Emergence and Evolution of Big Data Analytics in HIV Research: Bibliometric Analysis of Federally Sponsored Studies 2000-2019

Chen Liang[1*], Shan Qiao[2*], Bankole Olatosi[1], Tianchu Lyu[1], Xiaoming Li[2]

[1] Department of Health Services Policy and Management, Arnold School of Public Health, University of South Carolina, Columbia, SC, USA
[2] Department of Health Promotion, Education, and Behavior, Arnold School of Public Health, University of South Carolina, Columbia, SC, USA
[*] Equal contribution

Corresponding author:
Chen Liang, PhD
Department of Health Services Policy and Management
Arnold School of Public Health
University of South Carolina
915 Greene St, Suite 347
Email: cliang@mailbox.sc.edu
Tel: +1 803-777-8139
Fax: +1 803-777-1836

Keywords: HIV; AIDS; PLWH; Electronic Health Records; big data; bibliometrics; data mining

Abstract

Background: The rapid growth of inherently complex and heterogeneous data in HIV/AIDS research underscores the importance of Big Data analytics. Recently, there have been increasing uptakes of Big Data techniques in basic, clinical, and public health fields of HIV/AIDS research. However, no studies have systematically elaborated on the evolving applications of Big Data in HIV/AIDS research. We sought to explore the emergence and evolution of Big Data analytics in HIV/AIDS-related publications that were funded by the US federal agencies.

Methods: We identified HIV/AIDS and Big Data related publications that were funded by seven federal agencies (i.e., NIH, ACF, AHRQ, CDC, HRSA, FDA, and VA) from 2000 to 2019 by integrating data from NIH ExPORTER, MEDLINE, and MeSH. Building on bibliometrics and Natural Language Processing methods, we constructed co-occurrence networks using bibliographic metadata (e.g., countries, institutes, MeSH terms, and keywords) of the retrieved publications. We then detected clusters among the networks as well as the temporal dynamics of clusters, followed by expert evaluation and clinical implications.

Results: We harnessed nearly 600 thousand publications related to HIV/AIDS, of which 19,528 publications relating to Big Data were included in bibliometric analysis. Results showed that (1) the number of Big Data publications has been increasing since 2000, (2) US institutes have been in close collaborations with China, Canada, and Germany, (3) University of California system, MD Anderson Cancer Center, and Harvard Medical School are among the most productive institutes and started using Big Data in HIV/AIDS research early, (4) Big Data research was not active in public health disciplines until 2015, (5) research topics such as genomics, HIV comorbidities, population-based studies, Electronic Health Records (EHR), social media, precision medicine, and methodologies such as machine learning, Deep Learning, radiomics, and data mining emerge quickly in recent years.

Conclusions: We identified a rapid growth in the cross-disciplinary research of HIV/AIDS and Big Data over the past two decades. Our findings demonstrated patterns and trends of prevailing research topics and Big Data applications in HIV/AIDS research and informed fast evolving research areas including HIV comorbidities, genomics, secondary analysis of EHR, social media, machine learning, and Deep Learning as future directions of HIV/AIDS research.

1 Introduction

In the United States (US), approximately 1.2 million people are living with HIV (PLWH), nearly 15% of which are undiagnosed and unaware of it, posing substantial risks of HIV transmission. Despite an estimated decline of 13% new HIV diagnoses from 2010 to 2018, the fronts of HIV/AIDS research have seen new challenges as new infections continue to surge in subpopulations and subregions [1] and there have been surging disease burdens for HIV patients and healthcare systems [2].

To overcome these challenges, a trending research topic is the integration of data from clinical care and social care. In 2018, there were 37,968 new diagnoses in the US, among which 43% were Black/African American men and 26% were Hispanic/Latino. New diagnoses have increased by 71% in Native Hawaiian/Other Pacific Islander. White people who inject drugs have increased by 51% in new diagnoses since 2014. The Southern US has 38% of the population yet account for 51% of new diagnoses [3], with a 24% higher proportion of new diagnoses in suburban and rural areas as compared to other regions in the US [4]. These findings collectively suggest that factors influencing HIV transmission among vulnerable populations are inherently complex and involved with nuances from both clinical care and social care. Therefore, a novel approach to identifying the dynamics of HIV transmission through integrating PLWH's clinical data, social care data, social determinants of health (SDOH), and health administrative data is demanded.

Another research interest is HIV-associated comorbidities, coinfections, and complications that lead to worse clinical outcomes and high health expenditure [5–10]. Relevant studies were often conducted on small and single cohort with limited data, thus the findings were scattered [11]. Until recently, the detection of HIV-associated comorbidities, coinfections and complications, and providing evidence-based clinical decision support (CDS) through multi-centered and longitudinal medical records were becoming increasingly available as advances in EHR interoperability unleashed the potential of secondary use of large-scale clinical data. Nevertheless, there are untapped questions and a systemic understanding of the secondary use of EHR for HIV research remains absent. For example, statistical modeling on large-scale EHR has resulted in salient clusters of diagnoses, chronic conditions, and clinical outcomes for PLWH [5,9,10]. These findings can provide patient-level evidence to guide clinical treatment. When it comes to population-level evidence, however, large-scale EHR integration, data mining methods, and CDS theories tailored for HIV patients are critical but remain understudied.

Identifying socio-behavioral factors of PLWH appears to be interesting as well. The proliferation of social media, personal health records (PHR), and other forms of patient-generated health data (PGHD) in research has expanded our understanding that perceptions and risk behaviors are associated with worse clinical outcomes and HIV transmission. For example, stigma, mental status, social support, and SDOH are associated with HIV transmission, care continuum, and disease management [12–14], yet these data sources exhibit unsolved data quality and interoperability issues [15]. Additionally, recent studies found substance use disorder (SUD) to be associated with worse clinical outcomes and increased risk behaviors among PLWH, yet the

SDOH characteristics of risk populations and effective intervention strategies remain unknown [11,16,17]. To address these gaps, a research agenda is needed to harness socio-behavioral data and integrate them with existing EHR.

Over the last two decades, these emerging research topics continued to evolve and jointly call for a combined approach to leveraging Big Data analytics for integrating, analyzing, and understanding heterogeneous health data for HIV/AIDS research. Such an effort is consistent with the National Institutes of Health (NIH) initiative of Big Data to Knowledge (BD2K) funded in 2013 [18]. Despite an increasing number of publications and funded projects from US federal agencies [19–21], no existing studies have systematically reported how Big Data analytics has reshaped the landscape of HIV/AIDS research in terms of research grant support, collaborations, research outcomes, and trending research topics. We sought to address these questions by employing the bibliometric method to analyze publications sponsored by research grants from US federal agencies that focus on Big Data in HIV/AIDS research. The bibliographic analysis was rooted in information science and has been used for detecting networks, patterns, and trends among intricate and large-scale bibliographic metadata (e.g., titles, abstracts, authors, institutes, keywords, MeSH terms) [22]. Findings of the study may retrospectively reveal interesting research patterns and predictively inform trending and future research directions in Big-Data-driven HIV/AIDS research.

2 Materials and methods

2.1 Data extraction

2.1.1 Federally funded HIV/AIDS projects

We identified HIV/AIDS-related projects in NIH ExPORTER. The database consists of administrative data for research projects funded by seven federal agencies: NIH, the Administration for Children and Families (ACF), the Agency for Healthcare Research and Quality (AHRQ), Centers for Disease Control and Prevention (CDC), Health Resources and Services Administration (HRSA), U.S. Food and Drug Administration (FDA), and Veterans Affairs (VA).

We requested data from 2000 to 2019. To identify HIV/AIDS-related projects, we used regular expression (“HIV” OR “AIDS” OR “HIV/AIDS” OR “HIV-1” OR “human immunodeficiency virus” OR “acquired immunodeficiency syndrome”) to be applied in the fields of “abstract_text”, “project_terms”, “project_title”, and “study_section_name”. These four fields were selected from a total of 46 fields in ExPORTER. The regular expression was determined by an HIV expert (SQ) and a medical informatics expert (CL). The “project_IDs” was used as a unique identifier. Fields “project_terms” and “study_section_name” contain controlled text that was determined by NIH; “abstract_text” and “project_title” contain free text that was provided by the recipients of funded projects. Therefore, we performed lemmatization and punctuation removal on free-text fields prior to the regular expression. The data processing and extraction were performed using in-house Python programing codes.

2.1.2 Publications involved with Big Data

We extracted publications that were resulted from these funded HIV/AIDS projects by querying about "PMIDs" that were linked to the "project_IDs" we have identified. With these "PMIDs", we batch downloaded MEDLINE metadata by interacting with Entrez Programming Utilities (E-utilities) at the National Center for Biotechnology Information (NCBI).

To query Big-Data-related PMIDs, two domain experts (CL and SQ) selected a collection of Medical Subject Headings (MeSH) terms that represent themes of Big Data. See Table 1. These MeSH terms were used in Perl queries for E-utilities. The pseudo-query is as follows:

[HIV/AIDS-related PMIDs] AND ("big data" OR "data mining" OR "data science" OR "machine learning" OR "deep learning" OR "natural language processing" OR "information storage and retrieval" OR "data warehousing" OR "knowledge discovery" OR "artificial intelligence" OR "algorithms" OR "cloud computing")

The query was applied on the [TW] (text words), a field in MEDLINE metadata that includes free text in the title, abstract, other abstract, MeSH terms, MeSH Subheadings, Publication Types, Substance Names, Personal Name as Subject, Corporate Author, Secondary Source, Comment/Correction Notes, and Other Terms ([OT] field) typically non-MeSH subject terms (e.g., keywords).

Upon the completion of the query, we received bibliographic information of 19,528 publications in MEDLINE format. This total excludes less than 5% of data loss that is due to erroneously reported or downloaded PMIDs. Figure 1 shows the diagram of data extraction procedures.

Table 1. Big-Data-related MeSH terms.

MeSH terms	MeSH IDs
big data	D000077558
data mining	D057225
data science	D000077488
machine learning	D000069550
deep learning	D000077321
natural language processing	D009323
information storage and retrieval	D016247
data warehousing	D000073458
knowledge discovery	D063369
artificial intelligence	D001185
algorithms	D000465
cloud computing	D000067917

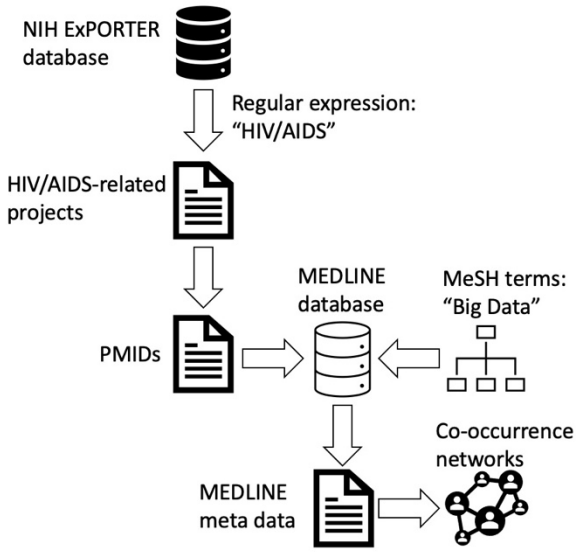

Figure 1. Data extraction flowchart.

2.2 Data analysis

We constructed co-occurrence networks of the bibliographic data using the following data elements available in MEDLINE metadata: countries, institutes, MeSH terms, and keywords. These co-occurrence networks exhibit frequency occurrences of the involved data elements over time (2000-2019). For example, we can explore international collaborations by assessing the co-occurrences of countries. Aside from the co-occurrence networks, we also generated clusters for MeSH terms and keywords, which were used to detect salient research topics and trends over time. Specifically, we employed the Latent Semantic Analysis (LSA) [23] and log-likelihood ratio (LLR) for analyzing the neighboring MeSH terms and keywords to be included in a cluster. Co-occurrence networks, clustering, and visualization were completed on CiteSpace [24].

3 Results

The number of relevant publications has constantly increased from 818 as of 2000 to 2,450 as of 2016 (Figure 2). There is a decline in 2017 (n=2,431), 2018 (n=2,176), and 2019 (n=1,586), which is likely because of the budget cut for US federal funding in HIV/AIDS research since FY 2017 as well as the ongoing integration of newly reported data in ExPORTER.

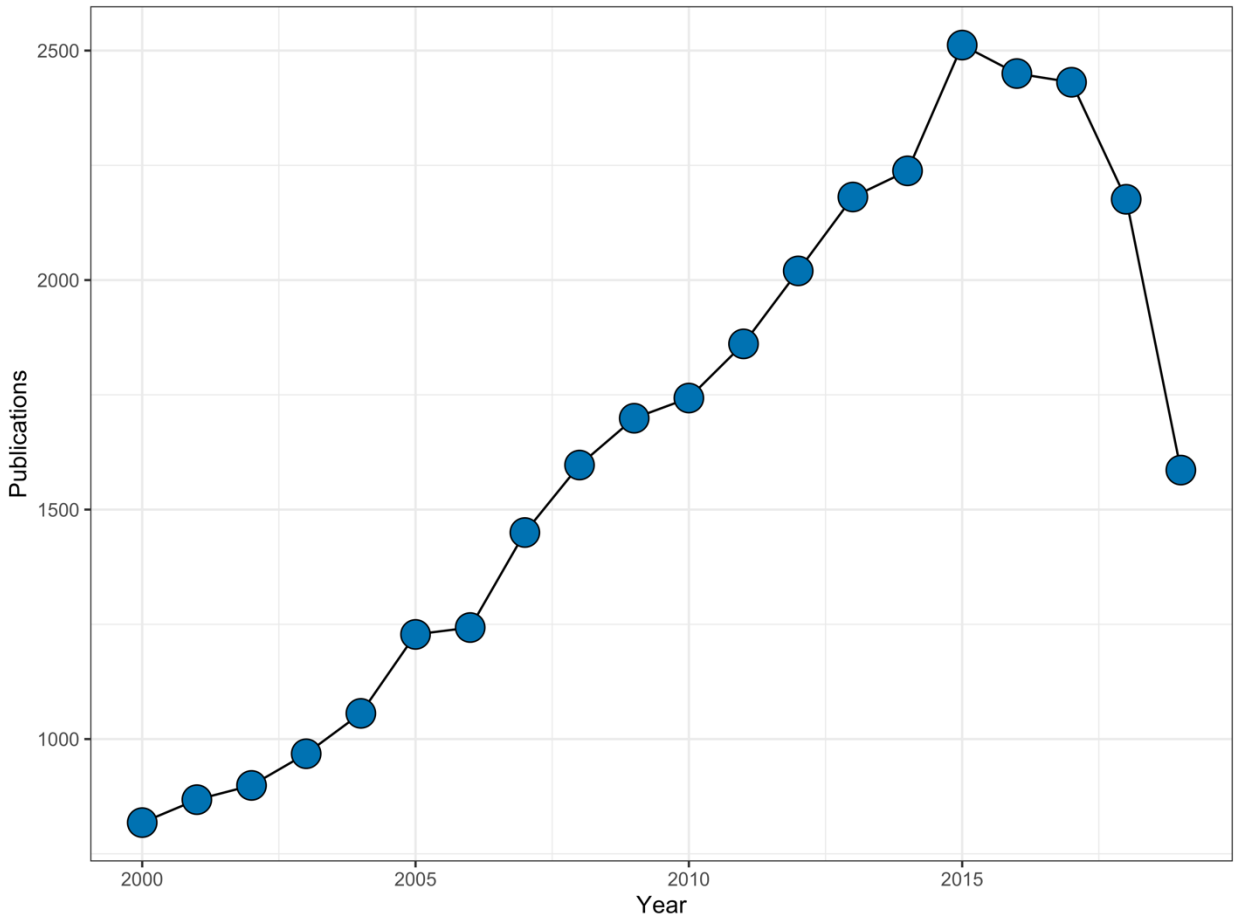

Figure 2. Number of publications by year.

3.1 International collaborations

There were 97% (n=18,937) of the publications that were published with at least one US institute. US institutes often collaborate with China (n=462, 2.4%), Canada (n=412, 2.1%), and Germany (n=326, 1.7%). See Table 2. The US began to apply Big Data analytics in HIV/AIDS research as early as 2000. China, Canada, and France initiated Big Data in HIV/AIDS research earlier than other non-US countries.

Table 2. Number of publications by countries.

Countries	No. of publications	Percentage
US	18,937	97.0%
China	462	2.4%
Canada	412	2.1%
Germany	326	1.7%
France	256	1.3%
Australia	190	1.0%

UK	177	0.9%
Switzerland	148	0.8%
Spain	142	0.7%
Italy	133	0.7%
Total	19,528	100%

3.2 Participation and productivity of institutes

As shown in Table 3, the University of California system (n=309) and Harvard University (n≥284) were among the most productive institutes. Note that research centers, schools, and research institutes affiliated with a university could be counted apart from their parent institutes in MEDLINE data format because authors do not constantly include their parent institutes in publications.

Table 3. Number of publications by institutes.

Institutes	No. of publications
University of California	309
MD Anderson Cancer Center	138
*Harvard Medical School	134
*Athinoula A. Martinos Center for Biomedical Imaging	103
University of Pennsylvania	77
University at Albany	53
Jude Children's Research Hospital	53
*Harvard T.H. Chan School of Public Health	47
Stanford Center for Biomedical Informatics Research	45
Emory University	44
Fred Hutchinson Cancer Research Center	43
Duke University Medical Center	30
Total	19,528
* Same parent institute	

We developed a time zone figure to visualize the participation and productivity of institutes over time (Figure 3). This figure shows institutes that were not only productive in Big-Data-related HIV/AIDS research but also entered the field early, such as the University of California system, MD Anderson Cancer Center, and Harvard Medical School. Several institutes were not at advantage in terms of the number of publications but started exploring the relevant research as early as between 2000 to 2008, including Johns Hopkins University School of Medicine, Drew University, Fred Hutchinson Cancer Center, Jude Children's Research Hospital, Massachusetts General Hospital, etc. Despite a fairly large number of HIV/AIDS research coming from the field of public health, the only public health institute that was active in Big-Data-related research was Harvard T.H. Chan School of Public Health as dated in 2015 (Figure 4), suggesting that in public health Big Data was introduced to HIV/AIDS research comparatively late.

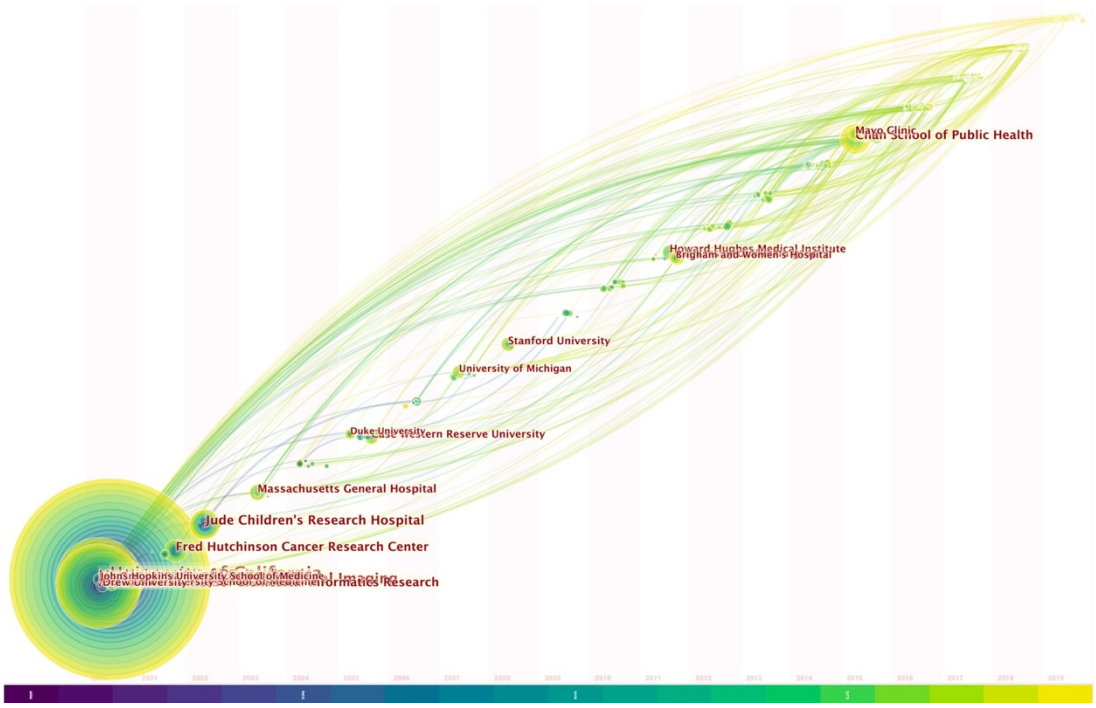

Figure 3. Time zone figure. The size of nodes indicates the frequency of relevant publications by institutes. The colors indicate the years by which an institute started to produce significant amounts of relevant publications, spanning 2000 (dark) to 2019 (light). The edges indicate the collaborations between institutes. For clarity of the figure, the thickness of the edges is unified and does not represent frequency differences. The time zone where an institute displays indicates the year the institute began to produce relevant publications.

3.3 Areas of study

Figure 4 demonstrated the top 100 bursts of MeSH terms and keywords, in which bursts are MeSH terms and keywords that rise sharply in frequency in the co-occurrence networks [25]. These bursts carried critical implications in terms of emerging research topics and methods in publications. In addition to MeSH terms and keywords that have been used in high-frequency through-out 2000-2019, we found several MeSH terms and keywords that did not emerge until recent years. To ease the description, we organized them into the following categories. (1) Machine learning methods: “deep learning”, “convolutional neural network”, “feature selection”, “random forest”, “clustering”, “LASSO”, “sensitivity”, “specificity”, “Markov Chain Monte Carlo”, “cross validation”, “support vector machine”, etc. Arguably, these methods and techniques have been increasingly used for automated classification and clustering of health data. (2) Emerging health data related terms: “data sharing”, “electronic medical record”, “missing data”, “imputation”, “social media”, etc. These terms provide insights on the growing number of electronic health data and multi-modal health data that have been used for Big Data analysis. (3) Biomedical techniques and related topics: “precision medicine”, “rna-seq” (RNA sequencing), “radiomics”, “parallel imaging”, “diffusion MRI”, “fMRI”, “genomics”, “tractography”, “network analysis”, “imaging analysis”, “next-generation sequencing”,

“compressed sensing”, “flow cytometry”, “signal processing”, “imaging processing”, “mass spectrometry”, etc. (4) HIV-associated comorbidities, coinfections, and complications: “glaucoma”, “cancer”, etc.

Subject Categories	Year	Strength	Begin	End	2000 - 2019
natural language processing	2000	**45.83**	2000	2019	
deep learning	2000	**24.64**	2017	2019	
information extraction	2000	**14.72**	2000	2019	
temporal reasoning	2000	**12.43**	2000	2019	
convolutional neural network	2000	**11.37**	2017	2019	
word sense disambiguation	2000	**11.14**	2000	2019	
relation extraction	2000	**11.14**	2000	2019	
patient classification	2000	**11.14**	2000	2019	
medical language processing	2000	**10.6**	2000	2019	
ontology	2000	**10.59**	2000	2019	
multi-label classification	2000	**10.16**	2000	2019	
clinical language processing	2000	**10.09**	2000	2019	
ecg	2000	**9.8**	2000	2019	
microbial ecology	2000	**9.54**	2000	2019	
adverse reaction	2000	**9.54**	2000	2019	
motion artifact	2000	**9.27**	2000	2019	
prognostic model	2000	**9.27**	2000	2019	
precision medicine	2000	**8.35**	2016	2019	
microbiome	2000	**8.17**	2000	2019	
data visualization	2000	**7.93**	2000	2019	
data mining	2000	**7.73**	2000	2019	
cochlear implant	2000	**7.12**	2000	2019	
rna-seq	2000	**6.6**	2017	2019	
feature selection	2000	**6.55**	2014	2019	
random forest	2000	**6.25**	2014	2019	
data sharing	2000	**6.09**	2016	2019	
personalized medicine	2000	**5.79**	2000	2019	
16s rrna gene sequencing	2000	**5.55**	2000	2019	
analytics	2000	**5.55**	2000	2019	
adolescence	2000	**5.55**	2000	2019	
3d iterative reconstruction	2000	**5.55**	2000	2019	
allergy	2000	**5.55**	2000	2019	
ammonia	2000	**5.55**	2000	2019	
radiomics	2000	**5.54**	2016	2019	
bayesian analysis	2000	**5.48**	2000	2019	
automatic lymphoma classification	2000	**5.44**	2000	2019	
annotation guideline	2000	**5.44**	2000	2019	
azoxymethane	2000	**5.44**	2000	2019	
argininosuccinate lyase	2000	**5.44**	2000	2019	
bayes classifier	2000	**5.44**	2000	2019	
asthma epidemiology	2000	**5.44**	2000	2019	
bandwidth effect	2000	**5.44**	2000	2019	
auc	2000	**5.34**	2000	2019	
parallel imaging	2000	**5.3**	2014	2019	
binary tree	2000	**5.23**	2000	2019	
biomedical cloud	2000	**5.23**	2000	2019	
bifactor model	2000	**5.23**	2000	2019	
bioluminescent imaging	2000	**5.23**	2000	2019	
bimodal hearing	2000	**5.23**	2000	2019	
diffusion mri	2000	**5.12**	2015	2019	

Subject Categories	Year	Strength	Begin	End	2000 - 2019
16s rrna gene	2000	**5.02**	2000	2019	
prediction	2000	**4.97**	2014	2019	
brain	2000	**4.92**	2014	2019	
diagnosis	2000	**4.92**	2014	2019	
binary classification	2000	**4.79**	2000	2019	
fmri	2000	**4.77**	2012	2019	
genomics	2000	**4.61**	2015	2019	
animal model	2000	**4.23**	2000	2019	
bayesian method	2000	**4.14**	2000	2019	
glaucoma	2000	**4.12**	2013	2019	
annotation	2000	**4.12**	2013	2019	
tractography	2000	**4.1**	2014	2019	
sensitivity	2000	**4.1**	2013	2019	
biomarker	2000	**4.09**	2014	2019	
gwa	2000	**4.09**	2015	2019	
network analysis	2000	**4.03**	2014	2019	
electronic medical record	2000	**4.02**	2016	2019	
africa	2000	**3.95**	2000	2019	
biomedical image processing	2000	**3.9**	2000	2019	
missing data	2000	**3.88**	2015	2019	
biomedical informatics	2000	**3.81**	2000	2019	
schizophrenia	2000	**3.77**	2014	2019	
image analysis	2000	**3.74**	2011	2019	
bayesian network	2000	**3.7**	2000	2019	
clustering	2000	**3.7**	2016	2019	
em algorithm	2000	**3.7**	2016	2019	
lasso	2000	**3.66**	2015	2019	
surveillance	2000	**3.57**	2016	2019	
next generation sequencing	2000	**3.47**	2017	2019	
compressed sensing	2000	**3.42**	2013	2019	
specificity	2000	**3.41**	2013	2019	
signal processing	2000	**3.41**	2011	2019	
image processing	2000	**3.4**	2014	2019	
flow cytometry	2000	**3.39**	2016	2019	
human	2000	**3.39**	2016	2019	
metabolomics	2000	**3.36**	2016	2019	
next-generation sequencing	2000	**3.36**	2016	2019	
markov chain monte carlo	2000	**3.33**	2010	2019	
mass spectrometry	2000	**3.31**	2015	2019	
social media	2000	**3.29**	2017	2019	
imputation	2000	**3.23**	2015	2019	
retina	2000	**3.23**	2017	2019	
multi-atlas segmentation	2000	**3.2**	2015	2019	
pediatric	2000	**3.2**	2015	2019	
proteomics	2000	**3.13**	2015	2019	
cancer	2000	**3.11**	2017	2019	
cross-validation	2000	**3.11**	2007	2019	
head and neck cancer	2000	**3.08**	2016	2019	
systems biology	2000	**3.07**	2015	2019	
support vector machine	2000	**3.06**	2013	2019	

Figure 4. Top 100 MeSH term and keyword bursts. The red bar indicates the course of bursts.

On top of the co-occurrence network of MeSH terms and keywords, we generated clusters of MeSH terms and keywords. Table 4 shows the valid clusters, surrounding terms included in the clusters, and cluster parameters. The validity of the clusters was measured by Silhouette scores, in which a high score indicates a high consistency of terms within a cluster [26]. These clusters were generated by LLR. The LAS algorithm failed to produce robust results. Figure 5 visualizes these clusters with colored nodes of MeSH terms and keywords. Note that the labeled names of clusters in Figure 5 are the top log-likelihood-ratio terms in a cluster and do not necessarily represent the meaning of a cluster. The meaning of a cluster is interpreted by domain experts

(CL, SQ) in review of the included terms. To the best of our understanding, cluster #7 is related to eye disorder, comorbidity with AIDS. Cluster #6 is related to hearing loss, another comorbidity with AIDS. Cluster #5 is related to clinical natural language processing (NLP) that is focused on integrating and mining clinical notes within the electronic health records (EHR). Cluster #4 is related to the methods of structural biology. Cluster #3 is related to breast cancer, one of the comorbid cancers with AIDS. Cluster #2 is related to genomics, a field of biology focusing on genomes. Cluster #1 is comparatively less coherent as it involves medical imaging, clinical NLP, and comorbidity. Cluster #0 is related to diagnostics and clinical care of AIDS. Note that HIV-associated comorbidities generally appear across most of the clusters.

Table 4. Clusters generated based on the co-occurrence network of MeSH terms and keywords.

Clusters	Size	Silhouette	Year (mean)	Top terms (LLR: log-likelihood ratio)
0	276	0.71	2007	Risk factors, diagnostic code, HIV incidence, racial disparities, care population, acute HIV infection
1	230	0.71	2007	Human brain, clinical text, optical properties, computer tomography, malignant breast tumor
2	177	0.79	2008	Genome-wide association studies, sequencing data, exome sequencing, genetic variant, humanexome beadchip
3	162	0.72	2008	Breast cancer, large b-cell lymphoma, breast MRI, clinical trial, diffusion MRI, epidermal growth factor receptor
4	159	0.80	2006	Tandem mass spectra, efficient recognition, low sequence identity, conservative application, cryo-electron microscopy
5	129	0.82	2006	Clinical text, clinical narrative, temporal expression, clinical document, coreference resolution, i2b2 challenge
6	129	0.82	2006	Local field potential, moderate hearing loss, working memory, neuronal oscillator, Bayesian approach, hearing loss
7	12	0.98	2012	Optical coherence tomography, correspondent cluster, visual field measurement, detecting progression, recognizing glaucomatous defect patterns

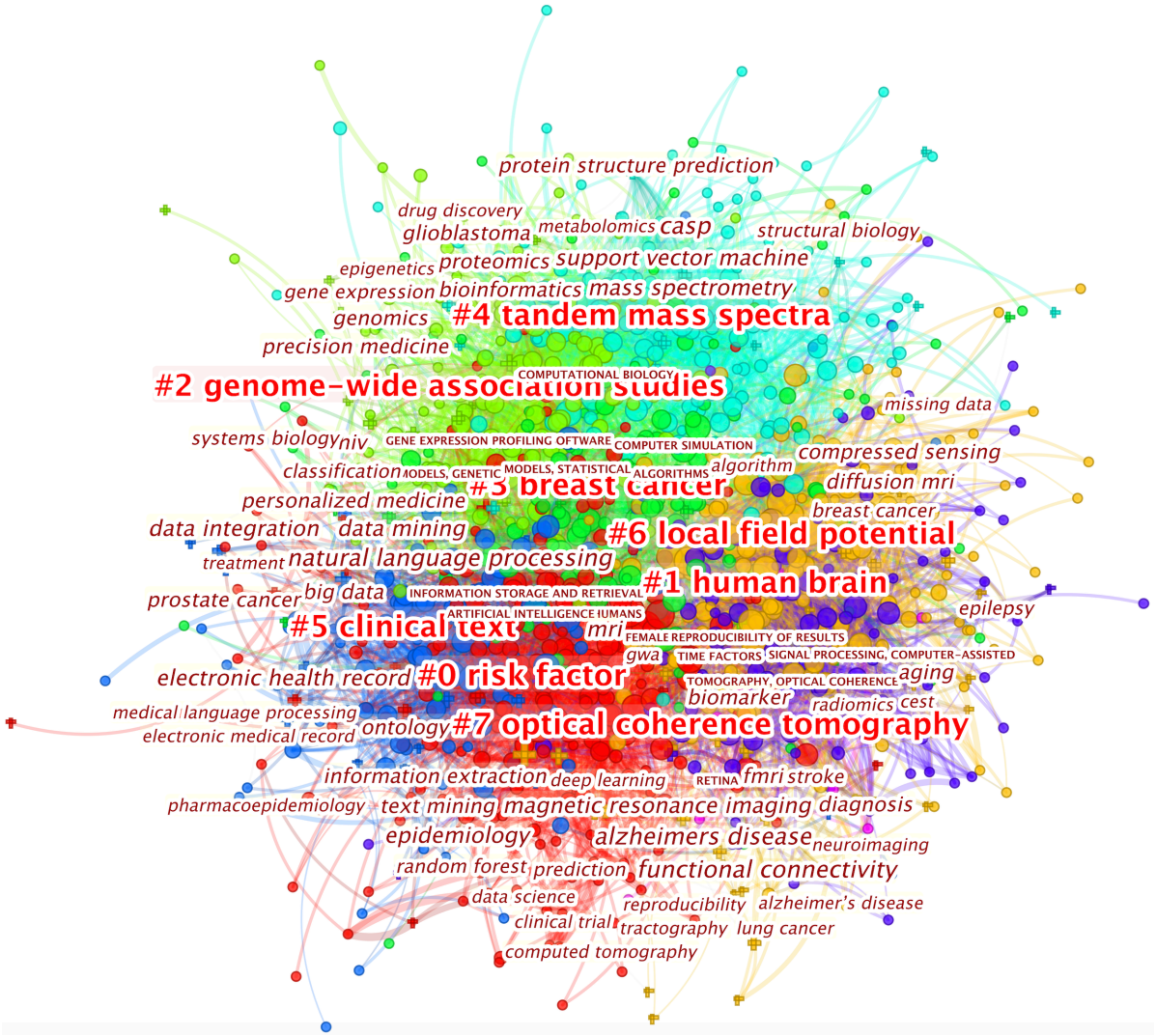

Figure 5. Visualization of clusters. Clusters are color labeled and annotated by a name (red) that refers to the term of top log-likelihood-ratio in a cluster. Other terms within the clusters are in maroon. The clusters are ranked by the Silhouette scores, in which a high value indicates well-coherent terms to its own cluster.

To visualize the temporal dynamics of the detected clusters and the surrounding terms, we projected the clusters onto a timeline view (Figure 6). This visualization demonstrated a number of findings that were not revealed in other analyses, e.g., terms within a cluster appearing at different times. Notably, machine learning algorithms, Deep Learning, and CDS raised rapidly in conjunction with the emergence of EHR-related and patient-data-related topics. Many of such combinations were related to HIV-associated comorbidities, coinfections, and complications. Among the eight clusters, at least three (cluster #7, #6, and #3) were directly related to HIV comorbidities, and cluster #7 was the only one that disappeared since 2017.

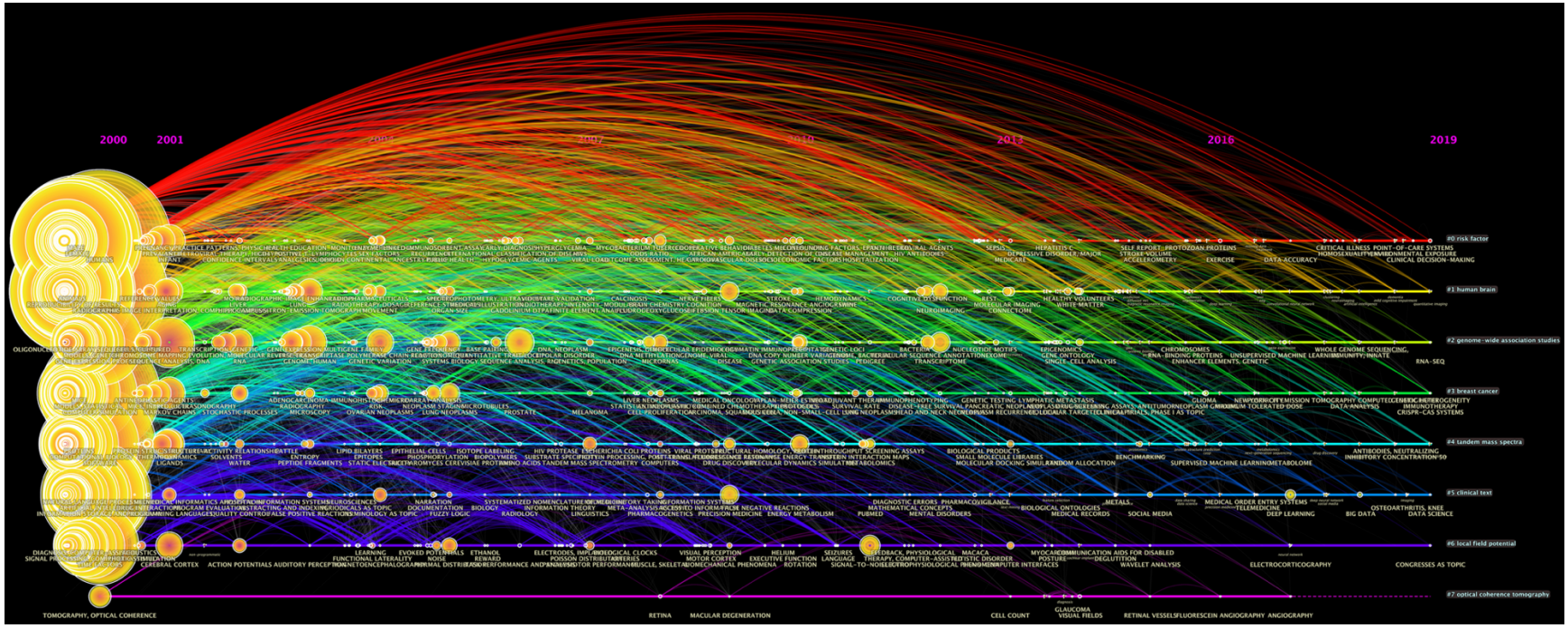
2000
2001
2007
2016
2019
#0 risk factor
#1 human brain
#2 genome-wide association studies
#3 breast cancer
#4 tandem mass spectra
#5 clinical text
#6 local field potential
#7 optical coherence tomography
TOMOGRAPHY, OPTICAL COHERENCE
RETINA
MACULAR DEGENERATION
CELL COUNT
GLAUCOMA
VISUAL FIELDS
RETINAL VESSELS
FLUORESCEIN ANGIOGRAPHY
ANGIOGRAPHY

Figure 6. Timeline view of clusters.

4 Discussion

WE extracted Big Data publications in HIV/AIDS research grants that were funded by the US federal agencies from 2000 to 2019. We demonstrated the emergence and evolution of diverse research topics, methods, data sources, and research collaborations in these publications. Our study captured a consistent upsurge in the number of both US federal funding support and the resulted publications for Big-Data-related HIV/AIDS research over time, suggesting an increasingly important role of Big Data in HIV/AIDS research. The exception is the decline from 2017 to 2019 that is likely related to the budget cut on federal funding in HIV/AIDS research. This trend is consistent with the US federal funding for HIV/AIDS research measured by total cost in the same period of time [27].

Analysis of international collaborations showed that these US-funded projects have fostered close and productive collaborations with countries such as China, Canada, Germany, and France, measured by the number of co-authored publications. These international collaborations started as early as 2000. With regard to institutional productivity, the University of California system, MD Anderson Cancer Center, Harvard University, University of Pennsylvania were among the most productive institutes. For temporal dynamics, we have three interesting findings. First, institutes (e.g., the University of California system, MD Anderson Cancer Center, Harvard Medical School) that applied Big Data in HIV/AIDS research as early as 2000 have been highly productive over the last two decades and have established wide collaborations with other institutes early. Some institutes (e.g., Johns Hopkins University School of Medicine, Drew University, Fred Hutchinson Cancer Center, Jude Children's Research Hospital, Massachusetts General Hospital) that initiated Big-Data-related HIV/AIDS research as early as around 2000 were comparatively less productive probably because of the diverse research strategies and relatively small setting as compared to large academic settings such as the University of California system. Second, most of the high-profile institutes were centered on biomedical and clinical research. Third, public health communities have been making substantial contributions to HIV/AIDS research over time but were not active in Big Data until a much later time marked by the emergence of Harvard T.H. Chan School of Public Health in 2015.

The emergence of Big Data in HIV/AIDS research was as early as 2000 but the research foci and context have changed substantially over time. The application of Big Data was found in the following cross-cutting themes: (1) HIV-associated comorbidities, coinfections, and complications. Big Data methods adopted in this theme varied by specific research topics. For example, Markov Chain and Stochastic processes were widely used in modeling cancer progression (e.g., breast cancer) as early as 2001. Survival analysis, genetic testing, and bioinformatics were introduced a few years later. In recent years, an increasing number of studies within the theme were found in conjunction with the secondary use of EHR. (2) System biology including genomics and bioinformatics. This theme focused on fundamental research

questions such as HIV/AIDS mechanism, pathogenesis, and antiretroviral therapy (ART) drug towards HIV cure and next-generation therapies. (3) Secondary use of EHR and analysis of PGHD. This theme widely interacted with HIV-associated comorbidities, coinfections, and complications in terms of modeling clinical characteristics based on large-scale EHR. It also interacted with epidemiological, behavioral, and social aspects of HIV/AIDS research. Although this theme is relatively small by number of publications, it grew rapidly in recent years. Additionally, there are other cross-cutting topics including pharmacy, therapies, medical imaging, etc.

Our findings provided novel implications to the state of HIV/AIDS research as well as research opportunities. Emerging opportunities in HIV/AIDS research are increasingly available in Big Data. One major source of Big Data is genomics data that underpin the discovery of genome functions and the genetic basis of AIDS. Since the completion of the Human Genome Project in 2003, next-generation sequencing data have become exponentially available in NCBI and have been used in HIV/AIDS research with a focus on HIV mechanism and pathogenesis [28]. The other major source of Big Data is EHR which was incentivized by the Health Information Technology for Economic and Clinical Health Act (HITECH) enacted in 2009 and the Medicare and Medicaid EHR Incentive Programs initiated in 2011 for enhancing EHR interoperability. A unique role of EHR lies in the capacity of supplying compressive patient data across the continuum of care, which is highly responsive to the continuum of AIDS care [29].

Opportunities and challenges are not apart. Health data exist in a wide variety of formats, with uneven quality and potential bias, which jointly complicate health information exchange and knowledge discovery. This challenge is critical for HIV/AIDS research because factors contributing to HIV transmission and medical care adherence are complicated by PLWH's health data from multiple sources spanning clinical, social care, and administrative data. However, efforts towards integrated data have been stalled by the lack of comprehensive data repositories [20]. The other challenge is the lack of tailored Big Data analytics. For example, there are limited but highly demanded studies in EHR modeling, particularly, modeling temporal clinical events [30,31], data extraction from clinical notes [31], and imputation for missing or sparse data [32]. Additionally, our findings suggested increased utilization of PGHD and social media in research, yet data normalization and mitigation strategies remained demanded to ensure acceptable data quality [15].

This study is subject to limitations. First, the search strategy for identifying relevant projects from ExPORTER favors minimizing miss-detected relevant projects and, as a result, may tolerate a small number of projects that include but do not primarily focus on HIV/AIDS. Second, our bibliographic data did not include citations because the data were from MEDLINE. Third, our method is limited in capturing meaningful clusters with an infrequent number of publications because bibliographic analyses heavily rely on co-occurrence probabilities. Despite limitations, this study is the first that systematically analyzed the emergence and evolution of Big Data in HIV/AIDS research, which provided critical implications on US federal research priorities, trending research topics, future directions, and challenges.

References

1. Gant Z, Johnson SD, Li J, et al. Diagnoses of HIV infection in the United States and dependent areas (updated), 2018. Published online 2020.
2. Ford N, Shubber Z, Meintjes G, et al. Causes of hospital admission among people living with HIV worldwide: a systematic review and meta-analysis. *lancet HIV*. 2015;2(10):e438--e444.
3. Linley L, Johnson AS, Song R, et al. Estimated HIV incidence and prevalence in the United States 2010--2016. Published online 2019.
4. Centers for Disease Control and Prevention. *HIV Surveillance in Urban and Nonurban Areas through 2017*.; 2017.
5. Yang X, Zhang J, Chen S, Weissman S, Olatosi B, Li X. Comorbidity patterns among people living with HIV: a hierarchical clustering approach through integrated electronic health records data in South Carolina. *AIDS Care*. Published online 2020:1-13.
6. Kozak MS, Mugavero MJ, Ye J, et al. Patient reported outcomes in routine care: advancing data capture for HIV cohort research. *Clin Infect Dis*. 2012;54(1):141-147.
7. Feller DJ, Zucker J, Yin MT, Gordon P, Elhadad N. Using clinical notes and natural language processing for automated HIV risk assessment. *J Acquir Immune Defic Syndr*. 2018;77(2):160.
8. Kim DJ, Westfall AO, Chamot E, et al. Multimorbidity patterns in HIV-infected patients: the role of obesity in chronic disease clustering. *J Acquir Immune Defic Syndr*. 2012;61(5):600.
9. McInnes DK, Shimada SL, Midboe AM, et al. Patient use of electronic prescription refill and secure messaging and its association with undetectable HIV viral load: a retrospective cohort study. *J Med Internet Res*. 2017;19(2):e34.
10. Goulet JL, Fultz SL, Rimland D, et al. Do patterns of comorbidity vary by HIV status, age, and HIV severity? *Clin Infect Dis*. 2007;45(12):1593-1601.
11. Altice FL, Kamarulzaman A, Soriano V V, Schechter M, Friedland GH. Treatment of medical, psychiatric, and substance-use comorbidities in people infected with HIV who use drugs. *Lancet*. 2010;376(9738):367-387.
12. Zhang C, Li X, Liu Y, et al. Stigma against people living with HIV/AIDS in China: does the route of infection matter? *PLoS One*. 2016;11(3):e0151078.
13. Zeng C, Li X, Qiao S, Yang X, Shen Z, Zhou Y. Anticipated stigma and medication adherence among people living with HIV: the mechanistic roles of medication support and ART self-efficacy. *AIDS Care*. Published online 2020:1-9.
14. Lowther K, Selman L, Harding R, Higginson IJ. Experience of persistent psychological symptoms and perceived stigma among people with HIV on antiretroviral therapy (ART): a systematic review. *Int J Nurs Stud*. 2014;51(8):1171-1189.
15. Fernandez-Luque L, Staccini P. All That Glitters Is not Gold: Consumer Health Informatics and Education in the Era of Social Media and Health Apps: Findings from the Yearbook 2016 Section on Consumer Health Informatics. *Yearb Med Inform*. 2016;(1):188.
16. Oldfield BJ, Muñoz N, McGovern MP, et al. Integration of care for HIV and opioid use disorder. *Aids*. 2019;33(5):873-884.
17. Cernasev A, Veve MP, Cory TJ, et al. Opioid Use Disorders in People Living with HIV/AIDS: A Review of Implications for Patient Outcomes, Drug Interactions, and Neurocognitive

Disorders. *Pharmacy*. 2020;8(3):168.
18. Bourne PE, Bonazzi V, Dunn M, et al. The NIH big data to knowledge (BD2K) initiative. *J Am Med Informatics Assoc*. 2015;22(6):1114.
19. Bushman FD, Barton S, Bailey A, et al. Bringing it all together: big data and HIV research. *AIDS*. 2013;27(5):835.
20. Olatosi B, Zhang J, Weissman S, Hu J, Haider MR, Li X. Using big data analytics to improve HIV medical care utilisation in South Carolina: A study protocol. *BMJ Open*. 2019;9(7):e027688.
21. Oliwa T, Furner B, Schmitt J, Schneider J, Ridgway JP. Development of a predictive model for retention in HIV care using natural language processing of clinical notes. *J Am Med Informatics Assoc*. Published online 2020.
22. Börner K, Chen C, Boyack KW. Visualizing knowledge domains. *Annu Rev Inf Sci Technol*. 2003;37(1):179-255.
23. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by latent semantic analysis. *J Am Soc Inf Sci*. 1990;41(6):391-407.
24. Chen C. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *J Am Soc Inf Sci Technol*. 2006;57(3):359-377.
25. Kleinberg J. Bursty and hierarchical structure in streams. *Data Min Knowl Discov*. 2003;7(4):373-397.
26. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math*. 1987;20:53-65.
27. Kaiser Family Foundation. U.S. Federal Funding for HIV/AIDS: Trends Over Time.
28. Collins FS, McKusick VA. Implications of the Human Genome Project for medical science. *Jama*. 2001;285(5):540-544.
29. Mugavero MJ, Amico KR, Horn T, Thompson MA. The state of engagement in HIV care in the United States: from cascade to continuum to control. *Clin Infect Dis*. 2013;57(8):1164-1171.
30. Tao C, Wongsuphasawat K, Clark K, Plaisant C, Shneiderman B, Chute CG. Towards event sequence representation, reasoning and visualization for EHR data. In: *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*. ; 2012:801-806.
31. Wang Y, Wang L, Rastegar-Mojarad M, et al. Clinical information extraction applications: a literature review. *J Biomed Inform*. 2018;77:34-49.
32. Goldstein BA, Navar AM, Pencina MJ, Ioannidis J. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J Am Med Informatics Assoc*. 2017;24(1):198-208.