Current HIV/AIDS Reports (2024) 21:208–219
https://doi.org/10.1007/s11904-024-00702-3

REVIEW

Challenges and Opportunities in Big Data Science to Address Health Inequities and Focus the HIV Response

Katherine Rucinski[1] · Jesse Knight[2,3] · Kalai Willis[4] · Linwei Wang[2] · Amrita Rao[4] · Mary Anne Roach[4] · Refilwe Phaswana-Mafuya[5,6] · Le Bao[7] · Safiatou Thiam[8] · Peter Arimi[9] · Sharmistha Mishra[2,3,10,11,12] · Stefan Baral[4]

Accepted: 31 May 2024 / Published online: 25 June 2024

Abstract

Purpose of Review Big Data Science can be used to pragmatically guide the allocation of resources within the context of national HIV programs and inform priorities for intervention. In this review, we discuss the importance of grounding Big Data Science in the principles of equity and social justice to optimize the efficiency and effectiveness of the global HIV response.

Recent Findings Social, ethical, and legal considerations of Big Data Science have been identified in the context of HIV research. However, efforts to mitigate these challenges have been limited. Consequences include disciplinary silos within the field of HIV, a lack of meaningful engagement and ownership with and by communities, and potential misinterpretation or misappropriation of analyses that could further exacerbate health inequities.

Summary Big Data Science can support the HIV response by helping to identify gaps in previously undiscovered or understudied pathways to HIV acquisition and onward transmission, including the consequences for health outcomes and associated comorbidities. However, in the absence of a guiding framework for equity, alongside meaningful collaboration with communities through balanced partnerships, a reliance on big data could continue to reinforce inequities within and across marginalized populations.

Keywords Big Data Science · HIV transmission dynamics · Health equity · Community HIV response · Key populations · Predictive modeling · Explanatory modeling

Denotes shared senior authorship between Sharmistha Mishra and Stefan Baral.

✉ Katherine Rucinski
rucinski@jhu.edu

[1] Department of International Health, Johns Hopkins School of Public Health, Baltimore, MD, USA

[2] MAP Centre for Urban Health Solutions, Unity Health Toronto, Toronto, ON, Canada

[3] Institute of Medical Science, University of Toronto, Toronto, ON, Canada

[4] Department of Epidemiology, Johns Hopkins School of Public Health, Baltimore, MD, USA

[5] South African Medical Research Council/University of Johannesburg Pan African Centre for Epidemics Research (PACER) Extramural Unit, Johannesburg, South Africa

[6] Department of Environmental Health, Faculty of Health Sciences, University of Johannesburg, Johannesburg, South Africa

[7] Department of Statistics, Pennsylvania State University, University Park, PA, USA

[8] Conseil National de Lutte Contre Le Sida, Dakar, Senegal

[9] Partners for Health and Development in Africa, Nairobi, Kenya

[10] Division of Infectious Diseases, Department of Medicine, University of Toronto, Toronto, ON, Canada

[11] Institute of Health Policy, Management and Evaluation & Division of Epidemiology, Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada

[12] ICES, Toronto, ON, Canada

Introduction

Increasing the Specificity of the HIV Response Means Explicitly Identifying and Addressing Health Inequities

We are at a pivotal moment in the global HIV response. Reductions in HIV incidence are at risk of losing momentum given a failure to address systematic gaps in prevention and treatment. HIV incidence is at a thirty-year low, yet 1.5 million people still acquire HIV each year, including about one million in countries across Sub-Saharan Africa [1]. Investments in HIV programming, along with advances in treatment and prevention, including antiretroviral therapy (ART) and pre-exposure prophylaxis (PrEP), have the potential to reduce transmission [2–5], but marked differences in access are limiting global progress and amplifying health inequities [6•]. Widespread ART rollout through national HIV programs has improved the availability of treatment for people living with HIV. However, only 68% of the estimated 38.4 million people living with HIV globally have been able to achieve viral suppression [1]. Further, despite large, multi-sector investments in PrEP scale-up over the last five years [7, 8], uptake and persistence have been lower than expected globally, and especially among individuals at high risk of HIV [9, 10].

Underlying gaps in treatment and prevention is the acknowledgement that HIV risks are not evenly distributed anywhere in the world, with heterogeneity in HIV acquisition and onward transmission well-established through both empirical studies and mathematical models [11•, 12]. This heterogeneity often reflects complex sexual and/or shared injecting networks that intersect with social, political, and economic marginalization [6•, 13•]. While heterogeneity in HIV acquisition has largely been used to guide the population-level interventions and national HIV programs that define the HIV response, heterogeneity in onward transmission has not been as consistently considered. This is particularly true in settings with a high prevalence of HIV, where the epidemic is largely considered "generalized", such as across countries in eastern and southern Africa [6•, 13•]. When there is also heterogeneity in onward transmission risks, health inequities are amplified if those who are most at risk of onward transmission are also least likely to be reached by programs and services [14–17•]. Understanding the intersections of heterogeneity in risk and heterogeneity in intervention coverage is fundamental to addressing health inequities within epidemics.

The current mismatch between the realities of the HIV pandemic and the existing response could continue to reinforce inequities within and across marginalized communities, ultimately undermining the progress made on the path to ends AIDS. In the context of the HIV epidemic, inequalities refer to heterogeneity in measures of: HIV acquisition, ART coverage, viral suppression, and PrEP uptake, along with other measures of prevention coverage. When these inequalities stem from preventable and modifiable mechanisms, they reflect larger systemic health inequities [18]. One such preventable mechanism relates to intervention access, including barriers within existing health-systems such as stigma and discrimination, institutionalized racism, homophobia, transphobia, ageism, and sexism. Communities most affected by these barriers are those that have been disproportionately impacted by HIV, including the LGBTQ community, sex workers, and people who inject drugs, among others. Thus, an equity-informed strategy for a given intervention, such as HIV treatment to achieve sustained viral suppression, involves addressing access barriers within health-systems and "meeting people where they are" with differentiated service delivery [19, 20].

Health equity is a priority for global public health [18, 21]. There are several frameworks for examining health equity in the context of public health. Chief among them is the health equity framework proposed by Peterson et al. which centers health outcomes at a population-level and across four spheres of influence. These spheres comprise relationships and networks, systems of power that determine access to resources, physiological pathways, and individual factors that shape one's response to their environment [22•]. The first two spheres, networks and systems of power, are particularly important in the context of infectious diseases. First, preventable mechanisms such as criminalization of sex work and barriers to access, can shape the networks within which transmission occurs. Second, preventable mechanisms can also act through systems of power, including decisions around resource allocation and power imbalances in the production policies that are informed by epidemiologic evidence. In the context of the HIV response, systems of power are particularly salient given four decades of community-engagement alongside community-based and community-led HIV research [23]. In addition to the health equity framework [22•], the modified socio-ecological model of HIV prevention and related multi-dimensional conceptual frameworks have also provided a foundation for examining and addressing heterogeneity in HIV risk [24•, 25]. These latter frameworks are similar to the health equity framework in centering population-level outcomes and networks, but with a greater focus on directed causal pathways that lead to HIV acquisition and to onward transmission. The unifying feature across these established conceptual frameworks is that measures of heterogeneity in onward HIV transmission risks represent downstream consequences of structural determinants manifesting as health inequities.

The application of Big Data Science in the field of HIV has expanded significantly over the last decade alongside machine learning algorithms [26], but often with limited attention to how these approaches may amplify or mitigate existing health inequities if results are used to guide the application of interventions and resources. At its core, Big Data Science comprises high-dimensional data, characterized by the "volume, variety, and velocity" of big data [27], along with the application of advanced statistical techniques and modeling approaches that capitalize on computational capacity for data storage and analytic speed [28]. Several frameworks have been proposed to manage challenges related to the social, ethical, and legal considerations of big data [29–31]. However, efforts to integrate disparate data sources and methodologies that are responsive to HIV epidemic heterogeneities have been less common [32]. The consequences include disciplinary silos within the field of HIV, a lack of meaningful engagement and ownership with and by communities, and potential misinterpretation or misappropriation of analyses that could further exacerbate health inequities.

Advancing the HIV response through Big Data Science therefore means systematically identifying the pathways that lead to health inequalities in HIV acquisition and transmission and aligning interventions to maximize health equity. In this paper, we draw on existing health equity frameworks and a modified socio-ecological model of HIV prevention to outline the potential risks of applying conventional Big Data Science approaches with respect to health equity. We then discuss potential opportunities for equity-informed Big Data Science to work towards a HIV response that is effectively aligned with individual, network, and structural risks that continue to drive HIV acquisition and transmission.

The Promise of Big Data Science

An effective HIV response requires continual reexamination of our knowledge about the pathways and prevention gaps that shape risks of onward transmission in the short- and long-term. Central to this reexamination are three collaborative principles. First, there is a need to meaningfully ground analyses in the lived experience and expertise of communities, particularly communities who face disproportionate risks of HIV acquisition and onward transmission and sustained barriers to healthcare access. Community leadership has long been enshrined as central to the HIV response, first codified with the Greater Involvement of People Living with HIV published in 1994 [33]. However, as prevention gaps become more and more "concentrated" among the most marginalized communities [34], better integrating the knowledge of these communities into the content and implementation of programs is increasingly central. Second, reexamination requires meaningful engagement with front-line programs and service providers, who are often also on the front-line of data collection. Third, there is a need for collaborations that transcend the disciplinary siloes that have historically guided decisions about funding and resource allocation [35]. This means harnessing the expertise of epidemiologists and social scientists alongside statisticians, computer scientists, and infectious disease modelers, among others.

The promise of Big Data Science in guiding the HIV response lies in its potential to help identify gaps in previously undiscovered or understudied pathways to HIV acquisition and onward transmission, including the consequences for health outcomes and associated comorbidities. Integral to Big Data Science for HIV is the concept that large and diverse data sources [28, 36, 37], can be leveraged to better understand HIV epidemics and the pandemic as a whole. These data sources consist of routinely collected data such as electronic health records and program/clinic registers [38, 39], surveillance data, and auxiliary data sources such as social media and digital data along with traditional research data collected through trials or observational studies. Using a Big Data Science approach, these data sources can support analyses that are critical to the HIV response, including the identification of key groups of individuals and priority geographic areas at high risk of HIV, setting program targets, the evaluations of progress towards goals, and the development of strategies for optimizing the impact of HIV prevention and treatment programs.

How Conventional Approaches in Big Data Science Could Amplify Health Inequities

There are three primary areas where conventional Big Data Science approaches may potentially undermine health equity including privacy, data biases, and opportunity costs [21].

First, privacy concerns need to be contextualized within legal and policy-environments that criminalize occupations, identities, and dependency including sex work, sexual and gender diverse communities, and people who use drugs [40•]. Privacy concerns also reflect power dynamics wherein data reporting, ownership, and governance moves away from communities and to governments, academics, and international policy-makers and donors. Many emerging data sources, such as routinely collected program data and social media data, are indexed at the individual-level and contain identifiable information. These data can include information about HIV status, as well as stigmatized and/or criminalized behaviors, such as buying and selling sex, same-sex practices, and substance use [41, 42]. Yet, repurposing these data for new analyses, including linkages to other datasets, may not be covered by appropriate oversight, informed consent, or principles of equitable data ownership [43]. Moreover, these data and other data sources not originally collected for the purposes of research may be subject to a lower threshold of regulatory oversight [44]. Thus, efforts to leverage these data may expose already vulnerable individuals to new privacy risks and could thereby erode trust and engagement in care [45]. Ensuring that Big Data Science advances equity and community ownership necessitates careful attention to real and perceived risks of privacy breaches. In the absence of this, Big Data Science could risk perpetuating existing power structures that undermine program effectiveness and potentiates health inequities.

Second, Big Data Science is subject to systematic data biases given its use and integration of existing data that are routinely collected through a range of data platforms. Some data collection is purposeful, with randomized trials or prospective surveys designed to capture detailed data and thereby support specific causal inference or predictive utility. However, most data within Big Data Science comprise observational and repurposed data which may be limited in their design, and which inherently introduce the risk for systemic biases such as selection bias, information bias, and analytic biases such as collider bias among others [46].

For example, household-based surveys may fail to reach key populations due to mobility, precarious housing, or congregate living arrangements such as barracks and brothels [47]. The resulting selection bias from these surveys has been shown to generate downwardly biased estimates of HIV prevalence due to missing data [48]. Similarly, people living with HIV who remain unaware of their HIV status or who have not yet initiated HIV treatment are inherently missing from programmatic records, as are individuals who have engaged in HIV services previously but who have subsequently become “lost to clinic” or “lost to care” [49]. Without tracing studies, confirmatory methods, or other potential adjustment to account for these missing observations, these data may perpetuate more optimistic estimates of ART use and viral suppression [50]. Individuals who are marginalized are also under-represented in social media and social networking data because they are less likely to access mobile phones, the internet and other mobile-based technologies [51, 52]. Even if members of vulnerable populations are reached by surveys, they may not disclose stigmatized and/or criminalized behaviors such as anal sex due to social desirability biases [53–56]. If these selection and reporting biases are ignored, the prevalence of such behaviors from large datasets may be underestimated. Further, when these data sources, with their systematic biases entrenched, comprise a large number of observations that minimize the potential for random error, narrow confidence intervals may artificially reinforce the credibility of these estimates and any resulting inferences [57–60].

Another type of systematic bias can arise in how we use data in algorithms for prediction or analyses for inference. Many big data algorithms, such as random forests, neural networks, and others are designed to identify associations and generate predictions, but are not designed to infer causation [61]. Yet without grounding these predictions in established causal theory, such algorithms are liable to draw erroneous conclusions about drivers of transmission. Specifically, such algorithms may suffer from inappropriate adjustment resulting in collider bias or unmeasured confounding, as well as context-specific limits to generalizability. Together, the downstream effects from these analyses could result in priorities for intervention and/or implementation that are inefficient and misaligned with the needs of marginalized communities [12, 62–65]. Even well-chosen causal inference methods are challenged by common nuances in infectious disease epidemiology, such as spillover effects or “interference”, wherein an individual’s infection status is affected by other individuals’ exposures through network effects [66]. Despite the promise of new machine learning algorithms to generate new insights [67], thorough and rigorous validation procedures are needed to ensure these systems avoid reinforcing pre-existing biases that disproportionately affect marginalized communities [68].

Third, there are potential tradeoffs for the time and resources that must be spent to develop data pipelines, implement new algorithms, and validate emerging data sources. Primary data are often missing among those at the highest risk of HIV acquisition and for transmission in settings with the highest HIV burden [69–71]. This is largely because the same underlying mechanisms (e.g., individual, network, structural) that increase HIV acquisition and transmission risks among marginalized populations challenge the ability to reach them, characterize their HIV burden, and evaluate their unmet HIV prevention and treatment needs [57, 72–75]. Further, while the overall amount of available data is growing in HIV-related research [26], the development of approaches that successfully leverage these data to root out and address health inequities have not kept pace. Simply stated, in its current form, Big Data Science cannot yet overcome missing primary data. And the time and resources spent to explore these algorithms and data gaps may reflect an opportunity cost, when compared with reinvesting in purposeful data collection to fill known data gaps and continued development of more conventional causal inference methodology.

Using Big Data Science to Focus the HIV Response Requires Collaboration and Engagement with Communities and Programs

In the context of the HIV response, both experimental and observational research data have historically been used to answer specific research questions in controlled settings. These data are less available for key populations and other marginalized communities for whom a reliable sampling frame cannot be constructed [76–78].

Moving forward, supplementing research data with emerging data sources including from programs and community-led monitoring approaches may facilitate the collation of data that is more nuanced and reflective of underlying transmission dynamics. Such sources comprise local data inputs, which can provide critical information regarding real-world uptake and engagement with HIV services and characteristics of ongoing inequities that continue to define the HIV response [11•]. While the systematic collection and collation of surveillance data for key populations have increased over the last decade, the use of routinely collected program data and community-led monitoring has traditionally been less utilized by institutions such as UNAIDS and others tasked with making epidemic projections [20, 79, 80]. For example, program data are routinely collected service-delivery data from implementing partners and community organizations, and may include client files, registries, and reporting indicators to support service management/implementation and evaluate routine program activities. Program data are typically collected and stored individually but reported in aggregate as a metric for appraisal against national service-delivery targets or epidemic transition metrics. Unlike experimental data and observational research studies, program data reflect the realities of day-to-day service delivery, and may often be messy and incomplete [81]. This limitation may inhibit the ability to link programmatic records with unique individuals over time [82, 83]. Moreover, aggregate-level data reported through HIV programs may implicate specific interpretations of HIV metrics [41]. Such challenges may discount these data for researchers and public health practitioners in the absence of local knowledge and contextual information that can inform analytic strategies to mitigate such biases. To mitigate these risks, meaningfully collaboration with communities involved in community-led monitoring provides an opportunity to collect and interpret a range of data sources, ultimately better serving networks of individuals at significant risk of HIV acquisition and transmission.

How Big Data Science Approaches Can Help Advance Health Equity and Specificity in the HIV Response

Emerging and equity-focused Big Data Science methods have the potential to improve data impact and program efficiency. Such efficiencies are critical for guiding the next phase of the HIV response, particularly given declining available resources from donors and government bodies [84]. These approaches draw on emerging data sources and necessitate investments in data integration, cleaning, and potential bias adjustments to support a range of analytic approaches comprising descriptive, predictive, explanatory, and simulation-based analyses (Fig. 1).

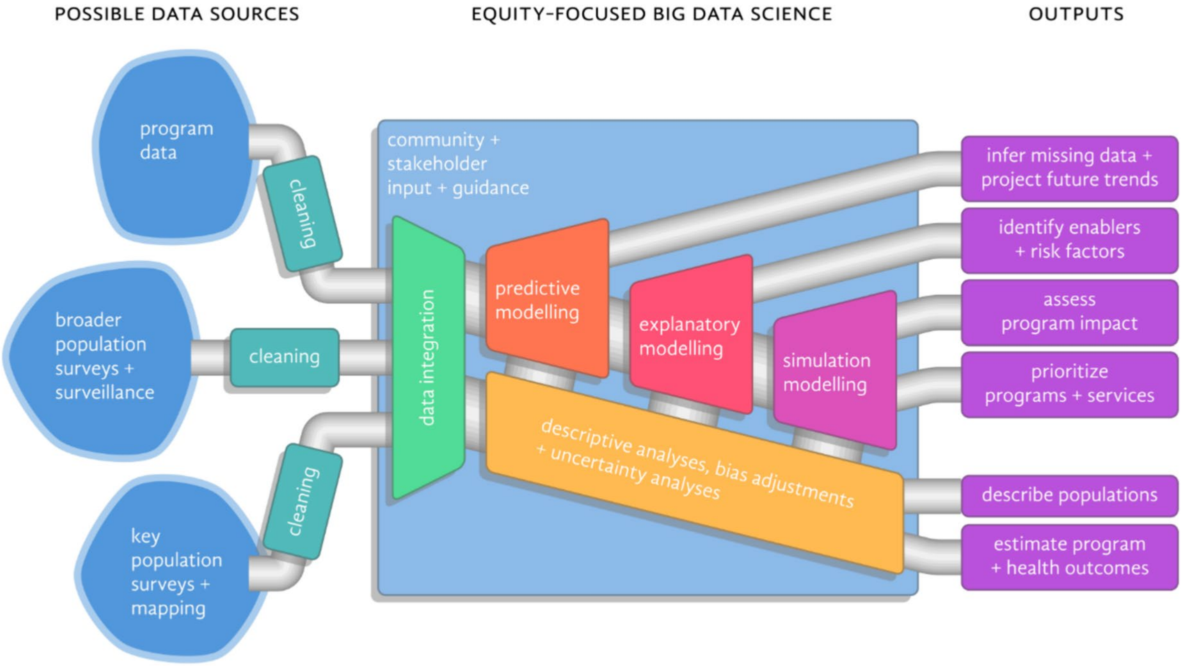
POSSIBLE DATA SOURCES
EQUITY-FOCUSED BIG DATA SCIENCE
OUTPUTS
program data
cleaning
broader population surveys + surveillance
cleaning
key population surveys + mapping
cleaning
community + stakeholder input + guidance
data integration
predictive modelling
explanatory modelling
simulation modelling
descriptive analyses, bias adjustments + uncertainty analyses
infer missing data + project future trends
identify enablers + risk factors
assess program impact
prioritize programs + services
describe populations
estimate program + health outcomes

Fig. 1 Equity-focused Big Data Science pipeline integrating emerging data sources

Data Integration and Linkage

Optimizing the contributions of emerging data sources in the spirit of informing an equitable HIV response requires attention to both data collation and data integration that will ultimately inform interventions, implementation, and other programmatic decisions. These processes draw on transdisciplinary expertise, whereby researchers, communities, other stakeholders work collaboratively to identify best-practices for combining disparate data sources [85]. Data integration includes three distinct and important elements: data cleaning, linkage, and validation. These integration steps ensure that data are being treated in a systematic and replicable way so that they can be used together and to minimize bias.

HIV-related data, including research data, data from health records, routine program registers, and surveillance systems, are often collected and stored using a wide range of approaches. Data collection may comprise electronic records and paper-based forms. As data are collected there may be real-time updates made to a central data warehouse, or updates to local files requiring manual data entry. Across both approaches, missing data, transcription errors, and incomplete files may result in "noisy" data that necessitate time-consuming cleaning processes [86]. Ultimately, each data file or system will have its own set of rules and challenges, and the process of identifying the rules, addressing the challenges, and harmonizing with an underlying repository is the task of data cleaning.

Once data are cleaned, the next major component to data integration is data linkage. Where unique identifiers are made available, linking records for the same individual over time or across data source types may be feasible [41, 69]. If unique identifiers are imperfectly recorded or unavailable, probabilistic linking algorithms can be utilized to support record linkages. Linking algorithms utilize probabilities of agreement and disagreement between a range of matching variables to link two or more files for the same individual [87]. Compared to a more deterministic approach, probabilistic linking is more forgiving of data entry errors and reporting inconsistencies that may be inherent to nontraditional data sources, particularly routinely-collected program data [88].

As a final step of data integration, data validation involves examining and evaluating the quality of data. The quality of implementation and reporting of research studies, surveillance, and routine program data collection can vary widely, challenging the comparison or integration of these disparate data sources. There are numerous tools and frameworks available to help assess the quality of existing evidence, including the evaluation of randomized controlled trials and cohort studies, measures of disease occurrence, and assessments of internal and external validity. More recent tools have been designed to specifically evaluate the quality of HIV epidemiologic evidence for populations in the absence of a reliable sampling frame, as is often the case for data collected with key populations and other marginalized communities [76]. However, these tools have not been widely implemented in the context of Big Data Science [89]. Data quality should also be assessed with input from front-line data collection teams and community members with lived experience, who can help identify and adjust for context-specific sources of bias and uncertainty.

Overall, integrating data sources allows an assessment of the differential outcomes among people living with HIV who are from historically marginalized communities, as well as provide insights into mitigating these risks. In Malawi, efforts to optimize routine data collected across community-led HIV programs have demonstrated the potential for using real-world data to identify strategies that promote linkage to care and ART uptake for key populations [20, 90]. Importantly, while much of the outreach for marginalized communities is led by community-based organizations, individuals living with HIV are usually referred to government-led treatment clinics upon diagnosis. Supporting effective data integration via intervention strategies that encompass these government-owned data sources would facilitate an understanding of differential outcomes in these clinics and determinants of those who the programs are failing to inform.

Description

Gaps in data for marginalized communities can be addressed across a continuum of analytic approaches, with descriptive epidemiology playing a critical role in defining characteristics of a target population across person, places, and time [91]. While causal inference methods draw from a potential outcomes framework to estimate the effect of a proposed intervention, descriptive epidemiology can be used to inform analytic priorities as well as targets for intervention [92]. In the context of Big Data Science, descriptive epidemiology can be useful for working with data sources to understand key elements of a given population, or to document observed or "factual" patterns of disease over time [93]. For example, describing the overall distribution of annualized HIV infections among and across differing identifies or occupations. Importantly, descriptive analyses are still subject to the myriad information and selection biases that plague causal contrasts. However, the manner in which these biases are explored or addressed in the context of Big Data Science has the potential to reinforce inequities in the absence of guidance from community members with context-specific knowledge. For example, overadjustment of key contextual variables such as geography or race can obfuscate important heterogeneity in disease patterns, thus undermining programmatic decisions and priorities for intervention.

Prediction

Prediction to inform policy and funding decisions is a central goal of Big Data Science. However, missing data on the most proximal determinants of HIV challenges the predictive accuracy and validity, and thus utility of these approaches. And ignoring proximal determinants in programs can undermine the potential impact of all other efforts at local epidemic control [94]. Thus, while predictive models have long been used to characterize HIV epidemics, including through the use of big data analytics, the underlying disparities in data availability may mean downstream model outputs further perpetuate inequities with consequences especially for marginalized communities [95].

Small area estimation approaches represent a tool in predictive modeling to advance equity in the HIV response. These approaches are less focused on the identification of counterfactuals, and instead are used to characterize epidemiologic patterns in a given place or time. Broadly, small area estimation is a set of statistical techniques used to estimate parameters for “small areas”. These techniques are generally applied when the “small area” of interest is part of a larger survey when empiric estimates may have unacceptably large standard errors or where there are no empiric data. This may be particularly valuable in the case of estimating population size estimates, or other HIV-related outcomes for key populations used to inform program planning and resource allocation for HIV programs. Here, a model-based approach uses a statistical model that “borrows strength” from other small areas or years in the sample survey or from auxiliary data at the small area level. For example, traditional extrapolation approaches to produce estimates where there are no empiric data for key populations relied on crude regression methods that assume homogeneity of all urban centers or rural areas—assumptions that are unlikely to hold true near mining industries, areas with high immigration, settlements along busy roadways, and fishing villages [96, 97]. Moreover, where data do exist, there is limited utilization to inform mathematical modeling and local programmatic or policy response [62, 72, 98, 99]. Efforts to mitigate epidemiologic gaps for marginalized communities have been successful in leveraging social media data to help tackle complex questions around population size in highly stigmatized settings, including as a way of informing direct programs for gay, bisexual, and other cisgender men who have sex with men [100]. Recent research in Namibia has also reinforced the programmatic importance of integrating consensus-building into size estimation methodologies [101], reflecting a larger trend of ensuring community are directly involved in the generation and application of size estimates to maximize relevance for policies and programs [102].

Explanatory Modeling

Data may be used to answer explanatory epidemiologic questions, including both descriptive and causal analyses. Whereas descriptive questions aim to characterize features of the target population, causal questions aim to isolate effects of key exposures and/or interventions on specified outcomes [91]. Answering explanatory epidemiologic questions has the potential to help us understand current and relevant implementation challenges, optimize interventions, and improve service delivery. The distinct advantage of using emerging data sources, including routinely collected data, rather than exclusively research data to answer explanatory questions is that even well-designed, community-based studies often do not adequately capture real-world conditions, including available resources, competing priorities of those implementing services or an intervention, and local context [103]. In South Africa, routinely collected data from a community-led HIV program has been successfully used to identify implementation strategies associated with PrEP uptake and persistence within a large cohort of women at high risk of HIV in South Africa [104, 105]. This ability to capture real-world conditions also lends itself to being able to answer additional implementation-related questions. These include assessing the effectiveness of distinct intervention components, fidelity assessments as a mechanism or mediatory of intervention success, as well as factors associated with intervention implementation. Routinely collected data from local program partners harnessed in conjunction with other related data can be used to conduct rigorous epidemiologic and implementation-related analyses to identify impactful interventions across programs for key populations.

Simulation Modeling

Compartmental models and agent-based models can be used to describe the heterogeneity in risk and also the differential impact of interventions not addressing this heterogeneity in risk [106, 107]. These models are defined by how the underlying processes (i.e. mechanisms) of transmission, disease progression (transitions between health-states), and/or an intervention’s causal effects are simulated. As such, these models capture the downstream impact of prevention among populations at highest risk of acquisition and transmission [108]. Common data inputs for these models include population size estimates, sexual behavior data, as well as biological parameters, and rates of access to interventions [62]. Data may be stratified so that they are reflective of heterogeneity in risks, including strata defined by age, sexual risk, and other population-level attributes. These models can then be used to estimate important outputs to guide HIV interventions and resource allocation, including stratified estimates of intervention coverage, HIV incidence and prevalence, and other related health outcomes [109, 110]. However, the model outputs are highly dependent on the parameters and model calibration to empirical estimates of model outputs, also known as "calibration targets". Efforts can be made to reflect and address information and selection biases in these models to minimize the risks of the perpetuation of inequities. For example, the duration and relative infectivity of acute phase HIV can be sampled from wide prior distributions [111]. Alternatively, the modeled population can be stratified into subgroups comprising those who have "never tested" or "ever tested" for HIV, with the idea of capturing a testing gap across more marginalized populations that would be invisible in an "average testing rate" [109]. Similarly, the proportion of highly mobile populations who are captured in household-based HIV prevalence data can be explicitly modeled [112].

Model calibration is an integral step in applied modeling that precedes counterfactual analyses. All too often the details of this laborious process are buried in the supplementary materials of peer-reviewed publications. Yet model calibration can offer stand-alone results, such as: the posterior (joint) distributions of model parameters, underlying transmission dynamics ("who infected whom"), and plausible short-term epidemic trajectories. Thus, a calibrated mathematical model reflects another framework for the synthesis of heterogeneous data, within a common set of modeling assumptions. Moving forward, including these complete details such as model design, parameterization, and calibration decisions, as well as exploring insights from the calibrated model, will be essential for ensuring transparency and accessibility in decision-making prior to analyses of counterfactual scenarios.

Conclusions

The risk and burden of HIV are not evenly distributed anywhere in the world, including in the most broadly generalized HIV epidemic settings. It has become increasingly clear that understanding the distribution of these risks is central to comprehensively addressing the needs of communities with sustained risks for HIV acquisition and transmission. Big Data Science can help identify and address these needs, but only if guided by an approach that leads with equity. Equity has long been an afterthought in the context of the HIV response, but leveraging an equity-lens is necessary for mitigating the potential harms of traditional big data approaches related to privacy, data biases, and opportunity costs, while simultaneously leveraging new methodologies that maximize the utility of current data and resources. As described above, these methodologies can integrate multiple data sources through data cleaning, linkage, and validation; generate new insights from emerging data sources; fill data gaps and adjust for biases through predictive modeling; synthesize all available data through transmission modeling; estimate the impacts of addressing unmet needs through explanatory modeling; and identify efficient interventions through economic analysis. Ultimately, these methodologies will help fill data blind spots and capture risk heterogeneities and intervention gaps which continue to shape the epidemic. Given declining resources for the HIV response globally, it is more important than ever to identify and comprehensively address the unmet needs of people at risk of and living with HIV.

Acknowledgements This study was funded, in part, by the National Institute of Allergy and Infectious Disease (R01AI170249; R01AI136664) and in part by the Canadian Institutes of Health Research (FN-13455). SM is supported by a Tier 2 Canada Research Chair in Mathematical Modeling and Program Science (CRC number 950-232643). KR also received funding from the National Institute of Mental Health (K01MH129226). The work reported herein was also made possible through funding by the South African Medical Research Council Project Code #57035 (SAMRC File ref no: HDID8528/KR/202) through its Division of Research Capacity Development under the Mid-Career Scientist Programme through funding received from the South African National Treasury. The content hereof is the sole responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the SAMRC. This publication also resulted (in part) from research supported by the Johns Hopkins University Center for AIDS Research, an NIH funded program (1P30AI094189), which is supported by the following NIH Co-Funding and Participating Institutes and Centers: NIAID, NCI, NICHD, NHLBI, NIDA, NIA, NIGMS, NIDDK, NIMHD.

Author Contributions K.R, S.M, and S.B conceptualized this work. K.R, J.K, K.W, L.W, A.R, M.A.R, S.M, and S.B drafted the original manuscript. J.K prepared Figure 1. All authors reviewed and edited the manuscript and revised it critically for important intellectual content. J.K was supported by the Ontario Ministry of Colleges and Universities (QEII-GSST)

Data Availability No datasets were generated or analysed during the current study.

Declarations

Informed Consent This article does not contain human subjects data, and thus informed consent was not obtained.

Competing Interests The authors declare no competing interests.

Research Involving Human Participants and/or Animals This article does not contain any studies with human or animal subjects performed by any of the authors.

References

Papers of particular interest, published recently, have been highlighted as: • Of importance

1. Joint united nations programme on HIV/AIDS; UNAIDS DATA 2023. 2023. Available at: https://www.unaids.org/en/resources/documents/2023/2023_unaids_data.
2. Swindells S, Andrade-Villanueva J-F, Richmond GJ, Rizzardini G, Baumgarten A, Masiá M, Latiff G, Pokrovsky V, Bredeek F, Smith G, et al. Long-acting cabotegravir and rilpivirine for maintenance of HIV-1 suppression. N Engl J Med. 2020;382:1112–23.
3. Phillips AN, Bansi-Matharu L, Cambiano V, Ehrenkranz P, Serenata C, Venter F, Pett S, Flexner C, Jahn A, Revill P, et al. The potential role of long-acting injectable cabotegravir–rilpivirine in the treatment of HIV in sub-Saharan Africa: a modelling analysis. Lancet Glob Health. 2021;9:e620–7.
4. Fonner VA, Ridgeway K, Van Der Straten A, Lorenzetti L, Dinh N, Rodolph M, Schaefer R, Schmidt H-MA, Nguyen VTT, Radebe M, et al. Safety and efficacy of long-acting injectable cabotegravir as preexposure prophylaxis to prevent HIV acquisition. AIDS. 2023;37:957–966.
5. Smith J, Bansi-Matharu L, Cambiano V, Dimitrov D, Bershteyn A, Van De Vijver D, Kripke K, Revill P, Boily M-C, Meyer-Rath G, et al. Predicted effects of the introduction of long-acting injectable cabotegravir pre-exposure prophylaxis in sub-Saharan Africa: a modelling study. The Lancet HIV. 2023;10:e254–65.
6. • Baral S, Rao A, Sullivan P, Phaswana-Mafuya N, Diouf D, Millett G, Musyoki H, Geng E, Mishra S. The disconnect between individual-level and population-level HIV prevention benefits of antiretroviral treatment. The Lancet HIV. 2019;6:e632-e638. **This study demonstrated the importance of focusing HIV prevention and treatment investments towards those at highest risk of onward HIV transmission, rather than a model solely focused on universal uptake of HIV prevention and treatment.**
7. Karim SSA, Baxter C. HIV pre-exposure prophylaxis implementation in Africa: some early lessons. Lancet Glob Health. 2021;9:e1634–5.
8. Irungu EM, Baeten JM. PrEP rollout in Africa: status and opportunity. Nat Med. 2020;26:655–64.
9. World Health Organization. Global data shows increasing PrEP use and widespread adoption of WHO PrEP recommendations. 2021. Available at: https://www.who.int/news-room/feature-stories/detail/global-data-shows-increasing-prep-use-and-widespread-adoption-of-who-prep-recommendations.
10. Joint united national programme on HIV/AIDS. Pre-exposure prophylaxis use expands, but not fast enough. 2022. Available at: https://www.unaids.org/en/resources/presscentre/featurestories/2022/january/20220117_preexposure_prophylaxis-use-expands.
11. • Mishra S, Silhol R, Knight J, Phaswana-Mafuya R, Diouf D, Wang L, Schwartz S, Boily M-C, Baral S. Estimating the epidemic consequences of HIV prevention gaps among key populations J Int AIDS Soc. 2021;24(Suppl 3):e25739. **This study outlines and discusses a conceptual framework for understanding and estimating the transmission population attributable fraction over time (tPAFt) via transmission modelling as a measure of onward transmission risk from HIV prevention gaps.**
12. Knight J, Kaul R, Mishra S. Risk heterogeneity in compartmental HIV transmission models of ART as prevention in Sub-Saharan Africa: a scoping review. Epidemics. 2022;40:100608.
13. • Makofane K, Van Der Elst EM, Walimbwa J, Nemande S, Baral SD. From general to specific: moving past the general population in the HIV response across sub-Saharan Africa. J Intern AIDS Soc. 2020, 23:e25605. **This study reflects on the usage of the general population construct in HIV, with a recommendation that the term be retired from the field's lexicon to promote efficiency and impact within the HIV response.**
14. Jin H, Restar A, Beyrer C. Overview of the epidemiological conditions of HIV among key populations in Africa. J Int AIDS Soc. 2021;24:e25716.
15. Lyons CE, Schwartz SR, Murray SM, Shannon K, Diouf D, Mothopeng T, Kouanda S, Simplice A, Kouame A, Mnisi Z, et al. The role of sex work laws and stigmas in increasing HIV risks among sex workers. Nat Commun. 2020;11:1–10.
16. Lyons CE, Twahirwa Rwema JO, Makofane K, Diouf D, Mfochive Njindam I, Ba I, Kouame A, Tamoufe U, Cham B, Aliu Djaló M, et al. Associations between punitive policies and legal barriers to consensual same-sex sexual acts and HIV among gay men and other men who have sex with men in sub-Saharan Africa: a multicountry, respondent-driven sampling survey. Lancet HIV. 2023;10:e186–94.
17. • Rucinski KB, Schwartz SR, Mishra S, Phaswana-Mafuya N, Diouf D, Mothopeng T, Kouanda S, Simplice A, Kouame A, Cham B, et al. High HIV Prevalence and Low HIV-Service Engagement Among Young Women Who Sell Sex: A Pooled Analysis Across 9 Sub-Saharan African Countries. JAIDS J Acquir Immune Defic Syndr. 2020;85:148-155. **This work presents evidence that addressing barriers to HIV service delivery among young women who sell sex is central to a comprehensive HIV response.**
18. World Health Organization. Health equity and its determinants. 2021. Available at: https://www.who.int/publications/m/item/health-equity-and-its-determinants.
19. Schwartz SR, Baral S. Remembering individual perspectives and needs in differentiated HIV care strategies. BMJ Qual Saf. 2019;28:257–9.
20. Rucinski K, Masankha Banda L, Olawore O, Akolo C, Zakaliya A, Chilongozi D, Schwartz S, Wilcher R, Persaud N, Ruberintwari M, et al. HIV testing approaches to optimize prevention and treatment for key and priority populations in Malawi. Open Forum Infect Dis. 2022;9:ofac038.
21. World Health Organization. Health equity. 2024. Available at: https://www.who.int/health-topics/health-equity#tab=tab_1.
22. • Peterson A, Charles V, Yeung D, Coyle K. The health equity framework: a science- and justice-based model for public health researchers and practitioners. Health Promot Pract. 2021;22:741–746. **The authors propose the Health Equity Framework, which comprises four, interacting spheres of influence that represent both categories of risk and protective factors for health outcomes as well as opportunities for strategies and interventions that address those factors.**
23. Simon-Meyer J, Odallo D. Greater involvement of people living with HIV/AIDS in South Africa. Eval Program Plann. 2002;25:471–9.
24. • Baral S, Logie CH, Grosso A, Wirtz AL, Beyrer C. Modified social ecological model: a tool to guide the assessment of the risks and risk contexts of HIV epidemics. BMC Public Health. 2013;13:482. **The authors propose a modified social**

ecological model (MSEM) to help visualize multi-level domains of HIV infection risks and guide the development of epidemiologic HIV studies.

25. Shannon K, Goldenberg SM, Deering KN, Strathdee SA. HIV infection among female sex workers in concentrated and high prevalence epidemics: why a structural determinants framework is needed. Curr Opin HIV AIDS. 2014;9:174–82.
26. Liang C, Qiao S, Olatosi B, Lyu T, Li X. Emergence and evolution of big data science in HIV research: bibliometric analysis of federally sponsored studies 2000–2019. Int J Med Inform. 2021;154:104558.
27. Kitchin R, McArdle G. What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. Big Data Soc. 2016;3:205395171663113.
28. Qiao S, Li X, Olatosi B, Young SD. Utilizing big data analytics and electronic health record data in HIV prevention, treatment, and care research: a literature review. AIDS Care. 2021. https://doi.org/10.1080/09540121.2021.1948499.
29. Crawford K, Schultz J. Big data and due process: toward a framework to redress predictive privacy harms. BC L Rev. 2014;55:93–128.
30. Devins C, Felin T, Kauffman S, Koppl R. The Law and Big Data. Cornell J L Public Policy. 2017;27:357.
31. Reed-Berendt R, Dove ES, Pareek M. UK-REACH Study collaborative group: the ethical implications of big data research in public health: “Big Data Ethics by Design” in the UK-REACH Study. Ethics Hum Res. 2022;44:2–17.
32. Olatosi B, Vermund SH, Li X. Power of Big Data in ending HIV. AIDS. 2021;35:S1–5.
33. Joint united nations programme on HIV/AIDS. The greater involvement of people living with HIV. 2007. Available at: https://www.unaids.org/sites/default/files/media_asset/jc1299-policybrief-gipa_en_0.pdf.
34. Garnett GP. Reductions in HIV incidence are likely to increase the importance of key population programmes for HIV control in sub-Saharan Africa. J Int AIDS Soc. 2021;24:e25727.
35. Choi BCK, Pak AWP: Multidisciplinarity, interdisciplinarity and transdisciplinarity in health research, services, education and policy: 1. Definitions, objectives, and evidence of effectiveness. Clin Invest Med. 2006, 29:351–364.
36. Young SD. A “big data” approach to HIV Epidemiology and Prevention. Prev Med. 2015;70:17–8.
37. van Heerden A, Young S. Use of social media big data as a novel HIV surveillance tool in South Africa. PLoS One. 2020;15:e0239304.
38. National AIDS and STI control programme: Key populations programme data collection tools: Reference Manual 2014. 2014. Available at: https://hivpreventioncoalition.unaids.org/sites/default/files/attachments/compressed_Kenya-KP-TOOLS-REFERENCE-GUIDE.compressed.pdf.
39. National AIDS and STI control programme: Key populations programme data collection tools: Revised Reference Manual - 2019. 2019. Available at: https://cquin.icap.columbia.edu/wp-content/uploads/2021/08/KP-Tools-Narrative_FINAL.pdf.
40. • Nkengasong J, Ratevosian J. Legal and policy barriers for an effective HIV/AIDS response. The Lancet 2023;401:1405–1407. **This commentary highlights the role of PEPFAR in addressing threats of structural barriers and punitive laws that result in stigma and discrimination and stand in the way of progress in the HIV/AIDS response.**
41. Rice B, Boulle A, Baral S, Egger M, Mee P, Fearon E, Reniers G, Todd J, Schwarcz S, Weir S, et al. Strengthening routine data systems to track the HIV epidemic and guide the response in Sub-Saharan Africa. JMIR Public Health Surveill. 2018;4:e36.
42. Weir SS, Baral SD, Edwards JK, Zadrozny S, Hargreaves J, Zhao J, Sabin K. Opportunities for enhanced strategic use of surveys, medical records, and program data for HIV surveillance of key populations: scoping review. JMIR Public Health Surveill. 2018;4:e28.
43. Xafis V, Schaefer GO, Labude MK, Brassington I, Ballantyne A, Lim HY, Lipworth W, Lysaght T, Stewart C, Sun S, et al. An ethics framework for big data in health and research. Asian Bioeth Rev. 2019;11:227–54.
44. Dokholyan RS, Muhlbaier LH, Falletta JM, Jacobs JP, Shahian D, Haan CK, Peterson ED. Regulatory and ethical considerations for linking clinical and administrative databases. Am Heart J. 2009;157:971–82.
45. Price WN, Cohen IG. Privacy in the age of medical big data. Nat Med. 2019;25:37–43.
46. Hammerton G, Munafò MR. Causal inference with observational data: the need for triangulation of evidence. Psychol Med. 2021;51:563–78.
47. Hakim AJ, MacDonald V, Hladik W, Zhao J, Burnett J, Sabin K, Prybylski D, Garcia Calleja JM. Gaps and opportunities: measuring the key population cascade through surveys and services to guide the HIV response. J Int AIDS Soc. 2018;21(Suppl 5):e25119.
48. Palma AM, Marra G, Bray R, Saito S, Awor AC, Jalloh MF, Kailembo A, Kirungi W, Mgomella GS, Njau P, et al. Correcting for selection bias in HIV prevalence estimates: an application of sample selection models using data from population-based HIV surveys in seven sub-Saharan African countries. J Int AIDS Soc. 2022;25:e25954.
49. Edwards JK, Lesko CR, Herce ME, Murenzi G, Twizere C, Lelo P, Anastos K, Tymejczyk O, Yotebieng M, Nash D, et al. Gone but not lost: implications for estimating HIV care outcomes when loss to clinic is not loss to care. Epidemiology. 2020;31:570–7.
50. Mirzazadeh A, Eshun-Wilson I, Thompson RR, Bonyani A, Kahn JG, Baral SD, Schwartz S, Rutherford G, Geng EH. Interventions to reengage people living with HIV who are lost to follow-up from HIV treatment programs: a systematic review and meta-analysis. PLoS Med. 2022;19:e1003940.
51. Flores L, Young SD. Ethical perspectives in using technology-enabled research for key HIV populations in rights-constrained settings. Curr HIV/AIDS Rep. 2023;20:148–59.
52. You WX, Comins CA, Jarrett BA, Young K, Guddera V, Phetlhu DR, Mulumba N, Mcingana M, Hausler H, Baral S, et al. Facilitators and barriers to incorporating digital technologies into HIV care among cisgender female sex workers living with HIV in South Africa. mHealth. 2020;6:15–15.
53. Langhaug LF, Sherr L, Cowan FM. How to improve the validity of sexual behaviour reporting: systematic review of questionnaire delivery modes in developing countries. Tropical Med Int Health. 2010;15:362–81.
54. Lowndes CM, Jayachandran AA, Banandur P, Ramesh BM, Washington R, Sangameshwar BM, Moses S, Blanchard J, Alary M. Polling booth surveys: a novel approach for reducing social desirability bias in HIV-related behavioural surveys in resource-poor settings. AIDS Behav. 2012;16:1054–62.
55. Béhanzin L, Diabaté S, Minani I, Lowndes CM, Boily M-C, Labbé A-C, Anagonou S, Zannou DM, Buvé A, Alary M. Assessment of HIV-related risky behaviour: a comparative study of face-to-face interviews and polling booth surveys in the general population of Cotonou. Benin Sex Transm Infect. 2013;89:595–601.
56. Fenton KA, Johnson AM, McManus S, Erens B. Measuring sexual behaviour: methodological challenges in survey research. Sex Transm Infect. 2001;77:84–92.

57. Viswasam N, Schwartz S, Baral S: Characterizing the role of intersecting stigmas and sustained inequities in driving HIV syndemics across low-to-middle-income settings. Current opinion in HIV and AIDS. 2020. https://doi.org/10.1097/coh.0000000000000630.

58. Kim H-Y, Grosso A, Ky-Zerbo O, Lougue M, Stahlman S, Samadoulougou C, Ouedraogo G, Kouanda S, Liestman B, Baral S. Stigma as a barrier to health care utilization among female sex workers and men who have sex with men in Burkina Faso. Ann Epidemiol. 2018;28:13–9.

59. Stangl AL, Lloyd JK, Brady LM, Holland CE, Baral S. A systematic review of interventions to reduce HIV-related stigma and discrimination from 2002 to 2013: how far have we come? J Int AIDS Soc. 2013;16:18734.

60 Relf MV, Holzemer WL, Holt L, Nyblade L, Ellis Caiola C. A review of the state of the science of HIV and stigma: context, conceptualization, measurement, interventions, gaps, and future priorities. J Assoc Nurses AIDS Care. 2021;32:392–407.

61. Raita Y, Camargo CA, Liang L, Hasegawa K. Big data, data science, and causal inference: a primer for clinicians. Front Med. 2021;8:678047.

62. Mishra S, Boily M-C, Schwartz S, Beyrer C, Blanchard JF, Moses S, Castor D, Phaswana-Mafuya N, Vickerman P, Drame F, et al. Data and methods to characterize the role of sex work and to inform sex work programs in generalized HIV epidemics: evidence to challenge assumptions. Ann Epidemiol. 2016;26:557–69.

63. Geng EH, Glidden DV, Bangsberg DR, Bwana MB, Musinguzi N, Nash D, Metcalfe JZ, Yiannoutsos CT, Martin JN, Petersen ML. A causal framework for understanding the effect of losses to follow-up on epidemiologic analyses in clinic-based cohorts: the case of HIV-infected patients on antiretroviral therapy in Africa. Am J Epidemiol. 2012;175:1080–7.

64. Baggaley RF, Fraser C. Modelling sexual transmission of HIV: testing the assumptions, validating the predictions. Curr Opin HIV AIDS. 2010;5:269–76.

65. Howe CJ, Dulin-Keita A, Cole SR, Hogan JW, Lau B, Moore RD, Mathews WC, Crane HM, Drozd DR, Geng E, et al. Evaluating the population impact on racial/ethnic disparities in HIV in adulthood of intervening on specific targets: a conceptual and methodological framework. Am J Epidemiol. 2018;187:316–25.

66. Igelström E, Craig P, Lewsey J, Lynch J, Pearce A, Katikireddi SV. Causal inference and effect estimation using observational data. J Epidemiol Community Health. 2022;76:960–6.

67. Bi Q, Goodman KE, Kaminsky J, Lessler J. What is machine learning? A primer for the epidemiologist. Am J Epidemiol. 2019;188:2222–39.

68. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health. 2021;3:e745–50.

69. Rice B, Sanchez T, Baral S, Mee P, Sabin K, Garcia-Calleja JM, Hargreaves J. Know your epidemic, strengthen your response: developing a new HIV surveillance architecture to guide HIV resource allocation and target decisions. JMIR Public Health Surveill. 2018;4:e18.

70. Leclerc-Madlala S, Broomhall L, Fieno J. The ‘end of AIDS’ project: Mobilising evidence, bureaucracy, and big data for a final biomedical triumph over AIDS. Glob Public Health. 2018;13:972–81.

71. Case KK, Ghys PD, Gouws E, Eaton JW, Borquez A, Stover J, Cuchi P, Abu-Raddad LJ, Garnett GP, Hallett TB, et al. Understanding the modes of transmission model of new HIV infection and its use in prevention planning. Bull World Health Organ. 2012;90:831-838A.

72. Shubber Z, Mishra S, Vesga JF, Boily M-C. The HIV Modes of Transmission model: a systematic review of its findings and adherence to guidelines. J Int AIDS Soc. 2014;17:18928.

73. Baral SD, Friedman MR, Geibel S, Rebe K, Bozhinov B, Diouf D, Sabin K, Holland CE, Chan R, Cáceres CF. Male sex workers: practices, contexts, and vulnerabilities for HIV acquisition and transmission. The Lancet. 2015;385:260–73.

74. Bien-Gund CH, Zhao P, Cao B, Tang W, Ong JJ, Baral SD, Bauermeister JA, Yang L-G, Luo Z, Tucker JD. Providing competent, comprehensive and inclusive sexual health services for men who have sex with men in low- and middle-income countries: a scoping review. Sex Health. 2019;16:320.

75. Kane JC, Elafros MA, Murray SM, Mitchell EMH, Augustinavicius JL, Causevic S, Baral SD. A scoping review of health-related stigma outcomes for high-burden diseases in low- and middle-income countries. BMC Med. 2019;17:17.

76. Rao A, Schwartz S, Viswasam N, Rucinski K, Van Wickle K, Sabin K, Wheeler T, Zhao J, Baral S. Evaluating the quality of HIV epidemiologic evidence for populations in the absence of a reliable sampling frame: a modified quality assessment tool. Ann Epidemiol. 2022;65:78–83.

77. Schwartz SR, Rao A, Rucinski KB, Lyons C, Viswasam N, Comins CA, Olawore O, Baral S. HIV-related implementation research for key populations: designing for individuals, evaluating across populations, and integrating context. J Acquir Immune Defic Syndr. 2019;82(Suppl 3):S206–16.

78. Hakim AJ, Johnston LG, Dittrich S, Prybylski D, Burnett J, Kim E. Defining and surveying key populations at risk of HIV infection: towards a unified approach to eligibility criteria for respondent-driven sampling HIV biobehavioral surveys. Int J STD AIDS. 2018;29:895–903.

79 do Nascimento N, Barker C, Brodsky I. Where is the evidence? The use of routinely-collected patient data to retain adults on antiretroviral treatment in low and middle income countries-a state of the evidence review. AIDS Care. 2018;30:267–77.

80. Munthali T, Musonda P, Mee P, Gumede S, Schaap A, Mwinga A, Phiri C, Kapata N, Michelo C, Todd J. Underutilisation of routinely collected data in the HIV programme in Zambia: a review of quantitatively analysed peer-reviewed articles. Health Res Policy Syst. 2017;15:51.

81. Sweeney P, DiNenno EA, Flores SA, Dooley S, Shouse RL, Muckleroy S, Margolis AD. HIV data to care-using public health data to improve HIV care and prevention. J Acquir Immune Defic Syndr. 2019;82(Suppl 1):S1–5.

82. Rao A, Lesko C, Mhlophe H, Rucinski K, Mcingana M, Pretorius A, Mcloughlin J, Baral S, Beyrer C, Hausler H, et al. Longitudinal patterns of initiation, persistence, and cycling on pre-exposure prophylaxis among female sex workers and adolescent girls and young women in South Africa. AIDS. 2023;37:977–86.

83. Hovaguimian F, Günthard HF, Hauser C, Conen A, Bernasconi E, Calmy A, Cavassini M, Seneghini M, Marzel A, Heinrich H, et al. Data linkage to evaluate the long-term risk of HIV infection in individuals seeking post-exposure prophylaxis. Nat Commun. 2021;12:1219.

84. Bekker L-G, Alleyne G, Baral S, Cepeda J, Daskalakis D, Dowdy D, Dybul M, Eholie S, Esom K, Garnett 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. 2018;392:312–58.

85. Horowitz CR, Shameer K, Gabrilove J, Atreja A, Shepard P, Goytia CN, Smith GW, Dudley J, Manning R, Bickell NA, et al. Accelerators: sparking innovation and transdisciplinary team science in disparities research. Int J Environ Res Public Health. 2017;14:225.

86. Gesicho MB, Were MC, Babic A. Data cleaning process for HIV-indicator data extracted from DHIS2 national reporting system: a case study of Kenya. BMC Med Inform Decis Mak. 2020;20:293.
87. Doidge JC, Harron K. Demystifying probabilistic linkage: Common myths and misconceptions. Int J Popul Data Sci. 2018;3:410.
88. Avoundjian T, Dombrowski JC, Golden MR, Hughes JP, Guthrie BL, Baseman J, Sadinle M. Comparing methods for record linkage for public health action: matching algorithm validation study. JMIR Public Health Surveill. 2020;6:e15917.
89. Enhancing the quality and transparency of health research: EQUATOR Network | enhancing the quality and transparency of health research. 2024. Available at: https://www.equator-network.org/.
90. Kate Rucinski PhD, MPH, Louis Masankha Banda, MSc: To close HIV prevention and treatment gaps in Malawi, study supports mix of approaches for key populations. 2022. Available at: https://www.idsociety.org/science-speaks-blog/2022/to-close-hiv-prevention-and-treatment-gaps-in-malawi-study-supports-mix-of-approaches-for-key-populations/#/+/0/publishedDate_na_dt/desc/.
91. Lesko CR, Fox MP, Edwards JK. A framework for descriptive epidemiology. Am J Epidemiol. 2022;191:2063–70.
92. Hernán MA, Hsu J, Healy B. A second chance to get causal inference right: a classification of data science tasks. Chance. 2019;32:42–9.
93. Fox MP, Murray EJ, Lesko CR, Sealy-Jefferson S. On the need to revitalize descriptive epidemiology. Am J Epidemiol. 2022;191:1174–9.
94. Joint united national programme on HIV/AIDS: The gap report. 2014. Available at: https://files.unaids.org/en/media/unaids/contentassets/documents/unaidspublication/2014/UNAIDS_Gap_report_en.pdf.
95. Marcus JL, Sewell WC, Balzer LB, Krakower DS. Artificial intelligence and machine learning for HIV prevention: emerging approaches to ending the epidemic. Curr HIV/AIDS Rep. 2020;17:171–9.
96. Edwards JK, Hileman S, Donastorg Y, Zadrozny S, Baral S, Hargreaves JR, Fearon E, Zhao J, Datta A, Weir SS. Estimating sizes of key populations at the national level: considerations for study design and analysis. Epidemiology. 2018;29:795–803.
97. Datta A, Lin W, Rao A, Diouf D, Kouame A, Edwards JK, Bao L, Louis TA, Baral S. Bayesian estimation of MSM population size in Côte d'Ivoire. Stat Public Policy. 2019;6:1–13.
98. Eaton JW, Johnson LF, Salomon JA, Bärnighausen T, Bendavid E, Bershteyn A, Bloom DE, Cambiano V, Fraser C, Hontelez JAC, et al. HIV treatment as prevention: systematic comparison of mathematical models of the potential impact of antiretroviral therapy on HIV incidence in South Africa. PLoS Med. 2012;9:e1001245.
99. Mishra S, Steen R, Gerbase A, Lo Y-R, Boily M-C. Impact of high-risk sex and focused interventions in heterosexual HIV epidemics: a systematic review of mathematical models. PLoS One. 2012;7:e50691.
100. Baral S, Turner RM, Lyons CE, Howell S, Honermann B, Garner A, Iii RH, Diouf D, Ayala G, Sullivan PS, et al. Population size estimation of gay and bisexual men and other men who have sex with men using social media-based platforms. JMIR Public Health Surveill. 2018;4:e9321.
101. Loeb T, Willis K, Velishavo F, Lee D, Rao A, Baral S, Rucinski K. Leveraging routinely collected program data to inform extrapolated size estimates for key populations in Namibia: small area estimation study. JMIR Public Health Surveill. 2024;10:e48963.
102 Viswasam N, Lyons CE, MacAllister J, Millett G, Sherwood J, Rao A, Baral SD. Global HIV Research Group: the uptake of population size estimation studies for key populations in guiding HIV responses on the African continent. PLoS One. 2020;15:e0228634.
103. Martin-Sanchez FJ, Aguiar-Pulido V, Lopez-Campos GH, Peek N, Sacchi L. Secondary use and analysis of big data collected for patient care. Yearb Med Inform. 2017;26:28–37.
104. Rao A, Lesko C, Mhlophe H, Rucinski K, Mcingana M, Pretorius A, Mcloughlin J, Baral S, Beyrer C, Hausler H, et al. Longitudinal patterns of initiation, persistence, and cycling on PrEP among female sex workers and adolescent girls and young women in South Africa, 2016–2021. AIDS. 2023. https://doi.org/10.1097/QAD.0000000000003500.
105. Rao A, Mhlophe H, Pretorius A, Mcingana M, Mcloughlin J, Shipp L, Baral S, Hausler H, Schwartz S, Lesko C. Effect of implementation strategies on pre-exposure prophylaxis persistence among female sex workers in South Africa: an interrupted time series study. The lancet HIV. 2023. Available at: https://doi.org/10.1016/S2352-3018(23)00262-X.
106. Garnett GP. An introduction to mathematical models in sexually transmitted disease epidemiology. Sex Transm Infect. 2002;78:7–12.
107. Garnett GP, Cousens S, Hallett TB, Steketee R, Walker N. Mathematical models in the evaluation of health programmes. The Lancet. 2011;378:515–25.
108. Mishra S, Pickles M, Blanchard JF, Moses S, Boily M-C. Distinguishing sources of HIV transmission from the distribution of newly acquired HIV infections: why is it important for HIV prevention planning? Sex Transm Infect. 2014;90:19–25.
109. Maheu-Giroux M, Marsh K, Doyle CM, Godin A, Lanièce Delaunay C, Johnson LF, Jahn A, Abo K, Mbofana F, Boily M-C, et al. National HIV testing and diagnosis coverage in sub-Saharan Africa: a new modeling tool for estimating the "first 90" from program and survey data. AIDS. 2019;33(Suppl 3):S255–69.
110. Johnson LF, Chiu C, Myer L, Davies M-A, Dorrington RE, Bekker L-G, Boulle A, Meyer-Rath G. Prospects for HIV control in South Africa: a model-based analysis. Glob Health Action. 2016;9:30314.
111. Bellan SE, Dushoff J, Galvani AP, Meyers LA. Reassessment of HIV-1 acute phase infectivity: accounting for heterogeneity and study design with simulated cohorts. PLoS Med. 2015;12:e1001801.
112. Correa-Agudelo E, Kim H-Y, Musuka GN, Mukandavire Z, Akullian A, Cuadros DF. Associated health and social determinants of mobile populations across HIV epidemic gradients in Southern Africa. J Migr Health. 2021;3:100038.