International Journal of Population Data Science

Journal Website: www.ijpds.org

Data Resource Profile: The GEMINI Database for Medical and ICU Inpatients Across Ontario, Canada

Jeruby Retnakanthan[1,†], Surain B Roberts[1,2,*,†], Sylvia Cheng[1], Anne Löffler[1], Afsaneh Raissi[1], Trong Shen[1], Ijere Ijere[1], Anis Sharafoddini[1], Nicole Yada[1,2], Chloe Chow[1], Fahad Razak[1,2,3,†], and Amol A Verma[1,2,3,†]

Submission History

Submitted:	17/03/2026
Accepted:	20/09/2026
Published:	18/11/2026

[1] Li Ka Shing Knowledge Institute, Unity Health Toronto - St. Michael's Hospital, Toronto, Ontario, Canada
[2] Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, Ontario, Canada
[3] Department of Medicine, University of Toronto Faculty of Medicine, Toronto, Ontario, Canada
[†] These authors contributed equally

Abstract

Introduction

With over 16 million people, Ontario is one of the largest unified health systems in North America, with a single public insurance system covering one of the most racially and culturally diverse populations in the world. GEMINI seeks to harness the unique data generated by this health system for research, by linking data from the EHR and administrative data generated by hospital admissions with longitudinal provincial data holdings.

Methods

GEMINI collects electronic health record (EHR) and administrative data extracted from hospital information systems for all patients admitted to any adult inpatient medical specialty or cared for by any adult intensive care service and discharged from one of 37 participating hospitals across Ontario. The administrative data are a subset of standardised data reported by hospitals to the Canadian Institute for Health Information. The EHR data include laboratory data, radiology reports, vital signs, blood transfusions, unstructured clinical notes, and medication orders. These data are extracted in batches by local hospital personnel and transferred securely to GEMINI, which is based at Unity Health Toronto, an academic health sciences centre. Data are de-identified, quality-checked, and standardised to an internal common data model before being provided to researchers. Provincial health insurance numbers are used to link patient data across encounters at multiple institutions and to facilitate linkage between GEMINI data and longitudinal population-based data, housed at the non-profit organisation, ICES.

Results

The GEMINI database contains over 20 billion data points and ∼2.5 million hospital admissions spanning from 1st April 2015–31st June 2024, held in high-performance computing environments under the governance of Unity Health Toronto.

Conclusion

Researchers can access GEMINI data for a wide range of epidemiological, clinical, and health services research applications. The data access process is described at https://geminimedicine.ca/access-gemini-data/ and questions can be directed to GEMINI.Research@unityhealth.to.

Key features

- **What is unique about the dataset:** The GEMINI database contains detailed clinical and administrative data from hospital admissions, with over 60% coverage of Ontario. With more than 15.5 million people, Ontario is one of the largest unified health systems in North America, with a single public insurance system covering one of the most racially and culturally diverse populations in the world.
- **Why the dataset was created:** Hospital information systems are not standardised in Canada, and data are often siloed across institutions and difficult to use for research, quality improvement, and other analytics. The GEMINI database was created to collect and standardise electronic health record (EHR) data from hospitals for research.
- **Details about the dataset:** GEMINI collects EHR and administrative data extracted from hospital information systems for all patients admitted to any adult inpatient medical specialty (e.g., general medicine, respirology, oncology, cardiology, etc.) or cared for by any adult intensive care service and discharged from one of the participating hospitals. The dataset includes ∼2.5M admissions spanning from 1st April 2015 – 31st June 2024, from 37 hospitals.
- **Description of any data linkage:** GEMINI data have been linked to a number of other datasets. Linkage to ICES data has enabled longitudinal analysis of patient trajectories before, during, and after hospitalisation, complementing the robust data holdings that are already available. GEMINI data have also been linked to Statistics Canada census data, allowing for in-depth neighbourhood-level analyses of sociodemographic information. Finally, GEMINI also conducts linkages for specific Research Ethics Board (REB) approved research projects to more richly characterise subpopulations within the cohort.
- **Main categories of data:** The GEMINI dataset can be broadly categorised into hospitalisation-level EHR data (i.e., data that are extracted from a hospital information system source) and administrative data (i.e., data that are already compiled by hospitals for other purposes). The administrative data in the GEMINI database are a subset of the standardised data reported by every hospital to the Canadian Institute for Health Information (CIHI). The EHR data include laboratory data, radiology reports, vital signs, blood transfusions, unstructured clinical notes, admission/discharge characteristics, and medication orders.
- **How to collaborate and access the dataset [contact details]:** To learn more about opportunities to collaborate or how to access the GEMINI dataset, please visit https://geminimedicine.ca/access-gemini-data/ or email GEMINI.Research@unityhealth.to with any questions.

*Corresponding Author:
Email Address: surain.roberts@unityhealth.to (Surain B Roberts)

https://doi.org/10.23889/ijpds.v8i6.3448
November 2026 © The Authors. Open Access under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/deed.en)

Background

For decades, Canada had been a world leader in research using routine healthcare data, which were historically derived from only "administrative sources", such as insurance claims and patient diagnoses collected to monitor hospital activities. However, health information systems have increasingly digitised in the last 20 years, such that administrative data now represent only a small fraction of the information that is generated in healthcare. Electronic health record (EHR) data are created from routine healthcare interactions (e.g., physician notes, laboratory tests, medical images, etc.) and can be found in these newly digitised health information systems. It represents many billions of data points every day and is of immense value for clinical and health systems research. Despite this exponential growth in the volume and diversity of EHR data, it is not shared well across the Canadian healthcare system. While other countries around the world have made great strides in harnessing EHR data for research and analytics (e.g., Cosmos [1], Clinical Practice Research Datalink [2], and MIMIC-IV [3]), it is often difficult to access in Canada due to fragmented computer systems, a maze of legal and privacy barriers, and a lack of robust infrastructure to harness clinical data in hospital information systems.

GEMINI, a not-for-profit research program based at Unity Health Toronto, aims to fill this gap and create a collaborative ecosystem of healthcare data and analytics to accelerate research and quality improvement. GEMINI collects and harmonises inpatient data from 35+ Ontario hospitals, provides meaningful quality metrics to physicians and hospitals, and leads world-class research focused on understanding and improving hospital care, with the ultimate goal of improving care for hospitalised patients. The program is dedicated to advancing open science and the responsible sharing of data. To that end, it enables dynamic and interdisciplinary communities of researchers to use GEMINI data to answer important questions alongside key knowledge users, including patients and caregivers, clinicians, and health system decision-makers.

GEMINI was initially established as a research study in 2014-15, with the purpose being to study the quality of care on general medicine wards using data collected from seven hospital sites in the Greater Toronto Area. An early GEMINI study showed that general medicine patients account for approximately 40% of emergency department admissions to hospital and 25% of hospital bed days [4]. Further, the number of general medicine hospital admissions increased by 32% between 2010 and 2015 [4]. Subsequent studies found considerable physician-level variation in processes of care, resource use and clinical outcomes in inpatient general medicine, representing practice differences that highlighted opportunities for quality improvement [5].

With the spread of COVID-19 across Ontario and the country in 2020, GEMINI received funding from the Canadian Institutes of Health Research (CIHR) and other sources to expand the program from seven to over thirty hospitals, and to expand beyond general medical beds to capture all medical admissions and the Intensive Care Unit. The purpose of this expansion was to further develop a large, detailed, clinical dataset that could support a wide breadth of research on COVID and non-COVID illness in hospitals. This represented a novel undertaking to harness Ontario's hospital EHR data and standardise it into one linked dataset that can be used to answer a wide array of research questions. This expansion enabled GEMINI data to be used in several major scientific contributions to Canada's COVID-19 pandemic response (see "Selected Research Applications" section below). In 2020, GEMINI also partnered with the Vector Institute, one of three Canadian National Artificial Intelligence (AI) Institutes, to enhance the secure computing environment and make it usable for the application of AI methods to address research questions relevant to GEMINI's purpose.

In alignment with the International Journal of Population Data Science's call for data resource profiles, the purpose of this paper is to describe GEMINI as a novel linked data resource that is available for researchers both across Canada and internationally.

Methods

Study Population & Sources of Data

GEMINI is based in Ontario, Canada, home to a population of over 16 million people. Ontario has one of the largest unified health systems in North America, with a single public insurance system and universally insured hospital services for its highly diverse population. The province's health care system is managed through six distinct Ontario Health regions [6], enabling a population-based health planning approach.

GEMINI collects detailed clinical EHR data extracted from hospital information systems for all patients admitted to any adult inpatient medical specialty (e.g., general medicine, respirology, oncology, cardiology, etc.) or cared for by any adult intensive care service at participating hospitals, from 1st April 2015 and onwards. While expansion to include data from paediatric acute care facilities and inpatient paediatric wards is currently in progress (see "Future Directions" section), the full GEMINI database (i.e., with all participating hospitals) is currently limited to data from adult inpatient care.

GEMINI's data are from a diverse range of academic teaching hospitals and large community hospitals across the province, providing comprehensive geographic coverage of all six Ontario Health regions. An evaluation of patient case-mix/complexity and differences in patient characteristics and clinical outcomes between academic and community hospitals has been previously described by Colacci et al in 2025 [7]. This study population contains over 2.5 million hospitalisations and represents >60% of all medical and Intensive Care Unit (ICU) inpatients in the province. GEMINI also collects administrative data, which are used to further enrich the detailed clinical data that are extracted from hospital-specific EHR. For admissions meeting these inclusion criteria, GEMINI data include all observations from emergency department triage to hospital discharge. GEMINI maintains and updates this dataset on a regular basis, with data intake occurring every three to twelve months, depending on the hospital.

Legal authority and data governance

GEMINI's activities are governed by Research Ethics Board-approved study protocols and data sharing agreements with

every participating organisation, in alignment with Ontario's *Personal Health Information Protection Act, 2004* [8] and the *Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans* [9], a joint policy of Canada's three federal research agencies. Under its current governance framework, GEMINI supports research that addresses the overarching objectives of developing methods to improve data quality, characterising populations of hospitalised patients, predicting/modelling patient outcomes and resource use, studying the effects of the COVID-19 pandemic, and quantifying associations between the organisational aspects of hospital care and resource use/clinical outcomes.

To ensure maximum benefit and use of the database, GEMINI aims for its data to be used as widely as possible while still implementing robust methods of data governance to protect patient privacy and ensure ethical use. It uses the "Five Safes" model [10] to ensure privacy protection — safe projects, people, settings, data, and outputs. For example, to meet the "safe people" requirement, GEMINI has developed an Approved Researcher Model to verify the legitimacy of lead researchers proposing to directly access and use GEMINI data. This model, as well as GEMINI's overarching governance policy, has been developed in collaboration with diverse stakeholders, including oversight committees, patient/family partners, and clinicians/researchers from participating hospitals and partner organisations. Another example can be found in the variety of disclosure controls that GEMINI implements in order to meet the "safe outputs" requirement. This includes, but is not limited to, strict oversight on what outputs can leave GEMINI's data environment, appropriate cell suppression methods (with cell sizes of 5 or smaller being suppressed) to reduce the risk of reidentification, and review of any study results by the project's designated GEMINI Investigator (see "Data Access" section below) prior to publication. GEMINI's governance and privacy practices have been reviewed by legal counsel and privacy officers from all participating hospitals and are aligned with other similar large data repositories [11], both at a national and international level.

Data collection and extraction

GEMINI has developed methods to work with staff from participating hospitals to extract clinical and administrative data from hospital information systems. The GEMINI team provides participating hospitals with a standardised Data Reference Model, which describes the data to be collected, along with providing detailed guidance on data elements and types. GEMINI provides guidance on formatting and file structure, but does not impose additional mapping or standardisation requirements on participating sites. Data are extracted by local hospital personnel, using programming scripts customised to the local hospital information systems.

These data are held within each participating hospital on institutional firewall-protected servers and transferred to St. Michael's Hospital using a secure and encrypted transfer service. Unprocessed data received from hospitals is stored on Unity Health Toronto's servers, physically located in the data centre of one of its sites, St. Michael's Hospital. Physical access to the data centre is restricted to specified St. Michael's Hospital IT staff using two factor authentication (card and PIN). The data centre also includes environmental controls, such as fire suppression systems and air conditioning, as well as a backup uninterruptible power supply and generator to maintain system availability during a power outage.

Data Processing Pipeline

Once transferred to Unity Health Toronto, data from individual hospitals are linked and harmonised by GEMINI staff into a Postgres relational database. The data pipeline has been previously described [12] and an updated description was published in March 2025 [13]. In brief, the data are de-identified by removing personally identifying variables from the dataset (medical record number, encounter number, date of birth, first/middle/last name, postal code, health card number, and emergency room registration number). Authorised GEMINI staff use the Ontario Health Insurance Plan (OHIP) number to link encounters for the same patient across multiple institutions. All residents of Ontario are covered under OHIP, which provides each resident a unique OHIP number. The SHA256 hash algorithm (a standard one-way cryptographic hashing function that is approved by the Government of Canada Communications Security Establishment for the protection of sensitive information) is applied to each OHIP number to create a unique identifier that allows for integration of various data tables at the central data repository and reidentification at the local hospital site for the purpose of manual validation. This approach is highly secure and privacy-protective, as only a small subset of GEMINI personnel are permitted to apply the hash algorithm and remove patient identifiers. The data undergo a series of manual and computational quality checks that are investigated by GEMINI staff and the local data extraction team at participating hospitals through an iterative process to ensure high data quality [13].

The data pipeline is divided into three main zones within the GEMINI system, to segment the infrastructure based on data sensitivity and restrict access to authorised users in each zone. Participating hospital's identifiable personal health information (PHI) remains exclusively in Zone 1, where direct identifiers will be removed before being transferred to Zone 2. Only a small number of GEMINI personnel can access Zone 1. Zone 2 is the data processing zone and contains de-identified data in various formats that are the result of internal GEMINI workflows to conduct data processing steps. Only GEMINI personnel can access Zone 2. Standardised and pseudonymised data that are ready for research is stored in Zone 3, which can be accessed by GEMINI personnel and by external researchers who are given limited access to a project-specific database.

Zone 3 is split into two high-performance computing environments, one hosted in the St. Michael's Hospital Data Centre primarily for GEMINI personnel, and one in the HPC4Health environment, a secure, cloud-based, high-performance platform hosted in the data centre of the Hospital for Sick Children (Toronto, ON), primarily for external researchers. The St. Michael's Hospital environment was developed in-house using open-source code (https://openondemand.org/) and provides a user-friendly web interface where users can seamlessly launch applications (e.g., RStudio Workbench, JupyterLab, and VScode IDE sessions), automate data pipelines, and access computational

resources. HPC4Health enables healthcare organisations to establish a "tenancy" in their high-performance computing cluster [14], and allows GEMINI to establish a secure, remotely-accessible environment for external researchers that is firewalled from GEMINI's internal data holdings, providing added levels of security. HPC4Health users are provided with a number of custom-built tools to make their research with GEMINI data more scalable, transparent, and reproducible (see "Research Software" section below). To ensure that data disclosure is limited to the specific data needed for each project, researchers are provided with project-specific databases matching their unique research questions.

A comprehensive summary of GEMINI's data flows and processes has been previously described by Fu & Shen et al. and is available in Figure1 and the corresponding description below (previously published by Fu & Shen et al.) [13].

Linkage to ICES

Linking GEMINI's detailed inpatient data with Ontario's rich longitudinal population-based datasets creates a unique data resource for research. GEMINI data are best-suited for research questions within the scope of the inpatient hospital setting; however, examination of system-wide healthcare utilisation, post-discharge patient outcomes and comprehensive lookback periods are not possible due to a lack of outpatient data. To address this important limitation, GEMINI data were linked to a vast inventory of longitudinal population-based datasets held at ICES in 2023. ICES is an independent not-for-profit corporation and charity, as well as a prescribed entity under Ontario's *Personal Health Information Protection Act, 2004* and the *Coroners Act* [15]. This linkage and GEMINI's ongoing partnership with ICES fills a significant gap by enabling more comprehensive research about trajectories of care, healthcare utilisation, and health outcomes before, during, and after hospitalisation. All researchers requesting access to linked GEMINI-ICES datasets through ICES' secure computing environment must be approved by both GEMINI and ICES, and adhere to both GEMINI and ICES' policies and procedures. Currently, there are around a dozen active research projects using the GEMINI-ICES linkage to provide a full picture of how a patient transitions through the healthcare system, with several more projects in development.

Other Linkages

GEMINI data have also been linked to a number of other datasets. These linkages range from recurring data transfers with publicly available sources and other health data repositories to one-off linkages needed to fulfil the research objectives of a particular project, as permitted by the REB and GEMINI's data sharing agreements.

GEMINI data are linked to 2016 and 2021 Statistics Canada census data, which include neighbourhood-level sociodemographic data collected across Canada. As these variables are not regularly collected in hospital computer systems, this linkage of census-dissemination-area-level data has allowed researchers to explore new dimensions of GEMINI data, provide insights on neighbourhood marginalisation, and facilitate equity-informed research [16]. These data are linked to hospitalisations using Statistics Canada PCCF+ postal code linkage [17]. Neighbourhood and sociodemographic data are periodically updated as GEMINI receives new data and postal codes. To better understand variations in patient care and outcomes at the physician-level, GEMINI has also conducted linkages with publicly available data on physician characteristics (e.g., medical school graduation year) held by physician licensing bodies.

Data linkages may also be conducted in order to fulfil specific research objectives, as detailed in separate protocols and following REB approval. If the external data contains PHI, it is de-identified immediately after linkage, which is performed only by authorised personnel working on GEMINI's secure system. Some examples of previous linkages done for the purpose of specific research studies include linkage of qualitative interview data [18], survey data [19], and resident call schedules [20].

Results

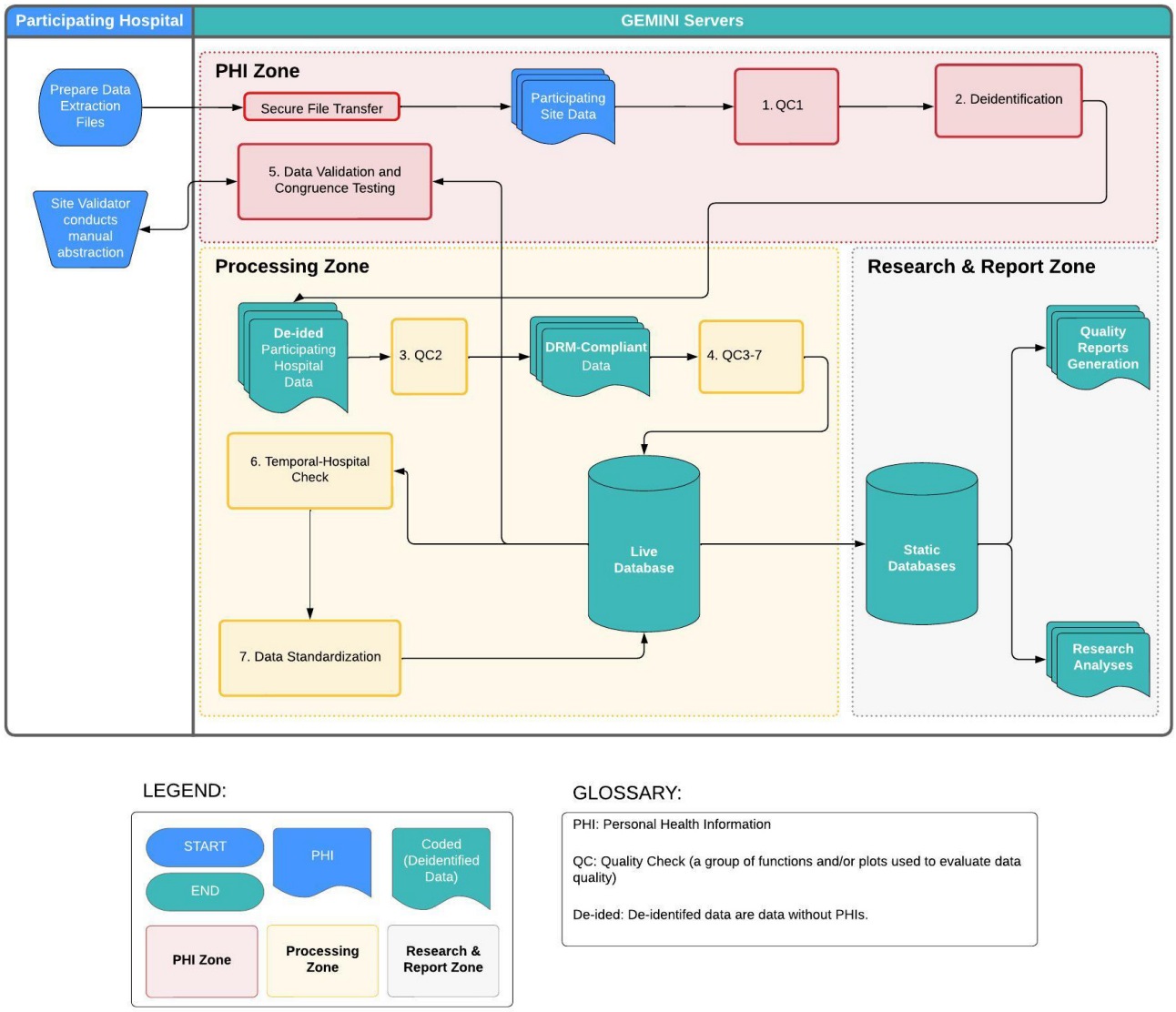
Description of Dataset

The GEMINI dataset is composed of granular hospitalisation-level clinical and administrative data from 37 hospitals across Ontario, with continuous expansion to additional hospitals over time. The clinical data that GEMINI collects from participating hospitals contains detailed variables, including but not limited to laboratory test results, radiology reports, admission/discharge characteristics, vital signs, blood transfusions, clinical notes, and medication orders. While specialised ICU data (e.g., ventilator settings) is not available, all of the other clinical data collected by GEMINI (e.g. medication orders, imaging reports, laboratory tests, blood transfusions, etc.) are collected for ICU patients.

GEMINI's administrative data include routinely collected variables, which hospitals report to the Canadian Institute for Health Information (CIHI) for the Discharge Abstract Database (DAD) and the National Ambulatory Care Reporting System (NACRS). These data include well-defined variables, such as patient demographics (e.g., age, sex), ICD-10-CA diagnosis codes, Canadian Classification of Health Intervention (CCI) codes, and select costing variables. Hospitals are mandated to collect these data annually for transfer to CIHI, which ensures that data are highly structured and standardised. The GEMINI database does not include explicit costing data beyond the Resource Intensity Weights that are already included in CIHI data. However, GEMINI data can be used to quantify the use of several costly resources, including imaging, transfusions, and other resource use captured by the electronic health record.

Detailed descriptions of GEMINI's cohort have been previously published [7], and GEMINI's Data Dictionary is publicly available at https://geminimedicine.ca/the-gemini-database/. A table detailing key features of the GEMINI dataset (e.g., geographic coverage, cohort size, date range, etc) in comparison to other similar databases can also be found in Appendix A.

Figure 1: Data flows and processes for the updated GEMINI platform[23]



The following steps are involved in preparing and adding data to the GEMINI database: (1) Data are extracted from the hospitals into the PHI Zone and undergo data quality assessment QC1 to check data coverage. (2) Data are de-identified and securely transferred to the Processing Zone. (3) Data undergo data quality assessment QC2 where variable names and formats are standardised and reformatted. (4) Data compliant with the formats described in the GEMINI Data Reference Model (DRM) then undergo the semi-automated data quality checks QC3-7, which are visually reviewed by a data analyst to flag potential issues. (5) Data are sampled from the central repository and sent back to the hospital site for reidentification and manual data validation. (6) Data issues are resolved, and the data are pushed into the live database for the Temporal-Hospital check. (7) Finally, data are standardised and added to the GEMINI database.

Data Quality

To ensure that GEMINI data are high quality, GEMINI performs rigorous data cleaning and validation processes for every batch of data it receives from participating hospitals. These quality checks aim to identify any potential errors in the data extraction work, including missing or incomplete variables, thus requiring ongoing engagement with hospital data teams to iteratively re-extract data. GEMINI has developed analytical processes to handle large volumes of data, along with workflows for removing personal identifiers, quality control, standardisation and validation. Detailed information on GEMINI's data flows, approach to evaluating and addressing issues with data quality and availability across key data elements, and comprehensive quality control processes have been previously published, and the methodology for ensuring data quality has been rigorously validated, demonstrating 98-100% agreement across key data elements compared with manual chart review of the electronic patient record [12]. Further information about data missingness is also available in the GEMINI Data Dictionary, as this is updated regularly and can be accessed here: https://geminimedicine.ca/the-gemini-database/.

In 2024, Fu & Shen et al. further described the maturation of quality assessment methods used for GEMINI data, including both computational data quality assessments and manual data validation [13]. This updated set of quality checks and the scalable, efficient, and comprehensive approach

to data quality assessment, as described in the paper, was necessary to adequately respond to GEMINI's substantial expansion of participating sites, data volume, and variety.

Data Standards

GEMINI data have been consistently standardised to an internal data model, based on our Data Reference Model. Research with this data model is supported with a wide range of tools available to users of the GEMINI research environments, including the Rgemini package (see subsequent section on software).

Research Software

GEMINI has developed a number of tools to improve and facilitate data standardisation, removal of personal identifiers, and analysis, many of which are available for external researchers conducting their work on HPC4Health. For example, Rgemini is a custom R package that contains a variety of functions used to perform standardised transformations and calculations with GEMINI data more efficiently. In keeping with GEMINI's commitment to open science, the Rgemini repository is publicly available on GitHub.

The *pyDeid* tool is a python-based customisation of the freely available, open-source de-identification software package "*deid*" and is used to de-identify the PHI in extracted clinical text notes [21]. Clinical text notes are an important component of the data GEMINI extracts from hospital electronic records, often containing valuable information about patients' diagnoses, medical history, clinical decision-making, and advanced care planning. However, free-text clinical notes may contain PHI (e.g., names, locations, date of birth, age, contact information and medical record numbers) that would otherwise be omitted from other data fields. Both *pyDeid* and the original *deid* program remove PHI from text, and where appropriate, replace it with realistic surrogate fictitious information. However, *pyDeid* is python-based and can process larger volumes of clinical notes (as collected by GEMINI) in a more timely manner and with higher accuracy than the original Perl-based deid software. With a deidentification speed of 0.48s per note, it outperformed other tools on various metrics, with a "best" accuracy value of 0.988, best precision of 0.889, best recall of 0.950, and best F1 score of 0.904 [21]. *pyDeid* can also be easily customised to a variety of locations and contexts outside of GEMINI's Ontario-specific clinical data.

The GEMINI-RxNorm system facilitates the processing of pharmacy data for a wide range of research questions. The tool is particularly useful for standardising unstructured pharmacy data in multisite data repositories. The GEMINI-RxNorm system automates the use of multiple RxNorm [22] tools in tandem with other datasets to identify drug concepts from unstructured medication data [23]. The user inputs a list of drugs to identify, GEMINI-RxNorm returns a sheet of predicted matches, and a clinical subject matter expert reviews that sheet to discard false positives. This process accurately returns 99.6% of pharmacy orders with >98% sensitivity and >90% positive predictive value, reducing the number of pharmacy orders for manual review by up to 99.99% and reducing the review time of those data from 30s to 5s per row of medication data.

GEMINI has also developed a tool to facilitate the semi-automated matching of country-specific ICD-10 codes to the Clinical Classification Software Refined (CCSR). CCSR groups the vast number of unique ICD-10 codes into clinically meaningful, mutually exclusive categories which can be used for a wide range of clinical research and quality measurement applications [24]. The GEMINI CCSR matching algorithm (available as an open-source package at https://github.com/GEMINI-Medicine/gemini-ccsr) expands the functionality of the original US-based CCSR tool [25] to (semi-)automatically categorise ICD-10 codes from other countries, and has been thoroughly validated on Canadian and Danish ICD-10 codes.

Further details on the tools above and GEMINI's other open-source software can be found at https://github.com/GEMINI-Medicine.

Discussion

Selected Research Applications

To date, GEMINI data have been used in over 100 publications and has supported 130+ grants across 13 institutions, cultivating a network of 1000+ scientists, students, and clinicians. Currently, there are over 85 active research projects using GEMINI data for a wide variety of research applications. Previous projects have made significant contributions to the COVID-19 pandemic response, including Canada's first and largest study describing hospitalisation for COVID-19 [26], equity-focused research on the disparate impact of COVID-19 on people with disabilities [27], the first study to examine critical drug shortages of the life-saving therapeutic tocilizumab [28], and development of machine learning models to predict the demand for personal protective equipment [29]. Other key areas of focus include examining variations in care in order to better identify quality improvement targets, developing and deploying clinical AI tools, and providing physicians and hospitals with confidential audit and feedback reports through Ontario Health's General Medicine Quality Improvement Network (GeMQIN) program [30]. These reports inform physicians and hospital leaders about their clinical care patterns and patient outcomes, including quality indicators spanning length of stay, readmission, in hospital mortality, routine bloodwork, advanced imaging, appropriate blood transfusion, and ordering of sedative-hypnotic medications.

Strengths/Limitations

One of GEMINI's core strengths can be found in the breadth and representativeness of the dataset. With more than 16 million people, Ontario is the largest unified health system in North America, with a single public insurance system (OHIP) covering one of the most racially and culturally diverse populations in the world. To date, GEMINI has collected data for ~2.5 million admissions (constituting over 20 billion data points) from 37 hospitals across all regions of Ontario. These richly detailed, multi-institutional hospital data are a unique system-wide resource, with demonstrated value for a variety of research and quality improvement initiatives. Recent linkage to the vast array of longitudinal population-based data holdings

hosted at ICES further strengthens the database, creating an even broader spectrum of contexts for which the data can be used.

Furthermore, GEMINI hosts a machine learning friendly computing environment, enabling the responsible development and deployment of AI and other data-driven solutions. Researchers are able to access a set of various tools and resources to facilitate more efficient research in GEMINI's environment and enhance user experience (see "Research Software" section above).

GEMINI's swift response to the various health system needs brought on by the COVID-19 pandemic exemplifies another key strength of the platform: its ability to deftly mobilise its data resources to generate critical research evidence in a timely manner. Researchers are able to rapidly initiate and carry out projects, without being delayed by a lengthy approval process (see "Data Access Process" section for further details). The agility of GEMINI's data platform makes it a vital resource for responding to urgent research needs and any future health system-wide threats.

Finally, GEMINI has also cultivated a dynamic and interdisciplinary community of knowledge users and researchers, including clinical teams, hospital administrators, and policymakers. These collaborations have allowed GEMINI to not only strengthen its existing capacity to use hospital data for research and quality improvement, but also provide novel opportunities to extend use of the GEMINI dataset to other diverse applications. For example, GEMINI's formal partnership with the Vector Institute, one of Canada's three AI Centres of Excellence, has made it one of the most widely used datasets in Canada for responsible development of AI in hospitals. GEMINI data are also being leveraged by CIHI for a wide range of initiatives, from creating tools to accelerate administrative health data coding, to improving the accuracy of health data, and deploying health AI. These tools will have a considerable impact on the timeliness and impact of CIHI data for research, quality improvement initiatives, and health system planning. These collaborations with the Vector Institute and CIHI are just two examples of the broad scope of applications for which the dataset can be used.

Despite the strengths outlined above and the unique benefits that GEMINI brings to Ontario's health data landscape, there are a few key limitations of the GEMINI dataset. Due to the time required for data transfer and preprocessing, as well as GEMINI's efforts to thoroughly harmonise and prepare the data for research, GEMINI data are typically available for research approximately 12 months behind real-time. Additionally, the GEMINI dataset contains limited individual-level equity variables. Although variables on disability, homelessness, age, and sex are available, GEMINI is not able to capture information on areas such as gender or individual-level race, as these variables are not available in routinely collected Canadian electronic hospital data. Additionally, GEMINI currently lacks Indigenous data and governance, though development of a separate, Indigenous-led work stream to govern how Indigenous data can be accessed and used to benefit Indigenous communities is in progress. Finally, GEMINI hospitals are mainly large urban and suburban hospitals, and thus smaller and rural hospitals are not well represented. Expansion of GEMINI's network and strengthening of internal capacity to address these shortcomings are well underway.

Future Directions

GEMINI continues to grow every year, expanding its network to enable more fulsome coverage of Canada's population. For example, GEMINI recently broadened its scope to start collecting data from all paediatric hospitalisations and emergency room presentations at existing GEMINI-participating hospitals, as well as several new specialty children's hospitals in Ontario. A small selection of pilot projects using paediatric data are currently underway, in preparation for broadening access to paediatric data to external researchers. These data will be described more fully in subsequent publications. Other key areas for expansion include engagement with smaller/rural hospitals to enhance geographic coverage and representativeness of the dataset, establishment of Indigenous community partnerships and Indigenous data governance, and development of multi-provincial hospital data partnerships to advance the platform nationally. These expansions will provide a larger, and more diverse, network of clinicians, health system leaders, and researchers with the ability to gain new insights from the rich data that GEMINI holds. Linkages with provincial administrative data centres, such as that which GEMINI currently has with ICES, will allow for a deeper understanding of patients and the health system.

Data Access

Researchers across Canada and internationally can request access to GEMINI data to answer research questions approved by the GEMINI Projects & Publications Committee, which is composed of representation across participating institutions and co-chaired by the GEMINI Principal Investigators (Razak/Verma). To ensure appropriate use of the GEMINI data, a GEMINI Investigator must be involved in all GEMINI research projects. GEMINI Investigators help researchers navigate GEMINI data access processes, provide advice based on their expert knowledge of the dataset, and ensure that projects are compliant with the GEMINI objectives and REB requirements. An updated list of GEMINI Investigators is available at: https://geminimedicine.ca/our-team/?profile_type=investigators.

Those interested in using GEMINI data are invited to submit a brief project proposal form to GEMINI.Research@unityhealth.to, using the standardised template document available on the GEMINI website. Once finalised and submitted to GEMINI.Research@unityhealth.to, project proposals are reviewed by two members of the GEMINI Projects and Publications Committee to ensure that the following criteria are met, with a final decision being communicated to the researchers within approximately 10-15 business days:

1. The project aligns with at least one of the core research objectives described in the GEMINI Hospital Cohort Study research protocol as approved by the REB of participating organisations, which may be amended from time-to-time.

2. The project does not duplicate an existing project that is underway.

3. The project is feasible, including that the relevant data are available through GEMINI, GEMINI-ICES, or other linkages approved by the REB and participating organisations.

4. The project has a clear objective, study plan, and deliverables.

5. Any disclosed conflicts of interest and industry ties adhere to GEMINI's policies on commercialisation.

If the project is not approved upon first review, researchers will have the opportunity to revise their proposal to address the reviewers' concerns. Once approved, GEMINI personnel will submit the project for REB approval, with Unity Health Toronto as the Board of Record. If the request for access to GEMINI data is part of a larger grant-funded study or if the study contains elements not covered by GEMINI's objectives/protocol, researchers may instead obtain independent REB approval from their home institution. In such cases, additional review and approval of the project by Unity Health Toronto's REB is not required. In all cases, research activities may not begin until the project has received REB approval.

For all projects, researchers are able to either hire services from a GEMINI Data Analyst to perform the analysis for the project and provide them with results, or directly access datasets for their project through the secure, remotely-accessible, high-performance computing environment at HPC4Health. If hiring services from a GEMINI Data Analyst, GEMINI personnel will work iteratively with the research team throughout the duration of the project to generate a data analysis plan, conduct the analyses, and provide results.

If researchers would like to directly access datasets and perform the analysis for their project themselves, GEMINI personnel will support them throughout the process of completing the required training/documentation, creating their credentials, and obtaining access to the GEMINI data platform through the secure HPC4Health cloud-based environment. In parallel, a GEMINI Data Analyst will create and provide the designated researcher(s) who will be accessing the data via HPC4Health with a de-identified datacut specific to their project objectives. All researchers accessing HPC4Health must adhere to GEMINI standard operating procedures regarding data access/use and privacy/security measures. Researcher access to HPC4Health will be controlled by GEMINI team members for the duration of the study.

All project proposals submitted to the GEMINI Projects & Publications Committee to date have been approved, with an average turn-around time of 12 days to initial project approval. The detailed process for accessing GEMINI data is described on our website at: https://geminimedicine.ca/access-data and any questions can be directed to GEMINI.Research@unityhealth.to.

For access to GEMINI-ICES linked datasets through ICES' secure computing environment, additional requirements must be met prior to starting research activities. All projects must also be approved through ICES' Information and Privacy Commissioner of Ontario (IPC)-approved policies and procedures, which requires submission of additional documentation to ICES.

GEMINI operates on a partial cost recovery model. Data access fees are minimised as much as possible and cover the costs of software licences for users in the secure research environment as well as the costs of project activation and dataset preparation. Analytical or methodological services, if required, are charged at an hourly rate. To obtain a project-specific quotation for data access or analysis, interested researchers may contact GEMINI.Research@unityhealth.to.

Conclusion

Harnessing electronic hospital clinical data is imperative to conduct high-impact research that addresses pressing healthcare challenges. GEMINI does this by collecting data from the EHR and administrative data generated by nearly 40 Ontario hospitals, resulting in a database covering over 20 billion data points and ~2.5 million hospital admissions. Dynamic communities of researchers, patients and caregivers, clinicians, and policymakers can use the essential data that GEMINI collects to answer important questions on clinical epidemiology, health systems planning, predictive analytics, and much more. Ultimately, GEMINI aims to improve hospital care by facilitating better access to large-scale, real-world acute care data.

Competing Interests

Amol Verma and Fahad Razak are part-time employees of Ontario Health beyond the scope of this work. Other authors report no disclosures. Amol Verma holds minority interests in Signal1, a health AI start-up company, for co-inventing the Chartwatch early warning system.

Acknowledgements and funding

This article was supported by GEMINI, a research program based out of Unity Health Toronto. The development of GEMINI's data platform has been supported by several funding partners, which can be found listed at https://geminimedicine.ca/partners. All analyses, results, conclusions, and opinions presented in this paper are solely those of the listed authors and are independent from GEMINI's funding sources. No endorsement by GEMINI's funding sources is intended, nor should it be inferred.

Ethics statement

Research ethics approval has been obtained from all participating GEMINI hospitals, as per the study protocols approved by Unity Health Toronto's REB (SMH REB: 20-216, CTO ID: 3344) and in full compliance with PHIPA and TCPS2.

Conflict of Interests Statement

AAV and FR are provincial clinical leads for Quality Improvement in General Internal Medicine with Ontario Health

(part-time employees), outside the submitted work. AAV holds minority interests in Signal1, a health AI start-up company, for co-inventing the Chartwatch early warning system. The other authors declare no conflicts.

Publication Consent

All authors confirm they have approved this data resource article for submission.

Data Availability Statement

Datasets from GEMINI are not permitted to be shared openly given they contain potentially sensitive patient information, based on the data governance policies of the GEMINI research network and its Research Ethics Board-approved study protocols. Data can be accessed in the secure GEMINI research environment. Information about data access can be obtained at https://geminimedicine.ca, by contacting GEMINI.Data@unityhealth.to, or upon request to the corresponding author.

AI Statement

The authors declare that no generative AI tools were used in the preparation of this manuscript.

References

1. Epic Cosmos. Accessed January 29, 2026. https://cosmos.epic.com/.

2. Clinical Practice Research Datalink (CPRD). Accessed January 29, 2026. https://ukhealthdata.org/members/clinical-practice-research-datalink-cprd/.

3. Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data*. 2023;10(1):1. https://doi.org/10.1038/s41597-022-01899-x

4. Verma AA, Guo Y, Kwan JL, et al. Patient characteristics, resource use and outcomes associated with general internal medicine hospital care: the General Medicine Inpatient Initiative (GEMINI) retrospective cohort study. *Can Med Assoc Open Access J*. 2017;5(4):E842-E849. https://doi.org/10.9778/CMAJO.20170097

5. Verma AA, Guo Y, Jung HY, et al. Physician-level variation in clinical outcomes and resource use in inpatient general internal medicine: an observational study. *BMJ Qual Saf*. 2021;30(2):123-132. https://doi.org/10.1136/bmjqs-2019-010425

6. Ontario Health Regions. Ontario Health. January 26, 2026. Accessed January 29, 2026. https://ontariohealth.ca/system/regions.

7. Colacci M, Loffler A, Roberts SB, et al. Patient Complexity, Social Factors, and Hospitalization Outcomes at Academic and Community Hospitals. *JAMA Netw Open*. 2025;8(1):e2454745. https://doi.org/10.1001/jamanetworkopen.2024.54745

8. Personal Health Information Protection Act, 2004, S.O. 2004, c. 3, Sched. A. Ontario. Accessed January 29, 2026. https://ontario.ca/laws/statute/04p03.

9. Government of Canada IAP on RE. Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans – TCPS 2 (2022). January 11, 2023. Accessed January 29, 2026. https://ethics.gc.ca/eng/policy-politique_tcps2-eptc2_2022.html.

10. UK Data. What is the Five Safes framework? UK Data Service. Accessed January 29, 2026. https://ukdataservice.ac.uk/help/secure-lab/what-is-the-five-safes-framework/.

11. Paprica PA, Crichlow M, Maillet DC, et al. Essential requirements for the governance and management of data trusts, data repositories, and other data collaborations. *Int J Popul Data Sci*. 2023;8(4):2142. https://doi.org/10.23889/ijpds.v8i4.2142

12. Verma AA, Pasricha SV, Jung HY, et al. Assessing the quality of clinical and administrative data extracted from hospitals: the General Medicine Inpatient Initiative (GEMINI) experience. *J Am Med Inform Assoc*. 2021;28(3):578-587. https://doi.org/10.1093/jamia/ocaa225

13. Fu A, Shen T, Roberts SB, et al. Optimizing the efficiency and effectiveness of data quality assurance in a multicenter clinical dataset. *J Am Med Inform Assoc*. 2025;32(5):835-844. https://doi.org/10.1093/jamia/ocaf042

14. High-Performance Computing for Healthcare. HPC4Health. 2026. Accessed February 5, 2026. https://hpcforhealth.ca/.

15. Our Organization. ICES. Accessed January 29, 2026. https://ices.on.ca/our-organization/.

16. Canadian Institute for Health Information. Measuring Health Inequalities: A Toolkit — Area-Level Equity Stratifiers Using PCCF and PCCF+. Published online 2018. https://cihi.ca/sites/default/files/document/toolkit-area-level-measurement-pccf-en.pdf.

17. Statistics Canada. Postal Code OM Conversion File Plus (PCCF+). March 6, 2017. Accessed January 29, 2026. https://150.statcan.gc.ca/n1/en/catalogue/82F0086X.

18. Srinivasan D, Wang R, Roberts SB, et al. Applying an implementation science lens to understand physician-level variation in patient length of stay in internal medicine. *BMC Health Serv Res*. 2025;25(1):1292. https://doi.org/10.1186/s12913-025-13304-5

19. Rawal S, Kwan JL, Razak F, et al. Association of the Trauma of Hospitalization With 30-Day Readmission or Emergency Department Visit. *JAMA Intern Med*. 2019;179(1):38. https://doi.org/10.1001/jamainternmed.2018.5100

20. Lam ACL, Tang B, Liu C, et al. Variation in Case Exposure During Internal Medicine Residency. *JAMA Netw Open*. 2024;7(12):e2450768. https://doi.org/10.1001/jamanetworkopen.2024.50768

21. Sundrelingam V, Parimoo S, Pogacar F, et al. pyDeid: an improved, fast, flexible, and generalizable rule-based approach for deidentification of free-text medical records. *JAMIA Open*. 2025;8(1):ooae152. https://doi.org/10.1093/jamiaopen/ooae152

22. RxNorm Technical Documentation. Accessed February 5, 2026. https://nlm.nih.gov/research/umls/rxnorm/docs/techdoc.html.

23. Waters R, Malecki S, Lail S, et al. Automated identification of unstandardized medication data: a scalable and flexible data standardization pipeline using RxNorm on GEMINI multicenter hospital data. *JAMIA Open*. 2023;6(3):ooad062. https://doi.org/10.1093/jamiaopen/ooad062

24. Malecki SL, Loffler A, Tamming D, et al. Development and external validation of tools for categorizing diagnosis codes in international hospital data. *Int J Med Inf*. 2024;189:105508. https://doi.org/10.1016/j.ijmedinf.2024.105508

25. Clinical Classifications Software Refined (CCSR). Healthcare Cost and Utilization Project. November 2025. Accessed February 5, 2026. https://hcup-us.ahrq.gov/toolssoftware/ccsr/ccs_refined.jsp.

26. Verma AA, Hora T, Jung HY, et al. Characteristics and outcomes of hospital admissions for COVID-19 and influenza in the Toronto area. *CMAJ Can Med Assoc J J Assoc Medicale Can*. 2021;193(12):E410-E418. https://doi.org/10.1503/cmaj.202795

27. Brown HK, Saha S, Chan TCY, et al. Outcomes in patients with and without disability admitted to hospital with COVID-19: a retrospective cohort study. *CMAJ*. 2022;194(4):E112-E121. https://doi.org/10.1503/cmaj.211277

28. Verma AA, Pai M, Saha S, et al. Managing drug shortages during a pandemic: tocilizumab and COVID-19. *CMAJ*. 2021;193(21):E771-E776. https://doi.org/10.1503/cmaj.210531

29. Furman E, Cressman A, Shin S, et al. Prediction of personal protective equipment use in hospitals during COVID-19. *Health Care Manag Sci*. 2021;24(2):439-453. https://doi.org/10.1007/s10729-021-09561-5

30. Ontario General Medicine Quality Improvement Network. Ontario Health. September 11, 2025. Accessed February 5, 2026. https://ontariohealth.ca/quality/hospital/general-medicine.

Abbreviations

AI:	Artificial Intelligence
CCI:	Canadian Classification of Health Intervention
CCSR:	Clinical Classifications Software Refined
CIHI:	Canadian Institute for Health Information
CIHR:	Canadian Institutes of Health Research
CPRD:	Clinical Practice Research Datalink
DAD:	Discharge Abstract Database
EHR:	Electronic Health Record
GeMQIN:	General Medicine Quality Improvement Network
ICU:	Intensive Care Unit
IPC:	Information and Privacy Commissioner of Ontario
NACRS:	National Ambulatory Care Reporting System
OHIP:	Ontario Health Insurance Plan
PHI:	Personal Health Information
QI:	Quality Improvement
REB:	Research Ethics Board