Noel A, Bartelt K. Cosmos: Real World Data Powered by the Health Care Community. *Journal of the Society for Clinical Data Management*. 2023; 3(4): 6, pp. 1–4. DOI: https://doi.org/10.47912/jscdm.246

OPINION PAPER

# Cosmos: Real World Data Powered by the Health Care Community

Andrea Noel* and Kersten Bartelt*

Cosmos is a rapidly expanding, real-world healthcare dataset comprised of 217M+ deduplicated patient records from 219 healthcare organizations. A primary goal for Cosmos is to produce generalized medical knowledge to advance the understanding of the causes, treatment, and prevention of disease and bring that knowledge to the hands of doctors and patients. In order to meet this goal, Cosmos has a number of features to ensure a robust, meaningful, well-organized, safe, and high-quality dataset that can be used to generate real-world evidence for research, perform predictive modeling, and create tools to dynamically impact medicine at the point of care.

**Keywords:** Health Care; Health Care Data Set; Research; Interoperability; Precision Medicine; Community Research Network; Learning Health System; Real-world Evidence

Medicine is full of questions. One of the most basic questions is, "What will give patients the best chance at the highest quality of life?" A randomized control trial is considered to be the gold standard of evidence to guide clinical decision making for a question such as this, but many times a randomized control trial is not available. Even when trials are available, the results may not be applicable to certain subpopulations of patients or to the individual patient in front of the clinician. Real-world evidence from sources such as electronic health care data can help to fill these gaps.

## What is Cosmos?

Cosmos is a rapidly growing, real-world health care data set that contains over 217 million individual patient records from 219 US health care organizations and one organization in Lebanon, with records linked and data de-duplicated between organizations.[1] Its primary goal is to advance the understanding of disease causes, treatment, and prevention, and to bring that knowledge to physicians and patients. Cosmos offers a robust, meaningful, well-organized, safe, and high-quality data set for generating real-world evidence, predictive modeling, and for creating tools to impact medicine in real-time at the point of care. Health care organizations using Epic's software for their comprehensive health care records can access the data set at no cost by contributing data to Cosmos. Additionally, the Cosmos team uses the data to conduct research studies and improve clinical workflows.

The data set includes a wide range of health data, including inpatient, outpatient, surgical, and emergency data from community and academic medical centers, Federally Qualified Health Centers, critical access hospitals, general practices, and specialty practices. Data types include, but are not limited to, clinical data with diagnoses, vitals, lab results, prescribed and administered medications, immunizations, family history, and cancer staging; procedural data, such as surgical, obstetric, and birth information; and demographic data including age, sex (legal sex, sex assigned at birth, and gender identity), race, ethnicity, social determinants of health, and social vulnerability index. Of note, birth data in Cosmos maintains the link between birthing parent and infants. Cosmos also includes data on billed codes, financial class, and patient digital engagement. New data types are constantly being added based on the needs of the Cosmos community.

## Interoperability facilitates large data sets

These different pieces of data can be incorporated into Cosmos because of standardization of health information and interoperability between health care organizations.[2] Since the mid-2000s, the health care industry has developed tools for data interoperability to meet certification requirements from The Office of the National Coordinator for Health Information Technology (ONC). National data standards, such as the United States Core Data for Interoperability (USCDI), define data types used by health information exchanges that health care systems using Epic participate in, like Car*e*quality.[3]

Health care organizations using Epic's interoperability network, Care Everywhere, exchange additional discrete

* Epic Systems Corporation, US

Corresponding authors: Andrea Noel, MD (anoel@epic.com); Kersten Bartelt, RN (kersten@epic.com)

datatypes where national standards have yet to be defined. These data are mapped to standard ontologies such as SNOMED, LOINC, RxNORM, and others whenever possible. Health care organizations send data to Cosmos via the Care Everywhere network and can send additional datatypes for research, which are also mapped to standard ontologies. Both national standards and standard ontologies enable data from Cosmos to be compared to other real-world data sets. An added benefit of using Care Everywhere for data exchange with Cosmos is that the patient linkage occurs natively without the need for a third-party token broker. The de-duplication processes happen both upstream and within Cosmos. Through Care Everywhere, de-duplication cleans up the composite record for duplicates such as encounters, diagnoses, histories, procedures, and results using confidence scoring algorithms or by matching directly on elements such as dates, times, and results. Within Cosmos, all data is de-duplicated with each new data load. With patient linking and de-duplication, Cosmos patient records can span decades and have greater data completeness.

Several other processes are in place both in health care systems using Epic and within Cosmos to facilitate data aggregation. All Cosmos source systems use Epic, so much of the data sent to Cosmos on a daily or bi-weekly basis is stored in local systems using standard ontologies with the same core data structure. This is a common data model across Epic sites. Data such as diagnoses, vitals, specialties, drugs, procedures, and results are connected in the same way to patients and encounters in every Epic system. This common structure minimizes the need for data transformation when data is sent to and stored in Cosmos which ensures high fidelity to the source data. In some special data types, the workflow to capture data may vary significantly between organizations (or other complexities exist). In these cases, Epic experts and contributing organizations work together to identify the correct data to map the concept appropriately. Cosmos also creates research concepts on stored data using definitions from the community, such as disease registries, COVID cases, and drug episodes, to help streamline common queries. Cosmos provides a detailed data dictionary that allows novel data elements like these to be tracked from their original creation point to the research database in Cosmos. Because of Epic standards for storing data, expert-driven workflow mapping, and concept mapping and phenotyping, the aggregate data in Cosmos has minimal transformations, and any transformation that is made is done to increase usability of data.

### Data privacy and responsible data use

Common challenges around safety in real-world data include data privacy, security, and ethics. Balancing research and development utility with patient privacy and security is crucial in Cosmos. The 16 categories of unique direct patient identifiers defined by the US Department of Health and Human Services are explicitly excluded from import, resulting in a HIPAA-defined[4] limited data set to maintain privacy. Cosmos also supports organizational privacy and consent policies, so certain types of data, such as substance abuse treatment, may not be sent to Cosmos if they are restricted from exchange at the organization, state, or country level as determined by the contributing healthcare system.

There are two main ways for users to access data in Cosmos, both of which minimize reidentification risk for data safety while still providing powerful tools for research and development. Basic research can be done quickly on the HIPAA limited data set via a secure web portal that includes data query and visualization tools that mask data for any cell counts under 11. Advanced research may be done on line-level data in a hosted virtual machine where data has been de-identified via the Expert Determination Method of De-identification described in the HIPAA Privacy Rule at 45 CFR 164.514(b)(1).[5] Raw data is programmatically prevented from being exported from the virtual machine for security, so these environments contain built-in statistical analysis tools like Python, R, and SQL to support various epidemiological research needs and machine learning tasks, such as building, training, and validating predictive models. Additional external datasets (like weather data) can be loaded into the virtual machine by the Cosmos team after review. New data types also undergo a review process that evaluates the risk of reidentification on a quarterly basis before being added to the de-identified data set. Because of the masking of low cell counts in the limited data set and the deidentification of the line-level data, neither data set qualifies as Human Subjects Research, and therefore does not require an Institutional Review Board (IRB) approval, but researchers may choose to pursue an IRB exemption if their institution deems it necessary. Researchers accessing either data set must undergo training on how to use the tools and interpret the data sets safely.

To support ethical data use, an elected governing council comprised of 15 representatives from pediatric, academic, and community health care organizations participating in Cosmos maintains a Rules of the Road document to guide ethics and data usage. Every participating organization, including Epic, agrees to these Rules of the Road before beginning their participation in Cosmos, and every user who accesses Cosmos signs a responsible use agreement upon their first access to the database. Each participating site designates a point person in their leadership to establish governance for Cosmos and to determine a process for providing access to their employees. Selling of Cosmos data, using data for advertising or comparing markets, and attempts to reidentify patients are strictly prohibited. Provisions exist regarding the use of Cosmos data for third-party funded research by Cosmos participants. All Cosmos queries are recorded, and users attest to their work in Cosmos every time they access data. The Cosmos team reviews all attested uses of the data and follows up with users or the Governing Council where necessary. These processes ensure that as the Cosmos data set continues to expand with time, data governance is guided by contributors to the data set.

## How is the community using Cosmos?

There are a lot of aspirations from the community for Cosmos use cases. Research and development are in progress to improve clinical characterization of disease, to estimate population-level effect for therapeutics, and to perform modeling for anomaly detection and patient outcome prediction. Early researchers compared Cosmos data to external data sources such as the US influenza surveillance system, Vaccine Adverse Event Reporting System (VAERS), and National Hospital Ambulatory Medical Care Survey. They found similar results and published their methods for research.[6] Validating the Cosmos data set against external databases and registries was a crucial step to establish data set reliability and representativeness and will continue in the future.

Several papers have been published using retrospective cohort studies to characterize disease, care patterns, and outcomes with Cosmos data. Completed studies have examined subjects such as maternal and neonatal outcomes related to maternal COVID infection, neonatal readmission risk related to hospital length of stay in birth encounter, fentanyl testing rates in emergency overdose care, and incidence of COVID infection and hospitalization rates among patients with alcohol use disorder.[7–8] Aggregate Cosmos data were used by Cosmos and the CDC for urgent public health evaluations, particularly around the COVID pandemic[9] and mpox.[10] In these publications, authorship remains with the researchers, and Cosmos is listed as the data source.

## How does Cosmos compare to other data sets?

Cosmos shares similarities with other large data sets like N3C[11] and AllOfUs,[12] such as the use of standard ontologies, limited and de-identified cuts of the data sets, linked patient records, required training for users, various data visualization tools, rules around data use, and safety mechanisms (eg, auditing). Cosmos differs from these data sets because all source systems use the same operational data model and transmit the HIPAA limited data directly to Cosmos for research. This means data are stored in Cosmos with minimal transformations and keep the fidelity of the data in the originally recorded state.

The primary differentiator of Cosmos is its ability to integrate real-world evidence and insights back into the electronic health records of participating organizations at the point of care. A series of different Cosmos-based software modules will help with a variety of clinical tasks. Clinicians will be able to explore treatment outcomes in real-time for common conditions in a precision cohort of patients similar to their own patient. Another Cosmos software module will connect clinicians caring for patients with similar rare conditions, like those defined in Orphanet.[13] An additional Cosmos feature is aimed at accelerating clinical trial recruitment by creating study cohorts with specific characteristics in Cosmos that can be pulled back into local systems by contributing health care organizations. Projects are also underway to investigate global predictive model development for long-term outcomes using Cosmos data, which may help clinical teams predict risks at local organizations for both common and rare conditions.

Another differentiator is the extent of expansion projected for Cosmos data elements. Participating organizations capture tens of thousands of discrete data elements, and only a portion have been aggregated into Cosmos so far. The data set will continue to expand over the coming years to support more sub-specialty data elements, risk scores, and rare disease research. To guide this growth, the Cosmos team regularly seeks feedback from the community by visiting contributing organizations, collaborating at conferences, and communicating directly with researchers and the governing council. To make full use of the real-world data, the Cosmos team encourages the Cosmos community to further advance medicine through research, scientific discovery, and the construction of real-world evidence tools with Cosmos.

## Competing Interests

The authors have no competing interests to declare.

## References

1. ***Epic Cosmos.*** Accessed October 20, 2023. https://cosmos.epic.com/.
2. **HealthIT.gov.** Interoperability standards and technology. Published November 4, 2022. Accessed January 17, 2022. https://www.healthit.gov/topic/interoperability/standards-and-technology.
3. **Carequality.** Carequality – interoperability framework. Published June 29, 2021. Accessed May 15, 2023. https://carequality.org/.
4. **US Department of Health and Human Services. Office for Civil Rights (OCR).** Disclosures for emergency preparedness – a decision tool: limited data set (LDS). Published August 29, 2007. Accessed May 17, 2023. https://www.hhs.gov/hipaa/for-professionals/special-topics/emergency-preparedness/limited-data-set/index.html.
5. **National Archives: Code of Federal Regulations.** Other requirements relating to uses and disclosures of protected health information. Accessed January 17, 2022. https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-C/part-164/subpart-E/section-164.514.
6. **Tarabichi Y, Frees A, Honeywell S,** et al. The Cosmos collaborative: a vendor-facilitated electronic health record data aggregation platform. *ACI Open.* 05(1). DOI: https://doi.org/10.1055/s-0041-1731004
7. **Handley SC, Gallagher K, Breden A,** et al. Birth hospital length of stay and rehospitalization during COVID-19. *Pediatrics.* 2022 Jan 1; 149(1): e2021053498. DOI: https://doi.org/10.1542/peds.2021-053498
8. **Youngs C, Gupta N, Emerman C.** Covid-19 immunization and disease burden for patients with alcohol use disorder evaluation through the

use of an electronic database, J. *Addiction Research and Adolescent Behaviour*. 5(4). DOI: https://doi.org/10.31579/2688-7517/051

9. **Shah MM, Joyce B, Plumb ID,** et al. Paxlovid associated with decreased hospitalization rate among adults with COVID-19 — United States, April–September 2022. *MMWR Morb Mortal Wkly Rep*. 2022; 71: 1531–1537. DOI: https://doi.org/10.15585/mmwr.mm7148e2
10. **Deputy NP, Deckert J, Chard AN,** et al. Vaccine effectiveness of JYNNEOS against mpox disease in the United States. *Engl J Med*. 2023; 388: 2434–2443. DOI: https://doi.org/10.1056/NEJMoa2215201
11. **National COVID cohort collaborative (N3C).** National center for advancing translational sciences. Published May 12, 2020. Accessed May 24, 2023. https://ncats.nih.gov/n3c.
12. **National Institutes of Health.** All of Us research program. Published June 1, 2020. Accessed May 24, 2023. https://allofus.nih.gov/.
13. **Orphanet.** Accessed May 24, 2023. https://www.orpha.net/consor/cgi-bin/index.php.

**How to cite this article:** Noel A, Bartelt K. Cosmos: Real World Data Powered by the Health Care Community. *Journal of the Society for Clinical Data Management*. 2023; 3(4): 6, pp. 1–4. DOI: https://doi.org/10.47912/jscdm.246

**Submitted:** 03 February 2023 **Accepted:** 12 September 2023 **Published:** 07 November 2023

JSCDM

*Journal of the Society for Clinical Data Management* is a peer-reviewed open access journal published by Society for Clinical Data Management.

OPEN ACCESS