BMJ Open

Using big data analytics to improve HIV medical care utilisation in South Carolina: A study protocol

Bankole Olatosi,[1] Jiajia Zhang,[2] Sharon Weissman,[3] Jianjun Hu,[4] Mohammad Rifat Haider,[5] Xiaoming Li[6]

To cite: Olatosi B, Zhang J, Weissman S, *et al*. Using big data analytics to improve HIV medical care utilisation in South Carolina: A study protocol. *BMJ Open* 2019;**9**:e027688. doi:10.1136/bmjopen-2018-027688

► Prepublication history for this paper is available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2018-027688).

Received 7 November 2018
Revised 28 March 2019
Accepted 4 June 2019

For numbered affiliations see end of article.

Correspondence to
Dr Bankole Olatosi;
olatosi@mailbox.sc.edu

ABSTRACT

Introduction Linkage and retention in HIV medical care remains problematic in the USA. Extensive health utilisation data collection through electronic health records (EHR) and claims data represent new opportunities for scientific discovery. Big data science (BDS) is a powerful tool for investigating HIV care utilisation patterns. The South Carolina (SC) office of Revenue and Fiscal Affairs (RFA) data warehouse captures individual-level longitudinal health utilisation data for persons living with HIV (PLWH). The data warehouse includes EHR, claims and data from private institutions, housing, prisons, mental health, Medicare, Medicaid, State Health Plan and the department of health and human services. The purpose of this study is to describe the process for creating a comprehensive database of all SC PLWH, and plans for using BDS to explore, identify, characterise and explain new predictors of missed opportunities for HIV medical care utilisation.

Methods and analysis This project will create person-level profiles guided by the Gelberg-Andersen Behavioral Model and describe new patterns of HIV care utilisation. The population for the comprehensive database comes from statewide HIV surveillance data (2005–2016) for all SC PLWH (N≈18000). Surveillance data are available from the state health department's enhanced HIV/AIDS Reporting System (e-HARS). Additional data pulls for the e-HARS population will include Ryan White HIV/AIDS Program Service Reports, Health Sciences SC data and Area Health Resource Files. These data will be linked to the RFA data and serve as sources for traditional and vulnerable domain Gelberg-Anderson Behavioral Model variables. The project will use BDS techniques such as machine learning to identify new predictors of HIV care utilisation behaviour among PLWH, and 'missed opportunities' for re-engaging them back into care.

Ethics and dissemination The study team applied for data from different sources and submitted individual Institutional Review Board (IRB) applications to the University of South Carolina (USC) IRB and other local authorities/agencies/state departments. This study was approved by the USC IRB (#Pro00068124) in 2017. To protect the identity of the persons living with HIV (PLWH), researchers will only receive linked deidentified data from the RFA. Study findings will be disseminated at local community forums, community advisory group meetings, meetings with our state agencies, local partners and other key stakeholders (including PLWH, policy-makers and healthcare providers), presentations at academic conferences and through publication in peer-reviewed articles. Data security and patient confidentiality are the bedrock of this study. Extensive data agreements ensuring data security and patient confidentiality for the deidentified linked data have been established and are stringently adhered to. The RFA is authorised to collect and merge data from these different sources and to ensure the privacy of all PLWH. The legislatively mandated SC data oversight council reviewed the proposed process stringently before approving it. Researchers will get only the encrypted deidentified dataset to prevent any breach of privacy in the data transfer, management and analysis processes. In addition, established secure data governance rules, data encryption and encrypted predictive techniques will be deployed. In addition to the data anonymisation as a part of privacy-preserving analytics, encryption schemes that protect running prediction algorithms on encrypted data will also be deployed. Best practices and lessons learnt about the complex processes involved in negotiating and navigating multiple data sharing agreements between different entities are being documented for dissemination.

Strengths and limitations of this study

- This study is among the first in the USA to accumulate individual level data from multiple sources for predictive model development and validation of health utilisation in a statewide population of persons living with HIV (PLWH).
- This study is unique in its ability to examine health utilisation for all South Carolina PLWH from initial diagnosis across different treatment points and times in the HIV treatment cascade.
- Obtaining data release agreement, collecting and merging HIV sensitive data from different entities require significant time and effort to ensure privacy and confidentiality of all subjects.
- Although missing data or incorrect data may be a problem for care status classification, mandatory statewide reporting of HIV diagnosis and laboratory markers (CD4 and viral load) provides confidence in data completeness.
- Machine learning techniques could yield a several combinations of factors that could be difficult to interpret, but in anticipation of this problem, we have constituted a clinician expert review panel, and plan to use chart abstractions for further validation.

INTRODUCTION

Ending the HIV epidemic requires focus on 'treatment as prevention' as a goal. Prevention through linkage and retention in HIV medical care are key objectives of Healthy People 2020 and National HIV/AIDS Strategy.[1 2] Advances in HIV medications have made living a healthy life possible for persons living with HIV (PLWH), with well-established associations between linkage to/retention in HIV medical care and viral load (VL) suppression.[3–17] Recent numbers and proportions of South Carolina (SC) PLWH engaged in each step of the HIV treatment cascade show that cumulatively, only 66% received any medical care, and of these, only 54% received continuous HIV medical care.[18] Similarly, national estimates show that 42%–59% of patients with HIV are not in HIV medical care.[4 12 19]

Health inequities persist in the HIV treatment cascade for PLWH in SC.[20] These health inequities are important since a lack of engagement and retention in HIV medical care increases the likelihood of HIV transmission due to poor VL suppression.[3–14 16 17] Neither access to health insurance nor early linkage to care have thus far predicted retention in and consistent use of HIV medical care. This study proposes using both statistical and machine learning techniques to identify new and important predictors of HIV cascade treatment outcomes for PLWH, that is, CD4 and VL.

A responsive healthcare system must engage all PLWH at all stages of the HIV treatment cascade from HIV testing, linkage to care, timely initiation of care, retention in care and adherence to antiretroviral treatment (ART) with repeated CD4/viral load (VL) testing. Adoption of electronic health records (EHR), combined with the use of big data science (BDS) techniques provides an opportunity to improve health outcomes and manage high-risk and high-cost PLWH.[21–24] This study will create a comprehensive dataset accumulated from multiple sources, and use innovative BDS techniques to analyse new and old predictors of retention in HIV medical care. General health utilisation behaviour pre-HIV diagnosis will be studied to identify where missed opportunities for HIV testing occurs. PLWH profiles based on HIV medical care seeking behaviour will be developed with concomitant identification of gaps in HIV care and missed opportunities. The resulting model from this study will be used by the State Department of Health and Environmental Control (DHEC) to improve HIV care, through individual targeted linkage to, and retention in care. This study extends beyond the traditional scope of most HIV research in proposing novel machine learning processes for developing and validating a model of HIV medical care utilisation. In this protocol, we describe the process for acquiring datasets, data linkages and methods we will use to establish a population cohort from 2005 to 2016. Linkage of different datasets at the individual level using unique identifiers (IDs) at the population level enables us to achieve the following specific aims:

1. Use five commonly used measures of retention (described elsewhere) in HIV care[25] to generate a profile and pattern of care-seeking behaviour for SC PLWHs.
2. Use data mining and predictive analytics to identify missed opportunities for HIV testing prior to HIV diagnoses for SC PLWHs using CD4 count/VL as laboratory-based markers of time.
3. Identify gaps in the treatment cascade for all SC PLWHs who were never in care, not in care or who transition in and out of care compared with those consistently in HIV medical care.
4. Develop and validate a predictive risk model useful for targeting HIV care linkage interventions to all SC PLWHs who are never in care, not in care, transition in and out of care and at risk for dropping out of care.

Purpose

The purpose of this study is to describe the framework and process for creating a comprehensive database of all SC PLWH and plans for using BDS to explore, identify, characterise and explain new predictors of missed opportunities for HIV medical care utilisation. Study findings will be integrated with ongoing efforts of the SC DHEC's Data-to-Care (DTC) Program, and the Ryan White Care Program to link and retain PLWH in care.

METHODS AND ANALYSIS

This project is a population cohort-based study aimed at improving HIV treatment for all SC PLWH. Analytic focus will be on classification to predict where missed opportunities occur, gap identification and utilisation management across the HIV treatment cascade. We will classify the PLWH to four care groups: (1) never in care, (2) transitioning in and out of care, (3) not in care and (4) consistently in care using CD4/VL laboratory biomarkers and other demographics. An investigation into why certain health utilisation behaviours occur within the care group will be done using segmentation analysis. We will use predictive data mining to score newly identified variables representing the probability of the individual behaviour (action) occurring in the future (in this case HIV medical care utilisation).[26–28] Findings will be validated through a triangulated process involving the use of a HIV clinician expert panel, chart abstraction review process and through the 'DTC' community advisory board to explain and interpret new patterns/characteristics.

Guiding conceptual framework

This study is informed by an adaptation of the Gelberg-Andersen Behavioral Model framework.[11 26] This model identifies factors affecting health services utilisation among vulnerable patient populations and measures domains relevant for elucidating health utilisation patterns.[29 30] Socioecological factors affecting engagement in HIV medical care model will help guide the interpretation of new clusters of predictors.[12] Studies have

identified factors referred to in both models as important to PLWH.[18 19 31–36]

Study area and population

This study will be conducted in SC located in the south-eastern region of the USA, with a population of 4 961 119 in 2016.[37] Current epidemiological profile shows the PLWH ((n=18998) as mostly African-American [69%], men (71%) and aged 30–49 years (41%)).[38]

Data sources

The proposed data linkage is complex, and the comprehensive dataset is large. To our knowledge, no such data have ever been linked to study HIV treatment outcomes at the population level in the USA. Novel application of BDS techniques using a population of PLWHs can break new grounds and provide additional tools for improving care outcomes through the estimation of individual risks. Data sources are described in greater detail below.

SC Office of Revenue and Fiscal Affairs (RFA) Integrated Data System

RFA collects individual health utilisation data based on state laws requiring mandatory data reporting. The state law Section 44-6-170 guides the RFA in collecting and releasing healthcare related data. Since 1996, the RFA receives reports on all diagnoses from emergency departments, hospital inpatient, ambulatory care and outpatient surgery facilities in the Uniform Billing form (UB-92) format.[39] Non-compliant facilities face stiff penalties which increases compliance. At the RFA, individual patient information will be linked using unique patient IDs such as name, birth date and social security number.

The RFA's integrated data structure is recognised as a great example of an integrated data system in the USA (figure 1).[40] Data from RFA includes, but are not limited to:

1. All payer healthcare inpatient database.
2. Medicaid services claims data (including demographic file, visits file and pharmacy file).
3. State Employee Health Services Plan data.
4. Department of Corrections data (crime rates, prison history, etc).
5. Department of Mental Health.

DHEC e-HARS and Ryan White RSR data

Confidential name-based reporting of HIV/AIDS in SC as a reportable disease began in February 1986 leading to the creation of DHEC's enhanced HIV/AIDS reporting system (e-HARS).[39 41] e-HARS is a laboratory-based reporting system to which all statewide CD4 and VL tests are reported since 1 January 2004 as mandated by the Code of Laws of SC Section 44-29-10: Regulation 61-20.[39 41] e-HARS is a collection of computer programs and data files developed by the Centers for Disease Control and Prevention to simplify the management and analysis of HIV/AIDS surveillance data. DHEC also provides linkage to Ryan White Care clinic data through the Ryan White HIV/AIDS Program Service Reports (RSR). The annual RSR captures information regarding the services provided by all Ryan White-funded entities.[42]

Health Sciences South Carolina (HSSC) Data

HSSC is a biomedical research collaborative consisting of six of the state's largest health systems namely the

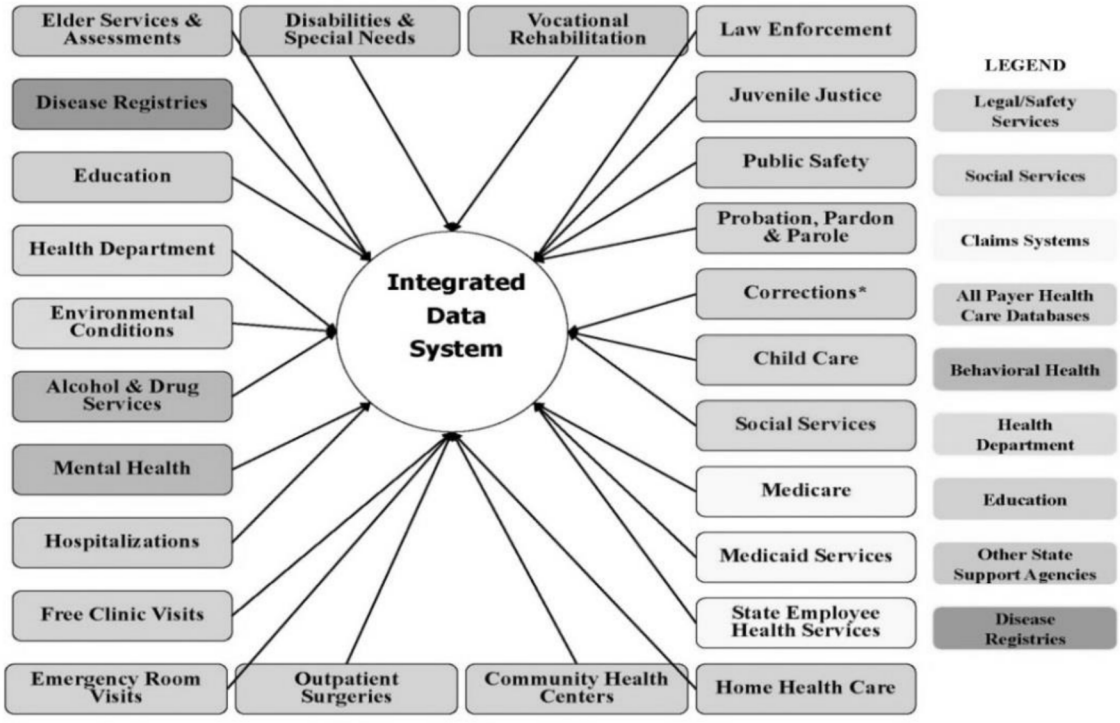

Figure 1 South Carolina Office of Revenue and Fiscal Affairs Integrated Data System.

Greenville Hospital System, University Medical Center, Palmetto Health, Spartanburg Regional Healthcare System, McLeod Health, AnMed Health and Self Regional Healthcare and the state's largest research-intensive universities Clemson University, the Medical University of SC and the University of SC. HSSC clinical data warehouse includes a Master Patient Index from multiple health systems that allows for the matching of clinical records across disparate information systems for a single patient.

Area Health Resources File (AHRF) and census tract data

The AHRF is a public dataset made available by Health Resources and Services Administration which contains data on healthcare professions, hospitals and healthcare facilities and US census population data. The AHRF provides health system-level information in areas such as healthcare professions, health facilities, hospital utilisation, expenditure and environment. Additional data on education, poverty, median income, employment, and so on, for different areas in SC will be extracted from the American Community Survey data.

Data linkage, release and security

RFA operations are guided by the SC data oversight council (DOC) as mandated by the state legislative assembly. The DOC oversees and regulates the collection and release of healthcare data in approved formats based on prevailing privacy laws. Data elements, which, when linked to other databases, can directly or indirectly identify a patient/healthcare professional, health insurer or healthcare facility are restricted. Protected health information such as patient name, address and social security numbers are never releasable; however, the RFA will use them to conduct final data linkages. The RFA will act as the honest data broker, deidentify the data and create unique IDs useful for research purposes. Table 1 demonstrates the extent of data linkage and highlights connections across different data sources. Detailed data release agreements are required and were secured from each data source before linkage. As a result of the scope and complexity of the data linkage, the state health department (DHEC) and the RFA created an intra-agency data sharing agreement to guide the data sharing process between both agencies. This intra-agency agreement provided specific terms and rules for data linkage to ensure confidentiality. In addition, SC Medicaid, departments of mental health, social services and corrections carefully reviewed the proposed data linkage to ensure stringent adherence to patient confidentiality. During data linkage, the RFA retained the right to create new variables in lieu of variables that could remotely identify any individual. Two examples are described here. First, the RFA scrambled the dates by introducing a modifier known only to them, while maintaining the original time between each date. This made it impossible for researchers to identify PLWH based on dates of service. In another example, rather than releasing spatial identifying data like zip codes or census block information, the RFA created the needed variables for the study, after carefully assessing the demand for the information. So instead of researchers receiving zip code data to calculate individual distance travelled for care, the RFA computed the distance travelled and included it in the final dataset as a computed variable. This workaround ensured that while the researchers did not get the zip code information, they still received the measure of interest (distance travelled to/from facility) computed by the RFA. A similar approach was deployed for all potential identifiable patient information. The state health department (DHEC) and the RFA hold the keys to the unique IDs created during the linkage. This will enable them to apply the models and algorithms created during the study, to individualised PLWH interventions in the future. All data transfers occur using secure file transfer protocols. Data are received from sources fully encrypted and stored fully encrypted with encryption software. The data are stored in a key-access only facility hosting servers. Network shares for users are created via a distributed file system and user access is controlled via user groups containing unique IDs that require complex passwords on facility site.

Population inclusion criteria

Only living PLWH whose residence at diagnosis was SC are included in the study. A key question when describing PLWH not in HIV medical care is outmigration. Census estimates show a positive net immigration trend for SC, and so we do not expect any problem.[43] Nevertheless, we will scrutinise the data carefully to interpret our data appropriately in case migration becomes an issue. Only cases with age ≥13 in the diagnosed prevalence year are included in the analysis (ie, age must be ≥13 in 2005 to be included in analysis). We chose 2005 since this was the year after the state law mandatory reporting of all CD4 and VL tests to e-HARS began. Twelve years (2005–2016) of HIV utilisation data is available for this study.

HIV MEDICAL CARE PATTERNS AND GELBERG-ANDERSEN MODEL VARIABLES

Gelberg-Andersen Model variables

Predisposing, enabling and need factors correspond to the wide array of HIV data associated with linkage to care, retention, re-engagement and ART monitoring in the HIV treatment cascade. Variables reflecting Gelberg-Andersen Model along with their corresponding stages along the HIV treatment cascade and their data sources are illustrated in figure 2. Focus will be placed on categorising variables under the appropriate factor predictive of their HIV medical care utilisation. These variables are available from the previously described data sources.

Detailed examples of specific Gelberg-Anderson predisposing variables available through the e-HARS database include patient name, birth date, social security number (restricted data elements for linkage only), date of first positive HIV test, AIDS diagnosis date, source of report

Table 1 HIV treatment cascade and corresponding variables data sources*†

HIV treatment cascade	Variables based on Gelberg-Andersen Model	Data sources
Diagnosis	Level at diagnosis ► CD4 ► Viral load	Department of Health and Environmental Control Enhanced HIV/AIDS Reporting System (DHEC e-HARS)
	Predisposing factors	
HIV linkage to care	Demographics	Revenue and Fiscal Affairs (RFA) Medicaid DHEC e-HARS
	Health beliefs	
	Vulnerable domains Location	RFA American Community Survey (ACS) Census Tract
	Criminal behaviour, violent status	Department of Corrections
	Mental Illness	Department of Mental Health
	Childhood characteristics	Department of Social Services (DSS)
Retention	**Enabling factors**	
	Regular source of care	Medicaid RFA Ryan White Service Reports (RSR)
	Social support, public benefits	DSS RSR
	Health services resources	ACS Census Tract CDC GPS data RFA Area Health Resource Files (AHRF) Palmetto Health Hospital data Greenville Health System data
Re-engagement	Case management	DHEC e-HARS RSR
	Community resources	ACS Census Tract
	Location variables (poverty, education, median income, employment)	ACS Census Tract
	Need factors	
	Evaluated health—diagnosis, comorbidities	RFA
	Perceived health	RSR
	Health behaviours	
	Personal health practices	RFA Health Sciences South Carolina (HSSC) DHEC e-HARS RSR
ART monitoring	Use of health services (HIV test dates, AIDS diagnosis dates)	RFA HSSC Palmetto Health Hospital data Greenville Health System data DHEC e-HARS

Continued

Table 1 Continued

HIV treatment cascade	Variables based on Gelberg-Andersen Model	Data sources
Viral suppression	**Outcomes** Viral load CD4 level	DHEC e-HARS

*All data linkage is conducted at the individual unit level using name, date of birth and social security number.
†All records from other datasets linked to the e-HARS cohort are available through the RFA and other data sources listed above.
ART, antiretroviral treatment; CDC, Centers for Disease Control and Prevention; GPS, Global Positioning System.

and transmission risk factor. Others include gender, race/ethnicity, county of residence, year of death, cause of death (International Classification of Diseases (ICD)-9, ICD-10 codes), poverty, education, median income, employment, and so on, all CD4 +T cell counts and VL values. Vulnerable predisposing domain variables, such as criminal behaviour, violent status, mental illness, and childhood characteristics, are provided through RFA data sources. Enabling factors such as regular source of care will be obtained from Medicaid/State Healthcare plan. Data related to social support and public benefits will be obtained from RSR through medical case management and RFA. Need factor variables will be obtained through inpatient claims data from HSSC and RFA. The inpatient data include patient demographics, source and type of admission, visits/encounters, diagnoses, procedures, laboratory results and length of stay. For HIV utilisation outcomes data, e-HARS will provide CD4 and VL measures as predictors of retention in care. The core variables for HIV utilisation, for example, missed visits, appointment adherence, constancy in 3-month or 4-month, 6-month interval (retention in care variables) are described in table 2. We will define visits/health encounters that lead care providers to suggest HIV testing, that is, to persons belonging to a high-risk group for HIV acquisition and who presented with HIV-related or non-HIV-related clinical conditions.

Antiretroviral medication and polypharmacy

EHR from patient encounters available through the RFA and HSSC capturing information about PLWH healthcare service utilisation and medications will be analysed during the study. EHR data from the RFA contain information related to encounter visits, diagnosis, laboratory services and medications (number prescribed, drug class, indication, strength and dosage). Binary variables will be created for ART status.

Data analysis plan

Data management, cleaning and mining

Data will be assessed for reliability to deal with issues related to missing, aberrant or extreme values. New variables and data inclusion criteria in each aim will be validated through chart review, HIV clinicians' expert panel and the DTC community advisory board.[26–28] We will reduce the dimension of the variables (number of variables) using autocorrelation, multicollinearity and principal component analysis as guided. The focus during data management will be on eliminating extreme outliers, and excluding irrelevant variables and discretising (binning) continuous variables. Appropriate methods will be deployed to prepare for classification and prediction.

Population segmentation

To achieve specific aim #1, supervised and unsupervised machine learning will be used to identify care patterns or HIV utilisation. Once an individual is assigned to a certain care group, it is necessary to discover a pattern that may lead them to this care group. Thus, we will try to classify them based on the individual and system factors that lead the recognition of the care group. Many classification tools are available in machine learning: logistic regression, naive Bayes classifier, support vector machine and random forest. Most can be implemented in R. Different distance methods (eg, Jaccard or Gower) will be deployed to measure similarities between HIV utilisation. Once the distance between variable is correctly measured, the clustering method, a method of unsupervised learning, will be operated over a distance matrix instead of the original data matrix. Both the K-means clustering and partitioning around medoids algorithms (PAM) will be applied to classify the data into 'k' groups.[26–28 40] For example, we will use cluster analysis to partition the health utilisation data into groups of similar health utilisation behaviour. Derived clusters will be interpretable based on relevant Gelberg-Andersen variables and as such can be assigned a description/class label. The result of a cluster analysis is a binary tree, or dendrogram, with n–1 nodes. We propose using tree pruning to adjust model complexity for the creation of an optimal model. The simplest model with the highest validation assessment will be considered the best model. Statistics for judging the model will depend on the type of prediction classifications, rankings or estimates. For prediction estimates, the squared error (difference between target and estimate) will be used to assess model performance.[2 26–28 40]

Predictive modelling

For specific aims #2–3, we will identify the associations between the missed opportunities and/or treatment cascade, such as timing to care and consistency in care with Gelberg-Andersen domain inputs such as gender, race, sexual orientation, crime history, location, and so

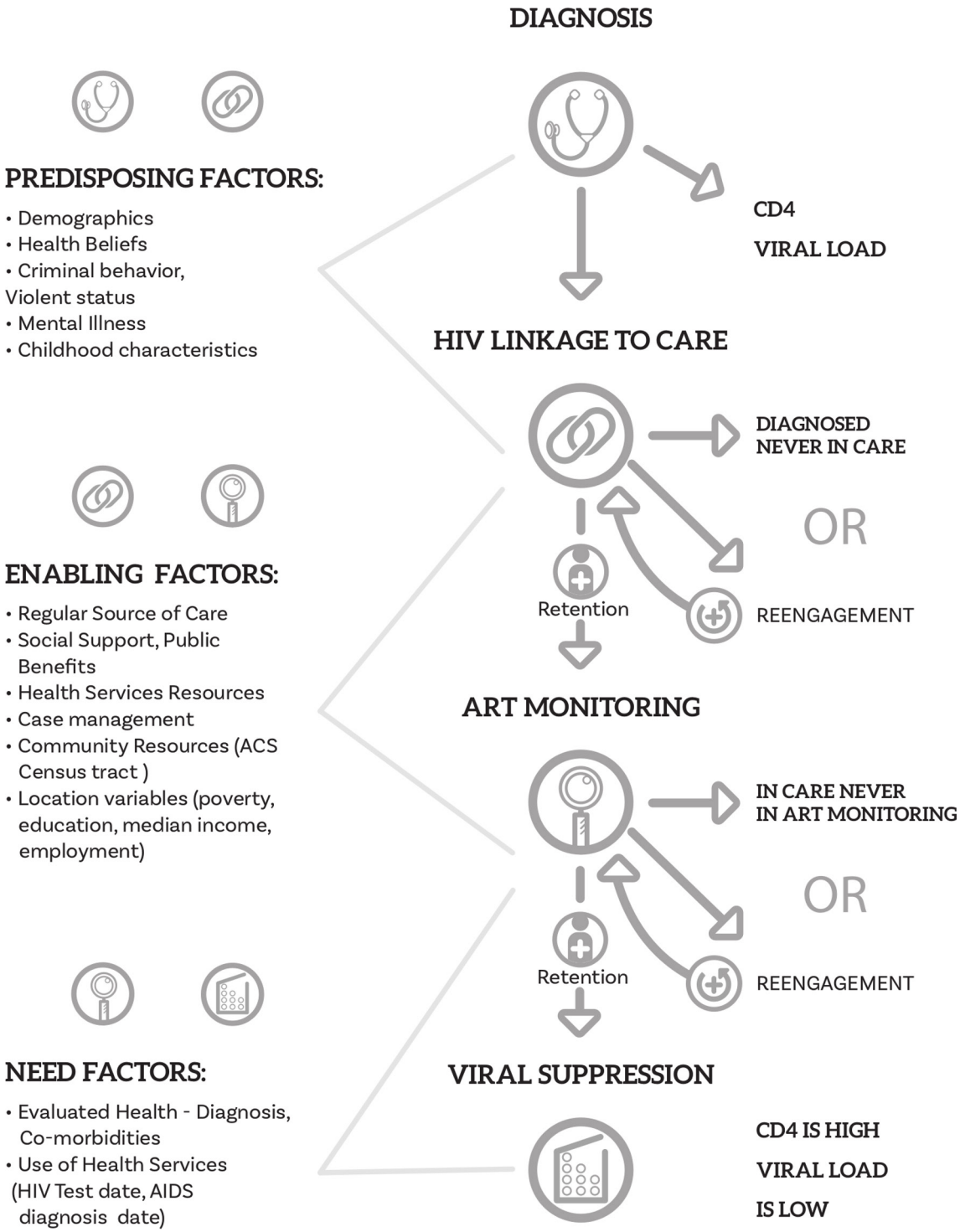

Figure 2 HIV treatment cascade including Gelberg-Andersen Model variables and data sources.

on. To achieve specific aim #4, we will build and validate appropriate predictive statistical models, such as multinomial logistic regression model, Cox proportional hazards model and generalised linear mixed model, to investigate the impact of care pattern, missed opportunity and Gelberg-Andersen variables to investigate how soon linkage to care, status in care and CD4/VL level changes. A HIV clinician expert panel and the DTC community group will provide guidance on the interpretation of new findings. Once the model is validated, trained and assessed using test data, it will be deployed and monitored for performance. An overview of the analytic plan is illustrated in figure 3.

Model development and validation

Once the pattern is identified such as health utilisation pattern, variable distributions will be summarised for pattern status (mean, SD and counts), and compared using the t-test, analysis of variance test and χ^2 test. Where test assumptions are not satisfied, non-parametric tests

Table 2 Selected variables for HIV care pattern determination*

Measure	Type of output; calculation overview	Observation time needed to calculate
Missed visit	Dichotomous; were there any missed visits in the interval? Count; number of missed visits in the interval	At least 6 months
Appointment adherence	Continuous; attended appointments divided by (attended appointments plus missed appointments)	Patient: at least 1 year; clinic: as short as 1 day
Constancy, 3-month or 4-month intervals	Categorical; number of 3-month or 4-month intervals with at least one attended visit	At least 6–8 months
Constancy, 6-month intervals	Categorical; number of 6-month intervals with at least one attended visit	At least 1 year
Constancy, 6-month intervals, longer term	Dichotomous; At least one attended visit in each 6-month interval with at least 60 days between visits	At least 2 years
Constancy, HIV/AIDS Bureau	Dichotomous; At least two attended visits in 12 months, separated by at least 90 days	At least 1 year
Gaps	Dichotomous; did the time between two contiguous attended visits exceed a threshold (eg, 6 months)? Continuous; what is the longest duration of time between two contiguous attended visits?	At least 1 year

*Definitions for visits have been previously described elsewhere.[25]

(Wilcoxon rank test and Kruskal-Wallis Test) will be applied. Results will be discussed among the investigation team to find the most appropriate pattern characteristic. Care patterns identified in aim 1 and the risk factors and missed opportunities identified in aim 2 will serve as the main exposure variables to identify the association between the HIV treatment cascade, HIV care patterns and missed opportunities. All variables in Gelberg-Andersen Behavioral Model, which can be identified in the linked dataset, will be used as the potential risk factors. With the patient characteristics, hospital information, care information and variables in Gelberg-Andersen Behavioral Model, a high dimension of predictors is expected. Cross validation will be used to evaluate the prediction performance and to compare the different risks prediction model. The data will be divided into training, testing and validation data. Potential interactions between variables will be examined, using the log-likelihood ratio test to determine statistical significance. Using machine learning techniques, we will create algorithms that can alert/flag high dropout risk PLWH. In addition, we will automate the algorithms to determine the best predictors of re-engagement into HIV medical care based on historical health utilisation patterns, CD4 counts and viral suppression. The algorithms will be shared with DHEC and RFA, who will apply these algorithms to the original data. The algorithms will also be shared with the Ryan White Program for use in monitoring and retaining PLWH in care.

Model monitoring and validation

While the promise inherent in data mining is in discovering new and useful patterns in big data, its true value is in responding to these patterns by acting on them. This

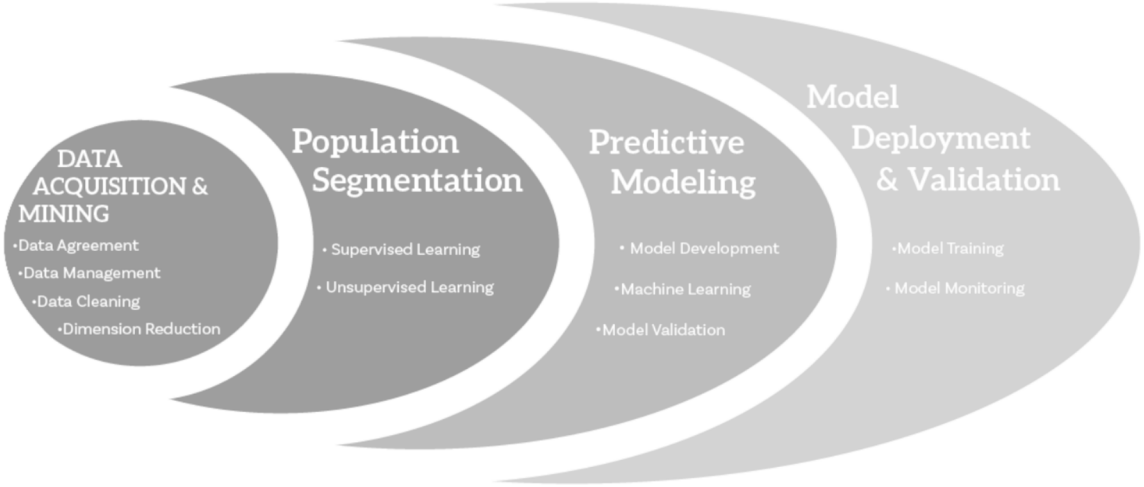

Figure 3 Analytic plan.

ultimately moves data into information, information into action and action into value. We anticipate developing 2–3 predictive models. After predictive models are developed, and validated, they will be deployed against new claims data showing healthcare utilisation. Since we have observations for PLWH who use HIV medical care consistently (an identified segment/cluster), we will create a unique normalised score for this population and set it as the norm for comparison. Profiles and predictors of care status for the not-in-care population will be assessed by the models deployed against new claims data (2017–2018). Model scores will be assessed to evaluate model performance and accuracy using a propensity score analysis.[2 26–28 40] Findings will be reviewed by the HIV clinician's expert panel as well as the DTC advisory group. They will help inform the identification of new variables and predictors by cross validating them based on abstracted charts. Validation rules will be established to guide this process.[19] Algorithms for the model will be created and embedded with DHEC's HIV surveillance system and Ryan White Program. Qualified state personnel will use this information to identify/flag those at risk for dropping out of care, those out of care and those not likely to engage in care after linkage. The goal in doing this is to help provide targeted assistance to such high-risk individuals. The algorithms will be reviewed intermittently based on the availability of new data to ensure good performance.

Patient involvement

Patients were not involved in preparing this study protocol.

DISCUSSION

The linkage of several databases capturing traditional clinical outcomes through the EHR and other health claims-based system, integrated to a social determinants of health data system under the purview of this study, holds significant promise for HIV medical care. Prior to this study, the confidential nature of HIV has limited translational research across the HIV treatment cascade to either clinical or social determinants studies, but rarely both. The challenges with finding such data sources are significant, as is the ability to measure variables at the individual level. BDS techniques offer the potential of opening new possibilities for managing complex health conditions like HIV.[43] The robust value inherent in using population-based cohorts for improving health outcomes and predicting future health utilisation is documented.[44 45] Successful examples of previous and ongoing application of BDS techniques can be found in automated ECG interpretation, automated detection of lung nodules from X-rays, and creation of Framingham Risk Score.[46] Other examples using unsupervised learning for pattern identification exist in the ongoing development of precision medicine,[47–51] systolic heart failure survival prediction[52] and the use of machine learning to automate diagnosis of acute brain infractions.[53] BDS techniques have successfully been applied to large-scale clinical studies such as the CArdiovascular disease research using LInked Bespoke studies and Electronic health Record study in the UK[44] and Comparison of Medical Therapy, Pacing, and Defibrillation in Heart Failure (COMPANION) trials in the USA.[45] Studies point to these examples as the future and real added value of BDS techniques like machine learning.[54] The combination of machine learning techniques with expert clinicians, case workers, stakeholders together has serious potential to improve the collective health of PLWH. This study is unique in its data linkage and population focus, giving it strengths and precision unavailable to previous studies using samples of PLWH for retention in care studies. This study also goes beyond prior studies by integrating both model development and model validation using BDS techniques and meet the checklists for reporting as recommended by the 'Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis' guidelines.[55] The novelty in this study is the integration of traditional and vulnerable domain factors with large troves of EHR data, and its tieback to actual linkage and retention in care. The state department of health (DHEC) surveillance programme, and Ryan White Program will be strengthened with the ability to flag PLWH at risk for not engaging in care after linkage, or those at risk at dropping out of care. This study plans to automate a process that had been previously manually done. It also improves the process by flagging individuals at high risk for not engaging in or dropping out of care for intervention. In addition, locations/providers where missed opportunities for HIV care re-engagement occurred will also receive strengthening to improve patient engagement. DHEC's legal mandates allow them to work closely with clinical providers in targeting individualised interventions to such at-risk PLWH. This will help improve engagement in, and future retention in care among PLWH.

Future improvement and application of the model

Plans for model improvement using future population HIV care utilisation data will improve model performance and external validation as we focus on evaluating the incremental value of specific predictors (new and old) for HIV care utilisation. Future data will be deployed to the models for model maintenance and improvement. The use of the health system by PLWH not in care represents missed opportunities for re-engaging them into HIV medical care. Without treatment as a form of prevention, ending the HIV epidemic is harder to achieve. Benefits from advances in HIV treatment are also lost for those not in care. However, investigating characteristics for those who are not in care is difficult and expensive for health departments. The application of BDS to this process will make substantive improvements, and allow clinicians, social workers and other stakeholders help re-engage this hard to reach population back into care.

Author affiliations
[1]Health Services, Policy and Management, Arnold School of Public Health, University of South Carolina, Columbia, South Carolina, USA
[2]Department of Epidemiology and Biostatistics, University of South Carolina, Columbia, South Carolina, USA
[3]Internal Medicine, School of Medicine, University of South Carolina, Columbia, South Carolina, USA
[4]Department of Computer Science & Engineering, College of Engineering, University of South Carolina, Columbia, South Carolina, USA
[5]Department of Health Promotion, Education & Behavior, Arnold School of Public Health, University of South Carolina, Columbia, South Carolina, USA
[6]Health Promotion Education and Behavior, Arnold School of Public Health, University of South Carolina, Columbia, South Carolina, USA

Acknowledgements The authors thank SC Department of Health and Environmental Control, SC Revenue and Fiscal Affairs office, Health Sciences South Carolina for providing the data on People Living with HIV in South Carolina.

Contributors BO, XL and MRH drafted the manuscript. XL, BO and JZ conceived and designed the study. JZ, SW, and JH made substantial contributions to the study design and manuscript editing. XL and BO are responsible for study coordination; BO, JZ, JH and MRH are responsible for data quality control, management and analysis. All authors contributed to the writing of the study protocol in an iterative manner and have read and approved the final manuscript.

Funding This study is supported by the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under Award Number 1R01AI127203-01A1. Li and Olatosi are the PI for the study.

Disclaimer The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. HealthyPeople.gov. HIV Objectives. https://www.healthypeople.gov/2020/topics-objectives/topic/hiv/objectives [Accessed 25 May 2018].
2. Linoff GS, Berry MJ. *Data mining techniques: for marketing, sales, and customer relationship management*: John Wiley & Sons, 2011.
3. Andrews JR, Wood R, Bekker LG, *et al*. Projecting the benefits of antiretroviral therapy for HIV prevention: the impact of population mobility and linkage to care. *J Infect Dis* 2012;206:543–51.
4. Celum C, Hallett TB, Baeten JM. *HIV-1 prevention with ART and PrEP: mathematical modeling insights into resistance, effectiveness, and public health impact*: Oxford University Press, 2013.
5. Cohen SM, Hu X, Sweeney P, *et al*. HIV viral suppression among persons with varying levels of engagement in HIV medical care, 19 US jurisdictions. *J Acquir Immune Defic Syndr* 2014;67:519–27.
6. Crawford TN. Poor retention in care one-year after viral suppression: a significant predictor of viral rebound. *AIDS Care* 2014;26:1393–9.
7. Edun B, Iyer M, Albrecht H, *et al*. The South Carolina HIV Cascade of Care. *South Med J* 2015;108:670–4.
8. Fleishman JA, Yehia BR, Moore RD, *et al*. Establishment, retention, and loss to follow-up in outpatient HIV care. *J Acquir Immune Defic Syndr* 2012;60:249–59.
9. Gardner EM, Young B. The HIV care cascade through time. *Lancet Infect Dis* 2014;14:5–6.
10. Hall HI, Gray KM, Tang T, *et al*. Retention in care of adults and adolescents living with HIV in 13 U.S. areas. *J Acquir Immune Defic Syndr* 2012;60:77–82.
11. Hall HI, Tang T, Westfall AO, *et al*. HIV care visits and time to viral suppression, 19 U.S. jurisdictions, and implications for treatment, prevention and the national HIV/AIDS strategy. *PLoS One* 2013;8:e84318.
12. Mugavero MJ, Amico KR, Horn T, *et al*. The state of engagement in HIV care in the United States: from cascade to continuum to control. *Clin Infect Dis* 2013;57:1164–71.
13. Rebeiro P, Althoff KN, Buchacz K, *et al*. Retention among North American HIV-infected persons in clinical care, 2000-2008. *J Acquir Immune Defic Syndr* 2013;62:356–62.
14. Weissman S, Duffus WA, Iyer M, *et al*. Rural-urban differences in HIV viral loads and progression to AIDS among new HIV cases. *South Med J* 2015;108:180–8.
15. Yehia BR, Ketner E, Momplaisir F, *et al*. Location of HIV diagnosis impacts linkage to medical care. *J Acquir Immune Defic Syndr* 2015;68:304–9.
16. Yehia BR, Rebeiro P, Althoff KN, *et al*. Impact of age on retention in care and viral suppression. *J Acquir Immune Defic Syndr* 2015;68:413–9.
17. Zanoni BC, Mayer KH. The adolescent and young adult HIV cascade of care in the United States: exaggerated health disparities. *AIDS Patient Care STDS* 2014;28:128–35.
18. Underwood C, Hendrickson Z, Van Lith LM, *et al*. Role of community-level factors across the treatment cascade: a critical review. *J Acquir Immune Defic Syndr* 2014;66:S311–18.
19. Zimmerman RS, Morisky DE, Harrison L, *et al*. Validity of behavioral measures as proxies for HIV-related outcomes. *J Acquir Immune Defic Syndr* 2014;66(Suppl 3):S285–92.
20. Hall HI, Frazier EL, Rhodes P, *et al*. Differences in human immunodeficiency virus care and treatment among subpopulations in the United States. *JAMA Intern Med* 2013;173:1337–44.
21. Bates DW, Saria S, Ohno-Machado L, *et al*. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Aff* 2014;33:1123–31.
22. Eggleston EM, Finkelstein JA. Finding the role of health care in population health. *JAMA* 2014;311:797–8.
23. Rosenbaum S. The Patient Protection and Affordable Care Act: implications for public health policy and practice. *Public Health Rep* 2011;126:130–5.
24. Shaw FE, Asomugha CN, Conway PH, *et al*. The Patient Protection and Affordable Care Act: opportunities for prevention and public health. *Lancet* 2014;384:75–82.
25. Mugavero MJ, Davila JA, Nevin CR, *et al*. From access to engagement: measuring retention in outpatient HIV clinical care. *AIDS Patient Care STDS* 2010;24:607–13.
26. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning. Data Mining, Inference and Prediction*. New York: Springer, 2008.
27. Tibshirani R, James G, Witten D, *et al*. *An introduction to statistical learning-with applications in R*. New York, NY: Springer, 2013.
28. Tufféry S. *Data mining and statistics for decision making*: Wiley Chichester, 2011.
29. Gelberg L, Andersen RM, Leake BD. The Behavioral Model for Vulnerable Populations: application to medical care use and outcomes for homeless people. *Health Serv Res* 2000;34:1273.
30. Stein JA, Andersen R, Gelberg L. Applying the Gelberg-Andersen behavioral model for vulnerable populations to health services utilization in homeless women. *J Health Psychol* 2007;12:791–804.
31. Eberhart MG, Voytek CD, Hillier A, *et al*. Travel distance to HIV medical care: a geographic analysis of weighted survey data from the Medical Monitoring Project in Philadelphia, PA. *AIDS Behav* 2014;18:776–82.
32. Oldenburg CE, Perez-Brumer AG, Reisner SL. Poverty matters: contextualizing the syndemic condition of psychological factors and newly diagnosed HIV infection in the United States. *AIDS* 2014;28:2763.
33. Qiao S, Li X, Stanton B. Social support and HIV-related risk behaviors: a systematic review of the global literature. *AIDS Behav* 2014;18:419–41.
34. Sagrestano LM, Clay J, Finerman R, *et al*. Transportation vulnerability as a barrier to service utilization for HIV-positive individuals. *AIDS Care* 2014;26:314–9.
35. Tomori C, Risher K, Limaye RJ, *et al*. A role for health communication in the continuum of HIV care, treatment, and prevention. *J Acquir Immune Defic Syndr* 2014;66(Suppl 3):S306–10.
36. Vaughan AS, Rosenberg E, Shouse RL, *et al*. Connecting race and place: a county-level analysis of White, Black, and Hispanic HIV prevalence, poverty, and level of urbanization. *Am J Public Health* 2014;104:e77–84.
37. *United States Census Bureau*. South Carolina: Quick Facts, 2018. Available: https://www.census.gov/quickfacts/fact/table/SC#viewtop [Accessed 25 May 2018].
38. Division of Surveillance and Technical Support BoDC, SC DHEC. *An Epidemiologic Profile of HIV and AIDS in South Carolina 2017*. Columbia SC, 2017.
39. Centers for Disease Control and Prevention (CDC). Missed opportunities for earlier diagnosis of HIV infection-South Carolina, 1997-2005. *MMWR Morb Mortal Wkly Rep* 2006;55:1269.

40. Truxillo C, Lucas B, Patetta M, *et al*. *Advanced Business Analytics*: SAS Institute, Cary, NC, 2012.
41. Olatosi BA, Probst JC, Stoskopf CH, *et al*. Patterns of engagement in care by HIV-infected adults: South Carolina, 2004-2006. *AIDS* 2009;23:725–30.
42. Health Resources and Services Administration (HRSA). Ryan White HIV/AIDS Program Services Report (RSR) [updated]. 2018. https://hab.hrsa.gov/program-grants-management/ryan-white-hivaids-program-services-report-rsr [Accessed 10 July 2018].
43. Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. *N Engl J Med* 2016;375:1216–9.
44. Denaxas SC, George J, Herrett E, *et al*. Data resource profile: cardiovascular disease research using linked bespoke studies and electronic health records (CALIBER). *Int J Epidemiol* 2012;41:1625–38.
45. Kalscheur MM, Kipp RT, Tattersall MC, *et al*. Machine Learning Algorithm Predicts Cardiac Resynchronization Therapy Outcomes: Lessons From the COMPANION Trial. *Circ Arrhythm Electrophysiol* 2018;11:e005499.
46. Kannel WB, Doyle JT, McNamara PM, *et al*. Precursors of sudden coronary death. Factors related to the incidence of sudden death. *Circulation* 1975;51:606–13.
47. Corren J, Lemanske RF, Hanania NA, *et al*. Lebrikizumab treatment in adults with asthma. *N Engl J Med* 2011;365:1088–98.
48. Deo RC. Machine Learning in Medicine. *Circulation* 2015;132:1920–30.
49. Gorodeski EZ, Ishwaran H, Kogalur UB, *et al*. Use of hundreds of electrocardiographic biomarkers for prediction of mortality in postmenopausal women: the Women's Health Initiative. *Circ Cardiovasc Qual Outcomes* 2011;4:521–32.
50. Hsich E, Gorodeski EZ, Blackstone EH, *et al*. Identifying important risk factors for survival in patient with systolic heart failure using random survival forests. *Circ Cardiovasc Qual Outcomes* 2011;4:39–45.
51. Woodruff PG, Modrek B, Choy DF, *et al*. T-helper type 2-driven inflammation defines major subphenotypes of asthma. *Am J Respir Crit Care Med* 2009;180:388–95.
52. Kwon JM, Lee Y, Lee Y, *et al*. An Algorithm Based on Deep Learning for Predicting In-Hospital Cardiac Arrest. *J Am Heart Assoc* 2018;7:e008678.
53. Beecy AN, Chang Q, Anchouche K, *et al*. A Novel Deep Learning Approach for Automated Diagnosis of Acute Ischemic Infarction on Computed Tomography. *JACC Cardiovasc Imaging* 2018;11:1723–5.
54. Chen JH, Asch SM. Machine Learning and Prediction in Medicine - Beyond the Peak of Inflated Expectations. *N Engl J Med* 2017;376:2507–9.
55. Collins GS, Reitsma JB, Altman DG, *et al*. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Med* 2015;13:1.