Protocol

A Multisource, Multilevel Contextual Database to Support Ending the HIV Epidemic Initiative: Protocol for Design, Construction, Harmonization, and Quality Assurance

Nikita Rao[1], PhD, MPH; Lingling Zheng[1]; Wyatt Deihl[2], MPH; Raul U Hernandez-Ramirez[3], PhD; Donna Spiegelman[4], ScD; J Lucian Davis[5], MD, MAS; Debbie Humphries[1], PhD, MPH, MA

[1]Rigorous, Rapid and Relevant Evidence aDaptation and Implementation (R3EDI) for Ending the HIV Epidemic (EHE), Center for Methods in Implementation and Prevention Science (CMIPS), Yale School of Public Health, Yale University, New Haven, CT, United States

[2]Partnerships and Development Department, CAN Community Health, Tampa, FL, United States

[3]Center for Interdisciplinary Research on AIDS (CIRA), Center for Methods in Implementation and Prevention Science (CMIPS), Yale School of Public Health, Yale University, New Haven, CT, United States

[4]Center for Methods in Implementation and Prevention Science (CMIPS), Yale School of Public Health, Yale University, New Haven, CT, United States

[5]Section of Pulmonary, Critical Care, and Sleep Medicine, Yale School of Medicine, Department of Epidemiology of Microbial Diseases, Yale School of Public Health, Yale University, New Haven, CT, United States

Corresponding Author:

Nikita Rao, PhD, MPH
Rigorous, Rapid and Relevant Evidence aDaptation and Implementation (R3EDI) for Ending the HIV Epidemic (EHE)
Center for Methods in Implementation and Prevention Science (CMIPS), Yale School of Public Health, Yale University
PO Box 208034, New Haven
New Haven, CT 06250
United States
Phone: 1 7707147619
Email: nikita.rao@yale.edu

Abstract

Background: The Ending the HIV Epidemic (EHE) initiative remains a national priority in the United States, aiming to reduce new HIV infections by 90% by 2030. As we cross the initiative's midpoint, there is renewed commitment to strengthening the HIV workforce's capacity to plan, implement, and sustain effective HIV prevention, treatment, and care interventions. Despite substantial improvement in HIV outcomes, uneven implementation of evidence-based interventions reflects persistent gaps between available evidence and its translation into locally actionable practice. Achieving EHE goals requires tailoring implementation to the diverse sociostructural conditions shaping HIV outcomes across jurisdictions. Research increasingly highlights the value of integrated, contextual data to strengthen public health decision-making. Linking indicators spanning multiple conceptual domains across various geographic levels can support a robust understanding of the distinct drivers of HIV outcomes, yet existing data systems remain fragmented. A harmonized, multisource, multilevel database is therefore essential to support targeted, needs-based, and data-driven implementation under the EHE initiative. We present a reproducible protocol detailing the transferable methodology of building and maintaining such a resource.

Objective: This project aims to (1) build a high-quality contextual database integrating multiple data sources using transparent, replicable, and updatable methods and (2) develop and document systematic workflows for ongoing updates and quality assurance (QA) and support future use aligned with open-science frameworks and standard data practices.

Methods: This project will follow best practices in data architecture, acquisition, standardization, and QA. For objective 1, we will integrate data across multiple geographic levels (eg, ZIP code, county) for 2020-2025, with measures categorized into conceptual domains (eg, epidemiologic and sociodemographic) guided by established theoretical frameworks. For objective 2, we will develop a tiered data structure using a GitHub repository to store documentation, processing scripts, and QA logs aligning with open science practices. Construction and QA methods were informed by targeted PubMed literature reviews. Data sources include existing data repositories, literature-identified datasets, and gray literature with consistent formatting and permissive terms of use suitable for web scraping. Stakeholder engagement will span all phases of database development, informing variable selection, usability, and validation to enable iterative refinement and revision.

Results: Funding was obtained in August 2025. Literature reviews from October to November 2025 informed the database construction methods, source identification, and protocol development. Data acquisition, standardization, and QA began in May 2026, with acquisition remaining iterative. Cross-source integration, final QA checks, and dashboard development are expected to conclude in October 2026. Study results are anticipated for publication in summer 2027.

Conclusions: This protocol will provide a reproducible, transferable model for building and maintaining a contextual data resource that could support public health planning and advance implementation science through more context-responsive decision-making under the EHE initiative.

International Registered Report Identifier (IRRID): DERR1-10.2196/100648

JMIR Res Protoc 2026;15:e100648; doi: 10.2196/100648

Keywords: HIV; implementation science; data science; exploration; data harmonization

Introduction

The Ending the HIV Epidemic (EHE) initiative, coordinated by the US Department of Health and Human Services (HHS), is a national effort involving multiple federal, state, and local partners that aims to reduce new HIV infections by 90% by 2030 [1]. Although decades of biomedical and public health advances have improved HIV treatment, prevention, and care, uneven implementation of evidence-based interventions continues to hinder national progress toward EHE goals [2]. This challenge is evident in the stark concentration of HIV burden in a small number of counties and in the disproportionate impact on specific populations [3]. The challenge of translating evidence into action may intensify during periods of shifting policy and funding priorities, as the HIV workforce must make implementation decisions under increasingly constrained operational conditions [4]. Compounding these pressures, interruptions in the availability and continuity of data sources may affect researchers' and practitioners' ability to access the evidence needed to inform local implementation. A systematically assembled, well-documented database that tracks source-level changes can help buffer against these disruptions and support more resilient, evidence-informed decision-making.

A growing body of literature identifies the availability of diverse, multilevel data as a cornerstone of effective public health decision-making. In particular, contextual research evidence has emerged as essential for understanding the multifaceted nature of health systems and how they shape community health, as demonstrated by a recent systematic review examining data use in government-led public health interventions [5]. This is further supported by evidence emphasizing the role of contextual factors in influencing health outcomes. For example, recent county-level analyses using composite measures estimate that socioeconomic factors alone account for nearly half of county-level variation in health outcomes, far exceeding the contribution of clinical care [6]. A systematic review similarly found that socioeconomic, cultural, and environmental contexts shape the effectiveness of public health interventions yet remain underreported in the current evidence base [7]. National surveillance data further document substantial disparities in social determinants of health (SDOH) across populations, affirming the importance of contextual factors in understanding risk exposure and health status [8]. SDOH represent one key domain within a broader contextual framework that also encompasses environmental, structural, and geographic influences on health. Throughout this protocol, we use the term contextual indicators to refer to measurable, area-level characteristics of the environments in which people live that shape their health outcomes, spanning epidemiologic, sociodemographic, economic, structural, and policy-related conditions across geographic levels. The need for integrated, multilevel evidence is further illustrated by syndemic processes, through which shared social and structural determinants simultaneously shape multiple epidemics. Alongside intersectionality frameworks, this perspective also highlights how overlapping systems of inequality shape both exposure to risk and access to care [9,10]. For example, the opioid crisis has reshaped HIV transmission dynamics by exploiting overlapping social vulnerabilities [11]. Because SDOH operate across multiple levels, with some rooted in individual circumstances and others clustering geographically at the community level, individual-level data are often both unavailable and insufficient for capturing these contextual nuances. Addressing these complex drivers of health would therefore be aided by a data infrastructure that aggregates, standardizes, and links existing publicly available data across systems and scales, improving accessibility of multilevel contextual information to support public health intervention planning and practice. The development of such an infrastructure, however, requires confronting persistent challenges in how data systems are currently built and integrated.

Advances in contextual database design show that integrating data across multiple levels is necessary to capture the macro-micro linkages through which broader contextual factors shape individual and population health trends [12,13]. Yet, existing data systems remain disconnected, developed for differing purposes, varying scales, and disparate knowledge domains [12], contributing to a siloed understanding of contextual factors in practice. Data harmonization therefore plays a key role in generating evidence that is comparable, analytically robust, and relevant for policy. A multisource, multilevel database could substantially advance targeted implementation efforts under the EHE initiative.

As part of a broader effort to enhance the rigor and impact of HIV implementation research across EHE priority jurisdictions [14], the primary objective of this project is to build a reproducible, high-quality, contextual database, constituting the second version of an existing database

currently housed on the Yale Center for Methods in Implementation and Prevention Science (CMIPS) website [15]. This database is distinguished from existing contextual data resources by its integration and harmonization of multiple sources into a single, HIV-focused, multilevel resource. While AIDSVu primarily visualizes HIV surveillance data alongside selected social determinants, County Health Rankings & Roadmaps ranks counties on broad health measures, and US Census data provide raw counts and individual demographic measurements, each is typically organized at a single geographic level and for a distinct purpose. In contrast, the present database links harmonized contextual indicators across multiple geographic levels and conceptual domains, tailored specifically to EHE priority jurisdictions, and documents source-level provenance and updates to support transparent, reproducible reuse.

A secondary objective is to create transparent and systematic workflows for ongoing updates, quality assurance (QA), and future analyses, thereby providing resources for generating reproducible and transparent knowledge to support EHE-aligned implementation research. In this protocol, we describe the systematic processes and phases guiding database development, articulate the anticipated value of the linked dataset, and discuss anticipated barriers and facilitators, along with plans for dissemination and future analytic work.

We present this work as a reproducible contextual data infrastructure protocol with transparent methodology for assembling, harmonizing, documenting, and maintaining a multilevel contextual data resource. Its central contribution is not the resulting dataset alone but the integrated and reproducible process of building and sustaining such a resource spanning data architecture, harmonization, QA, documentation, and stakeholder engagement.

Methods

We will use a structured, theory-informed approach to develop the contextual HIV database, using conceptual frameworks to guide domain organization, identifying relevant data sources through a targeted and pragmatic literature review, and applying standardized procedures for database architecture, data acquisition, harmonization, QA, and stakeholder engagement. Together, these components will support the development of a rigorous, reproducible, and scalable data resource, as described below.

Conceptual Framework

In this database, measures will be organized into conceptual domains informed by Ecological Systems Theory (EST) and the First Nations Perspective on Health and Wellness, 2 complementary frameworks that together address distinct aspects of how context shapes HIV outcomes (Figure 1). Rather than offering redundant accounts of the same phenomena, each framework answers a different organizing question: EST specifies *at what* levels contextual indicators operate, while the First Nations Perspective informs *what is counted as health-relevant* and whose perspective defines it. First proposed by Urie Bronfenbrenner [16], EST frames lifespan development as occurring within nested systems of influence, from immediate family environments to larger societal structures [17]. This nested structure maps directly onto the multilevel architecture of our database, in which contextual indicators are assembled across individual, neighborhood, county, state, and national scales, allowing the framework to organize variables according to the level at which each indicator operates. EST is well-suited to HIV research, where outcomes are shaped by interactions between individual-level influences and macrolevel factors [18,19].

The First Nations perspective complements EST by supplying a holistic and inherently collective concept of health, one that emerges from the interplay of social, cultural, and relational systems rather than residing solely within the individual [20,21]. This perspective broadens the contextual domains we capture to include community and relational conditions that many health-systems models tend to omit, but that are increasingly recognized as central to HIV prevention and care in the United States. Although this perspective originates in an Indigenous health context, its emphasis on collective and structural determinants is transferable to US HIV implementation settings, where disparities are similarly patterned by community-level and structural conditions rather than individual behavior alone.

Together, these frameworks provide a theoretically grounded structure for organizing contextual factors into conceptual domains, illustrating how data across multiple levels interact to influence health outcomes. Figure 1 presents this structure as a proximal-to-distal organization of domains, ranging from more immediate, individual-level determinants to broader structural and systemic influences. This arrangement is not ranked but simply shows domains in relation to one another, consistent with the manuscript's emphasis on interconnected, multilevel influences. Each conceptual domain was drawn from these frameworks and then translated into specific database variables. For example, the economic stability domain maps to indicators such as housing stability status, median household income, and median hours worked. These domains were reviewed with stakeholders, including implementation scientists, people with lived experience of HIV, a research librarian, and a potential end user of the database, who examined the planned variables captured within these domains and how well they relate to one another. This review led to refinements, such as the addition of policy-related indicators in the social and community context domain, including reentry program funding per capita, sex work criminalization and enforcement intensity, and drug paraphernalia laws. In the context of HIV research, this approach supports the development of comprehensive, culturally relevant implementation strategies that address layered vulnerabilities in HIV prevention and care.

Figure 1. The relational HIV ecosystem model.



Literature Search

To inform database construction, harmonization, and QA procedures and to ensure alignment with best practices in community engagement [22], we conducted 2 targeted, pragmatic literature reviews. Rather than aiming to provide an exhaustive synthesis of the literature, these reviews were intended to identify usable methods and candidate contextual variables sufficient to inform our build. We therefore prioritized relevance and feasibility over exhaustive retrieval. Working with implementation researchers and a librarian from the Yale School of Public Health, we refined search strategies for (1) a methodological literature search focused on database design and management approaches and (2) a separate literature search identifying proxy variables, defined as indicators used to approximate relevant contextual conditions that are not directly measured, to ensure that contextual indicators are captured. These 2 reviews served complementary purposes: the methodological review identified approaches for building and managing the database (data management, harmonization, linkage, QA, and web-scraping), while the proxy-variable review identified its contextual content (candidate variables and data sources). In consultation with the librarian, we focused on PubMed as the primary database, for its broad indexing of health and biomedical literature. Recognizing that methodological literature in fields such as data engineering, information technology, and geospatial science is indexed across additional platforms, we supplemented the searches by informally reviewing citations and references from included studies and by conducting complementary gray-literature searches. Searching was considered sufficient once it yielded methods and variables adequate to inform the database design. To document this process transparently, we provide full search strings, filters, dates, and result counts in Multimedia Appendix 1 and screening criteria in Multimedia Appendix 2.

We imported search results into their respective libraries using Zotero citation management software (version 7.0; Corporation for Digital Scholarship) [23] and subsequently into Rayyan [24], a free online screening tool for literature reviews. Based on predefined inclusion and exclusion criteria (Multimedia Appendix 2), one researcher (NR) reviewed the methodological literature search results and identified approaches related to data management, web-scraping, data integration, and data linkage. Because this review addressed a narrow question under strict eligibility criteria, 7 studies ultimately met the criteria and informed the framework, reflecting the specificity of the question and the review's pragmatic scope rather than a search limitation. A second researcher (LZ) conducted the proxy variable literature search, which further helped identify traditional datasets (eg, American Community Survey [ACS], AIDSVu [Rollins School of Public Health, Emory University, in partnership

with Gilead Sciences, Inc, and the Center for AIDS Research at Emory University]) for inclusion in our database. Complementary gray literature searches were also conducted to identify unstructured, web-based sources appropriate for web scraping. Screening for both reviews was conducted by a single reviewer, without independent dual screening or interrater reliability assessment, consistent with the targeted pragmatic design of this literature search. While this may introduce selection bias, its impact on the database build is expected to be limited.

Overview of Database Structure and Architecture

Database development will follow a structured approach guided by best practices in data architecture, acquisition, standardization, and QA to create a multisource, multiyear, and multilevel contextual resource to support HIV implementation research. We will integrate national-, state-, county-, and ZIP Code-level data for the years 2020-2025, accounting for the differing update cycles of contributing sources, including annual, biennial, and 5-year releases (eg, ACS 5-y estimates). All measures will be categorized into clearly defined conceptual domains informed by EST and the First Nations Perspective on Health and Wellness to facilitate future analyses.

We will use a 4-tier data architecture to ensure transparent and reproducible data management. Tier 1 will consist of raw and unmodified source files alongside their original documentation. In Tier 2, datasets will be cleaned and standardized by applying naming conventions, recoding data structures, aligning geographical identifiers, and tagging years. Tier 3 will integrate all standardized sources into a harmonized master file using predefined merging and conflict-resolution rules [25]. QA procedures will be applied across Tiers 2 and 3 to assess data completeness, consistency, and integration. Tier 4 will provide the final dataset, comprising curated variables, derived measures, and conceptually grouped indicators to support future analyses. The compiled database is structured as a single integrated, long-format file rather than a multitable relational database. Each row represents one county in a year, uniquely identified by its Federal Information Processing Standards (FIPS) code and year. Standardized identifier columns (FIPS code, state, county name, and year) appear first in every version of the file, followed by all substantive variables, each prefixed by the source (eg, acs_ and aidsvu_) to prevent naming conflicts. This structure is built in two stages: (1) datasets from the same source are first stacked longitudinally across years and (2) these datasets are then integrated across sources by joining on the shared county FIPS identifier. All versions of the integrated file are saved under version-controlled naming conventions to support reproducibility. This 4-tier system provides a stable foundation for ongoing maintenance and future scalability, reflecting standard data design and management frameworks [22,26]. To maintain version control and ensure transparency and traceability throughout all stages of database construction, we will implement a complementary documentation and version-control plan. During the prebuild phase, all project materials will be housed in a shared folder for centralized access. Excel logs will be used to maintain detailed records of search strategies, inclusion/exclusion decisions, and candidate database summaries. Once the database build is initiated, we will migrate all documentation and code to a shared GitHub repository with version control and issue tracking to align with open-science standards and support long-term reproducibility. In addition to GitHub-based version control of code and documentation, each public release of the harmonized dataset will be maintained as a fixed, versioned snapshot hosted on Yale-managed infrastructure, accompanied by documentation describing changes relative to the prior release. Prior versions will remain accessible so that external users can retrieve and cite the specific version of the database used in a given analysis, even as the database continues to be updated.

To document variables and sources, we will maintain a structured data dictionary aligned with the Dublin Core Metadata Element Set [27], capturing each source's title, origin, publisher, date, coverage, license, and description. Source-level provenance, including the originating dataset, access date, and source version or release, will be recorded in the dictionary and updated as sources are refreshed, providing a timestamped history on GitHub.

For each derived variable, the data dictionary will record its source variables, the transformation applied, and a reference to the script that generates it. The transformation code will be kept under version control in GitHub. The data dictionary will show how each variable is produced, while GitHub's commit history and issue tracking will provide a time-stamped record of source-level changes, derivations, and updates over time. Together, these practices align with current FAIR (Findability, Accessibility, Interoperability, and Reusability) data principles that guide our approach (Multimedia Appendix 3) [28].

Data Acquisition, Collection, and Harmonization

The final selection of datasets and web-based sources to be included in the contextual database will be derived from three sources: (1) data sources identified in the existing CMIPS database, (2) datasets identified in the proxy-variable literature search, and (3) reports and gray-literature sources with consistent formatting and permissive terms of use suitable for web scraping.

Consistent with the goal of building a comprehensive contextual database, our approach prioritizes inclusion rather than selective exclusion. Candidate sources will be retained unless they lack a permissive license or a format amenable to reliable extraction. To refine and validate this preliminary list, we will share it with stakeholders, including the AIDSVu development team and potential end users, who will review it for relevance and utility. Stakeholder input will serve in an advisory role to identify additional sources for consideration to expand and prioritize candidate data sources rather than to remove them, keeping in line with the database's comprehensive intent. Suggested sources will be incorporated where

they meet licensing requirements; if a source does not meet these requirements, this exclusion reason will be documented.

Because the database is intended to be inclusive, sources are retained even when they overlap or conflict. Rather than resolving cases by exclusion, we preserve the sources and document overlaps, discrepancies, and source information in the data dictionary and methods documentation, so that end users can determine usage for their own purposes. Similarly, incomplete data will be retained rather than discarded and flagged and described in the data dictionary and methods documentation alongside missing, suppressed, and otherwise limited values. All inclusion decisions, data sources, and data quality descriptions will be documented in a log, accompanied by justification notes to maintain transparency and reproducibility.

Full datasets will be manually downloaded directly from the source when available. If unavailable, semiautomated web-scraping methods adapted from the web scraping, text mining, and spatial overlay method [29] will be applied to extract raw data values from unstructured sources such as PDFs and web pages. Web scraping tasks will be performed using the rvest package for static data [30] or the RSelenium package [31] for dynamic webpages. This approach supports reproducible and transparent collection of nontabular data, ensuring broader coverage of relevant contextual indicators.

Individual dataset cleaning and harmonization will include applying standardized naming conventions, addressing missing and inconsistent values, converting measures into uniform units and scales, and using data crosswalks to map and align geographical elements (eg, FIPS codes, Core Based Statistical Areas, ZIP codes, etc). Geographic units that require translation to the county level, including sources that do not report FIPS codes directly, will be mapped using the US Department of Housing and Urban Development United States Postal Service (USPS) ZIP-County Crosswalk File to derive a standardized FIPS code prior to matching. We will link datasets at the level of EHE jurisdictions (US counties, states, and equivalent areas). Matching will be performed on the FIPS code, as it is a standardized geographic identifier common to all our sources following crosswalk translation. Because FIPS codes are exact, standardized identifiers, this deterministic approach yields distinct matches and avoids the ambiguity of name-based matching, and conflicting matches are not anticipated. Jurisdiction names will be used only as a secondary check if they differ across sources and will be manually reviewed to identify and correct any formatting or naming discrepancies, based on the recommendation of Cao et al [25]. For example, a county in both AIDSVu and ACS is matched on its 5-digit FIPS code, with its name checked for formatting differences. Sources with differing update frequencies are accommodated through the matching process. Annually updated sources are matched by both FIPS code and year, while sources that update less frequently (eg, ACS 5-y estimates) are matched by FIPS code alone and populate only the years corresponding to their actual data release; years without a new source update are left blank rather than carrying forward the prior value, so that users are not at risk of interpreting a repeated value as an independently observed annual measurement. Users conducting analyses that require complete year-by-year coverage are responsible for determining and applying their own approach to these blank years (eg, carrying values forward, interpolation, or restricting analysis to observed years), consistent with the assumptions appropriate to their specific research question. This will be noted in the data dictionary and data methods for each nonannual indicator, along with the source's typical update frequency, so users can identify affected variables before conducting longitudinal analyses. We cross-referenced the US Census Bureau's official record of substantial county boundary changes [32] against the 57 EHE priority jurisdictions and confirmed no overlap during the 2020-2025 study period (eg, Connecticut's 2022 transition to planning regions and county-equivalent boundary adjustments in Alaska and Virginia all fall outside EHE-designated counties and states), supporting the stability of FIPS-based geographic identifiers for the scope of this database.

QA

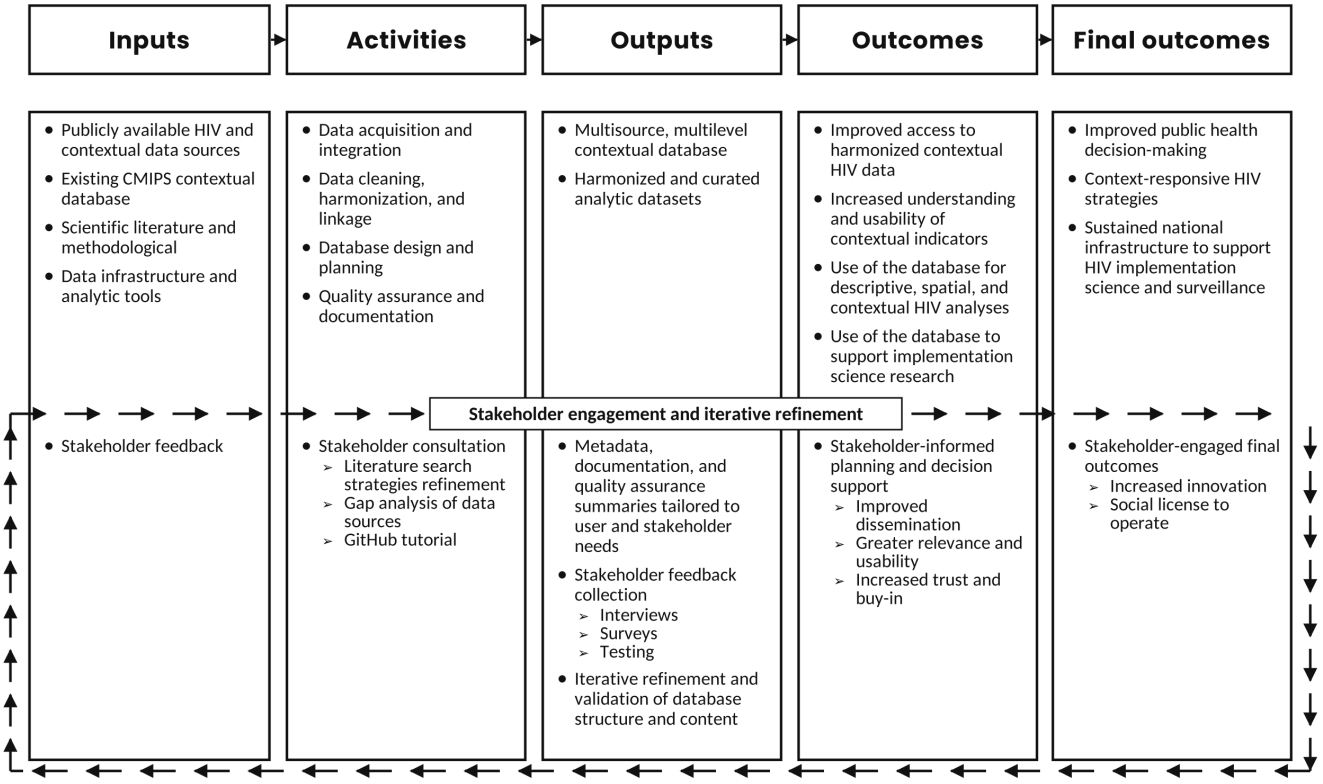
Our QA approach will follow the Bray and Parkin [26] framework, which assesses data quality in terms of completeness, comparability, timeliness, and validity across all data sources. In accordance with methods outlined by Palis et al [33], we will assess these benchmarks in Tiers 2 through 4 of database construction. Systematic reabstraction (including re-extracting key variables from original data sources) and re-derivation (independently reproducing derived variables and indicators) audits will be conducted manually on a stratified sample (eg, by HIV burden, region, and rurality). Audit scope, including variable selection and the sample size for each cycle, is determined based on the dataset under review. As a general guidance, audits sample at least 5% of records per stratum, with a minimum of 10 records per stratum, and are conducted by an auditor independent of the person who originally processed the data. A target of ≥95% agreement is consistent with reabstraction reliability standards reported for large national health registries such as the National Cancer Database [33] and with Landis and Koch [34]'s widely used interpretation of $\kappa \geq 0.8$ as indicating almost perfect agreement. When agreement or κ targets are not met, findings are escalated for team review to determine whether the discrepancy reflects a systematic processing error requiring correction and reprocessing of downstream steps, an isolated error requiring targeted correction, or an ambiguity in documented procedures requiring classification before reprocessing. These results will be summarized in a report for team discussion and action [35], in which each finding is classified as resolved, requiring corrective action with an assigned responsible party and target date, or accepted as a known limitation with the rationale documented. Insights from Zhang et al [36] will further inform our QA structure, including governance and data-security considerations relevant for future integration of sensitive or clinic-level data. Database updates will follow documented and version-controlled workflows encompassing data acquisition and harmonization alongside reapplication of QA checks to ensure consistency, transparency, reproducibility, and scalability over time. The update cadence will be

source-specific, aligning with each source's release schedule rather than a fixed timeline. Where feasible, automated update pipelines will be implemented for publicly available data on predictable schedules, while other sources will undergo manual review and updates to preserve data integrity.

Stakeholder Engagement

We will engage stakeholders during all phases of contextual database development, following principles from implementation science and registry-development frameworks [22,33,37,38]. Stakeholders were identified through purposive sampling, drawing on known experts and prospective users relevant to each phase of development; individuals were approached as their specific expertise or perspective became pertinent to the work at hand. During the planning and literature-search phases, the team engaged local implementation science researchers and a research librarian to refine the conceptual domains, search strategy, and inclusion criteria. Additional stakeholder involvement is planned throughout subsequent phases of the workflow, as illustrated in Figure 2.

Figure 2. Logic model for contextual database development, use, and stakeholder engagement. CMIPS: Center for Methods in Implementation and Prevention Science.



This figure presents a logic model linking project inputs, activities, outputs, outcomes, and anticipated final outcomes, with stakeholder engagement and iterative refinement integrated across all stages of database development. This includes engagement with data scientists, people living with HIV, and prospective end users such as implementation science researchers and other public health professionals. Consultations will be conducted iteratively rather than on a fixed schedule, tied to development milestones shown in Figure 2, with specific stakeholders engaged as relevant to each stage. For example, we consulted an expert in web-scraping and version control (GitHub) to inform our data-engineering approach and a prospective end user to identify additional contextual variables of practical value. Stakeholder feedback will inform iterative revisions to the identification of data sources, the data architecture, QA procedures, user interface features, and dissemination strategies. We will develop a stakeholder engagement matrix mapping key advisory groups, including clinicians, non-governmental organization workers, policymakers, researchers, and people living with HIV, to ensure comprehensive coverage. Stakeholders serve in an advisory capacity and provide input and recommendations, while final decisions regarding database design, content, and implementation rest with the research team. No stakeholder holds a formal governance, approval, or oversight role. Because stakeholders are engaged for distinct areas of expertise, their recommendations typically address different aspects of the project and rarely conflict; where recommendations do diverge, the research team will reconcile them considering the database's goals of comprehensiveness, feasibility, and user relevance and will document the rationale. Outreach is already underway, with structured consultations planned at key development milestones to ensure that the database is relevant, rigorous, responsive to user needs, and practice for real-world implementation.

Ethical Considerations

This study uses publicly available data aggregated at the ZIP code or higher geographic level and does not include individual-level or identifiable information. Data sources originate from federal, academic, state, or local public health agencies. Stakeholder engagement activities were intended to inform database development; no personal data were collected in these consultations, and this is not formal human subjects research. This study therefore qualifies as nonhuman subjects research under US federal regulations [39]. Institutional review is not required and therefore was not sought.

All data use will comply with applicable terms of use, licensing, and attribution requirements for each source. Should future expansions incorporate sensitive or restricted-access datasets, appropriate data-sharing agreements, governance structures, and institutional approvals will be pursued.

Results

This project was funded in August 2025. Our data-method search, combining the initial and supplemental searches described below, yielded 80 unique PubMed articles after duplicate removal. During the first-stage title and abstract screening, 16 studies were selected for full-text review, of which 7 met inclusion criteria following second-stage screening. Studies were excluded if they were reports or planned registries using data within clinical trials (Figure S1 in Multimedia Appendix 4). The proxy-variable search yielded 273 unique PubMed articles after duplicate removal. Of these, 63 studies were advanced to full-text review based on title and abstract screening, and 38 met inclusion criteria after second-stage screening. Studies were excluded if they were nonquantitative, duplicative, or relied on internal measures (Figure S2 in Multimedia Appendix 4).

As the initial data-methods search produced only 5 eligible studies, we conducted a supplemental structured literature search focused on database development methods to broaden our understanding of best practices and enhance methodological rigor. This supplemental search identified 2 additional studies, bringing the combined total reported above to 7. Detailed search strategies for all searches are provided in Multimedia Appendix 1.

Identification of the initial datasets for inclusion in the contextual database was finalized in April 2026. Data acquisition, standardization, and QA began in May 2026 and are ongoing, yielding to date 130 county-level datasets from 9 sources and 150 state-level datasets from 8 sources (Multimedia Appendix 5), with acquisition remaining iterative. Cross-source data harmonization, final QA checks, and dashboard development are expected to conclude in October 2026, with development of a public-facing domain including database access and additional tools, planned for a December 2026 release. Study results are anticipated for publication in summer 2027.

Discussion

This protocol outlines the planned development of a contextual database intended to support future research, implementation science, and public health decision-making under the EHE initiative. Beyond database construction, this work has potential implications for analytical applications, practical utility, and dissemination. These considerations collectively highlight the potential value of the database as a scalable and reproducible resource for contextual HIV research and implementation planning.

Analytic Implications

Our contextual database is designed to facilitate a wide range of analytic applications examining how multilevel social, structural, and demographic factors shape HIV outcomes, service availability, and burden across the United States. By integrating and harmonizing multisource and multilevel contextual indicators across SDOH domains, we anticipate that the database will provide a foundation for future analyses that inform epidemiological research, implementation strategies, and policy-relevant decision-making. It may be particularly useful for community-based organizations and other practitioners responsible for implementing EHE priorities, for whom analytic capacity and access to integrated contextual data may be limited. By reducing the burden of identifying, sourcing, and linking relevant datasets, the database may also support a more holistic understanding of SDOH by bringing together epidemiologic, demographic, and structural indicators that are often analyzed separately. The analytic applications described below are illustrative of the database's intended uses. The feasibility and validity of any specific analysis will depend on the completeness, temporal coverage, geographic consistency, and outcome availability of the underlying data, which vary across indicators and jurisdictions.

First, the database is intended to support descriptive and analytic epidemiological analyses, including examining spatial patterns in relation to contextual indicators, identifying clusters of high HIV burden associated with structural and community-level determinants, and characterizing temporal trends in HIV outcomes. Beyond description, analytic epidemiologic applications may include identifying determinants of HIV incidence and service usage, quantifying the contribution of structural and contextual factors to geographic variation in outcomes, and, where data completeness and longitudinal coverage permit, supporting causal inference approaches such as difference-in-differences or interrupted time series analyses where longitudinal data permit. These analyses may inform surveillance and help to set priorities for the EHE initiative and related efforts.

Second, the database may be relevant for implementation science applications, particularly those related to HIV epidemiology and service delivery. Contextual indicators spanning structural, health systems, and neighborhood domains could be used to evaluate moderators of prevention and treatment interventions, such as structural barriers, service availability, policy environments, and socioeconomic

conditions. Additionally, they may help identify geographic areas with high HIV burden, unmet needs, or service gaps. Subject to data availability, these data may enable the estimation of the incremental risk associated with specific epidemiologic and service delivery indicators across diverse contexts, providing a quantitative basis for directing resources, tailoring interventions, and informing policy. These applications align with the National Institutes of Health (NIH) priorities to end the HIV epidemic through implementation research on medical and behavioral interventions [40].

Finally, the database may support public health planning by enabling advanced analytic techniques, potentially including predictive and simulation modeling. For instance, machine learning prediction models and prioritization models could support resource allocation by accounting for higher-level health determinants, such as structural and community conditions [41], provided that data of sufficient quality, completeness, and temporal depth are available. Such applications would require careful validation before informing practice. We will also launch a public-facing website featuring interactive dashboards to help users visualize cross-domain contextual indicators, thereby supporting stakeholders in decision-making for EHE implementation. In doing so, the database aims to strengthen research and dissemination by creating reproducible, generalizable knowledge to advance EHE goals [14].

Dissemination

Dissemination of this work will occur through multiple channels, including peer-reviewed manuscripts describing the database architecture, methods, and analyses; presentations at scientific conferences and stakeholder meetings; a public-facing dashboard to support EHE-related planning and decision-making; public GitHub repositories containing analytic code, data-processing scripts, QA logs, and metadata; and comprehensive technical documentation and user guides.

Dedicated workshops and consultation sessions will also be conducted to facilitate technical proficiency in database navigation for people living with HIV, community organization staff, and the broader scientific community.

Strengths and Limitations

This project has several important strengths related to both database development and its anticipated utility. This database provides a comprehensive integration of multiyear, multilevel HIV and contextual indicators, assembled through a systematic search and data collection process. The reproducible workflow supports transparency through version control with GitHub, structured Excel logs, and metadata tracking. Robust QA procedures follow the Bray and Parkin [26] framework and are in accordance with established guidelines. Stakeholder engagement throughout all phases of database development enhances relevance, usability, and alignment with EHE priorities. The tiered architecture allows for scalability, enabling expansion to additional years, indicators, or geographies as new data become available, while supporting a range of analytic applications.

Several project-level limitations should be noted. Because the database relies on publicly available secondary data, source availability, update schedules, and completeness may vary across indicators and over time; these are tracked through systematic QA procedures, with findings recorded in the data dictionary and metadata to inform end users. Variability in data quality and completeness may require complex harmonization procedures; missing value patterns, functionally unusable data, and notable discontinuities in indicator values are systematically assessed during data cleaning and QA checks, with findings documented in the data dictionary to support transparency and informed analytic decision-making. Gaps in geographic coverage or changes in reporting practices may limit comparability across geographies, and transparent documentation of source-level characteristics is provided to help users interpret these differences accordingly. Public data constraints, such as suppressed values, may limit analytic precision and the ability to capture all implementation-relevant local contexts; these are similarly flagged in the data dictionary to inform analytic decisions. The current database focuses on EHE jurisdictions, which may limit generalizability; however, the tiered architecture is designed to support expansion to additional geographies as needs and resources allow. As with all analyses relying on aggregate data, findings reflect population-level patterns and do not capture individual-level heterogeneity, and results should be interpreted at the aggregate level.

Conclusions

As the EHE initiative crosses its midpoint, strengthening the capacity of the HIV workforce to make precise, context-sensitive decisions at the local level has become increasingly critical. Central to this effort is the development of tools to bridge the gap between available evidence and locally relevant implementation strategies. The central contribution of this work is a reproducible contextual data infrastructure protocol rather than the resulting dataset in isolation. This includes an integrated, transparent, and transferable methodology for assembling, harmonizing, documenting, and maintaining a multilevel contextual data resource. The HIV contextual database will provide a comprehensive, reproducible, and scalable resource to support multilevel public health research, implementation science, and strategic planning. By integrating multiyear and multilevel contextual indicators through systematic, transparent, and reproducible methods, the database will enhance the capacity to examine the multidimensional factors shaping HIV outcomes and burdens across the United States.

The tiered architecture, standardized procedures for obtaining, cleaning, and harmonizing data, and consistent QA checks ensure that the database can be maintained and expanded as new data become available and as emerging contextual measures become relevant. This platform aims to improve understanding of the structural and social determinants of HIV, support evidence-based decision-making, and contribute to efforts to reduce the HIV burden and advance health outcomes within the EHE initiative.

Acknowledgments

The authors wish to thank Dr Gregg Gonsalves for his critical review of the manuscript and Dr Patrick Sullivan, Kamaria Brisco, Gordon Le, Kate Nyhan, and Justin DeMayo for their valuable consultations on the methods.

Generative AI Use

No generative AI was used in the writing of this manuscript.

Funding

Research reported in this publication was supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health (NIH) under award number R24AI194905-01 (DH and JLD). This project is financed with 100% federal funding, totaling US $658,019. The funder had no involvement in the study design, data collection, analysis, interpretation, or writing of the manuscript. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.The contextual database builds on the existing Center for Methods in Implementation and Prevention Science (CMIPS) database and is hosted at Yale University; no author derives financial benefit from CMIPS or the proposed database, and the planned public dashboard will be made freely available without commercial interest or restricted access. Responsibility for maintaining, curating, and updating the database, as well as decisions regarding data access and publication, rests with the research team rather than with any external governing body; Yale University's role is limited to hosting and institutional affiliation and does not extend to control over the database's content, access, or dissemination. Several authors are affiliated with Yale University and contribute to the CMIPS database, which represents a scholarly and institutional interest rather than a competing financial interest.

Data Availability

All data generated through this project will be made publicly available upon completion of the database. The finalized, harmonized database will be released for public use via the Yale CMIPS website in alignment with open science practices. Original data sources, data-processing scripts, and documentation used for data cleaning, harmonization, and quality assurance will be archived in a public GitHub repository. Any additional materials not included in the repository may be made available from the corresponding author upon request.

Authors' Contributions

DH and JLD conceived and designed the study, building on foundational ideas developed by DS. NR led all aspects of project implementation, including database design and construction, data acquisition and harmonization, and manuscript writing. LZ assisted with data work, figure preparation, and manuscript editing. RUHR and WD provided critical feedback on the manuscript. DH and JLD supervised all phases of the project.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Search strategies.
[DOCX File (Microsoft Word File), 20 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Screening criteria for literature search.
[DOCX File (Microsoft Word File), 20 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Implementation of FAIR data principles.
[DOCX File (Microsoft Word File), 16 KB-Multimedia Appendix 3]

Multimedia Appendix 4

PRISMA 2020 flow diagrams for literature searches.
[DOCX File (Microsoft Word File), 68 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Data sources included in the database to date.
[DOCX File (Microsoft Word File), 20 KB-Multimedia Appendix 5]

References

1. EHE overview. HIV.gov. URL: https://www.hiv.gov/federal-response/ending-the-hiv-epidemic/overview [Accessed 2026-10-16]
2. Li DH, Gaines Lanzi R, Pyra M. Progress and priorities to end the HIV epidemic by 2030: combined lessons from 120 US implementation research projects. J Acquir Immune Defic Syndr. 2025;98(5S):e1-e9. [doi: 10.1097/QAI.0000000000003636] [Medline: 40163050]
3. Sullivan PS, Johnson AS, Pembleton ES, et al. Epidemiology of HIV in the USA: epidemic burden, inequities, contexts, and responses. The Lancet. Mar 2021;397(10279):1095-1106. [doi: 10.1016/S0140-6736(21)00395-0] [Medline: 33617774]
4. Armstrong WS. The human immunodeficiency virus workforce in crisis: an urgent need to build the foundation required to end the epidemic. Clin Infect Dis. May 4, 2021;72(9):1627-1630. [doi: 10.1093/cid/ciaa302] [Medline: 32211784]
5. Dam JL, Nagorka-Smith P, Waddell A, Wright A, Bos JJ, Bragge P. Research evidence use in local government-led public health interventions: a systematic review. Health Res Policy Syst. Jul 3, 2023;21(1):67. [doi: 10.1186/s12961-023-01009-2] [Medline: 37400905]
6. Whitman A, De Lew N, Chappel A, Aysola V, Zuckerman R, Sommers BD. Addressing social determinants of health: examples of successful evidence-based strategies and current federal efforts. US Department of Health and Human Services (HHS), Office of the Assistant Secretary for Planning and Evaluation (ASPE); 2022. URL: https://aspe.hhs.gov/reports/sdoh-evidence-review [Accessed 2026-10-16]
7. Rodgers M, South E, Harden M, Whitehead M, Sowden A. Contextual factors in systematic reviews: understanding public health interventions in low socioeconomic status and disadvantaged populations. Arch Public Health. Jun 13, 2025;83(1):153. [doi: 10.1186/s13690-025-01644-x] [Medline: 40514736]
8. Town M, Eke P, Zhao G, et al. Racial and ethnic differences in social determinants of health and health-related social needs among adults - behavioral risk factor surveillance system, United States, 2022. MMWR Morb Mortal Wkly Rep. Mar 7, 2024;73(9):204-208. [doi: 10.15585/mmwr.mm7309a3] [Medline: 38451870]
9. Kline N. Syndemic statuses: intersectionality and mobilizing for LGBTQ+ Latinx health equity after the Pulse shooting. Soc Sci Med. Feb 2022;295:113260. [doi: 10.1016/j.socscimed.2020.113260] [Medline: 32792206]
10. Mendenhall E, Singer M. What constitutes a syndemic? Methods, contexts, and framing from 2019. Curr Opin HIV AIDS. Jul 2020;15(4):213-217. [doi: 10.1097/COH.0000000000000628] [Medline: 32412998]
11. Hodder SL, Feinberg J, Strathdee SA, et al. The opioid crisis and HIV in the USA: deadly synergies. The Lancet. Mar 2021;397(10279):1139-1150. [doi: 10.1016/S0140-6736(21)00391-3] [Medline: 33617769]
12. Caporali A, Kluesener S, Neyer G, Krapf S, Grigorieva O, Kostova D. The Contextual Database of the Generations and Gender Programme: concept, content, and research examples. DemRes. 2016;35:229-252. [doi: 10.4054/DemRes.2016.35.9]
13. Ferrari G, Weber R. A contextual database at the district level using French census and other administrative data. French Institute for Demographic Studies (INED); 2024. URL: https://ideas.repec.org/p/idg/wpaper/1i8etzibo4xb9gqodfdt.html [Accessed 2026-10-16]
14. Humphries DL, Audet CM, Balasubramanian BA, et al. Building the road to end the HIV epidemic through coordinating and Capacity-building Hubs to Enhance the Science of HIV Implementation REsearch (CHESHIRE): reflections and directions. J Acquir Immune Defic Syndr. Apr 15, 2025;98(5S):e17-e27. [doi: 10.1097/QAI.0000000000003628] [Medline: 40163052]
15. Databases. Center for Methods in Implementation and Prevention Science (CMIPS), Yale School of Public Health. URL: https://ysph.yale.edu/cmips/research/databases/ [Accessed 2026-10-16]
16. Bronfenbrenner U. Toward an experimental ecology of human development. Am Psychol. 1977;32(7):513-531. [doi: 10.1037/0003-066X.32.7.513]
17. Eriksson M, Ghazinour M, Hammarström A. Different uses of Bronfenbrenner's ecological theory in public mental health research: what is their value for guiding public mental health policy and practice? Soc Theory Health. Nov 2018;16(4):414-433. [doi: 10.1057/s41285-018-0065-6]
18. Menza TW, Hixson LK, Lipira L, Drach L. Social determinants of health and care outcomes among people with HIV in the United States. Open Forum Infect Dis. Jul 2021;8(7):ofab330. [doi: 10.1093/ofid/ofab330] [Medline: 34307729]
19. Patel D, Clark HA, Williams WO, Taylor-Aidoo N, Wright C. CDC-funded HIV testing services outcomes and social determinants of health in Ending the HIV Epidemic in the U.S. jurisdictions. AIDS Behav. Apr 2024;28(4):1152-1165. [doi: 10.1007/s10461-023-04133-w] [Medline: 37479920]
20. First Nations perspective on health and wellness. First Nations Health Authority. 2025. URL: https://fnha.ca/wellness/spirit-and-culture/first-nations-perspective-on-health-and-wellness/ [Accessed 2026-10-16]

21. Gallagher J. Indigenous approaches to health and wellness leadership: A BC First Nations perspective. Healthc Manage Forum. Jan 2019;32(1):5-10. [doi: 10.1177/0840470418788090] [Medline: 30354687]
22. Bebbington E, Krayer A, Lea A, et al. Self-harm and Violence Presenting to Emergency Care Registry (SAVER) project: protocol for a mixed-methods study. BMJ Open Qual. Sep 14, 2025;14(3):e003463. [doi: 10.1136/bmjoq-2025-003463] [Medline: 40953885]
23. Zotero. 2023. URL: https://www.zotero.org/ [Accessed 2026-10-16]
24. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev. Dec 5, 2016;5(1):210. [doi: 10.1186/s13643-016-0384-4] [Medline: 27919275]
25. Cao Y, Yang JA, Nara A, Jankowska MM. Designing and evaluating a hierarchical framework for matching food outlets across multi-sourced geospatial datasets: a case study of San Diego county. J Urban Health. Feb 2024;101(1):155-169. [doi: 10.1007/s11524-023-00817-9] [Medline: 38167974]
26. Bray F, Parkin DM. Evaluation of data quality in the cancer registry: principles and methods. Part I: comparability, validity and timeliness. Eur J Cancer. Mar 2009;45(5):747-755. [doi: 10.1016/j.ejca.2008.11.032] [Medline: 19117750]
27. DCMI Usage Board. Dublin Core™ Metadata element set, version 1.1: reference description. Dublin Core. 2012. URL: https://www.dublincore.org/specifications/dublin-core/dces/2012-06-14/ [Accessed 2026-10-16]
28. Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data. Mar 15, 2016;3:160018. [doi: 10.1038/sdata.2016.18] [Medline: 26978244]
29. Galvez-Hernandez P, Gonzalez-Viana A, Gonzalez-de Paz L, Shankardass K, Muntaner C. Generating contextual variables from web-based data for health research: tutorial on web scraping, text mining, and spatial overlay analysis. JMIR Public Health Surveill. Jan 8, 2024;10:e50379. [doi: 10.2196/50379] [Medline: 38190245]
30. Wickham H. rvest: easily harvest (scrape) web pages. GitHub. 2025. URL: https://github.com/tidyverse/rvest [Accessed 2026-10-16]
31. Harrison J, Kim J, Völkle J. RSelenium: R bindings for Selenium Remote WebDriver. GitHub. 2023. URL: https://github.com/ropensci/rselenium [Accessed 2026-10-16]
32. Substantial changes to counties and county-equivalent entities: 1970-present. United States Census Bureau. URL: https://www.census.gov/programs-surveys/geography/technical-documentation/county-changes.html [Accessed 2026-10-16]
33. Palis BE, Janczewski LM, Browner AE, et al. The National Cancer Database conforms to the standardized framework for registry and data quality. Ann Surg Oncol. Sep 2024;31(9):5546-5559. [doi: 10.1245/s10434-024-15393-8] [Medline: 38717542]
34. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. Mar 1977;33(1):159-174. [Medline: 843571]
35. Liu X, Wang Y, Luo Z, et al. Automated cohort database system for cardiopulmonary physiotherapy: a comprehensive tool supporting research on cardiac surgery patients-framework design, development and validation. Comput Methods Programs Biomed. Aug 2025;268:108825. [doi: 10.1016/j.cmpb.2025.108825] [Medline: 40347620]
36. Zhang L, Jia W, Ji N, et al. Construction of the National Brain Tumor Registry of China for better management and more efficient use of data: a protocol. BMJ Open. Jan 2021;11(1):e040055. [doi: 10.1136/bmjopen-2020-040055] [Medline: 33455925]
37. Brownson RC, Shelton RC, Geng EH, Glasgow RE. Revisiting concepts of evidence in implementation science. Implement Sci. Apr 12, 2022;17(1):26. [doi: 10.1186/s13012-022-01201-y] [Medline: 35413917]
38. Smith S, Brand M, Harden S, et al. Development of an Australia and New Zealand lung cancer clinical quality registry: a protocol paper. BMJ Open. Aug 29, 2022;12(8):e060907. [doi: 10.1136/bmjopen-2022-060907] [Medline: 36038161]
39. 45 CFR 46. US Department of Health and Human Services, Office for Human Research Protections. 2025. URL: https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/index.html [Accessed 2026-10-16]
40. Advancing NIH's mission through a unified strategy. US Department of Health and Human Services, National Institutes of Health (NIH). 2025. URL: https://www.nih.gov/about-nih/nih-director/statements/advancing-nihs-mission-through-unified-strategy [Accessed 2026-10-16]
41. Wu H, Lu X, Wang H. The application of artificial intelligence in health care resource allocation before and during the COVID-19 pandemic: scoping review. JMIR AI. 2023;2(1):e38397. [doi: 10.2196/38397]

Abbreviations

ACS: American Community Survey
CMIPS: Center for Methods in Implementation and Prevention Science
EHE: Ending the HIV Epidemic
EST: Ecological Systems Theory
FAIR: Findability, Accessibility, Interoperability, and Reusability
FIPS: Federal Information Processing Standards

HHS: United States Department of Health and Human Services
NIH: National Institutes of Health
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
QA: quality assurance
SDOH: social determinants of health
USPS: United States Postal Service

Edited by Javad Sarvestan; peer-reviewed by Arushi Singh; submitted 07.May.2026; final revised version received 16.Sep.2026; accepted 28.Sep.2026; published 03.Nov.2026

Please cite as:
Rao N, Zheng L, Deihl W, Hernandez-Ramirez RU, Spiegelman D, Davis JL, Humphries D
A Multisource, Multilevel Contextual Database to Support Ending the HIV Epidemic Initiative: Protocol for Design, Construction, Harmonization, and Quality Assurance
JMIR Res Protoc 2026;15:e100648
URL: https://www.researchprotocols.org/2026/1/e100648
doi: 10.2196/100648

© Nikita Rao, Lingling Zheng, Wyatt Deihl, Raul U Hernandez-Ramirez, Donna Spiegelman, J Lucian Davis, Debbie Humphries. Originally published in JMIR Research Protocols (https://www.researchprotocols.org), 03.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Research Protocols, is properly cited. The complete bibliographic information, a link to the original publication on https://www.researchprotocols.org, as well as this copyright and license information must be included.