Development of a large urban longitudinal HIV clinical cohort using a web-based platform to merge electronically and manually abstracted data from disparate medical record systems: technical challenges and innovative solutions

Alan E Greenberg¹, Harlen Hays², Amanda D Castel¹, Thilakavathy Subramanian², Lindsey Powers Happ¹, Maria Jaurretche¹, Jeff Binkley², Mariah M Kalmin¹, Kathy Wood², and Rachel Hart², on behalf of the DC Cohort Executive Committee*

ABSTRACT

Objective Electronic medical records (EMRs) are being increasingly utilized to conduct clinical and epidemiologic research in numerous fields. To monitor and improve care of HIV-infected patients in Washington, DC, one of the most severely affected urban areas in the United States, we developed a city-wide database across 13 clinical sites using electronic data abstraction and manual data entry from EMRs.

Materials and Methods To develop this unique longitudinal cohort, a web-based electronic data capture system (Discovere⁶) was used. An Agile software development methodology was implemented across multiple EMR platforms. Clinical informatics staff worked with information technology specialists from each site to abstract data electronically from each respective site’s EMR through an extract, transform, and load process.

Results Since enrollment began in 2011, more than 7000 patients have been enrolled, with longitudinal clinical data available on all patients. Data sets are produced for scientific analyses on a quarterly basis, and benchmarking reports are generated semi-annually enabling each site to compare their participants’ clinical status, treatments, and outcomes to the aggregated summaries from all other sites.

Discussion Numerous technical challenges were identified and innovative solutions developed to ensure the successful implementation of the DC Cohort. Central to the success of this project was the broad collaboration established between government, academia, clinics, community, information technology staff, and the patients themselves.

Conclusions Our experiences may have practical implications for researchers who seek to merge data from diverse clinical databases, and are applicable to the study of health-related issues beyond HIV.

Keywords: electronic medical record, EMR, HIV, cohort, DC Cohort

BACKGROUND AND SIGNIFICANCE

Electronic medical records (EMRs) are being increasingly utilized to conduct clinical and epidemiologic research in numerous fields including breast,¹, prostate,², pancreatic,³,⁴ and population-based cancer research,⁵ cardiovascular disease,⁶ dialysis,⁷ genomics,⁸ primary care health,⁹ cost effectiveness,¹⁰ and HIV/AIDS.¹¹ There are a number of distinct advantages of using EMR data, including access to extensive clinical, laboratory, and treatment data¹,¹² in a relatively timely fashion,⁹ improved accuracy of information,¹³ capacity to merge and analyze data from disparate clinical sites,¹⁰ and ability to collect data without extensive and costly patient interactions.¹³ Challenges exist with using EMR data as well, namely that data are collected for clinical practice rather than research purposes,¹⁴ lack of standardization across EMR platforms causing difficulties merging variables across databases,¹,¹² under-reporting of care received at other institutions,¹ information written in free-text notes that is difficult to extract electronically,¹,¹³ missing or conflicting data,⁴,⁸ and errors in data linkage.⁶

Several clinical observational cohorts of HIV-infected persons in the United States have been established that combine data from multiple clinics and use varying approaches for data collection, including those that collect data at regular 6-month intervals,¹⁵,¹⁶ several that use a common EMR platform,¹⁷,¹⁸ data merged from clinics at academic medical centers,¹⁹ patients enrolled in randomized clinical trials of antiretroviral therapy,²⁰ patients seen at multiple outpatient clinics,²¹ and the largest of all, North American AIDS Cohort Collaboration on Research and Design (NA-ACCORD), which incorporates data from numerous HIV cohorts in North America.²²

The ability to establish an EMR-based HIV cohort in Washington DC presented a unique opportunity to better characterize HIV in a city that has one of the most severe HIV epidemics in the United States. Washington DC has an estimated 2.5% of adults and adolescents, and more than 16,000 persons, living with HIV/AIDS.²³ In response to this epidemic, in 2009 the National Institutes of Health (NIH) partnered with the DC Department of Health (DC DOH) to create the “Partnership for AIDS Progress” (PFAP) to support local HIV prevention and treatment research and inform the local epidemic response.²⁴ HIV care in DC is decentralized and provided at various community-based, academic and government clinics, and private physician offices. Accordingly, the PFAP supported the creation of the “DC Cohort” to
develop a combined electronic database for 13 of the largest HIV clinical sites, with the plan to consider expansion to several other clinics and private providers in the years ahead. The goals of the DC Cohort are to monitor and improve HIV care and treatment at individual sites and city-wide; to develop a powerful and real-time data platform that can be used to facilitate epidemiologic, clinical, and translational scientific research on HIV/AIDS and its co-morbidities; and to establish a semi-automated electronic data abstraction technology that can be used to assess the clinical evolution of a large HIV-infected cohort as it matures and as new therapeutic interventions become available.

There are several novel aspects of the DC Cohort that differ from other established HIV cohorts: the DC Cohort has a “city-wide” focus enrolling patients from multiple types of clinics across Washington DC; data are merged from numerous EMRs into a centralized web-based database using electronic data transfer and manual data entry for more than 200 data variables covering major areas of study interest such as demographics, medications, diagnoses, laboratory tests, pathology and clinical procedures, and resistance data; and the DC Cohort database is merged semi-annually with the DC DOH HIV/AIDS surveillance database in order to enhance the completeness of both data sources and assist in identifying patients receiving care at multiple sites.

OBJECTIVE
The objective of the present paper is to summarize the technical challenges encountered in the development of this unique cohort and the innovative solutions that were developed to address them. Our experiences may have practical implications for researchers who seek to merge data from diverse clinical databases, and are applicable to the study of health-related issues beyond HIV.

MATERIALS AND METHODS
Organizational Structure
Given the numerous organizations and clinics involved in the DC Cohort, it was critical to define the working relationships of the collaborating entities. As shown in Figure 1, the DC Cohort organizational structure was created with the foundation of 13 collaborating clinics each led by a site principal investigator (PI); a data and statistics coordinating center (DSCC) comprised of the study PIs and staff from The George Washington (GW) University and the Cerner Corporation; the executive committee (EC) serving as the DC Cohort central governing body, comprised of the site PIs, the GW, and Cerner PIs, a representative from the DC DOH HIV/AIDS, Hepatitis, STD, Tuberculosis Administration, two representatives from the National Institutes of Health (NIH), and two representatives of the patient Community Advisory Board (CAB); the CAB itself, comprised of one patient representative per site, who advise the EC; and the PFAP EC that included senior NIH and DC DOH scientists providing oversight and guidance to the EC. The GW PI (A.D.C.) is responsible for the scientific, administrative and fiscal oversight of the Cohort, and is the primary point-of-contact with the Site PIs, the DSCC, and the CAB; the GW Co-PI (A.E.G.) works closely with the GW PI in overall project management and is the primary point-of-contact with the PFAP EC; and the Cerner PI (H.H.) oversees the data management and IT aspects of the Cohort and is the primary point-of-contact for the Cerner Corporation.

Communication
Numerous personnel were involved in the design and implementation of the DC Cohort including research assistants (RAs), information technology (IT) staff, fiscal staff, and clinical care providers at the site level; and epidemiologists, biostatisticians, clinical research coordinators, and data management and analysis staff at the DSCC level. Regular calls and meetings were scheduled among these various groups to ensure ongoing and effective communication. In addition, the DSCC provided two reports to the site PIs; a quarterly progress report outlining each site’s progress on recruitment, data entry, and implementation and routinization of the data export program that provides the staff with an accurate understanding of their site’s performance; and a semi-annual benchmark report to enable the site PIs to assess and
review their individual clinic data, to have a synergistic view on quality of care provided at their site compared to the aggregate city-wide data, and to help identify patients who are out-of-care or have detectable viral loads. We plan to continue to distribute these valuable reports to the sites on a routine basis, and are currently considering how the generation and distribution of these reports can be automated to reduce workloads over time.

Development Methodology
In order to successfully coordinate the data export processes for the DC Cohort across multiple sites, an Agile software development methodology was implemented (Figure 2). Four different EMR platforms were used among the 13 sites: Allscripts, eClinical Works, GE Centricity and VISTA-CPRS (Table 1). A fifth EMR (Cerner) was recently implemented at two smaller sites at a single institution, and we are beginning the process of creating electronic exports for these sites as well. Though often the same platform was used at multiple sites, the EMR at individual sites varied due to site-specific modifications, customizations and implementation strategies among sites.26,27 There was also great variability in the length of time an EMR had been used at the various sites: some had EMRs in place for many years and others were still in the process of transitioning from paper records to an electronic platform. The EMR’s data hosting process also varied for many sites and includes cloud-based, client-hosted, and IT vendor-hosted systems.

Another major difference between sites was related to the IT personnel responsible for creating the export, as some sites had their own IT staff while others relied on IT subcontractors; all IT staff had competing clinical priorities and variable levels of familiarity and expertise with the process of developing export programs. Additionally, several sites changed their EMR platforms and/or their IT staff while the export programs were being implemented, creating further challenges and delays to the creation of a working export for the DC Cohort.

To address these challenges, the DSCC created an adaptive export specification document outlining all variables in the data abstraction process, the variable requirements, and the data quality standards. The DSCC also designed individual site management plans to implement technical specifications, fostered regular communication channels, and worked with the IT staff to iteratively refine the export program.
Addressing variability in site-specific data

The amount and types of clinical data that could be electronically abstracted varied greatly among sites (Table 1). Some clinical data elements were located in discrete fields within the EMR, while other data elements were found in text form, in administrative databases, or in pdf files that could not be exported electronically and therefore needed to be extracted manually. Also, EMR systems were dynamic and were changed as new data were added and systems were updated, so the location of data elements and the type of information available were not always consistent. The amount of data that were available to be extracted also differed between sites. For example, the EMR data for the clinics within hospitals were more extensive, as the hospital institutions provide care for multiple sub-specialties (e.g., cardiology, neurology). In contrast, community-based clinic sites do not generally provide sub-specialty care, so these data were not available to be included in the exports. Furthermore, there were differences in the data within sites; some clinicians provided only HIV-related care whereas other clinicians provided primary care in addition to HIV-related care. The quality and level of documentation between staff, as well as the historic data available on each patient, also varied.

To address these challenges, the DSCC focused on collecting a limited number of defined demographic, social, clinical, treatment, laboratory, and procedure variables deemed significant to the DC Cohort. Established data codes were used whenever possible, including the International Classification of Diseases 9th revision (ICD-9) codes for diagnoses, the National Drug Codes for medications and treatments, and standardized laboratory values. Because some data were not able to be exported, a combination approach to data entry was implemented. As shown in Table 1, at the time of enrollment, site RAs at all sites conducted medical record abstraction and manually entered data into a centralized web-based database (Discovere®, housed at the Cerner Corporation), selected baseline variables (e.g., antiretroviral therapy at consent date, selected co-morbidities, baseline CD4, and viral load). For sites where specific data elements could be extracted electronically in various modules (e.g., encounters, treatments, diagnoses, laboratory, pathology, and procedures), export programs were developed for these variables. Baseline data on demographic, social, and HIV risk factors and other HIV information (e.g., date of diagnosis, nadir CD4 count) were also entered manually into Discovere when necessary. Manual abstraction was performed to collect antiretroviral resistance profiles at all sites. At several sites, information on pathology and procedures were collected with both manual and electronic data abstraction. The typical workflow of participant data entry into Discovere is shown in Figure 3. Lastly, to improve the quality of the data, sites were sent monthly query lists for data that appeared to be missing, discrepant, or inaccurate that the site RA reviewed and resolved based on the data available in the EMR.

After the initial export program was completed, the DSCC continued to work with the site RAs and PIs and IT staff to refine the data export programs in order to produce more accurate data extractions and to...
address changes within EMRs. A representative from Cerner performed routine on-site audits at each site to confirm that study procedures were being followed correctly, and both manually entered and electronically exported data were being captured accurately and completely. The DSCC developed a clinical validation procedure of the export programs during which the electronically exported data for a sample of participants were compiled for each site, and the RAs compared the data received via the export to the source data in the EMR. The results of the validation were then used to refine the export programs as needed.

Addressing variability in site-level research experience

There was a considerable range of previous experience with human subjects research (e.g., patient consenting, Institutional Review Board (IRB) approval, maintaining research logs) across the sites, ranging from clinics with extensive research experience to those that had not previously participated in human subjects research prior to the DC Cohort. As a result, site initiation was strategically sequenced over three years, beginning with the more research-experienced sites, in order to permit the DSCC to gain experience with DC Cohort procedures, the creation of the export programs, and understanding the factors needed for a site to participate successfully in the Cohort. To ensure uniformity in manual data abstraction, the DSCC developed and deployed a manual of policies and procedures that is used for training all site RAs. In addition, the site RAs had monthly calls with the DSCC staff to review any data issues that had arisen, review new data queries that were implemented, and allow RAs to resolve questions related to data entry or data procedures. The site RAs were also provided with an annual refresher training on data entry and research procedures to ensure there was standardization among the RAs across the DC Cohort sites.

Concerns about Confidentiality

Based on discussions with CAB members, the primary concern of potential DC Cohort research participants was the confidentiality of their health information. This concern was addressed in several ways. First, all patients were consented for participation in the Cohort by site-based staff (RAs, nurses, or physicians, depending on the site) and informed that no identifying information (name, social security number, etc.) would be shared with anyone other than those persons who would ordinarily have access to their medical records, namely site-based staff and DC DOH HIV/AIDS surveillance personnel. Written consent for study participation was obtained and consent forms were stored at each site in secure and locked locations. The protocol was reviewed and approved by multiple IRBs including the GW University (which also served as the IRB of record for sites that did not have their own IRBs), the DC Department of Health, Georgetown University, Howard University, the Veterans Affairs Medical Center, and Washington Hospital Center. Moreover, the Discovere electronic data capture system was designed such that any personal identifiers (e.g., name, social security number, or medical record number) that were manually entered in order to register the participant were logically separated from the centralized database and were only visible to the staff at the site. During this registration process, Discovere created a system-generated DC Cohort Participant ID number that was used to uniquely identify participants in the DC Cohort database. The DC Cohort data were stored by Cerner in a remote host facility in Kansas City, Missouri that meets all HIPAA security guidelines. Access to both the export program and the data files produced were restricted to the site IT and site DC Cohort staff to further ensure that the participants’ confidentiality was maintained. Additionally, the DSCC had checks and systems in place to ensure that variables containing identifying information could not be electronically exported.
RESULTS
We have successfully facilitated the creation of a semi-automated electronic data abstraction system among the multiple sites participating in the DC Cohort study. Since enrollment began in 2011, the DC Cohort has approached more than 8700 patients and more than 7000 have been consented at 13 participating clinics, and longitudinal clinical data is now available on all participants. The numbers of patients approached for participation and consented over time is shown in Figure 4. Enrollment in the Cohort is dynamic and ongoing, with the sites offering enrollment to all eligible patients.

Development of Site-specific Export Programs
Clinical informatics experts (CIE) from Cerner worked with IT personnel from each site to abstract data electronically from each respective site’s EMR. The team worked with DC Cohort clinicians to develop data specifications to meet study requirements, and these specifications were then revised to ensure consistency across sites. Data specifications and data variable type and structural requirements were discussed with each individual site’s IT staff, and a data abstraction system was developed that was unique to each site yet with common data elements abstracted from all sites. The process of electronic data abstraction was conducted on a monthly basis whereby each site’s DC Cohort data were exported as text delimited files from the secure source system (EMR) to a secure, password-protected file transfer protocol (FTP) site. The Cerner personnel processed the files through an extract, transform, and load process to import the data into the Discovereelectronic data capture system, a proprietary web-based data entry system developed and supported by the Cerner Corporation. The use of Discovere has been effective, as it permits and facilitates the RAs at the individual site’s to make any needed manual changes or amendments to the data that have been electronically abstracted from that site. Once a baseline electronic data export was conducted for each participant (from date of consent forward), any new clinical data since the last data extraction were identified during the monthly export process. Through a “match and append” process that is based on the participant identification number, Discovere identified those data elements that were not present in the existing database and appended them to the DC Cohort database.

Technical Challenges
Several central technical challenges in data processing needed to be addressed to ensure the successful implementation of the DC Cohort. First, large amounts of EMR data needed to be extracted, uploaded monthly, and then merged with manually entered data, a process that became more complicated and time intensive as the overall size of the dataset grew. Second, for each data upload, procedures needed to be implemented to ensure that records for newly enrolled participants were correctly created and combined with records of those participants who had already been enrolled, and that any new data collected for individual participants who previously consented were appended to their previously created data record. Third, processes needed to be developed to ensure that data were clean and verified. To ensure data quality and verification of site-level data, clinical validation of electronically extracted data was conducted on a sample of 10 participants by the RA at each site before the routine electronic data abstraction program was initiated. Sites often went through several clinical validations in order to allow the export programs to be modified as needed until their accuracy was verified.

Quality Assurance
To ensure that consistent data would be available both for reporting back to the sites and for research analysis purposes, quality assurance, and data processing measures were performed before any DC Cohort data were analyzed. On a monthly basis, a copy of the DC Cohort database was extracted by the Cerner staff to create a limited dataset with no personal identifying information. The limited data undergo processing to create additional variables of analytic interest and to identify suspicious,
were audio-recorded with minutes maintained by GW). Problems were identified or modifications needed, these were dis-

of data from EMRs required constant vigilance and periodic modifica-

tions. The systematic collection of data from the exported data to the EMR itself by site-based staff for 10 randomly selected participants; and updates to the EMRs (e.g., the introduction of ICD-10 codes) and the Discovere system itself. When problems were identified or modifications needed, these were discussed on weekly IT “stand-up” calls (for which a written action items log was maintained by Cerner) and broader weekly DSCC calls (which were audio-recorded with minutes maintained by GW).

To quantify the magnitude of potential errors encountered for each module, we conducted an analysis of all queries that were attributed to incorrect data from June to August 2015. Overall, an average of 1099 queries were generated each month, with the majority of these in the treatment (22.9%), laboratory (21.8%), risk factors (20.6%), social factors (17.6%), and HIV information (13.8%) modules, with much smaller percentages in the pathology (1.6%), encounters (0.9%), demographic (0.5%), diagnosis (0.3%), and procedures (0.2%) modules.

Analysis and Reporting
On a quarterly basis, Cerner processed the data and sent a final version of the DC Cohort limited dataset as Statistical Analysis Software files to GW to be used for analysis and reporting. Routine reports were produced semi-annually for each site enabling each site PI to compare their participants’ socio-demographic information, clinical status, treatments, and outcomes to the aggregated summaries from all DC Cohort sites. In order to present the cleanest data possible and ensure there is adequate time for the site RAs to address data queries, data in these reports were censored at the cutoff date for the previous quarter. After the DC Cohort and DOH linkage was completed, these reports have been presenting summary data for enrolled participants both before and after the linkage. Additionally, the benchmark reports contain programmatic data for sites to use to assess the quality of HIV care being provided, including providing lists of PIDs that have detectable viral loads or have been out of care for >6 months.

In addition to receiving data in the semi-annual benchmark reports, site PIs may request data to be used in analyses for grant submissions, manuscripts, scientific conference abstracts, as well as for internal use (e.g., newsletters, internal reports). Lastly, the DSCC can provide analytic resources (scientific, statistical, and editorial assistance) for these analyses at the request of individual investigators.

DISCUSSION
The data collection system described in this report has a number of important technical strengths. The systematic collection of data from multiple sites and varying EMR systems that are integrated into a single database allows for congruity across data elements for large numbers of participants. This scalability allowed us to quickly deploy the data abstraction system to new sites, rapidly capture all data elements for consented patients, and adaptively change data requirements. The quality assessment by clinical validation of electronically abstracted data demonstrated the concordance of the DC Cohort database with the source data in the sites’ medical records. The electronic export programs allowed for reduction of error associated with manual data entry, which improves time and resource efficiency. Being a dynamic study, to facilitate novel developments in HIV care, electronic export of data from multiple EMRs gave us leverage to update data on specific data elements more efficiently, improving the scientific and analytical value of composite data for the entire study.

As an observational study of clinical care provided in an outpatient setting, there are a number of important technical limitations to the approach described as well. First, the data that are abstracted are limited to the discrete elements contained in the EMR, and consequently key information needed for research and programmatic efforts may sometimes be missing. These discrete elements can be augmented through manual data entry of elements that are captured in physician notes and other open text fields, however, this significantly increases the effort required by the on-site research staff. Second, technical challenges may limit collection of some data elements; and despite quality improvement measures, data queries, and clinical validation efforts, overall data quality cannot attain the same level available in other studies such as clinical trials. Third, changes in IT policies and procedures at individual participating sites can impact the data quality and the ability to extract all data elements of interest. For example, for selected diagnostic and clinical procedures, regulatory policies may require the results obtained to be in an electronically scanned document, rendering the data less accessible for electronic data export.

A critical advantage of this study is the ability to capture data from multiple sites and link participant information through the DC DOH HIV surveillance database. In this manner, participants who are lost to follow-up at one site and who may be seeking and receiving care at other sites can be identified and their clinical data from two or more sites merged, leading to greater completeness of the overall data set.

From a technical perspective, we have demonstrated that the merging of clinical data from disparate EMRs on a city-wide basis is feasible, but requires a long-term commitment and investment of human and fiscal resources. In addition to the central costs of the DSCC, each site received a fixed amount of support for the IT costs associated with developing and maintaining the electronic export system and for the effort of the Site PI, and variable support for research staff depending on the size of the HIV patient population and the amount of information that needed to be abstracted manually. Extensive and regular communication is critical, and the diverse expertise of IT specialists, clinicians, administrative staff, RAs, and epidemiologists and biostatisticians is required. The importance of limiting the collection of data elements to those that are most critical cannot be overstated; but this is balanced by the technical ease with which pre-coded and standardized variables, such as all ICD-9 and National Drug Codes, can be extracted.

We have also found that the analysis of the DC Cohort database for scientific purposes has been highly instructive. Although this present paper focuses on the data management and IT methods used to create the DC Cohort, we have begun to explore the clinical and epidemiological characteristics of this cohort as well. We have already conducted more than 20 separate site-specific and cohort-wide analyses on diverse topics such as disparities in viral suppression rates.
monitoring HIV care using national standards. 30 hepatitis C, 31 HIV drug resistance, chronic renal disease, retention in HIV care, statin use, sexually transmitted infections, and the impact of a patient navigation program. Examples of the types of issues that have been uncovered through the conduct of scientific analyses include 1) mapping errors, whereby different variables in the EMR were being exported to the same fields in Discovere; 2) log transformation errors, whereby viral load values reported in the EMR were being erroneously transformed when they were imported into Discovere; and 3) encounter errors, whereby information on patient visits that were not specifically for HIV care (e.g., dental visits) was being imported into the database. Once identified, these types of issues were addressed and resolved by Cerner and the Site IT staff working collaboratively to modify the electronic export programs.

Other valuable lessons learned include the importance of garnering the support and participation of numerous types of organizations. The successful implementation of the DC Cohort necessitated the collaboration of the federal government National Institutes of Health (NIH) that provided scientific oversight and fiscal support; local government District of Columbia Department of Health (DC DOH) that provided convening power, established the DC Cohort as a local public health priority, and facilitated access to local surveillance data; academic institutions that provided access to investigators, university-based clinical databases, and human subjects protection oversight; community clinics that provided access to investigators, clinical data on large numbers of participants, many of whom hail from underserved and diverse communities, and critical community support and logistic expertise for the Cohort; IT organizations, both those internal to the clinic sites and those that were subcontracted to develop and maintain the export programs, and the Cerner Corporation itself; and the unique patient CAB, which provided critical input into study design, enrollment practices, and the dissemination of Cohort-related data.

CONCLUSION

In summary, we have successfully developed and implemented a comprehensive electronic data capture system that should provide clinical, epidemiologic, and scientific research benefits for years to come. Clinically, investigators can monitor patient outcomes at their sites and compare their patients to a representative sample of persons receiving HIV care in Washington, DC, as reports are provided regularly summarizing key outcomes (such as laboratory values and antiretroviral regimens) that can contribute to monitoring changing practice patterns. Epidemiologically, the DC Cohort data set is being analyzed to assess numerous issues such as viral suppression rates, retention in care, and gender, race, and age-related disparities in HIV-related comorbidities and care. Lastly, the DC Cohort database is now being used as a scientific platform from which to launch related scientific investigations, including treatment programs for hepatitis C infection and studying the evolving molecular epidemiology of HIV in Washington, DC. The identification of numerous technical challenges and the development of a myriad of innovative solutions to address these challenges were the key elements to ensuring the development of this unique public health population-based cohort.

FUNDING

This research was supported by the National Institute of Allergy and Infectious Diseases at the National Institutes of Health (Grant UM1AI069503). This publication was facilitated (in part) by the infrastructure and services provided by the District of Columbia Center for AIDS Research, an NIH funded program (P30AI117970), which is supported by the following NIH Co-Funding and Participating Institutes and Centers: NIAID, NCI, NICHD, NHBLI, NIDA, NIMH, NIA, FIC, NIGMS, NIDDK, and OAR. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

ACKNOWLEDGEMENTS

The authors would like to thank the many Information Technology specialists who contributed to the development of the electronic export programs at the participating DC Cohort sites, including Michael Smoot (DC Primary Care Association, Washington, DC), Stephen Rader and Ankush Oberoi (Unity Health Care, Washington, DC), Robert Williamson (Veterans Affairs Medical Center, Washington, DC), Aamer Ahmed (Medstar Health, Washington, DC), Maneesh Ponnu (George Washington Medical Faculty Associates, Washington, DC), Sheelat Subbanwad (EClinical Works), and Brandon Smith (CHMB Inc). We would also like to thank the site PIs, RAs, the CAB, the patients themselves, the DC Department of Health and the National Institutes of Health for their numerous contributions to the DC Cohort.

*DC Cohort Executive Committee: Debra Benator, MD, Veterans Affairs Medical Center; Princy Kumar, MD, Georgetown University; Richard Elion, MD, Whitman-Walker Health; David Parenti, MD, George Washington Medical Faculty Associates; Maria Elena Ruiz, MD Washington Hospital Center; Angela Wood, MSW, Family and Medical Counseling Service; Lawrence D’Angelo, MD, MPH, Burgess Adolescent Clinic, Children’s National Medical Center; Natella Rakhimanina, MD, PhD, Special Immunology Service Pediatric Clinic, Children’s National Medical Center; Sohail Rana, MD, Pediatric Clinic, Howard University Hospital; Maya Bryant, MD, Adult Infectious Disease Clinic, Howard University Hospital; Annick Hebou, MD, MetroHealth; Ricardo Fernandez, MD, La Clinica Del Pueblo; Stephen Abbott, MD, Unity Health Care; Michael Kharfen, BA, HIV/AIDS, Hepatitis, Sexually Transmitted Diseases, Tuberculosis Administration (HAHSTA), DC Department of Health; Henry Masur, MD, National Institutes of Health.

REFERENCES

9. Stewart M, Amardeep T, Terry A, Chevendra V, Marshall J. Implementing and maintaining a researchable database from electronic medical records: a per-


AUTHOR AFFILIATIONS

1Department of Epidemiology and Biostatistics, Milken Institute School of Public Health, The George Washington University, Washington, DC, USA.

2Cerner Corporation, Kansas City, Missouri, USA.


