Article published online: 2021-06-30

The Cosmos Collaborative: A Vendor-Facilitated Electronic Health Record Data Aggregation Platform

Yasir Tarabichi[1,2,3] Adam Frees[4] Steven Honeywell[4] Courtney Huang[4] Andrew M. Naidech[5] Jason H. Moore[6] David C. Kaelber[1,7]

[1] Center for Clinical Informatics Research and Education, The MetroHealth System, Cleveland, Ohio, United States
[2] Division of Pulmonary and Critical Care Medicine, Department of Internal Medicine, The MetroHealth System, Cleveland, Ohio, United States
[3] School of Medicine, Case Western Reserve University, Cleveland, Ohio, United States
[4] Epic, Verona, Wisconsin, United States
[5] Department of Neurology, Northwestern University. Chicago, Illinois, United States
[6] Institute for Biomedical Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, United States
[7] Departments of Internal Medicine, Pediatrics, and Population and Quantitative Health Sciences, School of Medicine, Case Western Reserve University, Cleveland, Ohio, United States

Address for correspondence Yasir Tarabichi, MD, MSCR, Center for Clinical Informatics Research and Education, The MetroHealth Medical System, 2500 MetroHealth Drive, Cleveland, OH 44109, United States (e-mail: yxt277@case.edu).

ACI Open 2021;5:e36–e46.

Abstract

Objective Learning healthcare systems use routinely collected data to generate new evidence that informs future practice. While implementing an electronic health record (EHR) system can facilitate this goal for individual institutions, meaningfully aggregating data from multiple institutions can be more empowering. Cosmos is a cross-institution, single EHR vendor-facilitated data aggregation tool. This work aims to describe the initiative and illustrate its potential utility through several use cases.

Methods Cosmos is designed to scale rapidly by leveraging preexisting agreements, clinical health information exchange networks, and data standards. Data are stored centrally as a limited dataset, but the customer facing query tool limits results to prevent patient reidentification.

Results In 2 years, Cosmos grew to contain EHR data of more than 60 million patients. We present practical examples illustrating how Cosmos could further efforts in chronic disease surveillance (asthma and obesity), syndromic surveillance (seasonal influenza and the 2019 novel coronavirus), immunization adherence and adverse event reporting (human papilloma virus and measles, mumps, rubella, and varicella vaccination), and health services research (antibiotic usage for upper respiratory infection).

Discussion A low barrier of entry for Cosmos allows for the rapid accumulation of multi-institutional and mostly de-duplicated EHR data to power research and quality

Keywords
- electronic health record
- data aggregation
- research network
- health information exchange
- collaboration

received
February 14, 2021
accepted after revision
April 13, 2021

DOI https://doi.org/10.1055/s-0041-1731004.
ISSN 2566-9346.

Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

improvement queries characteristic of learning healthcare systems. Limitations are being vendor-specific, an "all or none" contribution model, and the lack of control over queries run on an institution's healthcare data.

Conclusion Cosmos provides a model for within-vendor data standardization and aggregation and a steppingstone for broader intervendor interoperability.

Introduction

Facilitated by the introduction of the Health Information Technology for Economic and Clinical Health Act of 2009, electronic health records (EHRs) have become ubiquitous across the United States.[1,2] In digitizing paper records and processes, healthcare systems gained the potential for immediate access to the data that they needed to analyze and refine their practices and improve their outcomes. This step is arguably an essential one for the development of a true "learning healthcare system."[3,4]

A natural extension of the learning healthcare system framework involves leveraging the collective experiences of numerous healthcare systems through multisite collaborations. This is the driving principle behind several successful quality initiatives, such as the American College of Surgeon's National Surgical Quality Improvement Program (NSQIP).[5–7] Unfortunately, EHR data tend to be siloed within institutions, and efforts for wider intersystem aggregation are often hampered by regulatory, technical, and financial barriers.[8] Despite these limitations, many initiatives have been successful, including several national research collaboratives,[9–11] surveillance networks,[12] and regional public health initiatives.[13,14]

Tarabichi et al recently described a federated, vendor-facilitated (Epic, Verona, Wisconsin, United States) EHR data aggregation initiative known as the Aggregate Data Program (ADP).[15] The ADP was a proof-of-concept disease-specific registry that reduced the barrier of entry to collaborators by leveraging native EHR tools for the periodic submission of aggregated EHR data to a central repository. The success of that initiative laid the groundwork for Cosmos. Like the ADP, Cosmos is vendor-facilitated with robust customer input into its design and implementation. Cosmos goes further by leveraging standard health information exchange infrastructures to continuously and automatically retrieve, harmonize, and collate a greater variety of discrete data points from participating organizations. In addition, Cosmos empowers its contributors with a web-based query building interface that allows users to go beyond a priori determined questions available in the ADP. Here, we describe Cosmos as it exists at the time of publication and provide examples for how the data and platform may be used to further public health surveillance, quality improvement, and research initiatives. This manuscript is intended to be the first formal description of this initiative, and these use cases were selected to demonstrate Cosmos' structure, functions, and capabilities.

Methods

Program Governance and Structure

Cosmos is managed by the Epic corporation (Verona, Wisconsin, United States), with guidance from elected representatives of their customer community (the Governing Council). The council currently has 11 members, consisting of executives, researchers, and clinicians from the organizations that participate in Cosmos. Council terms are 3 years, and members cannot serve consecutive terms but can be renominated after one election cycle. The council is responsible for promoting best practices and advising the vendor on the direction of the collaboration.

Cosmos is an opt-in service for Epic EHR customers. To participate, organizations must agree to the Cosmos guidelines (called the Rules of the Road). These guidelines are codeveloped by the Epic corporation and the Governing Council, and enforcement is ensured by both entities. The guidelines are not publicly published but are made available to all users of the Epic EHR platform.

Data Structure and Quality Control

Cosmos contains a variety of data points per patient, spanning many discrete data variable types (►Table 1). While Cosmos does use Epic's own proprietary data model to store the data, it favors direct linkage to standardized ontologies over custom ones, relying mostly on Uniform Medical Language System data models such as Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT), Logical Observation Identifiers Names and Codes (LOINC), RxNorm (standardized nomenclature for drugs in the United States), and CVX (vaccine administration codes).[16] Other data models leveraged include the National Uniform Claim Committee health care provider taxonomy,[17] National Uniform Billing Committee discharge codes (NUBC FL 17),[18] and International Classification of Disease revisions 9 and 10.[19] The majority of data transmitted to Cosmos must conform to the aforementioned ontologies before transmission. In rare circumstances, Cosmos curates additional nonstandard data when the common standards are deemed insufficient, such as in the case of documentation of birth control classification methodologies not well characterized in SNOMED-CT.

For most items, Cosmos uses the same mapping process that Epic customers need to complete for standard clinical health information exchange. Most institutions will have adopted their native data models to reflect and/or link to standard ontologies at the outset, but some additional mapping or manual corrections may need to be addressed by individual sites before data submission. A small subset of

Table 1 Cosmos data variables as of June 2020

Concept	Discrete Data Variables
Demographics	Legal sex; gender identity; birth date; race; ethnicity; zip, county and state of patient; date of death; status of patient (alive or deceased); cause of death; gestational age at birth; language (spoken and preferred)
Encounter details	Start/end date and time; type, specialty; reason for visit; age at encounter; pregnancy status at encounter; place of service (zip, county and state); mode of arrival; discharge disposition; organization type
Problems	Diagnosis, including date noted and resolved
Diagnoses	Encounter based admission and discharge diagnoses; surgical diagnoses; visit (encounter) diagnoses; billing diagnoses
Surgical history	Procedure, date/time
Social history	Smoking status, duration and intensity; smoking start/stop dates; sexual activity, alcohol usage status; illegal drug usage status
Family history	Problem or pertinent negative; relationship to patient, age of onset, sex and status (living or deceased)
Outpatient medications	Medication name, type, dose, unit, route, frequency, dispense quantity, refills, and start/end date; indications of use
Allergies	Date noted; allergen; reaction; reaction severity; last updated instance
Immunizations	Immunization; administration date; route, dose; unit
Vital signs	Date/time; blood pressure; pulse; temperature; respiratory rate; oxygen saturation; height; weight; body mass index; head circumference.
Results	Procedure; date/time; specimen source; value and units; abnormal flag; reference range Microbiology organism, sensitivity and testing method if applicable
Procedure	Start/end date; procedure instant; billed procedure; provider specialty
Inpatient medications	Medication name, type, dose, unit, route, and start/end date
Birth data	APGAR score at 1, 5, and 10 min; nourishment method; delivery method; hospital days; birth count and order (if multiple)
Social determinants of health	Social connections; physical activity, stress; education; food insecurity, financial resource strain; intimate partner violence
Insurance	Medicaid, Medicare, privately insured or self-insured status

Note: Variables are grouped by concept.
Abbreviation: APGAR, appearance, pulse, grimace, activity, and respiration.

nonstandardized data types, such as race, ethnicity, and reason for visit, are mapped by the vendor.

Once an institution agrees to contribute to Cosmos, the vendor provides a feedback loop between Cosmos and its contributors through periodic data quality reports. The reports include metrics on mapping completeness, identifying most frequently received unmapped values, as well as potential data irregularities, including date outliers, laboratory results with missing units or the reception of low rates of documentation of an important birth metric. Data completeness for important variables is considered and scored, dependent on the variable type. Anomalies in the frequency of data variables are monitored in longitudinal fashion, with attention to large relative changes in count data. Such changes are flagged and prompt manual review to determine if they are expected (such as increases in influenza vaccination rates in the fall). The vendor also assesses laboratory data distributions to detect potentially incorrect LOINC mapping.

Data Submissions and Triggers

The data submitted to Cosmos can be divided into two broad categories: "backload" data and "triggered" data. Backload data consists of records that existed prior to an organization's involvement in Cosmos. Triggered data are prospectively accumulated and submitted to Cosmos based on event-driven triggers, such as encounter closure, a result being filed, or a chart being corrected. Encounter record submissions are triggered by 7 days of inaction, even if the encounter has not been closed. This prevents the delay of transmission of discrete objective data, such as completed laboratory results, due to incomplete documentation which would not otherwise be transmitted to Cosmos.

Backload and triggered data are prioritized for submission via the Cosmos Queue (►Fig. 1). Priority is given to more recent events. To alleviate computational strain on both the submitting organization and on Cosmos, the backload is often performed in stages, limiting submissions to the most recent few years. Background processes advance

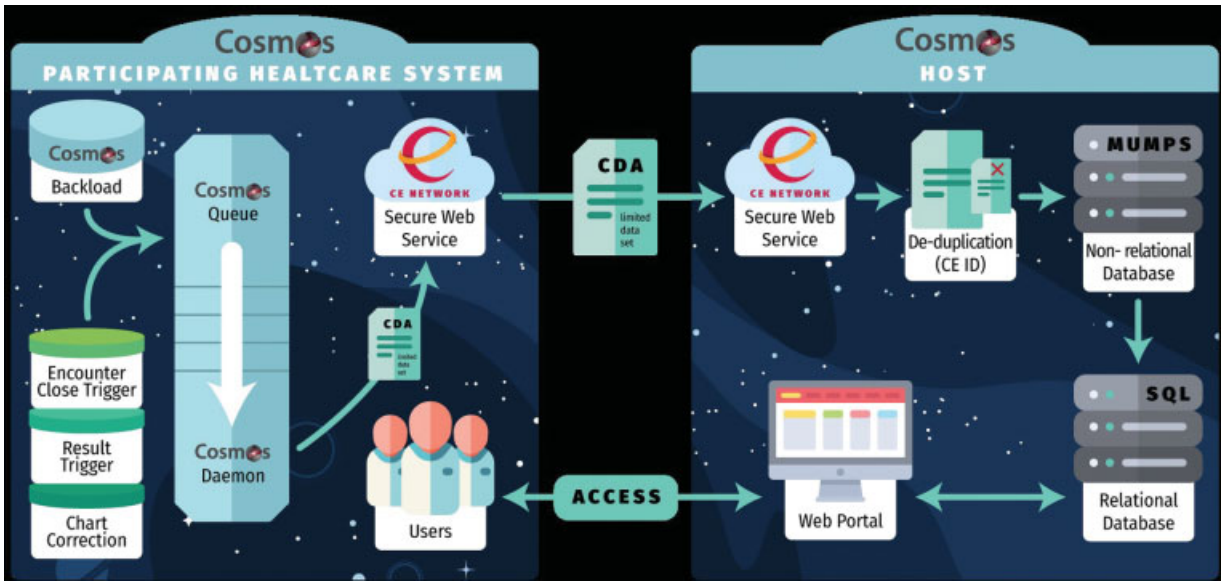

Fig. 1 Schematic for the Cosmos architecture. Backload and triggered data move onto the Cosmos queue, where it is processed by the Cosmos daemon. Data are transmitted to the Cosmos host as encrypted HL7 C-CDA documents over the Care Everywhere Network. Patient de-duplication is performed by using a hashed copy of a Care Everywhere ID, after which data are filed in a Massachusetts General Hospital Utility Multi-Programming System nonrelational database, and then a search query language relational database. All participating healthcare systems communicate with the same Cosmos host. Users can access the data via a web portal.

through the queue and prepare each record for transmission to Cosmos.

Data Privacy and Transmission

Patient privacy and data security are core concerns for the EHR vendor and all contributing customers. Data are transmitted to Cosmos through encrypted health level 7 Consolidated-Clinical Document Architecture (C-CDA) documents over a secure existing clinical health information exchange platform (the Care Everywhere network). Care Everywhere is a "point-to-point" or nonfederated peer-to-peer network health information exchange mechanism, which has been used since 2008 to transmit hundreds of millions of patients' charts between organizations that use the Epic EHR for clinical care purposes (►Fig. 1).[20]

Cosmos contains a limited dataset, as defined by the Health Insurance Portability and Accountability Act of 1996.[21] The initiative has been designed with safeguards to prevent submission of protected health information, with exceptions including dates of birth, dates of service/testing, 5 digit zip code, and a unique internal identifier (Care Everywhere [CE] ID). The CE ID is used to identify the same patient across multiple EHR instances and does not encompass any patient information.[22] Before being transmitted to Cosmos, the CE ID is hashed via the SHA-256 cryptographic function so that it cannot be used to reveal a patient's identity.[23] Studies have shown that the CE ID correctly identifies the same patient in different EHRs at least approximately 85% of the time.[22,24] A more recent study in Los Angeles revealed no false positives and an estimated false negative rate of close to 3%.[25] By leveraging the CE ID, most information about the same patient in multiple healthcare systems contributing to Cosmos is combined into a single patient record. This reduces double counting, and the provides a more temporally complete patient record when care is fractured between different institutions.

Free text data from notes or comments are not submitted to Cosmos. Free text data from laboratory results, however, are submitted after passing a strict inclusion list filter.

Data Access

Individuals from healthcare systems contributing data to Cosmos can query data in Cosmos through a secure web application. The web application leverages a graphical user interface that allows users to build modular queries without writing any code. Queries return cohort counts or summary measures (such as minimum, maximum, mean, or standard deviations of included continuous variables). Users can use Boolean logic to combine their criteria and place relative temporal restrictions on queries. For example, a query could retrieve the average age and the number of patients who received a certain vaccine and received a follow-up booster within 12 months. Because Cosmos only returns population level data and obscures counts <11 patients (only reported as "10 or fewer"), data returned from Cosmos queries do not constitute human subjects research. As a result, end-users do not require institutional review board approval for research purposes.

Current State of the Registry

Cosmos has been accepting data since 2018, with historical (backloaded) data extending as far back as 2005. As of August 2020, Cosmos has data from more than 60 million unique patients (►Fig. 2A), with representation from all 50 states. These contributions come from 75 participating sites (25 academic medical centers, 50 nonacademic medical centers, and 5 children's hospitals).

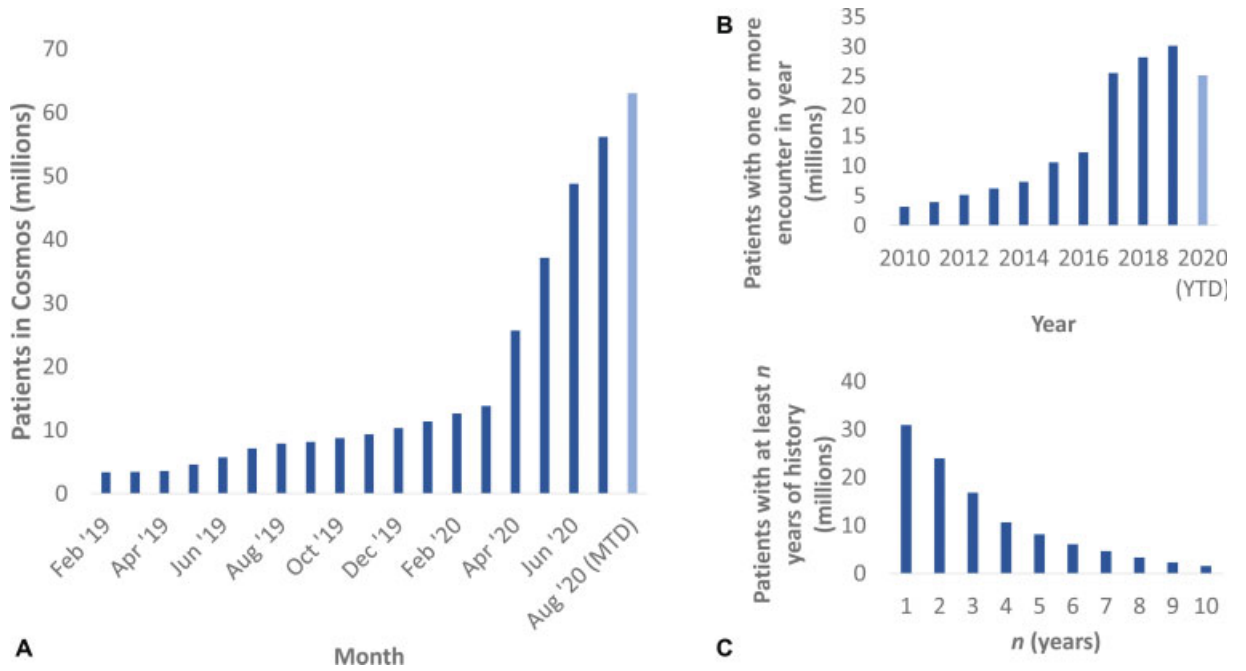

Fig. 2 Cosmos characteristics as of August 2020. (**A**) Cumulative number of unique patients in Cosmos as a function of time. (**B**) Number of unique patients with an encounter in Cosmos by year. (**C**) Length of time between first and latest encounter in Cosmos per unique patient. To generate this query, available laboratory results that included "influenza" or "severe acute respiratory syndrome" in their titles were screened to determine a rapid diagnostic test, as opposed to an antibody study (the resulting Logical Observation Identifiers Names and Codes are noted in the supplementary material). MTD, month to date; YTD, year to date.

The backload submissions for many contributing organizations are ongoing, and therefore, most of the records in Cosmos reflect recent information. Despite this, for each year in the past decade Cosmos contains encounter records from that year for millions of patients (►Fig. 2B). Additionally, over 15 million patients have at least 3 years of medical history in Cosmos, and over 1 million have at least 10 years (►Fig. 2C).

Statistical Approach to Sample Use Cases

Raw count data for the use cases were obtained directly from the native Cosmos query interface. Measures of prevalence are limited by EHR documentation completion. Descriptive statistics were provided for most measures and comparisons to alternative data sources were qualitative. Confidence intervals were calculated by using the binomial exact method, and proportions compared with Chi-square testing where applicable. All analyses were conducted in R (version 3.5.1) and figures generated with ggplot2.[26,27]

No institutional review board approval was needed due to the aggregate de-identified nature of the data that was accessed. This publication and its contents were approved by the Cosmos Governing Council.

Results

Chronic Disease Surveillance: Asthma and Obesity

One of the major functions of the Centers for Disease Control (CDC) is to measure and monitor important public health trends. The intersection of asthma and obesity, for instance, is of recent interest.[28] Cosmos enables combining elements of administrative data, vital signs, and demographics to study EHR asthma prevalence, and the likelihood of a clinically noted exacerbation along strata of sex and body mass index (BMI) (►Fig. 3). The ability to query vital sign data such as BMI makes the latter assessment more reliable than relying on diagnosis data alone.[29,30] In our analysis for the year of 2019, the prevalence of asthma was significantly greater in morbidly obese woman than morbidly obese men (14.0 vs. 8.1%, $p < 0.001$), consistent with recent data from the CDC.[28] In addition, our data show that morbidly obese asthmatic women were significantly more likely to experience clinically significant exacerbations compared with morbidly obese asthmatic men (21.7 vs. 18.9%, $p < 0.001$), mirroring evidence from a growing literature describing this phenomenon.[31,32]

Syndromic Surveillance: Seasonal Influenza and the Novel Coronavirus

The CDC collects frequent data on positive influenza testing nationwide, providing critical epidemiologic information to public health and healthcare officials every season.[33] A simple query of available laboratory data in Cosmos reveals distribution patterns of influenza A and B subtypes during the 2019 to 2020 flu season (►Fig. 4). Adding positive testing for the severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) reveals the timing of this pandemic at the end of the typical flu season. Similar to CDC findings, Cosmos data revealed that the 2019 season began with an atypical preponderance of the influenza B subtype, later superseded by the A sub-type, and then a rapid rise in SARS-CoV-2

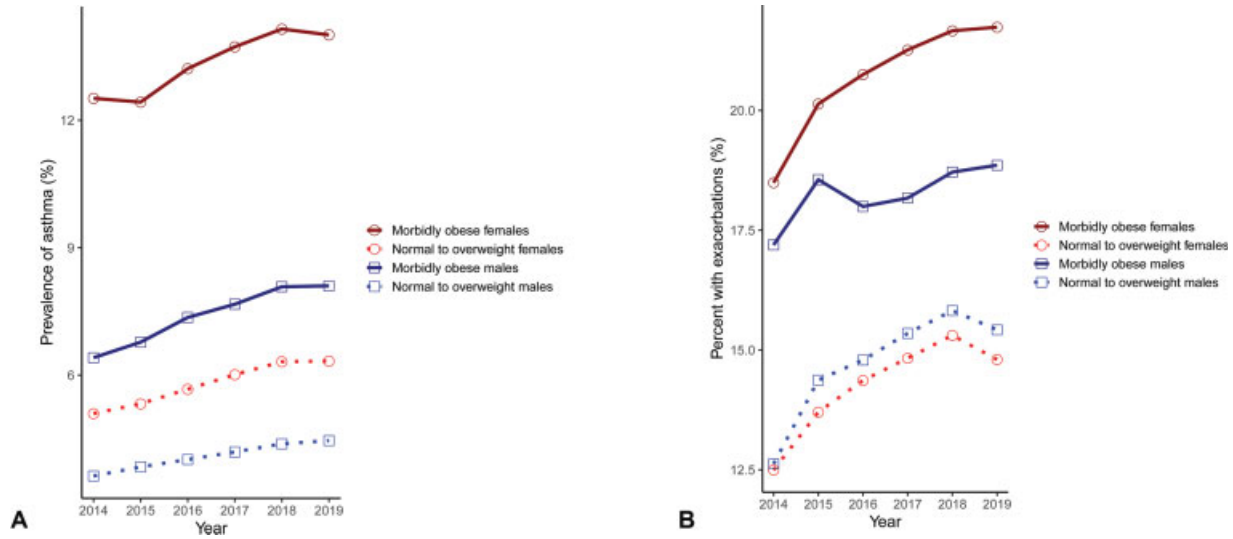

Fig. 3 Relationships between asthma and obesity in Cosmos. (**A**) The prevalence of asthma within different obesity classes, stratified by sex. (**B**) The percentage of asthmatics who have at least one encounter with a SNOMED diagnosis of asthma exacerbation, stratified by sex, and weight class. Annual asthma prevalence was defined as an encounter or problem list diagnosis during that year that mapped to a SNOMED diagnosis of asthma (SNOMED-CT 195967001). Asthma exacerbations were indicated by the presence of an encounter or problem list diagnosis that mapped to the SNOMED "exacerbation of asthma" concept (SNOMED-CT 281239006). Normal to overweight was defined as a BMI <30 kg/m^2, obese as a BMI of 30 to <40 kg/m^2, and morbidly obese as a BMI ≥ 40 kg/m^2 during each calendar year. BMI, body mass index; SNOMED-CT, Systematized Nomenclature of Medicine-Clinical Term.

positivity (►Fig. 4). As Cosmos collects patient zip code, the SARS-CoV-2 pandemic can be monitored both temporally and spatially at the national, state, county, and zip-code level. An assessment of the weekly number of patients testing positive for SARS-CoV-2 in four geographically distant states demonstrates the power of such an approach (►Fig. 5).

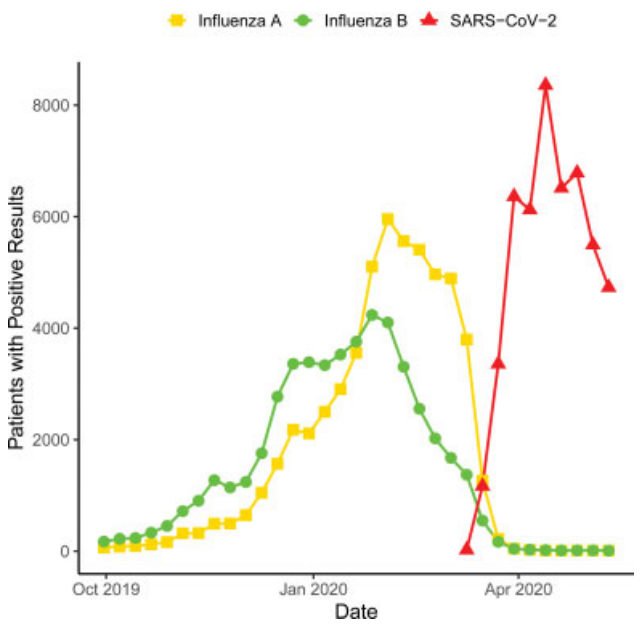

Fig. 4 Counts of positive influenza A and B assays, as well as severe acute respiratory syndrome coronavirus-2 assays in Cosmos per week. The leveraged Logical Observation Identifiers Names and Codes are noted in ►Appendix A. As noted in the text, any counts under 10 (including 0) are obscured by rounding up to 10.

Immunization Utilization and Adherence Reporting: HPV Vaccination Adherence

While Cosmos can easily retrieve vaccination rates among different demographic cohorts, the availability of temporal inclusion operators can create more meaningful queries for vaccine series adherence (►Table 2). For analysis on human papilloma virus (HPV) vaccination adherence, we queried initial and follow-up vaccination rates among adolescents with at least one outpatient visit. Our data revealed higher

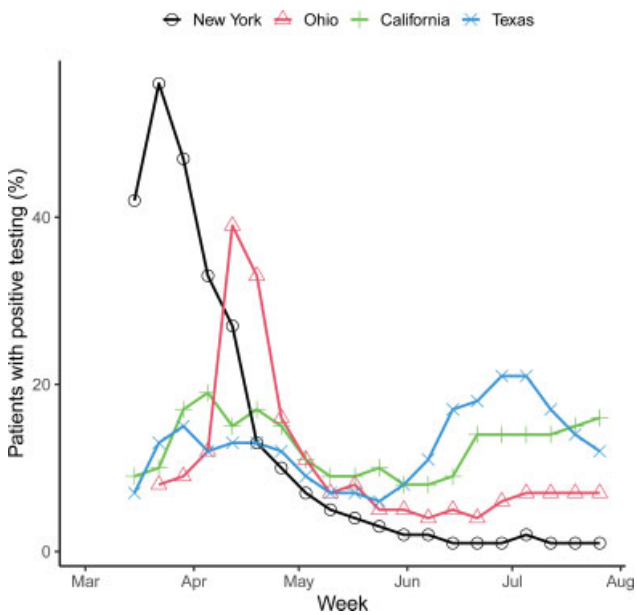

Fig. 5 Weekly proportion of patients testing positive for severe acute respiratory syndrome coronavirus-2 in four selected states. The leveraged Logical Observation Identifiers Names and Codes are noted in ►Appendix A.

Table 2 Human papilloma virus vaccination and completion rates in patients between the ages of 9 and 14, stratified by race

Race	Total eligible[a]	HPV vaccine[b] at least once	Patients vaccinated once who completed series within 12 mo
White	431,393	109,422 (25.4%)	62,918 (57.5%)
Black or African American	93,290	37,784 (40.5%)	15,302 (40.5%)
Asian	14,075	4,705 (33.4%)	2,521 (53.6%)
American Indian	3,063	898 (29.3%)	518 (57.7%)
Native Hawaiian	1,515	421 (27.9%)	230 (54.6%)

Abbreviation: HPV, Human papilloma virus.
[a]Eligibility was defined as any patient with a documented racial identity having at least one outpatient encounter between the ages of 9 and 14, between January 1, 2014 and December 31, 2019.
[b]Vaccination included bivalent HPV vaccination (CVX 118), quadrivalent HPV vaccination (CVX 62), 9-valent HPV vaccination (CVX 165), and "unspecified" HPV (CVX 137) vaccination.

first dose adherence in 9- to 14-year-old Black patients compared with White patients (40.5 vs. 25.4%, $p < 0.001$), but poorer two-series vaccine completion rates in Black compared with White patients (40.5 vs. 57.5%, $p < 0.001$). These disparities are consistent with others' findings.[34]

Vaccine Adverse Event Reporting: MMRV and Febrile Seizures

The Vaccine Safety Datalink (VSD) is a valuable resource for vaccine related adverse event monitoring.[35] Data from the VSD showed that children receiving the measles-mumps-rubella-varicella (MMRV) vaccine had a greater risk of febrile illness compared with measles-mumps-rubella with a separate varicella vaccine (MMR + varicella; 4.3 cases per 10,000 doses).[36] Using Cosmos, we queried 11 to 23 months old who received either an MMRV (CVX 03) or MMR vaccine (CVX 94), in the past 10 years, then limited that cohort to patients who had an encounter or problem list diagnosis of febrile convulsion (SNOMED-CT 41497008) within 10 days of vaccination. In total, 55 of the 110,644 (0.049% or 5.0 per 10,000) receiving the MMRV had a febrile convulsion within 10 days, as opposed to 312 of the 1,041,705 (0.030% or 3.0 per 10,000) receiving the MMR. This represented a statistically significant excess risk of 2.0 per 10,000 doses ($p < 0.001$), in keeping with the published literature.[36]

Health Services Research: Antibiotic Usage for Upper Respiratory Infections

The presence of encounter and medication prescription details in Cosmos allows for a wide range of queries to understand and quantitate the extent of important health services practices, such as potentially inappropriate use of antibiotics for upper respiratory infections (URI).[37,38] In the emergency department (ED) setting specifically, antibiotic prescription frequency for URIs has decreased over time at the national level.[39] We queried the number of patients per year who had an ED visit with an encounter diagnosis of URI (SNOMED-CT 54150009), and then stratified the result by age (<18 and ≥18) and any antibiotic prescription during that encounter. However, 2014 to 2017 estimates in adults and children fell within the confidence intervals of National Hospital Ambulatory Medical Care Survey (NHAMCS) data (►Table 3).[39] Though NHAMCS estimates were limited to data from 2017, we extended our query into 2019, revealing a

Table 3 Number of patients seen at least once in an emergency room setting for a diagnosis of "upper respiratory infection," with the number and percentage of those receiving an antibiotic during the encounter

	2010–2013	2014–2017	2018–2019
Age 18+			
All patients with ED visit for URI	18,881	95,598	179,140
Number of patients with ED visits for URI with antibiotic ordering (%, 95% confidence interval)	8,694 (46.1%; 45.3–46.8)	27,001 (28.2%; 28.0–28.5)[a]	38,327 (21.3%; 21.2–21.6)
Age < 18			
All patients with ED visit for URI	70,499	244,563	333,714
Number of patients with ED visits for URI with antibiotic ordering (%, 95% confidence interval)	13,377 (19.0%; 18.7–19.3)	30,826 (12.6%; 12.5–12.7)[b]	32,482 (9.7%; 9.7–9.9)

Abbreviations: CI, confidence interval; ED, emergency department; NMHACS, National Hospital Ambulatory Medical Care Survey; URI, upper respiratory tract infection.
[a]Comparable estimates based on NHAMCS data are 32.0 (95% CI: 22.0–43.5).
[b]Comparable estimates based on NHAMCS data are 10.1 (95% CI: 7.4–13.9).
Note: Antibiotic usage was based on 1,861 RxNorm codes that code for antibiotics (noted in ►supplementary material).

continuing reduction in the percentage of patients receiving an antibiotic prescription frequency for URI over time in both adults (28.2 vs. 21.3, $p < 0.001$) and children (12.6 vs. 9.7, $p < 0.001$).

Discussion

Cosmos is a rapidly growing EHR vendor-facilitated data collaboration that has accumulated data from over 60 million patients in under 2 years. The collaborative achieves a low barrier of entry for eligible healthcare facilities by leveraging existing clinical health information exchange infrastructure and data standards. Cosmos empowers customers with a user-friendly, flexible query building interface to generate new knowledge and insights.

Through several basic examples, we have shown how an intraplatform collaboration can potentially facilitate the evolution of institutions in their journey toward the goal of becoming learning healthcare systems. In mapping their data for submission to Cosmos, healthcare systems continue to standardize their data into common formats. Doing so allows for the meaningful aggregation of their data with other healthcare systems, and the ability to contrast their practices to those from other institutions. Such cross-institution data aggregation is a cornerstone of the success of other data collaboratives such as the NSQIP or the Mini-Sentinel initiative. In the case of NSQIP, the program lead to substantial improvements in surgical outcomes within the Veteran's Association (VA) system, which lead to its rapid and international growth.[40] With Mini-Sentinel, the volume of data needed to conduct certain analyses, such as assessing the risk of angioedema from different antihypertensive drugs, could only be realized with multi-institution aggregated health data.[41] The current implementation of Cosmos is focused on data harmonization and aggregation, but the query and analytics tools are being rapidly evolved to promote similar initiatives.

The use cases presented serve to further broad initiatives such as chronic disease surveillance, syndromic surveillance, immunization adherence, and adverse event reporting and health services research. While many more advanced queries and research questions are possible, we elected to demonstrate simple queries with historical precedent as a proof of concept to focus our discussion on the collaboration itself. The queries presented can be modified or extrapolated to generate new insights necessary for learning healthcare systems to evolve. Chronic diseases surveillance can encompass myriad conditions beyond asthma or obesity. Vaccination adherence and adverse event monitoring can be extended to novel vaccines and vaccine formulations. Health disparities can be revealed and monitored, as demonstrated in our HPV vaccine example. New outbreaks, such as COVID-19, can be assessed nationwide. Finally, healthcare systems or even national organizations can implement practice changing education and processes that can be monitored at a national level, as suggested by antibiotic utilization in URIs example. Most of the use cases presented benefit from a greater power and generalizability achieved with larger datasets comprised of pooled data from different sources. The MMRV case use, for instance, measures a rare event that may not be easily captured a using one institution's experience. The SARS-CoV-2 case also highlights interregional analyses that could not otherwise be done without multiregional collaboration.

Cosmos' breadth of data types and filters allows for more nuanced cohort identification. Unlike claims-based systems such as Marketscan, Cosmos can leverage more than diagnoses for better disease phenotyping.[42] This was broadly demonstrated in our obesity and asthma example, where reliance on diagnoses without measured BMI would greatly underestimate obesity. Many of the subject matter expert crafted "EHR phenotypes," such as those from eMERGE consortium, could be reproduced by leveraging the multitude of data types and temporal parameters in Cosmos' native query interface.[43] Importantly, all Cosmos users have access to the same query building platform (called SlicerDicer) that they can use with their own identifiable data. This allows institutions to both generate and validate criteria-based phenotyping locally before adapting them to the larger, de-identified Cosmos dataset.

By design, Cosmos is an opt-in, all or none service for users of the same EHR platform. Since the initiative leverages preexisting technology infrastructures and health information exchange standards, customers need only agree to contribute data, and continue mapping their data items to established standards if they have not already done so. While several similar nationwide data aggregation collaborations exist, many require the involvement of separate third parties, and often, the need for an appliance behind a customer's firewall.[9,11] Such arrangements require new relationships, data-use agreements, and software or hardware implementations (with their associated costs) that may be a barrier to entry for some healthcare systems. Even regional networks such as the MDPHnet and NYC Macroscope require a great deal of external funding and multilevel collaborations that may not be available or accessible everywhere.[13,44–47] To maintain accessibility and ensure Cosmos' success, the Epic Corporation does not currently charge its customers any additional costs for participating in Cosmos.

A potential downside of an "all or none" collaboration from the contributing healthcare system's perspective is concern regarding the loss of control over data contributed. Unlike other collaborations such as the Mini-Sentinel program or PCORnet, data contributions cannot be partial, and queries do not require approval at each site.[48] To address this concern, contributions of a single patient or institution cannot be identified in Cosmos. This is done primarily by withholding line level data and limiting query results to >10 patients. Another potential concern for Cosmos contributors is the housing of an institution's health information by the vendor in a central repository. To address this, the data transfer and storage processes are conducted securely and held to industry standards (the same standards used for health information exchange for clinical care). The Epic corporation owns and operates secure data centers under a comprehensive security program, which includes administrative, physical, and technical safeguards that follow industry best practices.[49]

Cosmos queries are only as strong as the available EHR data elements on which they are based. Data deduplication is not perfect, and data types are not harmonized across different Epic system. While the described mapping process may attenuate this heterogeneity, it is unlikely to remove it completely. In addition, true disease prevalence estimates in particular are notoriously difficult to ascertain from EHR data.[50] In being limited to EHR data from a single vendor, Cosmos invariably comes with selection bias that can only partially be overcome by data accrual from tens of millions of patients.[15] The non-randomly ascertained nature of the data and the bias associated with hospital systems that tend do use this specific EHR platform remain a major limitation for generalizability. The authors of this manuscript, from their perspective, strongly support cross-vendor collaboration and perceive within-vendor data harmonization as a requisite pre-step. Finally, the inability to re-sample or re-weight populations by demographics hinders data representation and generalizability.[51]

Cosmos provides a model for data standardization and aggregation that places the development and maintenance impetus on the EHR vendor, leveraging customers' existing health information exchange networks and standards used for clinical health information exchange to minimize effort required to participate. Since Cosmos relies on existing clinical health informatics exchanges, C-CDAs, and generally accepted data standards, future versions of this endeavor could conceivably incorporate data from healthcare systems using different EHRs.

Conclusion

Cosmos is a rapidly evolving resource that can assist institutions in their collective journeys toward becoming more complete learning healthcare systems. The initiative promotes within-vendor data standardization and aggregation, and presents a potential model for future customer driven inter-vendor EHR-based data collaborations.

Clinical Relevance Statement

Seamless data aggregation and population-level analytics between healthcare systems may be an important step in the evolution of all learning healthcare systems. A cross-institution, vendor-facilitated and health information exchange dependent data aggregation and analysis platform shows promise in facilitating this goal among users of the same EHR. Our experience shows the power and scalability of such an "all or none" contribution model, while highlighting numerous concerns that may be associated with such an approach.

Protection of Human and Animal Subjects

Since data returned from Cosmos is de-identified and presented in aggregate, Cosmos queries do not constitute human subjects research and so do not require institutional review board approval for research purposes.

Authors' Contributions

All listed authors provided substantial contributions to the conception of the work, as well as the analysis and interpretation of data for the work. All listed authors were involved in drafting and approving the final manuscript, and agree to be accountable for all aspects of the work.

Funding

Y.T. and D.K. report support by the Clinical and Translational Science Collaborative (CTSC) of Cleveland which is funded by the National Institutes of Health (NIH), National Center for Advancing Translational Science (NCATS), Clinical and Translational Science Award (CTSA) grant, UL1TR002548. The content is solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

Conflict of Interest

C.H., A.F., and S.H. reports other from Epic during the conduct of the study; employed by this company and developed the tool described as part of the role in this corporation. A.M.N. reports grants from NIH, from null, during the conduct of the study; other from Bedside Intelligence, LLC, from null, outside the submitted work.

References

1 H.R. 1 -American Recovery and Reinvestment Act of 2009. Accessed September 8, 2018 at: https://www.congress.gov/bill/111th-congress/house-bill/1
2 Henry J, Pylypchuk S, Searcy Y, Patel V. Adoption of electronic health record systems among U.S. non-federal acute care hospitals. : 2008–2015. Accessed September 8, 2018 at: https://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/Certification.html
3 McGinnis JM, Aisner D, Olsen L. The Learning Healthcare System: Workshop Summary. Washington, D.C.: National Academies Press; 2007
4 Friedman C, Rubin J, Brown J, et al. Toward a science of learning systems: a research agenda for the high-functioning learning health system. J Am Med Inform Assoc 2015;22(01):43–50
5 Cohen ME, Liu Y, Ko CY, Hall BL. Improved surgical outcomes for ACS NSQIP hospitals over time. Ann Surg 2016;263(02):267–273
6 Hall BL, Hamilton BH, Richards K, Bilimoria KY, Cohen ME, Ko CY. Does surgical quality improve in the American College of Surgeons National Surgical Quality Improvement Program: an evaluation of all participating hospitals. Ann Surg 2009;250(03):363–376
7 Khuri SF, Henderson WG, Daley J, et al; Principal Investigators of the Patient Safety in Surgery Study. Successful implementation of the Department of Veterans Affairs' National Surgical Quality Improvement Program in the private sector: the Patient Safety in Surgery study. Ann Surg 2008;248(02):329–336
8 Lenert L, Sundwall DN. Public health surveillance and meaningful use regulations: a crisis of opportunity. Am J Public Health 2012;102(03):e1–e7
9 Topaloglu U, Palchuk MB. Using a federated network of real-world data to optimize clinical trials operations. JCO Clin Cancer Inform 2018;2:1–10
10 Califf RM. The Patient-Centered Outcomes Research Network: a national infrastructure for comparative effectiveness research. N C Med J 2014;75(03):204–210
11 Kaelber DC, Foster W, Gilder J, Love TE, Jain AK. Patient characteristics associated with venous thromboembolic events: a cohort study using pooled electronic health record data. J Am Med Inform Assoc 2012;19(06):965–972
12 Psaty BM, Breckenridge AM. Mini-Sentinel and regulatory science–big data rendered fit and functional. N Engl J Med 2014;370(23):2165–2167

13 Vogel J, Brown JS, Land T, Platt R, Klompas M. MDPHnet: secure, distributed sharing of electronic health record data for public health surveillance, evaluation, and planning. Am J Public Health 2014;104(12):2265–2270

14 Buck MD, Anane S, Taverna J, Amirfar S, Stubbs-Dame R, Singer J. The Hub Population Health System: distributed ad hoc queries and alerts. J Am Med Inform Assoc 2012;19(e1):e46–e50

15 Tarabichi Y, Goyden J, Liu R, Lewis S, Sudano J, Kaelber DC. A step closer to nationwide electronic health record-based chronic disease surveillance: characterizing asthma prevalence and emergency department utilization from 100 million patient records through a novel multisite collaboration. J Am Med Inform Assoc 2020;27(01):127–135

16 Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004;32 (Database issue, suppl_1):D267–D270

17 American Medical Association National Uniform Claim Committee. Health care provider taxonomy. Accessed July 16, 2020 at: http://www.nucc.org/index.php/code-sets-mainmenu-41/provider-taxonomy-mainmenu-40

18 National Uniform Billing Committee. Official UB-04 data file. Accessed July 16, 2020 at: https://www.nubc.org/license

19 National Center for Health Statistics. Classification of diseases, functioning, and disability. Accessed July 16, 2020 at: https://www.cdc.gov/nchs/icd/index.htm

20 Dolin RH, Alschuler L, Beebe C, et al. The HL7 clinical document architecture. J Am Med Inform Assoc 2001;8(06):552–569

21 The Health Insurance Portability and Accountability Act of 1996. In. Vol Pub. L. 104–191. Stat. 1936. 1996. Accessed 1996 at: https://www.govinfo.gov/content/pkg/PLAW-104publ191/pdf/PLAW-104publ191.pdf

22 Winden TJ, Boland LL, Frey NG, Satterlee PA, Hokanson JS. Care everywhere, a point-to-point HIE tool: utilization and impact on patient care in the ED. Appl Clin Inform 2014;5(02):388–401

23 Dang QH. Secure hash standard. Federal Inf. Process. Stds. (NIST FIPS) - 180–4. Accessed 2015 at: https://www.nist.gov/publications/secure-hash-standard

24 Kaelber DC, Waheed R, Einstadter D, Love TE, Cebul RD. Use and perceived value of health information exchange: one public healthcare system's experience. Am J Manag Care. 2013;19(10 Spec No):SP337–343

25 Ross MK, Sanz J, Tep B, Follett R, Soohoo SL, Bell DS. Accuracy of an electronic health record patient linkage module evaluated between neighboring academic health care centers. Appl Clin Inform 2020;11(05):725–732

26 R: A Language and Environment for Statistical Computing [computer program]. Vienna, Austria: R Foundation for Statistical Computing; Accessed 2013 at: http://softlibre.unizar.es/manuales/aplicaciones/r/fullrefman.pdf

27 Wickham Hggplot2. Elegant Graphics for Data Analysis. Springer-Verlag New York; Available at: 2016

28 Akinbami LJ, Fryar CD. Current asthma prevalence by weight status among adults: United States, 2001-2014. NCHS Data Brief 2016;2016(239):1–8

29 Ma J, Xiao L, Stafford RS. Underdiagnosis of obesity in adults in US outpatient settings. Arch Intern Med 2009;169(03):313–314

30 Bardia A, Holtan SG, Slezak JM, Thompson WG. Diagnosis of obesity by primary care physicians and impact on obesity management. Paper presented at: Mayo Clinic Proceedings Accessed 2007 at: https://www.mayoclinicproceedings.org/article/S0025-6196(11)61333-5/abstract

31 Hasegawa K, Tsugawa Y, Lopez BL, Smithline HA, Sullivan AF, Camargo CA Jr. Body mass index and risk of hospitalization among adults presenting with asthma exacerbation to the emergency department. Ann Am Thorac Soc 2014;11(09):1439–1444

32 Schatz M, Zeiger RS, Zhang F, Chen W, Yang S-J, Camargo CA Jr. Overweight/obesity and risk of seasonal asthma exacerbations. J Allergy Clin Immunol Pract 2013;1(06):618–622

33 Centers for Disease Control and Prevention. U.S. Influenza Surveillance System: Purpose and Methods. Accessed 2020 at: https://www.cdc.gov/flu/weekly/overview.htm#:~:text=The%20U.S.%20influenza%20surveillance%20system,%2C%20clinics%2C%20and%20emergency%20departments

34 Jeudin P, Liveright E, Del Carmen MG, Perkins RB. Race, ethnicity, and income factors impacting human papillomavirus vaccination rates. Clin Ther 2014;36(01):24–37

35 Chen RT, Glasser JW, Rhodes PH, et al; The Vaccine Safety Datalink Team. Vaccine Safety Datalink project: a new tool for improving vaccine safety monitoring in the United States. Pediatrics 1997;99(06):765–773

36 Klein NP, Fireman B, Yih WK, et al; Vaccine Safety Datalink. Measles-mumps-rubella-varicella combination vaccine and the risk of febrile seizures. Pediatrics 2010;126(01):e1–e8

37 Nyquist A-C, Gonzales R, Steiner JF, Sande MA. Antibiotic prescribing for children with colds, upper respiratory tract infections, and bronchitis. JAMA 1998;279(11):875–877

38 Gonzales R, Steiner JF, Sande MA. Antibiotic prescribing for adults with colds, upper respiratory tract infections, and bronchitis by ambulatory care physicians. JAMA 1997;278(11):901–904

39 Ashman JJQuickStats. Percentage of Emergency Department Visits for Acute Viral Upper Respiratory Tract Infection at Which an Antimicrobial Was Given or Prescribed, by Age — United States. , 2010–2017. MMWR Morb Mortal Wkly Rep. 2020 (69:174). Accessed 2010 at: https://www.cdc.gov/mmwr/volumes/69/wr/pdfs/mm6906a6-H.pdf

40 Fuchshuber PR, Greif W, Tidwell CR, et al. The power of the National Surgical Quality Improvement Program–achieving a zero pneumonia rate in general surgery patients. Perm J 2012;16(01):39–45

41 Toh S, Reichman ME, Houstoun M, et al. Comparative risk for angioedema associated with the use of drugs that target the renin-angiotensin-aldosterone system. Arch Intern Med 2012; 172(20):1582–1589

42 Adamson DM, Chang S, Hansen LGJNYTH. Health research data for the real world: the MarketScan databases. Accessed 2008 at: http://patientprivacyrights.org/wp-content/uploads/2011/06/Thomson-Medstat-white-paper.pdf:b28

43 Kirby JC, Speltz P, Rasmussen LV, et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. 2016;23(06):1046–1052

44 Birkhead GS. Successes and continued challenges of electronic health records for chronic disease surveillance. Am J Public Health 2017;107(09):1365–1367

45 Klompas M, Cocoros NM, Menchaca JT, et al. State and local chronic disease surveillance using electronic health record systems. Am J Public Health 2017;107(09):1406–1412

46 Newton-Dame R, McVeigh KH, Schreibstein L, et al. Design of the New York City macroscope: innovations in population health surveillance using electronic health records. EGEMS (Wash DC) 2016;4(01):1265

47 Perlman SE, McVeigh KH, Thorpe LE, Jacobson L, Greene CM, Gwynn RC. Innovations in population health surveillance: using electronic health records for chronic disease surveillance. Am J Public Health 2017;107(06):853–857

48 Curtis LH, Brown J, Platt R. Four health data networks illustrate the potential for a shared national multipurpose big-data network. Health Aff (Millwood) 2014;33(07):1178–1186

49 Data Center Micro-Segmentation. A Software Defined Data Center Approach for a "Zero Trust" Security Strategy. Accessed August 26, 2020 at: https://blogs.vmware.com/networkvirtualization/files/2014/06/VMware-SDDC-Micro-Segmentation-White-Paper.pdf

50 Kharrazi H, Gonzalez CP, Lowe KB, Huerta TR, Ford EW. Forecasting the maturation of electronic health record functions among US hospitals: retrospective analysis and predictive model. J Med Internet Res 2018;20(08):e10458

51 Casey JA, Schwartz BS, Stewart WF, Adler NE. Using electronic health records for population health research: a review of methods and applications. Annu Rev Public Health 2016;37:61–81

Appendix A Logical Observation Identifiers Names and Codes

LOINC codes for influenza A	LOINC codes for influenza B	LOINC codes for SARS-CoV-2
44564–3	46083–2	94500–6
46082–4	44573–4	94314–2
44561–9	44574–2	94309–2
44558–5	80383–3	94306–8
80382–5	44572–6	94534–5
44559–3	5867–7	41458–1
5863–6	76080–1	41459–9
76078–5	82170–2	94531–1
48310–7	38382–8	
82166–0	40982–1	
31858–4	44575–9	
44563–5	44577–5	
43874–7	43895–2	
31859–2	31864–2	
5864–4	49534–1	
44560–1	5866–9	
5861–0	92976–0	
5862–8	85478–6	
49531–7	5865–1	
38381–0		
34487–9		
92977–8		
85477–8		
22827–0		
40891–3		

Abbreviations: LOINC, Logical Observation Identifiers Names and Codes; SARS-CoV-2, severe acute respiratory syndrome coronavirus-2.