JAMIA Open, 2024, **7(3)**, ooae069
https://doi.org/10.1093/jamiaopen/ooae069
Research and Applications

Research and Applications

Using publicly available, interactive epidemiological dashboards: an innovative approach to sharing data from the Rakai Community Cohort Study

Kevin Footer, MBA[1], Camille M. Lake, PhD[1], Joshua R. Porter, PhD[1], Grace K. Ha, PhD[1], Tanvir Ahmed, BASc[1], Alex Glogowski, MBA[1], Anthony Ndyanabo, MSc[2], M. Kate Grabowski, PhD[3], Larry W. Chang, MD[4,5], Joseph Ssekasanvu, MSc[4], Joseph Kagaayi, PhD[2], David M. Serwadda, MMed[2], Jackie Mckina, MA[2], Christopher Whalen, MA[1], Lloyd Ssentongo, BSc[1], Ivan Nsimbi, BSc[1], Benedicto Kakeeto, BSc[1], Godfrey Kigozi, PhD[2], Robert Ssekubugu, MSPH[2], Tom Lutalo, PhD[2], Maria J. Wawer, MD[4], Ronald H. Gray, MD[4], Steven J. Reynolds, MD[2,5,6], Alex Rosenthal, MS, MBA*[,1], Thomas C. Quinn, MD[4,6], Michael Tartakovsky, MS[1]

[1]Office of Cyber Infrastructure & Computational Biology, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Rockville, MD 20892, United States, [2]Rakai Health Sciences Program, Kalisizo, Uganda, [3]Department of Pathology, Johns Hopkins School of Medicine, Baltimore, MD 21287, United States, [4]Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, United States, [5]Johns Hopkins School of Medicine, Baltimore, MD 21205, United States, [6]Laboratory of Immunoregulation, Division of Intramural Research, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Rockville, MD 20892, United States

*Corresponding author: Alex Rosenthal, MS, MBA, Office of Cyber Infrastructure & Computational Biology, National Institute of Allergy and Infectious Diseases, National Institutes of Health, 5601 Fishers Lane, Rockville, MD 20852, United States (alexr@niaid.nih.gov)

Thomas C. Quinn and Michael Tartakovsky contributed equally to this work.

Abstract

Objectives: Public sharing of de-identified biomedical data promotes collaboration between researchers and accelerates the development of disease prevention and treatment strategies. However, open-access data sharing presents challenges to researchers who need to protect the privacy of study participants, ensure that data are used appropriately, and acknowledge the inputs of all involved researchers. This article presents an approach to data sharing which addresses the above challenges by using a publicly available dashboard with de-identified, aggregated participant data from a large HIV surveillance cohort.

Materials and Methods: Data in this study originated from the Rakai Community Cohort Study (RCCS), which was integrated into a centralized data mart as part of a larger data management strategy for the Rakai Health Sciences Program in Uganda. These data were used to build a publicly available, protected health information (PHI)-secured visualization dashboard for general research use.

Results: Using two unique case studies, we demonstrate the capability of the dashboard to generate the following hypotheses: firstly, that HIV prevention strategies ART and circumcision have differing levels of impact depending on the marital status of investigated communities; secondly, that ART is very successful in comparison to circumcision as an interventional strategy in certain communities.

Discussion: The democratization of large-scale anonymized epidemiological data using public-facing dashboards has multiple benefits, including facilitated exploration of research data and increased reproducibility of research findings.

Conclusion: By allowing the public to explore data in depth and form new hypotheses, public-facing dashboard platforms have significant potential to generate new relationships and collaborations and further scientific discovery and reproducibility.

Lay Summary

Public sharing of biomedical data promotes collaboration between researchers and accelerates the development of disease prevention and treatment strategies. However, open-access data sharing presents challenges to researchers who need to protect the privacy of study participants and ensure that data are used in an ethical way. This article presents an approach to data sharing which addresses the above challenges by using a publicly available dashboard with de-identified participant data from a large HIV surveillance cohort. Using two unique case studies, we demonstrate the capability of the dashboard to investigate how HIV treatment and prevention strategies have differential impacts on various communities in the Rakai and surrounding districts in south-central Uganda. By allowing the public to explore data in depth and form new hypotheses, public-facing dashboard platforms have significant potential to generate new relationships and collaborations and further scientific discovery and reproducibility.

Key words: human immunodeficiency virus (HIV); dashboard; Rakai Community Cohort Study (RCCS); Rakai Health Sciences Program (RHSP); de-identification.

Received: October 13, 2023; **Revised:** April 17, 2024; **Editorial Decision:** July 9, 2024; **Accepted:** July 15, 2024

Introduction

Background

In 2020, 38.0 million people around the world were living with HIV, with sub-Saharan Africa carrying the burden of more than half of the world's HIV infections.[1] While there is a general worldwide decline in HIV-related deaths, the number of new infections remains disproportionately high in sub-Saharan African countries, particularly among women.[2–4] The exploration of targeted HIV prevention strategies in areas of high HIV prevalence remains a high-priority research objective.[5,6]

Global research collaborations have accelerated the pace of HIV research in low- and middle-income countries by promoting shared expertise, funding, and research infrastructure. Many of these collaborations are long-standing, involve multiple institutes and sites, and focus on diverse facets of the HIV epidemic, ultimately facilitating world-class, high-impact work with direct translational benefits. Examples of successful collaborations include the Yale Institute for Global Health,[7] Long-Acting/Extended Release Antiretroviral Research Resource Program (LEAP),[8] and Rakai Health Sciences Program (RHSP).[9] RHSP was founded in 1987 and involves collaborations between multiple international institutions, including Columbia University, Johns Hopkins University, the Ugandan Virus Research Institute, and Makerere University, Kampala. For almost 30 years, the RHSP has conducted a population-based HIV surveillance cohort called the Rakai Community Cohort Study (RCCS) with thousands of participants, providing invaluable insight into HIV epidemiology, prevention, and disease pathogenesis.[10–12]

Despite these efforts, significant hurdles remain to advancing HIV research, including disseminating research findings to the broader scientific community and public. Funding organizations, researchers, and regulators promote open data sharing to advance progress[7,8]; however, publishing journal articles, particularly where protected health information (PHI) is involved, requires strict adherence to technical, fiduciary, and academic regulations that often limit the accessibility of the primary research. At the heart of the issue is balancing the competing priorities of maximizing research accessibility while maintaining the privacy of study participants.[9,10]

To address these challenges, the RHSP collaborated with the National Institute of Allergy and Infectious Diseases Office of Cyber Infrastructure and Computational Biology (NIAID/OCICB) to develop a method of data sharing that is built on publicly available, interactive data visualizations using de-identified, aggregate participant information. These visualizations, hosted on a publicly available platform, allow users to explore HIV infection trends in Uganda[13] while protecting individual participant privacy. The dashboard leverages RCCS survey data that has been integrated into the RHSP Data Mart,[14] including sociodemographic, behavioral health, laboratory, and HIV testing and service utilization data.[15] All data available for exploration in the dashboard have undergone significant de-identification and aggregation processes described herein to protect individual participants. The dashboard is free to use and can be found at https://www.rhsp.org/research/rccs/explore-rccs-data.

Significance

The open science movement promotes greater access to data and code, which has allowed the public to engage with the scientific community, facilitated greater trust in the scientific process, and promoted the generation of new ideas. Still, barriers remain to reproducing results or generating new insights from data due to lack of resources or knowledge. For example, acquiring, wrangling, and visualizing the raw data from the RHSP data mart would take considerable time, specialized knowledge, and computational resources. The RCCS dashboard described herein helps remove some of these barriers and makes data analysis and visualization relatively simple for any member of the public. For seasoned researchers, this work improves their ability to quickly analyze data for outliers, anomalies, and patterns. Finally, for researchers that would like to access the source data, this work empowers them to make more targeted and informed data requests while reducing the burden on data managers. Since the dashboard went live in 2019, it has had more than 3000 views, highlighting its potential to empower the public, facilitate research in progress, and foster new collaborations among scientific institutions.

Objectives

The objective of this study was to design a dashboard that mitigates the risk of participant identification while maximizing relevant epidemiological information. Using this framework, we demonstrate a sustainable and reproducible model for data sharing, which enables researchers to disseminate their work to the broader community through engaging visualizations based on de-identified, aggregate data.

Methods

Data summary

Data source

The RCCS is an open, population-based cohort of consenting participants aged 15-49 years in 34 communities in the Rakai and surrounding districts in south-central Uganda. Each survey "round" lasts approximately 18 months, with the survey team working continuously during each round, and visiting each community in the same order from round to round. The open cohort design enrolls all consenting new in-migrants who have moved into RCCS communities between survey rounds. The study tracks the incidence and prevalence of HIV and the impact of HIV treatment and prevention strategies, including antiretroviral therapy (ART) and male circumcision. Data from the RCCS have been utilized in over 400 publications and have contributed to knowledge ranging from HIV biology and transmission dynamics to translational prevention, care, and treatment.[16] Data from this cohort study (13 survey rounds in total) were utilized to build the dashboard herein described.

Data preparation

Data from the RCCS were included in the dashboard if:

1) The research findings were published or part of a manuscript in progress,
2) The data could be explored by different demographics of interest, and
3) The study samples were sufficiently large to minimize the risk of participant re-identification when the data were stratified in the dashboard.

Table 1 summarizes the demographic inclusion criteria and calculations used to visualize each outcome measure in the

Table 1. Definition of inclusion criteria and outcome measures in the RCCS dashboard.

Inclusion criteria	Specific criteria
Residency status	• *Agrarian and trading*: community permanent resident (≥6 months) in 1 of 30 continuously surveyed communities • *Fishing*: community permanent or transient resident (≥1 weeks and ≤6 months) in 1 of 4 fishing communities (excludes peri-fishing communities)
Age (years)	15-49
HIV status	Known (positive or negative)

This table lists the inclusion criteria for the RCCS sub-population used in the dashboard, as previously described.[2,3]

RCCS dashboard. The summary table is included as a separate tab in the published dashboard to promote a common understanding of the calculations used to generate the variables in the analytic view and dashboard.[13]

The outcomes included in Table 2—HIV prevalence, incidence, ART coverage, and circumcision coverage—were specifically chosen for dashboard development because each is quantified within the RCCS as a primary outcome of the study with programmatic relevance to HIV.[16] These measures were stratified by five primary demographics of interest, including geographic stratum (agrarian, trading, and fishing communities), sex, age group, marital status, and religion.[2]

To date, the RCCS dashboard contains data for approximately 66 300 persons who participated in the 13 survey rounds over a 17-year timespan included in the present analysis, encompassing about 190 000 study visits—a number which continues to expand.[9]

Anonymization of cohort participants

All identifiers directly referencing individuals, including names and participant identification numbers, were removed from the data set to protect participant privacy. In addition, specific communities and ages were generalized to broader community and age groups to reduce the precision of the attributes and the risk of re-identification of individual study participants. To further mitigate risk and protect participant privacy while maximizing the utility of the information in the dashboard, we assessed the sample sizes that could be generated from combinations of five demographic variables. It was found that first selecting an epidemiological measure and prevention strategy, then grouping participants by sex, community, and one additional demographic (either age group, marital status, or religious group) resulted in a minimum number of 30 participants per outcome group, a threshold which was agreed upon by study administrators to adequately minimize the potential for re-identification for any single participant for any combination of groups and study rounds. Thus, data about individuals were de-identified by both the Safe Harbor method and the Expert Determination method in accordance with the HIPAA Privacy Rule.[17,18] The dashboard is thus designed to display aggregate data, with the first two demographic stratifications as community and sex, while the user selects the third demographic (age, religion, or marital status). Because of these measures to carefully present data in aggregate, this method of data presentation is similar in nature to presenting aggregated data in a publication, a mechanism of data sharing to which study participants agree during the informed consent process.

Validation

We compared the HIV incidence and prevalence calculations presented in the dashboard against previously published epidemiological findings,[2,3] both of which are based on data that is now housed in the RHSP data mart.[14] Differences in calculated rates of HIV incidence and prevalence, stratified by community, were negligible and did not alter any scientific interpretation of the outcomes.

Data availability

Custom images created using the data available in the dashboard are available for download. The dashboard does not provide means of accessing individual-level data or of exporting individual-level or aggregate-level data from the dashboard. Some data are available to researchers upon request; interested parties may contact RHSP at datarequests@rhsp.org.

Interactive dashboard

Technical architecture and development

We previously developed the RHSP Data Mart to store RCCS data in a modernized system that allows users to create complex queries to access the data, fostering data sharing and reproducibility of results.[14] Survey (sociodemographic, behavioral, health and HIV service utilization information) and laboratory data from the RCCS were integrated into the RHSP Data Mart using Microsoft SQL Server as previously described.[14] The RCCS dashboard was developed in Tableau Desktop using analytic views from the data mart. Data were pulled into Tableau from the data mart using native Tableau Data Connectors and stored as a Tableau Data Extract file. Rates of HIV incidence, HIV prevalence, ART, and circumcision coverage were calculated in Tableau using the definitions in Table 2. For testing, the dashboard and data extract were published to an internal QA environment, where members of the research team compared the data available on the dashboard with previously published calculations (**Figure S1**).[2,3] All differences between the dashboard data, the source of which undergoes cleaning and updating procedures regularly, and static publication data were considered minimal by study officials.

Upon completion of testing, the RCCS dashboard was published to Tableau Public, a free platform that allows registered users to explore, create, and publicly share data visualizations online.[13,19] Tableau Public's infrastructure supports millions of viewers and offers a global user community which promotes learning and collaboration. The dashboard was published with an extract of the analytical view which can be refreshed to provide users with up-to-date data. A full view of the data pipeline can be found in Figure 1. The published dashboard was embedded in the RHSP's website using an inline frame to provide better visibility to users interested in the RHSP's work.

Dashboard design

The dashboard was designed to enable users to drive their own analyses and to optimize interactivity of the visualizations. Wireframes were created using Tableau Desktop and modified accordingly during design feedback sessions between collaborating institutions. Downloads of the

Table 2. Defines each of the outcome measures in the RCCS dashboard, including prevalence, incidence, antiretroviral therapy coverage, and circumcision coverage.

Measure	Definition	Calculation
Prevalence	Percentage of people living with HIV among all participants during a survey round	$\frac{\text{Number of HIV-positive cases in current survey round}}{\text{Total population tested in current survey round}} \times 100$
Incidence	Percentage of initially HIV-negative population that becomes HIV-seropositive between the preceding and current survey round (or prior 2 survey rounds if HIV status in prior round is unknown) divided by person-years accrued between surveys expressed per 100 person years	$\frac{\text{Number of new HIV infections observed in current survey round among individuals who were HIV}-\text{negative in the prior survey round}^{*}}{\text{Person}-\text{years accrued by individuals who were HIV}-\text{negative in the prior survey round}^{*} \text{ and who were re}-\text{surveyed in the current survey round}} \times 100$
Antiretroviral therapy coverage	Percentage of people living with HIV who self-report use of antiretroviral therapy	$\frac{\text{Number of people living with HIV who self-report use of antiretroviral therapy}}{\text{Total population of people living with HIV}} \times 100$
Circumcision coverage	Percentage of males who self-report being circumcised	$\frac{\text{Number of males who self-report being circumcised}}{\text{Total population of males}} \times 100$

* Or prior 2 survey rounds if HIV status in prior round is unknown.

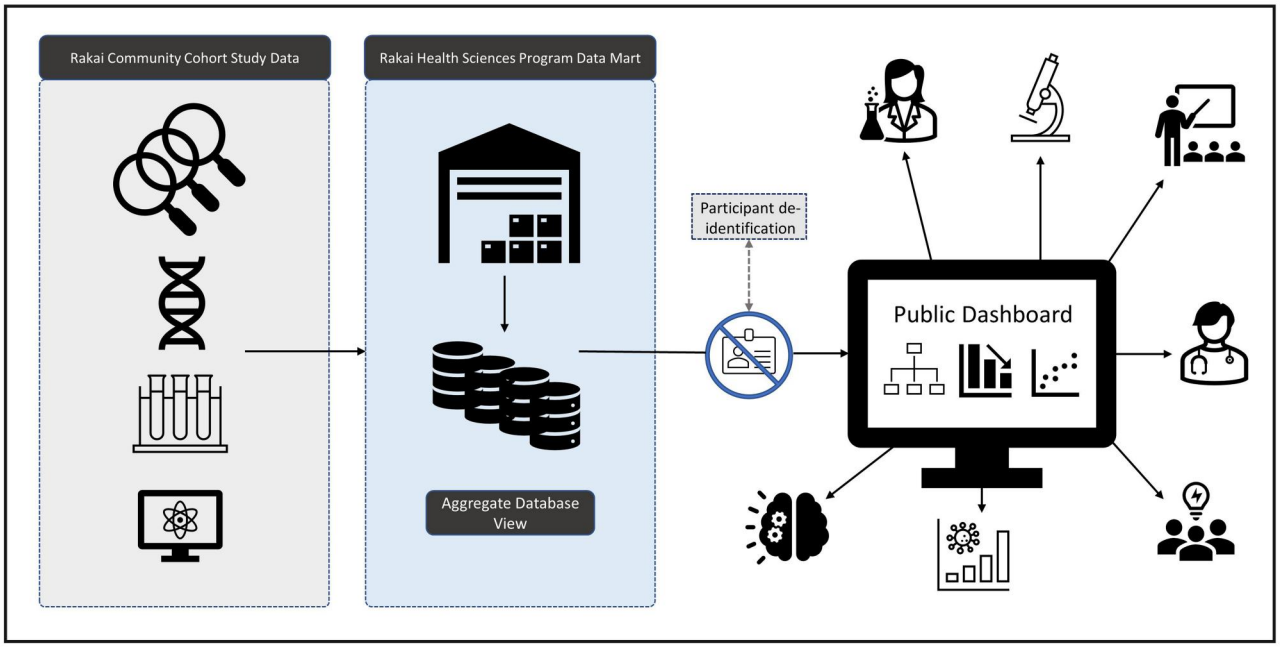

Figure 1. Data pipeline graphical abstract. Data in this study were sourced from the Rakai Community Cohort Study (RCCS) as part of the Rakai Health Sciences Program (RHSP). Data stored in the RHSP Data Mart as previously described[14] were de-identified, aggregated, and pulled into Tableau for public access and utilization.

underlying data and Tableau workbook were restricted through the Tableau Public account. The dashboard hosted on Tableau Public was also embedded on the RHSP website to make it easier for users to access.

Results

Dashboard utilization

The RCCS dashboard, hosted on Tableau Public,[19] allows users to explore HIV incidence and prevalence trends in the same subset of RCCS participants described in previous publications.[2,3,14] The dashboard was initially rolled out during an RHSP conference in Kampala in 2019 for stakeholders from government, academic, and research organizations, and subsequently at a webinar for RHSP personnel collaborators. The RCCS dashboard has also been shared on social media by members of the RHSP community and is accessible from the RHSP website,[9] demonstrating different means to increase dashboard accessibility and visibility.

Figure 2 shows the dashboard embedded in the RHSP website. Users of the dashboard can interact with a subset of RCCS data by first selecting a measure of HIV infection (incidence or prevalence), then stratifying measures by different demographic groups. From the "HIV Measure by Demographic" tab on the dashboard, users can visualize HIV incidence or prevalence trends across the survey period by selecting the appropriate measure, as demonstrated in Figure 2A. HIV infection trends can be stratified by different

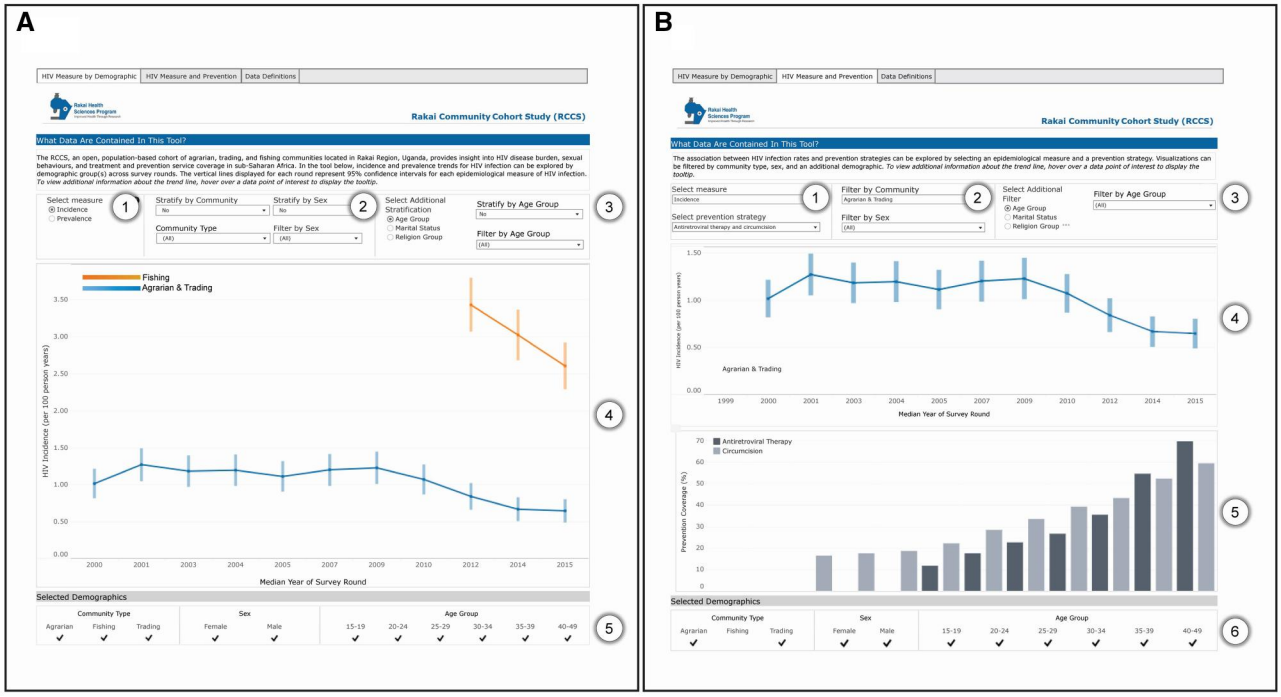

Figure 2. Dashboard design and overview. (A) Depicts the interactive features of the dashboard from the "HIV Measure by Demographic" tab, including: (1) HIV infection measure selector, (2) stratification and filtering actions by community, sex, and community type, (3) additional stratification and filtering based on age group, marital status, or religious group, (4) a line chart visualization of HIV infection trends by survey round in different community groups, and (5) selected demographics legend. (B) depicts the interactive features of the dashboard from the "HIV Measure and Prevention" tab, including: (1) HIV infection measure and prevention strategy selectors, (2) filtering actions by community and sex, (3) additional filtering based on age group, marital status, or religious group, (4) a line chart visualization of HIV infection trends by survey round, (5) a bar chart visualization of prevention coverage percentages stratified by intervention strategy and survey round, and (6) selected demographics legend. Singular data points represent the mean value for the specified year of survey round; vertical bars represent 95% confidence intervals.

demographics using the "Stratify by" drop-down selections along the top of the dashboard. These actions enable a user to visualize a trend line by community, sex, age group, marital status, and religion group. For example, the HIV incidence trend line for the "Agrarian & Trading" community group can be further stratified to display agrarian and trading communities separately by selecting "Yes" in the "Stratify by Community" drop-down filter. An additional tab in the dashboard, labeled "HIV Measure and Prevention," allows users to view HIV incidence and prevalence trends together with ART and male circumcision coverage rates across population groups, as shown in Figure 2B.

The RCCS dashboard allows users with little technical expertise to understand HIV epidemic trends and the impact of HIV prevention strategies in different contexts; previously, this information would have had to be evaluated directly from the original sources.[2,3] Additionally, the data tables in these published papers show a more limited set of demographic combinations than what is available in the dashboard, which is necessary due to space limitations. The following case studies highlight the dashboard's capability to allow users to drill down to highly specific demographic groups and combine different demographics to generate novel hypotheses from the data in different research contexts.

Case study #1: investigation of declines in HIV incidence in the agrarian and trading communities by marital status

From the "HIV Measure by Demographic" tab of the dashboard, users can view incidence and prevalence trends among different demographic groups over time. The following case study demonstrates the ability to stratify by marital status to gain new insights from the data about specific communities.

By selecting Measure = "Incidence," Stratify by Community = "No," Community Type = "Agrarian" and "Trading" (shown in the dashboard as "Multiple values"), Stratify by Sex = "No," and Filter by Sex = "All," users can see that overall HIV incidence rates in the agrarian and trading communities (combined) were steady until about 2009, after which HIV incidence declined as ART and circumcision interventions were scaled up (viewable on the "HIV Measure and Prevention" tab) (Figure 3A). If users increase the stratification further by choosing Select Additional Stratification = "Marital Status," Stratify by Marital Status = "Yes," and Filter by Marital Status = "All," this population can be broken down by reported marital status, including groups "Not Married; Previously Married," "Married," and "Never Married" (Figure 3B). The additional stratification highlights the increase and subsequent steep decline in HIV incidence among the unmarried/previously married group; the other groups declined modestly in comparison. From viewing these data, users may hypothesize that the HIV prevention strategies ART and circumcision had the highest impact on the unmarried/previously married group among the three demographic profiles displayed, which may suggest that future intervention strategies can be targeted towards specific groups for more effective HIV reduction in the agrarian and trading communities, though these postulations warrant further investigation.

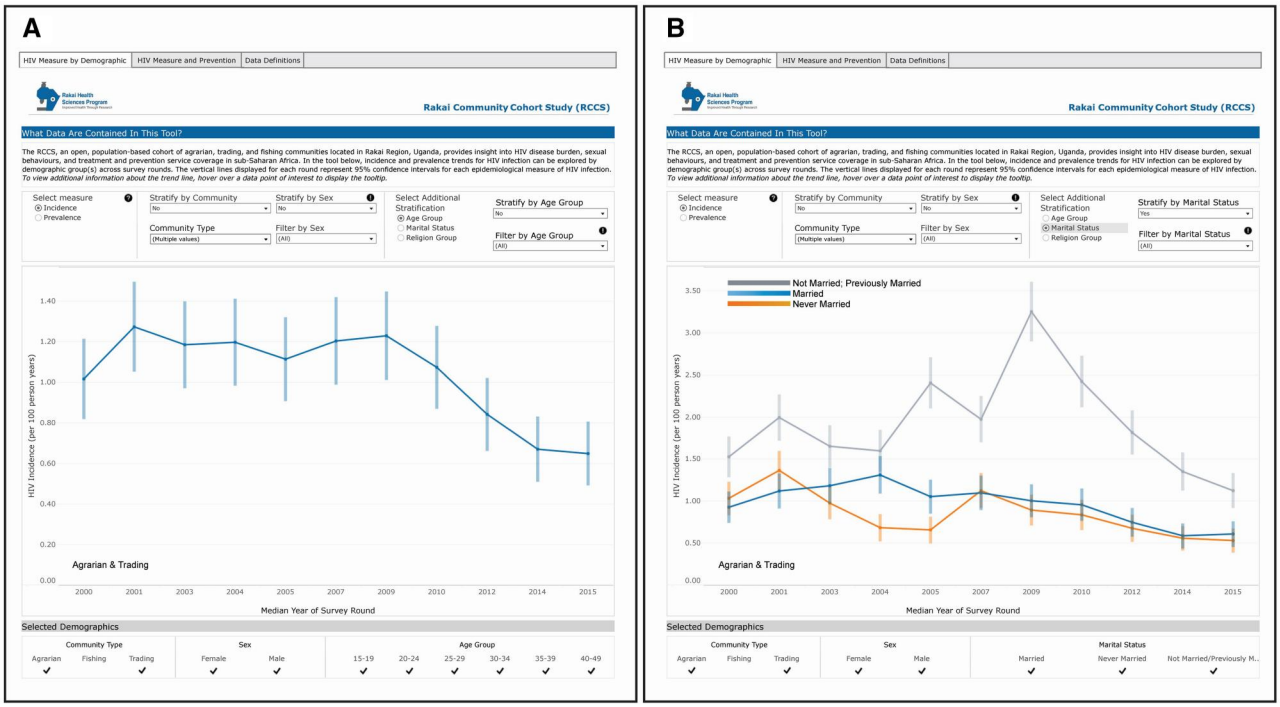

Figure 3. Case study #1: HIV incidence trends in agrarian and trading communities stratified by marital status. Using the "HIV Measure by Demographic" tab of the dashboard, users can view overall HIV incidence among the agrarian and trading communities (grouped) by selecting the following filters and stratifications: Select measure = "Incidence," Stratify by Community = "No," Community Type = "Agrarian" and "Trading," Stratify by Sex = "No," and Filter by Sex = "All" (A). To further stratify by marital status, users can update the right-hand side filters as follows: Select Additional Stratification = "Marital Status," Stratify by Marital Status = "Yes," and Filter by Marital Status = "All" (B). Singular data points represent the mean value for the specified year of survey round; vertical bars represent 95% confidence intervals.

Case study #2: exploration of HIV incidence trends by prevention strategy and religion group

From the "HIV Measure and Prevention" tab of the dashboard, users can view incidence and prevalence trends over time along with rates of the HIV prevention measures ART and circumcision. The following case study demonstrates the ability to stratify by religious group to gain new insights from specific communities.

By choosing Select measure = "Incidence," Select prevention strategy = "Antiretroviral therapy and circumcision," Filter by Community = "Agrarian & Trading," Filter by Sex = "Male," Select Additional Filter = "Religion Group," and Filter by Religion Group = "Non-Muslim," users can see a marked decline in HIV incidence among non-Muslim males in agrarian and trading communities (combined) which inversely mirrors the scale-up of both ART and circumcision in this population (Figure 4A). By changing Filter by Religion Group to "Muslim," users can see that the though the circumcision coverage is nearly 100% in this population (and therefore, controlled for in this group), the trend in HIV incidence over time is similar to that in the non-Muslim group (Figure 4B). Users may hypothesize that this similarity in HIV incidence trends between the two groups demonstrates the success of ART on its own as an interventional strategy among males in agrarian and trading communities, though this insight warrants further investigation.

These examples showcase the ability of this dashboard to facilitate epidemiological hypothesis generation by allowing finer-grained demographic filtering than what is often available in published data tables. The dashboard serves as an interactive complement to a review of published literature that can provide deeper insight into a topic with a relatively low barrier to entry.

Discussion

Here we demonstrate an innovative approach for sharing research findings through publicly available, interactive visualizations that allow users to explore HIV incidence and prevalence trends in the Rakai region of Uganda using an aggregated, de-identified data set. This work is the first to our knowledge to use interactive visualizations on a public-facing dashboard platform to disseminate epidemiological research findings. The benefits of this work for RHSP researchers and the broader scientific community and public, as well as technical limitations of this work, are highlighted below.

Benefits

Accessible exploration of research data

While typical data sharing platforms serve as repositories for data, these files must be cleansed and structured to perform analyses and construct useful visualizations manually, putting the onus on researchers to investigate and perform these duties without guidance. The RCCS dashboard requires no technical knowledge of SQL to explore the data and provides simple, easy-to-interpret graphics which aid in data digestion and interpretation. The user-centric approach of the dashboard allows smooth interaction with the data, aiding in analysis and interpretation while fostering hypothesis generation.

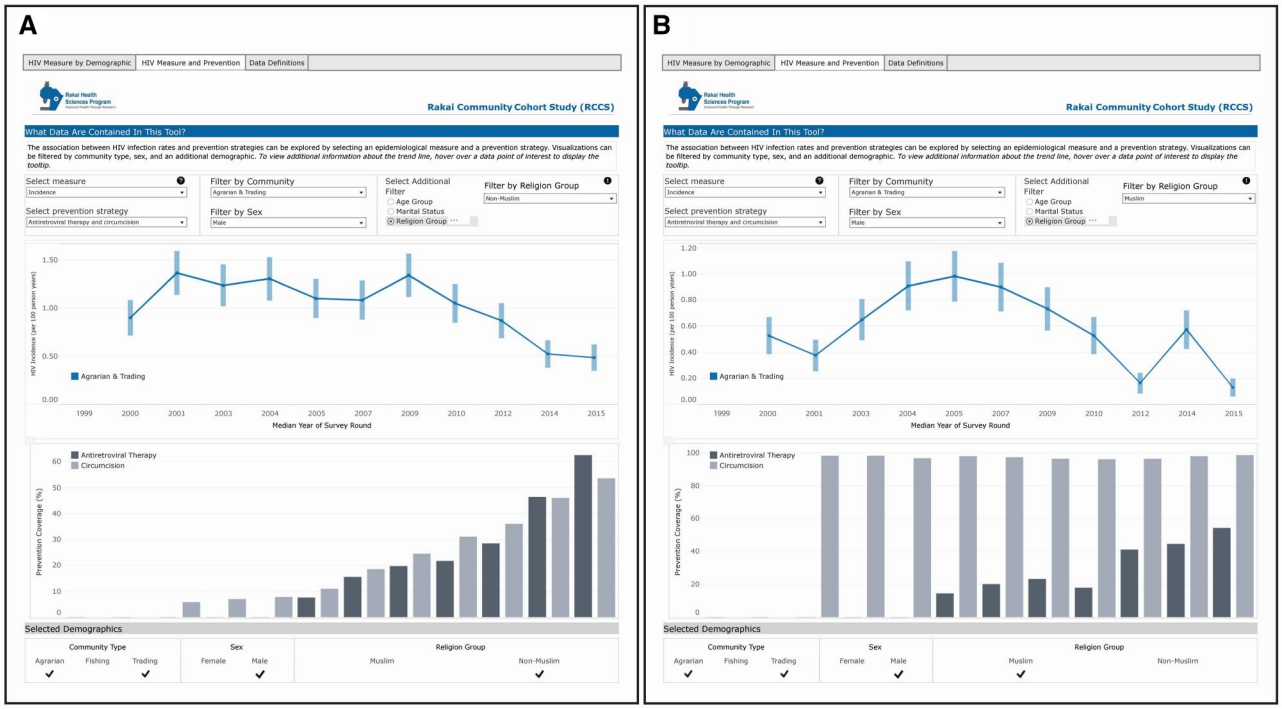

Figure 4. Case study #2: HIV incidence and prevention strategy trends among men in agrarian and trading communities stratified by religious group. Using the "HIV Measure and Prevention" tab of the dashboard, users can view HIV incidence and corresponding temporally aligned ART and circumcision coverage percentages among non-Muslim men in agrarian and trading communities (grouped) by selecting the following filters and stratifications: Select measure = "Incidence," Select prevention strategy = "Antiretroviral therapy and circumcision," Filter by Community = "Agrarian & Trading," Filter by Sex = "Male," Select Additional Filter = "Religion Group," and Filter by Religion Group = "Non-Muslim" (A). To view these trends among Muslim men in the same communities, users can change the Filter by Religion Group to "Muslim" (B). The top graph displays HIV incidence (per 100 person years) by median year of survey round, while the bottom graph displays prevention strategy coverage percentages for the same years in those communities. Singular data points represent the mean value for the specified year of survey round; vertical bars represent 95% confidence intervals.

Reproducible research findings from a validated data pipeline

Ensuring the integrity of research findings is critical to any scientific endeavor; this can be facilitated in part by using common, transparent data standards and definitions. Adherence to these standards promotes trust in the research community and allows for increased reproducibility across data platforms and collaborations. In this context, as the RCCS continues to grow and data collection for new survey rounds is completed, additional data may be added to the dashboard by using the data pipelines that were developed in the RHSP Data Mart. The criteria that were defined for incident cases, as well as the calculations for incidence, prevalence, and prevention strategy coverage rates, can be reused for future analyses due to the consolidated nature of the data pipeline, from storage in the RHSP Data Mart[14] to uploading and analysis of the data on the dashboard. Given the extensive validation that was done to ensure that outcomes aligned across the dashboard and the data sources following participant de-identification, we believe that the establishment of a common repository of validated data definitions and logic, as done herein, can benefit researchers specifically through increased data integrity and reproducibility. The approach to standardize the data flow and analysis pipeline saves time, promotes transparency, and ensures that researchers use the same standard rules for defining cohorts and other key variables needed in their analyses to generate reproducible results.

Publicly available data aggregated to protect privacy

With the growing need for research data to be made more widely accessible, the approach for anonymizing sensitive personal health and social data will depend on predefined context-dependent risk thresholds. The risk of participant re-identification necessitates maintaining appropriate safeguards to anonymize the data.[20] When considering data sharing initiatives, an evaluation of re-identification risk using statistical methods can provide guidance to research groups and organizations who want an unbiased approach to ensuring the privacy of their study participants. In this study, the dashboard and underlying dataset were designed to keep the risk of participant re-identification below a defined threshold, primarily using generalization techniques. Although the dashboard is publicly available, access to the underlying data in aggregate is restricted, following the principles of a managed access data model. However, researchers can initiate requests to the RHSP Data Management Team for additional data beyond what is made publicly available in the dashboard. While this restriction does prevent researchers from conducting secondary analyses without having to seek permission, it also protects the privacy of study participants, allows transparency for the data team into data flow, and fosters communication with potential collaborators. Future work may explore the use of additional data anonymization techniques (eg, k-anonymity and encryption keys) to control the risk of re-identification of study participants using a statistical approach.

Limitations

As described in "Materials and Methods", data exploration was conducted using the dashboard to investigate whether

there were differences between the dashboard data and previously published findings, both of which are sourced from data that is now collectively housed in the RHSP data mart.[2,3] The data in the dashboard matched previous findings closely with only minor variations (eg, the published[2] value of HIV incidence in the agrarian and trading communities in 2014 was 0.66 incident cases per 100 person-years; the dashboard value of the same year is 0.67). We believe that some date fixes in the RHSP data mart partly explain the minor variations in the computation of person time for HIV incidence. Additionally, the RCCS data contain a small number of instances of tracked individuals sero-reverting (initially HIV seropositive becoming seronegative with further testing), which is expected in a study of this size and may lead to minor changes in incidence and prevalence calculations over time. Importantly, none of the differences observed were large enough to alter interpretation of the data.

Conclusion

As data sharing continues to evolve, standardized data sharing approaches across every step in the data management and visualization pipeline will become increasingly critical to maintaining study integrity and reproducibility. Publicly available dashboards displaying aggregate, de-identified data are a user-friendly mechanism by which data usage can be standardized and maintained, while fostering hypothesis generation and research collaboration. In this study, we demonstrate that data sharing can be accomplished in a way that benefits the researchers who produce the data, collaborators, and study participants using a de-identifying data pipeline and public visualization platform. We encourage the inclusion of Tableau,[19] Power BI,[21] and other visual analytics platforms in the growing toolkit of technologies available to researchers to quickly gain insight into their data while promoting the visibility of their work to the scientific community.

Acknowledgments

We thank the Research and Education Network of Uganda (RENU) for improving the reliability of networks and facilitating the collaborative development environment.

Author contributions

Anthony Ndyanabo, Tanvir Ahmed, Alex Glogowski, Lloyd Ssentongo, and Grace K. Ha are the developers for this work. Kevin Footer is the solution architect. Joseph Ssekasanvu is a subject matter expert. Christopher Whalen, Tom Lutalo, Ivan Nsimbi, and Jackie Mckina solutioned and supplemented the development of the website infrastructure, and Christopher Whalen advised on the project overall. Grace K. Ha, Camille M. Lake, and Joshua R. Porter contributed to writing and editing the article. M. Kate Grabowski, Larry W. Chang, Godfrey Kigozi, Robert Ssekubugu, Joseph Kagaayi, Maria J. Wawer, Ronald H. Gray, Steven J. Reynolds, and Thomas C. Quinn established the RCCS and/or continue cohort data collection, provided the information used in the data mart, and contributed to dashboard testing and interpretation of results. Jackie Mckina and Benedicto Kakeeto collected underlying data and provided initial input and continuous feedback on the dashboard during the development process. David M. Serwadda, Alex Rosenthal, and Michael Tartakovsky are the project sponsors. All authors made substantial contributions to the work, reviewed and approved the publication, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Supplementary material

Supplementary material is available at *JAMIA Open* online.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. However, this project has been funded in part with federal funds from the National Institute of Allergy and Infectious Diseases (NIAID), National Institutes of Health, Department of Health and Human Services under contracts HHSN316201200018W/75N98119F00012 to Deloitte Consulting LLP and HHSN316201200160W/D13PD01160 and D14PD00002 to NET ESOLUTIONS CORPORATION. Additional funding was provided from the NIAID Division of Intramural Research.

Conflicts of interest

None declared.

Data availability

The data underlying the dashboard described in this article cannot be shared publicly due to privacy reasons and restrictions from data owners. However, some de-identified data can be provided to interested parties subject to completion of the RHSP data request form and signing of a Data Transfer Agreement. Interested parties may contact RHSP at datarequests@rhsp.org. The dashboard can be accessed at https://www.rhsp.org/research/rccs/explore-rccs-data.

References

1. Global HIV & AIDS statistics—Fact sheet. Accessed May 18, 2023. https://www.unaids.org/en/resources/fact-sheet
2. Grabowski MK, Serwadda DM, Gray RH, et al., Rakai Health Sciences Program. HIV prevention efforts and incidence of HIV in Uganda. *N Engl J Med*. 2017;377(22):2154-2166.
3. Kagaayi J, Chang LW, Ssempijja V, et al. Impact of combination HIV interventions on HIV incidence in hyperendemic fishing communities in Uganda: a prospective cohort study. *Lancet HIV*. 2019;6(10):e680-e687.
4. Kharsany ABM, Karim QA. HIV infection and AIDS in sub-Saharan Africa: current status, challenges and opportunities. *Open AIDS J*. 2016;10:34-48.
5. Chang LW, Mbabali I, Kong X, et al. Impact of a community health worker HIV treatment and prevention intervention in an HIV hotspot fishing community in Rakai, Uganda (mLAKE): study protocol for a randomized controlled trial. *Trials*. 2017;18(1):494.
6. Ratmann O, Kagaayi J, Hall M, et al., Rakai Health Sciences Program and the Pangea HIV Consortium. Quantifying HIV transmission flow between high-prevalence hotspots and surrounding communities: a population-based study in Rakai, Uganda. *Lancet HIV*. 2020;7(3):e173-e183.
7. Yale Global HIV/AIDS Research Network. Accessed May 22, 2023. https://medicine.yale.edu/yigh/faculty-support-initiative/faculty-networks/yale-garner/

8. LEAP | Home. Accessed May 22, 2023. https://longactinghiv.org/
9. About—History | RHSP Background and History. Accessed May 18, 2023. https://www.rhsp.org/index.php/about-us/our-history
10. Nakigozi G, Makumbi F, Reynolds S, et al. Non-enrollment for free community HIV care: findings from a population-based study in Rakai, Uganda. *AIDS Care*. 2011;23(6):764-770.
11. Wagman JA, Gray RH, Campbell JC, et al. Effectiveness of an integrated intimate partner violence and HIV prevention intervention in Rakai, Uganda: analysis of an intervention in an existing cluster randomised cohort. *Lancet Glob Health*. 2015;3(1):e23-e33.
12. High prevalence of liver fibrosis associated with HIV infection: a study in rural Rakai, Uganda. Accessed May 18, 2023. https://doi.org/10.3851/IMP1783
13. HIV incidence and prevalence trends observed in Uganda. Tableau Public. Accessed May 18, 2023. https://public.tableau.com/app/profile/rhsp.dashboard.users/viz/RCCSHIVInfectionandPrevention/HIVMeasureandPrevention
14. Ndyanabo A, Footer K, Ahmed T, et al. Establishing a centralized data mart from the Rakai community cohort study to improve HIV research in Rakai, Uganda. *JAMIA Open*. 2022;5(2):ooac032.
15. Chang LW, Grabowski MK, Ssekubugu R, et al. Heterogeneity of the HIV epidemic in agrarian, trading, and fishing communities in Rakai, Uganda: an observational epidemiological study. *Lancet HIV*. 2016;3(8):e388-e396.
16. Rakai Community Cohort Study—RCCS. Accessed May 22, 2023. https://www.rhsp.org/research/rccs/rccs-overview
17. Simon GE, Shortreed SM, Coley RY, et al. Assessing and minimizing re-identification risk in research data derived from health care records. *EGEMS (Wash DC)*. 2019;7(1):6.
18. Rights (OCR) O for C. Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) privacy rule. HHS.gov. 2012. Accessed July 19, 2023. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html
19. Tableau Public. Tableau. Accessed May 18, 2023. https://www.tableau.com/products/public
20. data-policy-models-for-funding-bodies-eagda.pdf. Accessed May 19, 2023. https://wellcome.org/sites/default/files/data-policy-models-for-funding-bodies-eagda.pdf
21. Data Visualization | Microsoft Power BI. Accessed May 19, 2023. https://powerbi.microsoft.com/en-us/

JAMIA Open, 2024, 7, 1–9
https://doi.org/10.1093/jamiaopen/ooae069
Research and Applications