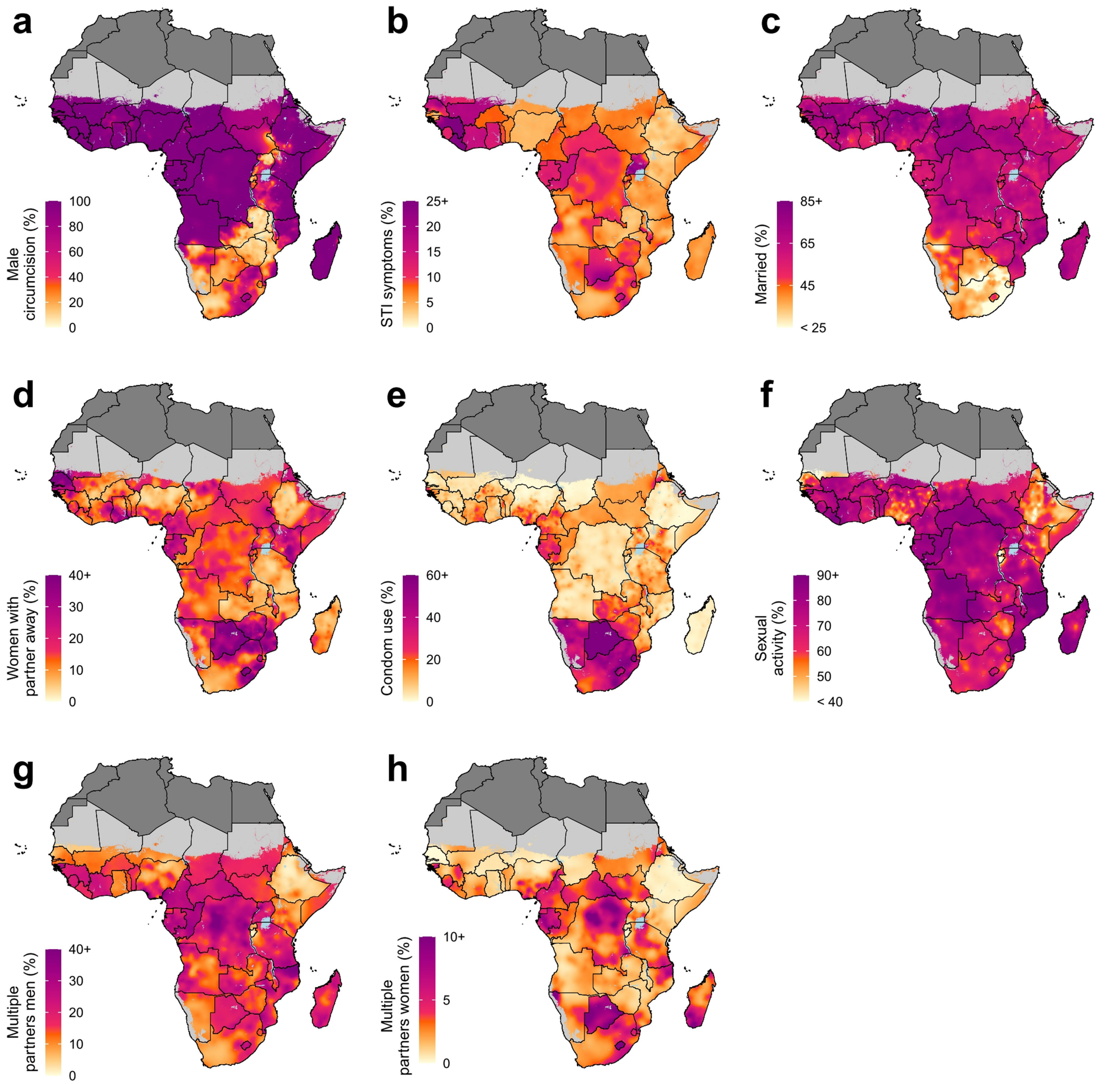
Extended Data Fig. 3 | ANC sentinel surveillance data coverage by year. A data point is defined as an ANC sentinel surveillance site used in the analysis for the given year. A site may be a hospital, city or town, or administrative region. There are a total of 9,794 ANC data points from 2000 to 2017. Countries in white have no available ANC data in the given year. Countries in dark grey were not included in the analysis.



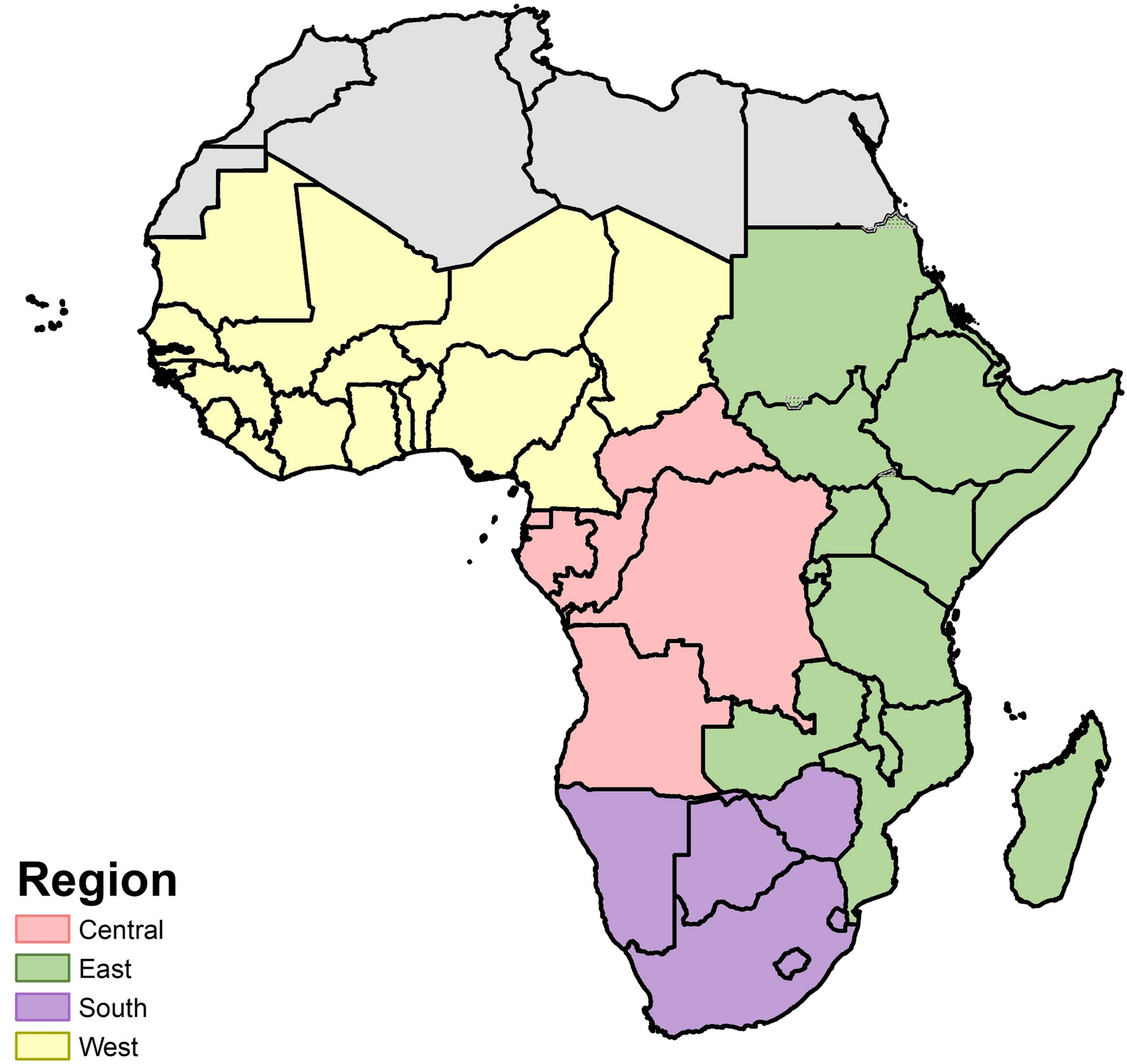
Extended Data Fig. 4 | Relative uncertainty in HIV prevalence in adults aged 15–49 in 2017. Overlapping population-weighted quartiles of HIV prevalence and relative 95% uncertainty in 2017 at the 5 × 5-km grid cell level. Relative uncertainty is defined as the ratio of the width of the 95% uncertainty interval to the mean estimate. Maps reflect administrative boundaries, land cover, lakes and population; grid cells with fewer than 10 people per 1 × 1 km, and classified as barren or sparsely vegetated, are coloured light grey[25,26,37–40]. Countries in dark grey were not included in the analysis.


HIV covariate survey microdata
Extract and subset microdata
Geolocate data
Collapse data
HIV covariate survey reports
Geolocate data
Crosswalk non-standard age groups
Compile dataset
Resample polygons
Point-level compiled data
Spatio-temporal Gaussian process model
Pre-existing covariates
HIV-specific covariates
HIV prevalence survey microdata
Extract and subset microdata
Geolocate data
Collapse data
HIV prevalence survey reports
HIV prevalence literature data
Geolocate data
Crosswalk non-standard age groups
ANC sentinel surveillance data
Geolocate data
Compile dataset
Resample polygons
Point-level compiled data
Boosted regression tree model
Lasso regression
Generalized additive model
Stacking
Out-of-sample stacking predictions
In-sample stacking predictions
Spatiotemporal Gaussian process model
Posterior prevalence predictions
GBD national prevalence estimates
Populations
National-level calibration
5x5-km prevalence estimates
Aggregation
Admin 1, 2 prevalence estimates
5x5-km PLHIV estimates
Legend
Input
Process
Results
Data Processing
Modelling
Post-model processing

Extended Data Fig. 5 | Analytic process overview. The process used to produce HIV prevalence estimates among adults in sub-Saharan Africa involved three main parts. In the data-processing steps (green), data were identified, extracted and prepared for use in the HIV prevalence model and in covariate models. In the modelling phase (orange), we used these data and covariates in a stacked generalization ensemble model and spatiotemporal Gaussian process model. In the post-processing phase (blue), we calibrated the prevalence estimation to match GBD 2017 estimates at the national level, aggregated prevalence estimates to the first- and second-level administrative subdivisions in each country and calculated the number of people living with HIV.



Extended Data Fig. 6 | Prevalence of covariates at 5 × 5-km grid cell level in 2017. a–h, Maps of HIV-specific covariates in 2017 include prevalence of male circumcision (**a**), prevalence of signs and symptoms of sexually transmitted infections (**b**), prevalence of marriage or living as married (**c**), prevalence of partner living elsewhere among women (**d**), prevalence of condom use during the most recent sexual encounter (**e**), prevalence of sexual activity among young women (**f**), prevalence of multiple partners among men in the past year (**g**) and prevalence of multiple partners among women in the past year (**h**). Maps reflect administrative boundaries, land cover, lakes and population; grid cells with fewer than 10 people per 1 × 1 km, and classified as barren or sparsely vegetated, are coloured light grey[25,26,37–40]. Countries in dark grey were not included in the analysis.



Extended Data Fig. 7 | Modelling regions. Modelling regions were defined as the four GBD regions in sub-Saharan Africa: central, east, south and west. Sudan was included in the east sub-Saharan Africa region for this analysis (in GBD, it is included in the North Africa and Middle East region). Countries in grey were not included in the analysis.

nature research

Corresponding author(s): Simon I. Hay

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	No primary data collection was carried out for this analysis.
Data analysis	This analysis was carried out using R version 3.5.0. The main geostatistical models were fit using R-INLA version 18.07.12. All code used for these analyses is publicly available online at https://github.com/ihmeuw/lbd/tree/hiv-africa-2019.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The findings of this study are supported by data that are available in public online repositories, data that are publicly available upon request from the data provider, and data that are not publicly available due to restrictions by the data provider and which were used under license for the current study. A detailed table of data sources can be found in Supplementary Tables 2, 4-6. More information about each data source is available on the Global Health Data Exchange (http://ghdx.healthdata.org/), including information about the data provider and links to where the data can be accessed or requested (where available).

Administrative boundaries were retrieved from the Global Administrative Unit Layers (GAUL) dataset, implemented by FAO within the CountrySTAT and Agricultural Market Information System (AMIS) projects [37]. Land cover was retrieved from the online Data Pool, courtesy of the NASA EOSDIS Land Processes Distributed Active Archive Center (LP DAAC), USGS/Earth Resources Observation and Science (EROS) Center, Sioux Falls, South Dakota [38]. Lakes were retrieved from the Global Lakes and Wetlands Database (GLWD), courtesy of the World Wildlife Fund and the Center for Environmental Systems Research, University of Kassel [39,40]. Populations were retrieved from WorldPop [25,26].

All estimates produced as part of this analysis are publicly available from the Global Health Data Exchange (http://ghdx.healthdata.org/ihme-data/africa-hiv-prevalence-geospatial-estimates-2000-2017) and via a user-friendly data visualization tool (https://vizhub.healthdata.org/lbd/hiv).

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

Sample size was calculated as the number of unique data source-location pairs with observations of HIV prevalence. This sample size is reported in the methods section: "We compiled a dataset of 29,103 data points from 134 seroprevalence surveys in 41 countries and 9,794 data points from sentinel surveillance of antenatal care clinics (ANC data) in 46 countries." This is an observational study with no hypothesis testing and the sample size was not pre-specified. We evaluate the overall performance of our modelling strategy, given the available data, as part of a validation exercise reported in the Supplementary Information (Section 4.3).

Data exclusions

Surveys that did not contain all relevant variables (HIV blood test results or at least one of the covariates) or that did not contain subnational geographic detail or could otherwise not be geolocated were excluded as not relevant for this analysis. Surveys that did not sample from the general population or did not sample both males and females (with the exception of surveys contributing to covariates that are sex-specific) were excluded as they were not representative of the population of interest. Surveys that did not contain information about the sample size or confidence intervals associated with a prevalence estimate were excluded as information about sample size (which can be derived from confidence intervals) was required by our modeling strategy. These exclusions were all pre-specified. In addition, a number of surveys were identified as poor quality or inconsistent with other data sources and were subsequently excluded (a list of these surveys and justification for their exclusion is reported in Supplementary Table 7). Antenatal care clinic sentinel surveillance data that could not be geolocated or that overlapped with an alternate source that were found to be more consistent were excluded; this exclusion criteria was pre-specified. In addition, a small number of site-years were identified as outliers and subsequently excluded (as described in Supplementary Information Section 2.4.2).

Replication

This is an observational study using many years of survey and surveillance data and in principle could be replicated. Due to the time required to extract, process, and geo-located all data, as well as to run the statistical models, we have not undertaken an explicit replication analysis.

Randomization

Randomization was not relevant to this study. This analysis is an observational mapping study and there were no experimental groups.

Blinding

Blinding was not relevant to this study, as it was an observational study using survey and surveillance data.

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description

Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study).

Research sample

State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For studies involving existing datasets, please describe the dataset and source.

Sampling strategy

Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed.

Data collection

Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper, computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether *the researcher was blind to experimental condition and/or the study hypothesis during data collection.*

Timing

Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort.

Data exclusions

If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.

Non-participation

State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.

Randomization

If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled.

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description

Briefly describe the study. For quantitative data include treatment factors and interactions, design structure (e.g. factorial, nested, hierarchical), nature and number of experimental units and replicates.

Research sample

Describe the research sample (e.g. a group of tagged Passer domesticus, all Stenocereus thurberi within Organ Pipe Cactus National Monument), and provide a rationale for the sample choice. When relevant, describe the organism taxa, source, sex, age range and any manipulations. State what population the sample is meant to represent when applicable. For studies involving existing datasets, describe the data and its source.

Sampling strategy

Note the sampling procedure. Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.

Data collection

Describe the data collection procedure, including who recorded the data and how.

Timing and spatial scale

Indicate the start and stop dates of data collection, noting the frequency and periodicity of sampling and providing a rationale for these choices. If there is a gap between collection periods, state the dates for each sample cohort. Specify the spatial scale from which the data are taken

Data exclusions

If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.

Reproducibility

Describe the measures taken to verify the reproducibility of experimental findings. For each experiment, note whether any attempts to repeat the experiment failed OR state that all attempts to repeat the experiment were successful.

Randomization

Describe how samples/organisms/participants were allocated into groups. If allocation was not random, describe how covariates were controlled. If this is not relevant to your study, explain why.

Blinding

Describe the extent of blinding used during data acquisition and analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.

Did the study involve field work? □ Yes □ No

Field work, collection and transport

Field conditions

Describe the study conditions for field work, providing relevant parameters (e.g. temperature, rainfall).

Location

State the location of the sampling or experiment, providing relevant parameters (e.g. latitude and longitude, elevation, water depth).

Access and import/export

Describe the efforts you have made to access habitats and to collect and import/export your samples in a responsible manner and in compliance with local, national and international laws, noting any permits that were obtained (give the name of the issuing authority, the date of issue, and any identifying information).

Disturbance

Describe any disturbance caused by the study and how it was minimized.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Unique biological materials

Policy information about availability of materials

Obtaining unique materials

Describe any restrictions on the availability of unique materials OR confirm that all unique materials used are readily available from the authors or from standard commercial sources (and specify these sources).

Antibodies

Antibodies used

Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.

Validation

Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)

State the source of each cell line used.

Authentication

Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.

Mycoplasma contamination

Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.

Commonly misidentified lines
(See ICLAC register)

Name any commonly misidentified cell lines used in the study and provide a rationale for their use.

Palaeontology

Specimen provenance

Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information).

Specimen deposition

Indicate where the specimens have been deposited to permit free access by other researchers.

Dating methods

If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals

For laboratory animals, report species, strain, sex and age OR state that the study did not involve laboratory animals.

Wild animals

Provide details on animals observed in or captured in the field; report species, sex and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.

Field-collected samples

For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.

Human research participants

Policy information about studies involving human research participants

Population characteristics

Describe the covariate-relevant population characteristics of the human research participants (e.g. age, gender, genotypic information, past and current diagnosis and treatment categories). If you filled out the behavioural & social sciences study design questions and have nothing to add here, write "See above."

Recruitment

Describe how participants were recruited. Outline any potential self-selection bias or other biases that may be present and how these are likely to impact results.

ChIP-seq

Data deposition

□ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

□ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links
May remain private before publication.

For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.

Files in database submission

Provide a list of all files available in the database submission.

Genome browser session
(e.g. UCSC)

Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.

Methodology

Replicates

Describe the experimental replicates, specifying number, type and replicate agreement.

Sequencing depth

Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.

Antibodies

Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.

Peak calling parameters

Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.

Data quality

Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.

Software

Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.

Flow Cytometry

Plots

Confirm that:

□ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

□ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

□ All plots are contour plots with outliers or pseudocolor plots.

□ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.

Instrument

Identify the instrument used for data collection, specifying make and model number.

Software

Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.

Cell population abundance

Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.

Gating strategy	*Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.*

□ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type	*Indicate task or resting state; event-related or block design.*
Design specifications	*Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.*
Behavioral performance measures	*State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).*

Acquisition

Imaging type(s)	*Specify: functional, structural, diffusion, perfusion.*
Field strength	*Specify in Tesla*
Sequence & imaging parameters	*Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.*
Area of acquisition	*State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.*

Diffusion MRI □ Used □ Not used

Preprocessing

Preprocessing software	*Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).*
Normalization	*If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.*
Normalization template	*Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.*
Noise and artifact removal	*Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).*
Volume censoring	*Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.*

Statistical modeling & inference

Model type and settings	*Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).*
Effect(s) tested	*Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.*

Specify type of analysis: □ Whole brain □ ROI-based □ Both

Statistic type for inference (See Eklund et al. 2016)	*Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.*
Correction	*Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).*

Models & analysis

n/a	Involved in the study
□	□ Functional and/or effective connectivity
□	□ Graph analysis
□	□ Multivariate modeling or predictive analysis

Functional and/or effective connectivity	*Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).*
Graph analysis	*Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).*
Multivariate modeling and predictive analysis	*Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.*