Article

An institutional framework to support ethical fair and equitable artificial intelligence augmented care

Check for updates

Steven Dykstra[1,2], Matthew MacDonald[1,2], Rhys Beaudry[1,2], Dina Labib[1,2], Melanie King[1,2], Yuanchao Feng[1,2,3], Jacqueline Flewitt[1,2], Jeff Bakal[3], Bing Lee[3], Stafford Dean[3], Marina Gavrilova[4], Paul W. M. Fedak[1,2] & James A. White [1,2,5] ✉

Coordinated access to multi-domain health data can facilitate the development and implementation of artificial intelligence-augmented clinical decision support (AI-CDS). However, scalable institutional frameworks supporting these activities are lacking. We present the PULSE framework, aimed to establish an integrative and ethically governed ecosystem for the patient-guided, patient-contextualized use of multi-domain health data for AI-augmented care. We describe deliverables related to stakeholder engagement and infrastructure development to support routine engagement of patients for consent-guided data abstraction, pre-processing, and cloud migration to support AI-CDS model development and surveillance. Central focus is placed on the routine collection of social determinants of health and patient self-reported health status to contextualize and evaluate models for fair and equitable use. Inaugural feasibility is reported for over 30,000 consecutively engaged patients. The described framework, conceptually developed to support a multi-site cardiovascular institute, is translatable to other disease domains, offering a validated architecture for use by large-scale tertiary care institutions.

Cardiovascular disease (CVD) remains the leading cause of mortality worldwide, contributing ~19 million deaths in 2020, a 19% increase from 2010[1]. Progressive growth in the prevalence of CVD is leading to an unsustainable expansion in healthcare costs, with CVD spending rising from $212 billion in 1996 to $320 billion in 2016 and projected to exceed $800 billion by 2036 in North America[2]. To address this, institutions must pursue innovative approaches for delivering personalized care, a field increasingly reliant on AI-based techniques to improve disease detection and predict clinical outcomes for individual patients[3–7]. The development, surveillance, and maintenance of these tools, collectively termed AI-augmented clinical decision support (AI-CDS), rely on achieving ethical access to high-quality matched data resources representative of an institution's local community. While seminal programs to accomplish this have been implemented by notable institutions, such as the Mayo Clinic[8] and Stanford[9,10], or have been established for sizeable multi-center research initiatives[11–14], there is a rapidly expanding need for individual institutions to support the development of AI-CDS using data locally sourced from their community.

In this paper, we describe a centrally governed program developed by the Libin Cardiovascular Institute (LCI) to address core barriers surrounding the ethical collection, contextualization, and curation of digital health data resources to support AI-CDS. Our work was focused on establishing a comprehensive, scalable, and interoperable platform for the longitudinal collection and surveillance of matched structured and unstructured (e.g., diagnostic imaging, ECGs, or dictated text) electronic health data in "model ready" form. Core requirements for this program, branded the PULSE program, included the: (i) routine and transparent engagement of patients for informed consent to engage their data resources for the evaluation of AI-CDS interventions, (ii) longitudinal capture of patient-reported social determinants of health (SDOH) and patient reported outcome measures (PROMs) for contextualization of AI prediction models, (iii) automated transformation of structured health data into task-relevant data schemas, (iv) automated pseudonymization and pre-processing of unstructured (or semi-structured) diagnostic data into model-ready assets, (v) shared governance

[1]Department of Cardiac Sciences, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada. [2]Libin Cardiovascular Institute, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada. [3]Alberta Health Services, Calgary, AB, Canada. [4]Department of Computer Science, University of Calgary, Calgary, AB, Canada. [5]Department of Diagnostic Imaging, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada. ✉e-mail: jawhit@ucalgary.ca

and operational data management services for all platform stakeholders, inclusive of clinicians, researchers, and data scientists.

Results
Not applicable

Discussion
In this paper, we summarize the cumulative knowledge, experience, and recommendations gained from developing a scalable institutional framework to support ethical AI-CDS. Our presented framework and methodology addresses core barriers routinely encountered by institutions aiming to leverage digital health data resources for research, innovation, and AI-augmented care. Built on foundational principles prioritizing patient engagement and consent, the PULSE program was designed to complement established healthcare information technology teams and services to enable longitudinal access to curated data products meeting rigorous acceptance standards for use in AI-CDS, including their contextualization to SDOH and PROMs.

The PULSE framework was conceived for and developed to service the needs of a cardiovascular institute. However, its core principles and design have been established to broadly support diverse domains of healthcare. Its modular infrastructure and patient-centered governance model was constructed to maximize adaptability to unique clinical environments and specialties. Furthermore, the PULSE program can be scaled from small (single site) to large (multi-site) hospital systems through the consistent use of interoperable cloud-based infrastructure, centralized data services and transformations, and common operational workflows. By addressing core barriers to healthcare data usage, PULSE offers a validated and versatile framework to guide the ethical and equitable integration of AI-enabled CDS into clinical practice. These collective processes and recommendations are presented with a desire to catalyze and support the adoption of similar platform-based approaches by other healthcare systems.

Methods
The Libin Cardiovascular Institute and Alberta Health Services
The LCI is located in Calgary, Alberta, Canada and is a joint entity of the University of Calgary and Alberta Health Services (AHS). Alberta Health Services (AHS) is a government-funded, publicly administered health authority. Cardiovascular patients are managed across four adult acute care hospitals in the Calgary region, supported by EPIC's centralized electronic health record (EHR) service. The cardiac sciences clinical department supports approximately 25 cardiovascular clinics and manages over 40,000 unique patients annually.

Guiding principles of the PULSE program
While the benefits of AI applications in healthcare have been shown across most clinical domains[4], real world adoption has remained slow relative to other industries. This is related to core barriers inclusive, but not limited to ethical, technological, regulatory, workforce-related, social, and patient safety concerns. A comprehensive systematic review highlighting each of these unique barriers was published by Ahemed, et al.[15]. The PULSE Program was designed to maximally address these barriers by establishing a scalable infrastructure to promote electronic health data's transparent and ethical utilization for personalized care innovation. Program development was driven by a local need to overcome barriers to the ethical abstraction, transformation, and use of multi-domain electronic health data to develop and explore AI-CDS. In this context, four guiding principles were established: (i) the usage of healthcare data resources to evaluate AI-CDS requires transparent and informed patient consent, (ii) prediction models should be routinely contextualized by SDOH and PROMs to mitigate bias, (iii) health data resources intended for use in AI-CDS should be purposely engineered to meet task-specific requirements and be longitudinally monitored for stability, and (iv) AI-CDS models should undergo iterative surveillance

for bias across demographic sub-groups. PULSE was designed to deliver an appropriate infrastructure, standard operational procedures, and governance to address these guiding principles.

PULSE program recommendations
To define infrastructure needs, extensive consultation was conducted with broadly representative stakeholders, including healthcare executives, administrators, clinical program leadership, clinical researchers, data science team leads, hospital information technology specialists, and institutional privacy and research ethics boards. This collaborative process established a list of 9 program recommendations aimed at addressing core barriers surrounding the implementation of AI-augmented decision support. These core barriers, and corresponding guiding principles are described in Table 1. These principles have been developed with a focus on the local needs of the institution, while ensuring the long-term sustainability and adaptability to evolving technologies and regulatory landscapes.

Patient engagement and consent for prospective data collection and usage
By prioritizing the transparent declaration of data collection and usage to the patient community and establishing permissions to evaluate validated models in clinical care pathways, PULSE adopts a voluntary "opt-in" consent for program participation. Upon arrival at any clinic environment of the institute, patients are invited to review a recruitment poster describing the PULSE program, accompanied by a QR code. Tablet devices are also made available in clinics for those patients without personal devices. Scanning of the QR code directs the patient to an informed consent document for the PULSE program, followed by clinic-configured health questionnaires. Digital copies of the consent and survey responses in the form of a report are automatically sent to the patient by email. Consent is provided for the automated abstraction, pseudonymization, and surveillance of multi-domain digital health data with iterative contextualization to electronic health surveys, as illustrated in Fig. 1. We also obtain permission for patient contact to participate in future intervention studies leveraging trained models.

Completion of program enrollment triggers the automated abstraction and pre-processing of pre-defined structured and unstructured data resources from institutional archival servers for a retrospective period of 25 years and a prospective period of 10 years under data disclosure agreements executed with the data custodian and approval of the REB. As a core, institutionally endorsed program, data management processes adhere to pre-defined data protection guidelines and common ethical and legal frameworks, avoiding duplicate and/or competing efforts encountered by project-specific approaches. Pre-processed (i.e., model-ready) data resources are maintained for centralized access by institutional investigators using a commonly agreed-upon set of usage principles, monitored in partnership with the legal data custodians.

Social Determinants of Health (SDOH) and Patient Reported Outcome Measures (PROMs)
While broadly recognized to provide strong contextual value for predicting cardiovascular outcomes[16-18] and of critical importance for the fair and equitable delivery of AI models[19-21], SDOH and PROMs remain the least consistently captured variables by electronic health records. This primarily reflects a lack of configurable and patient-facing tools that provide contextualized adjustment to patient described health features. PULSE emphasizes patient involvement for routinely including SDOH and PROMs in personalized prediction models, permitting iterative assessments and calibration of model performance across diverse populations. To achieve this, standardized instruments were engineered to deliver relevant content across clinical domains, including gender, ethnicity, employment, and access to healthcare resources, while obtaining periodic assessments of patient-reported health status and quality of life. As shown in Figs. 1, 2, instruments are deployed upon program enrollment, followed by periodic

Table 1 | PULSE program recommendations for addressing core barriers surrounding the development and implementation of AI-augmented Clinical Decision Support (AI-CDS)

| Identified barrier	PULSE program recommendations
The secondary use of health data resources for purposes that may influence future patient care decisions requires appropriate justification and informed patient consent	Transparently develop and disclose a justifiable need to patients for the secondary use of their health data to support AI-CDS, inclusive of informed patient consent and research ethics board (REB) oversight.
Inconsistent data access practices can lead to non-compliant data usage and inappropriate data migration	Centralize data access under common legal data sharing agreements (DSA) established with the data custodian. Establish custodian approved environments for investigator and/or partner access to curated data resources with appropriate compute resources.
Lack of clear governance for data protection, usage, and monitoring	Establish objective governance structures for establishing and maintaining policies for data protection, acceptable usage, access security, and use monitoring.
Inconsistent or unreliable de-identification and data migration processes can threaten patient privacy	Develop and validate scalable, privacy-compliant, and auditable data de-identification and migration pathways for structured and unstructured data to destination servers, inclusive of data lineage and provenance logs.
Lack of structured data stewardship and oversight processes for post-access data monitoring	Establish data stewardship processes for the monitoring and reporting of data resource usage across academic, clinical and external partner activities.
Poorly validated or inconsistent data transformations limits real-world implementation of trained AI-CDS	Establish, validate and version-lock data transformations, quality assurance processes, and pre-processing pipelines for the generation of reproducible data products appropriate for model training and clinical inference. Execute longitudinal surveillance for concept drift.
Lack of data interoperability prevents the adoption and external validation of trained AI-CDS	Develop and maintain common data models as versioned data schemas that are compliant or contextualized by validated data ontologies, ensuring interoperability with international and National Institute of Standards and Technology (NIST) data standards.
Poor and bureaucracy-ladened access processes to health data prolongs innovation cycles and prevents appropriate model surveillance	Provide hierarchical, administratively controlled, and auditable access to regularly refreshed and curated de-identified data products for approved investigators or strategic partners within intuitive and cloud-accessible environments.
Lack of systematic documentation and tracking of program outputs limits support for ongoing investment	Implement tracking and reporting of program outputs to assess impact on scientific productivity, intellectual property development, technology transfer, and healthcare outcomes.

Patient engagement Data abstraction Clinical Decision Support (CDS)
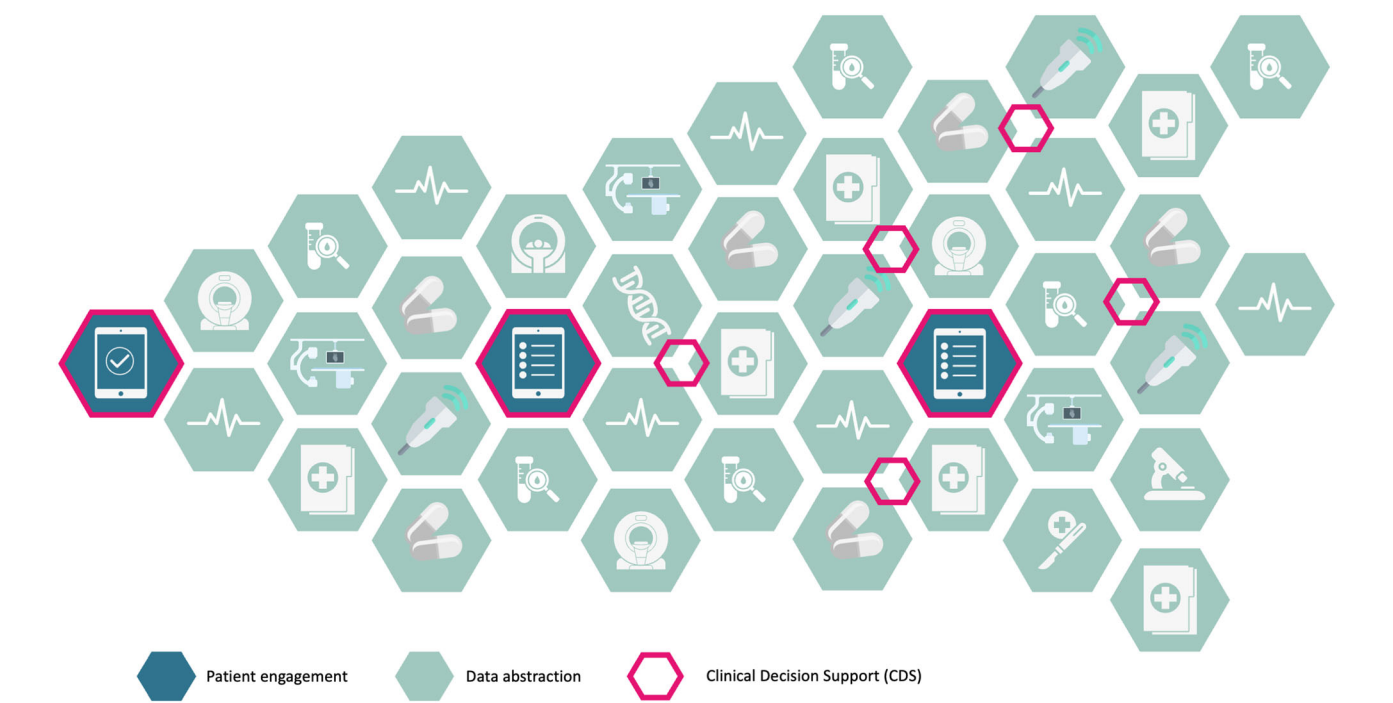
Fig. 1 | Multi-Domain Data Growth for AI Augmented Clinical Decision Support. Illustration of patient level data resource growth initiated by informed consent for data abstraction, capture of SDOH and longitudinal collection of PROMs. Automated data abstraction performed by pre-defined data schemas. Clinical decision support (CDS) interventions deployed based on iterative surveillance of combined data resources.

updating of health features at minimum time intervals by automated e-link surveys or at the time of repeat clinic visits.

An inherent benefit of establishing pre-defined data schemas inclusive of SDOH and PROMs is the capacity to monitor for population drift and to conduct iterative surveillance of model calibration within demographic sub-groups. Temporal shifts in local population demographics should prompt targeted re-assessment of model performance to identify emerging bias and, as required retrain or refine models to balance performance across sub-cohorts, this being a pre-defined role of the Data Advisory Group in their mandate to maximize fairness and equity. It is,
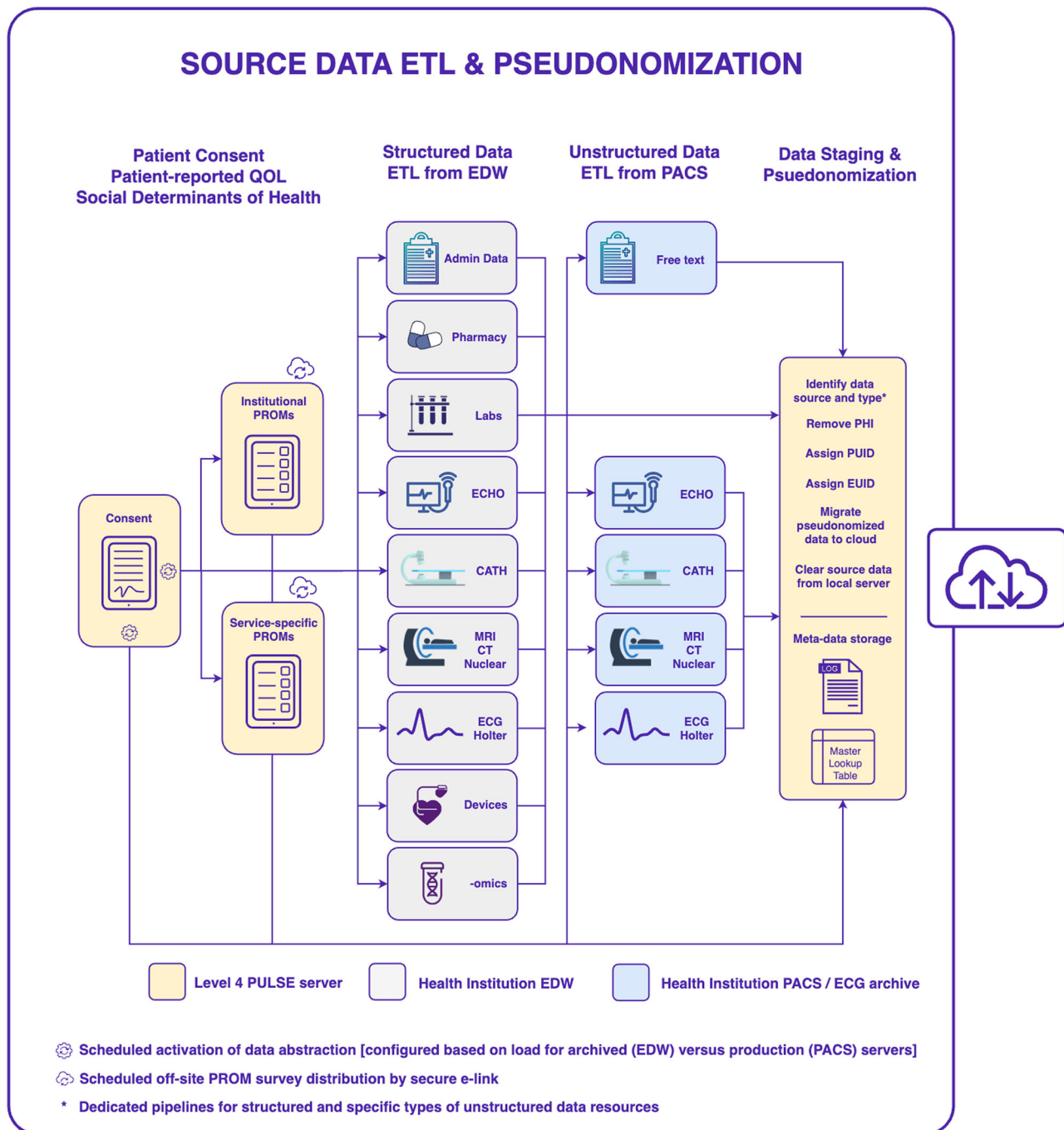
Fig. 2 | Overview of PULSE program data extraction, transfer, and loading (ETL) process. Architectural design for automated and iterative extraction, transformation, and loading of structured and unstructured cardiovascular data resources for consented program participants using centralized pseudonymization. Following patient self-guided program enrollment and consent in clinics, pan-institutional and service-specific patient-reported outcome measures (PROMs) are collected in addition to social determinants of health (SDOH) using point-of-service and/or remote (e-link) electronic surveys. Patient consent status is monitored by a Level 4 (PHI identified) server for automated query-retrieve of eligible data resources from institutional data repositories. Resource-specific pipelines for data pseudonymization are then executed prior to their migration to a cloud data lake for authorized end-user activities.

however, recognized that participation bias remains uniquely challenging in the context of consent required programs, limiting access to potentially beneficial AI-CDS in those not comfortable having their data used for such purposes.

Retrospective data enrichment

In addition to patient consent for prospective data usage and surveillance, PULSE endeavours to ethically access foundational institutional data resources for model development and discovery. While lacking the contextual features delivered by prospective enrollment, retrospective data resources are critical to support the training of foundational models. To facilitate this, PULSE adopted a REB-approved waiver of consent method to access the Institute's historical data resources in partnership with and under the supervision of the local health authority. These data resources are processed using identical de-identification and pre-processing pipelines to establish a curated core data resource for model training that is then

iteratively enriched through prospective data collection (accompanied by SDOH and PROMs).

Compliance with patient privacy regulations

The regulatory landscape of personal data protection and privacy in healthcare is anticipated to evolve rapidly over the next several years[22,23]. Recent landmark announcements from the European Union[24] and the United States[25] focus on the imminent introduction of enforceable boundaries for data usage in AI applications, expanding on established, regionally enforced legislation surrounding the management of protected health information (PHI). While the US Health Information Portability and Protection Act (HIPAA) delivers core recommendations surrounding PHI[26], regional adherence to PIPEDA in Canada[27] and global compliance with GDPR for members of the EU (https://eur-lex.europa.eu/eli/reg/2016/679/oj) contribute to complex requirements that will soon be compounded by AI-specific regulations. Accordingly, scalable and adaptable architectures for adherence to migratory regulatory mandates are required. To address this, PULSE developed bespoke and configurable solutions to permit efficient and adaptable pseudonymization services, substantially reducing barriers to migrating data resources to cloud-hosted environments. This service, installed on a virtual machine behind the healthcare institution's firewall, ingests and processes incoming data assets from archival servers and applies asset-configurable transformations to remove HIPAA[26]/PIPEDA protected patient identifiers (inclusive of meta-data fields accompanying unstructured data), replacing all patient level identifiers with a single 128-bit universally unique identifier (UUID). Each unique data asset (e.g., lab test, imaging study, ECG, etc.) is similarly encoded with its own UUID to permit provenance tracking as an independent resource.

Diagnostic testing data assets are uniquely managed by dedicated microservices for pseudonymization, metadata abstraction, and removal of pixel-encoded private health information (e.g., for ultrasound images), as required. These processes strictly adhere to validated processes for source hardware and software versions, mandating core testing and approval for newly encountered sources. This testing includes confirming appropriate performance on destination platforms (e.g., DICOM image viewers and analysis software).

Guidance in HIPAA posits that maintaining dates beyond the year alone may identify patients when accompanied by contextual data. However, the absence of dates limits data assets' capacity to express temporal associations. PULSE universally shifts record dates at the unique patient level, including their date of birth, by a fixed but random number of days, maintaining within-record temporal associations. This shift (±200 days to ensure a 1-year random period) is generated using a deterministic algorithm (i.e., one-way-hash) inclusive of the patient's randomly generated PUID, allowing for future de-coding by authorized personnel.

Cloud migration of pseudonymized data assets

Server and compute infrastructures for large-scale institutional data platforms have migrated significantly over the past decade, adopting cloud-hosted solutions that deliver best-in-class cyber-security, scalable storage and compute resources, and extensible access to bespoke solutions for healthcare data management[28]. The PULSE cloud architecture follows a "lake-house" design where pseudonymized data assets are migrated to multiple data storage lakes upon completion of on-premises processing, with structured data subsequently duplicated to an online analytical processing (OLAP) data warehouse. OLAP databases store data in a columnar format for efficient combinatorial processes, which comprise most of the analytical workload. Unstructured imaging data resources are mounted to a web-DICOM service with an extensible open-source DICOM viewer (Open Healthcare Imaging Framework, OHIF) to allow for rapid access, visualization, and data labeling[29]. An architectural diagram of this cloud environment is provided in Fig. 3. In addition to using unstructured data for

model training, feature extraction from these resources can be stored in the relational database to expand structured data resources, as shown in Fig. 4.

Semantic and temporally sensitive data schema design

All data assets of PULSE are components of a purposely engineered data product designed to support AI-CDS in cardiovascular care and are composed of pre-defined schemas. The conceptual design of PULSE was inspired by a desire to consider the semantic meaning and temporal relationships of patient interactions with the healthcare system. Defining each clinical encounter (e.g., hospital or clinic visit), procedure (e.g., surgery or intervention), or diagnostic test (e.g., imaging test, ECG) by a standardized data schema enables the institution to deliver data products that can be certified for use in AI-CDS. In the context of PULSE, we combine these clinical schemas with patient-reported SDOH and PROMs, establishing a data model purposely engineered to support fair and equitable healthcare AI.

Data schemas are systematically developed, approved, versioned, and migrated to production environments for prioritized clinical targets, as graphically illustrated in Fig. 5. This activity is centrally managed by a Data Advisory Group charged with prioritizing AI-CDS targets for the Institute and ensuring each schema is developed inclusive of relevant stakeholders, inclusive of clinical domain experts (e.g., diagnostic test schemas) and patient representatives (e.g., to inform clinic-specific health questionnaires to mitigate bias). This is operationalized as a longitudinal consensus-based activity inclusive of defined stages of schema development, as shown in Fig. 5.

Special considerations for structured versus unstructured data resources

A list of core institutional data resources, structured and unstructured, routinely abstracted from institutional data archival servers, is presented in Fig. 2. Structured data are considered any data provided in a format readily managed by relational databases (i.e., categorical or numerical values). Most data resource formats held by healthcare institution electronic data warehouses (EDWs) are structured, except for free text (e.g., dictated notes). This includes laboratory, pharmacy, radiology information systems, electronic health record captured variables, and administratively coded hospital diagnoses, outcomes, and procedures. The latter are coded by locally applicable ICD-10 standards (ICD-10-CA and Canadian Classification of Health Interventions (CCI)).

The abstraction of structured data resources is executed through automated extract, transformation, and load (ETL) services developed using programmatic query-retrieve mechanisms (SQL SELECT queries), followed by the transfer of processed (schema) data resources into the OLAP data warehouse. Similar ETL and migration services for non-EDW data sources are feasible for resources deemed of value for personalized care discovery (e.g., phenomics, genomics, proteomics, metabolomics, wearables, etc.).

Data transformations

Before migrating structured or unstructured data resources to production environments in model-ready formats, common transformations must be considered. These transformations are essential to ensure that matched resources are available for model training and future clinical deployment (inference). A summary of transformations commonly applied by PULSE is summarized in Fig. 6, defined during schema development and approval. For example, structured data must be routinely cleaned to harmonize data inputs, correct errors, or re-code to clinically relevant variable definitions. Unstructured data resources typically require common and task-specific pre-processing before classification or prediction models are used. These processes, each critical for appropriate and efficient development and deployment of AI-CDS tools, extend beyond the operational scope of conventional health data analytics teams and become core justification for centralized healthcare AI services.
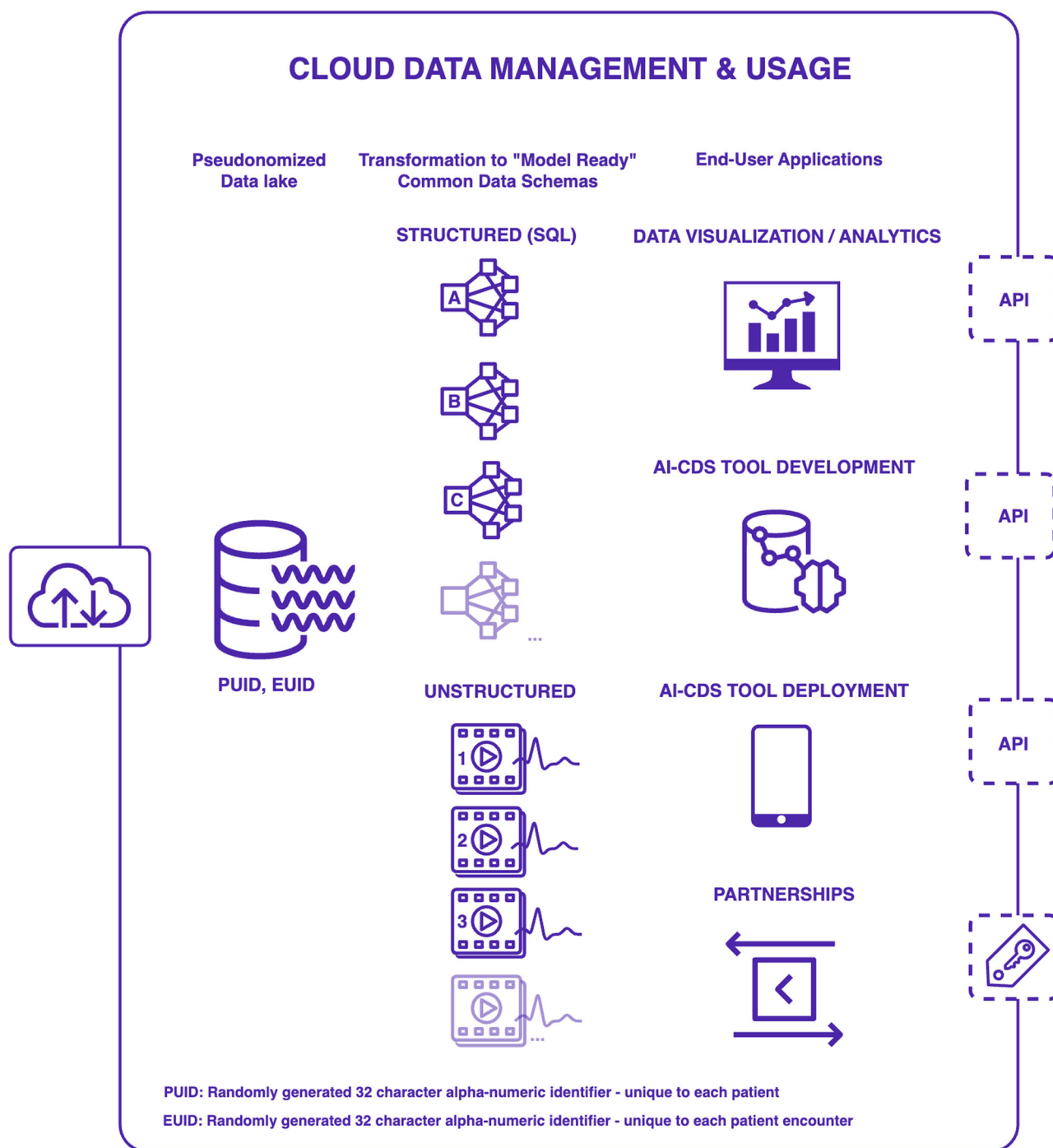
Fig. 3 | Overview of PULSE program cloud infrastructure. Coordinated migration of certified and validated data schemas from structured (SQL relational database) and unstructured (web-DICOM server and non-relational databases) resources by common, subject-specific PUIDs. Data schemas are made available by their combined attachment to project-specific data packages that are provisioned to virtual machine (VM) environments on the host server (accessed by secure tokens) accompanied by appropriate access to GPU and CPU compute resources. Application program interface (API) data exchange can be supported for relevant software applications enabling iterative tabular data visualization (e.g., Tableau, Microsoft Power BI, etc.) and unstructured data analysis (e.g., image or signal processing).

Data freshness and lineage

PULSE is designed to deliver an iteratively refreshed, centrally governed resource for AI-CDS research and innovation. In this context, the achievable cadence of data freshness varies based on its source and its dependency on clinically active (i.e., production) servers. Our institution's adoption of routine archival services for relevant data resources reduces latency for these activities, providing a dedicated "source schema" within the institutional EDW for structured data. Unstructured data resources (e.g., imaging) are considered sufficiently large in storage requirements to preclude routine archival duplication to the EDW and are therefore migrated by scheduled batch processing from PACS production servers.

PULSE supports data lineage tracking through automated logging of data ingestion and pre-processing tasks, coded with their versioned transformation operations. Best practices for development operations (DevOps) are followed such that any changes to ETL codebases (typically implemented in SQL or Python) are version-controlled with traceability of contributions by developers. Schema design is parameterized into configuration files so any changes (e.g., variable renamings) are versioned. Additionally,
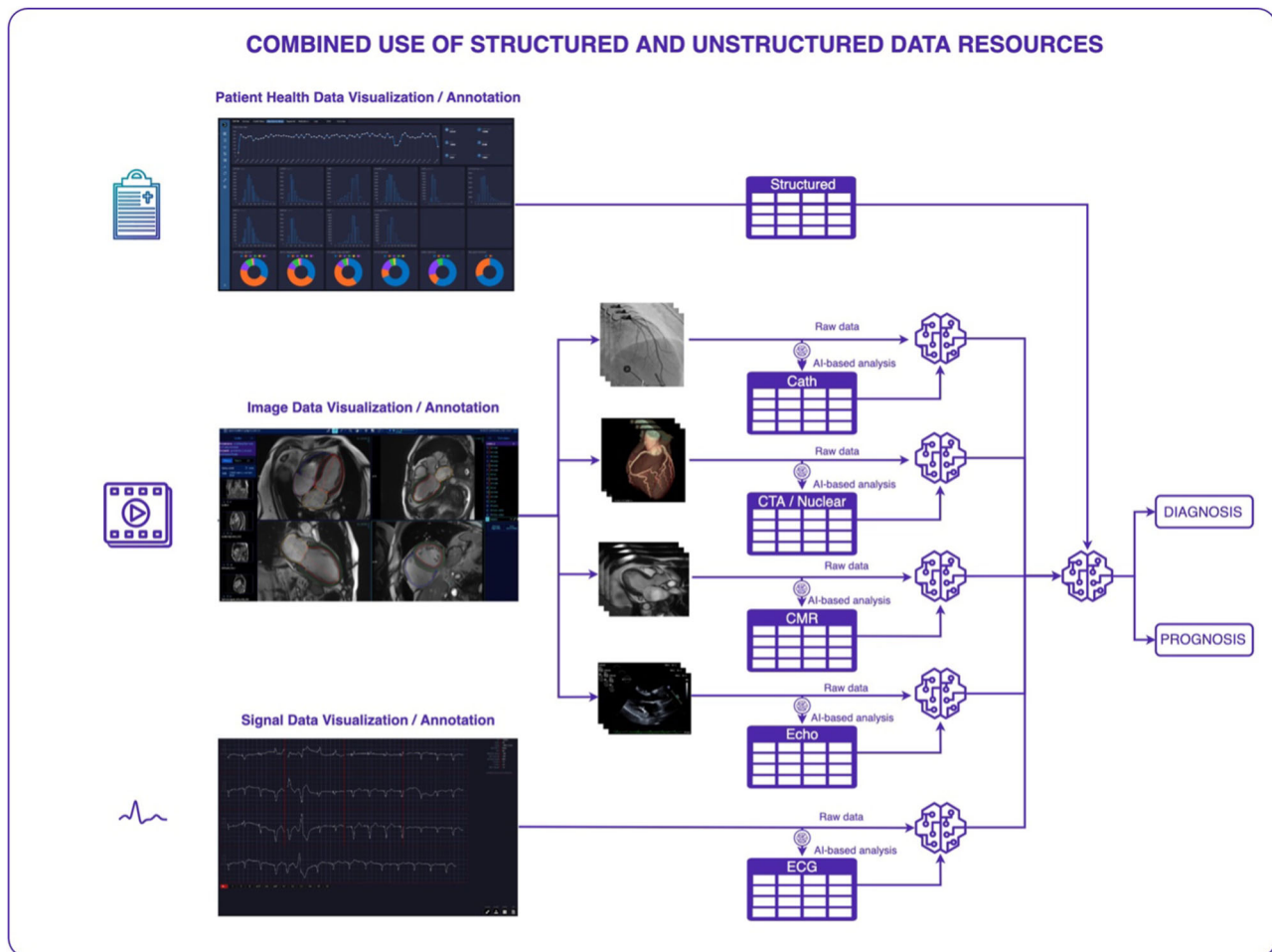
Fig. 4 | Approaches for the combined use of multi-domain data resources to deliver personalized care. Patient-reported and electronic health record abstracted data are routinely delivered as structured data resources for model training and inference. Image and signal data can be modeled directly (as unstructured data resources) or passed through validated AI-enabled analytics pipelines for feature extraction, then entered as structured data resources for prediction modeling. Composite data resources from each discrete data source can then be considered by ensemble-based prediction models to maximize prediction accuracy from available input models.

ingestion and delivery data pipelines operate in distinct development and production environments, adhering to rigorous deployment protocols that facilitate regression testing before their release. Software testing is automated using unit and system-level tests that run before deployment. These best practices help identify issues early and ensure data products planned for use in AI-CDS are certified. Finally, time-stamped ingestion, delivery, and usage logs are maintained across all data pipelines, including structured and unstructured assets. These facilitate internal audit and quality assurance processes essential to system compliance with institutional policies.

Data governance, accessibility, and usage
The delivery of large-scale curated digital health data resources presents numerous opportunities for engaging in research and innovation activities. However, accompanying these are inherent concerns regarding the sensitive nature of health data usage in the context of regional and federal laws governing its use. These concerns were highlighted by the recent American Heart Association (AHA) recommendations for health information collection, sharing, and use[30]. Therefore, establishing scalable and reproducible approaches to evaluate, approve, and monitor the appropriate usage of data (e.g., data stewardship) is essential. PULSE adopts a centralized and standardized intake service for all data resource requests. Requests are reviewed by a Steering Committee for compliance with program-defined mandates, and approved requests are granted cloud access to pre-defined (schema-based)

resources. This access is provided using a "data under glass" approach, where all data is securely mounted to a provisioned virtual machine to permit in-cloud processing by research investigators and/or partners. Cloud hosting permits access to scalable GPU or CPU compute resources to support model training and inference without data transfer or relocation.

Delivery of outbound CDS to clinical care teams
Current regulatory pathways for delivering CDS in clinical settings consider the intended use, degree of transparency and explainability, and potential risk to patient safety. These were clarified in the 2022 FDA Guidance Document for Clinical Decision Support Software, available at https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-decision-support-software. Based on these guidelines, a tool designed to summarize health information to identify eligibility for guideline-recommended care may not require regulatory approval. In contrast, models designed to diagnose disease or predict benefit from therapeutic interventions are more likely to be considered a regulated medical device given their intention to alter clinical judgment.

Regulatory exempt CDS commonly use limited tabular variables routinely available from the EHR and lack need for AI support, making deployment from EHR-based workflows appropriate. In contrast, AI-CDS focusses on early diagnosis and/or therapeutic guidance by leveraging multi-domain source data, such as medical images, ECGs, or other
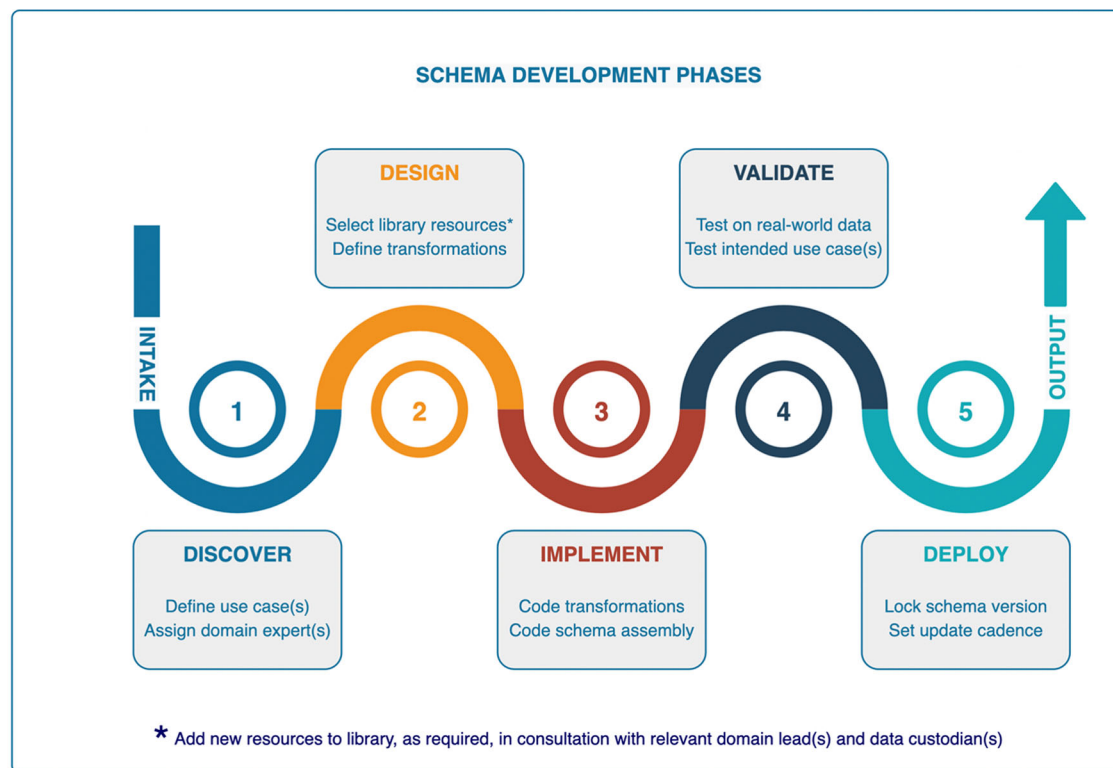
Fig. 5 | Schema development. Phases of schema development leveraging curated data resources of the PULSE program library. All schema requests are registered and assigned a unique ID for data lineage tracking and versioning. The PULSE data analyst works with designated domain expert(s) to define use case(s) prior to entering the design phase. Design is executed by selection of relevant variables from the existing library with identification of required transformations (establishing new variables in the library), followed by sourcing new variables from other sources (also added to the library). Schema constructed by the data engineer are then passed back to the data analyst for validation using real-world data and feasibility of use case deployment. Upon validation schema configuration is locked, versioned, and scheduled for desired data generation schedule.

wearable technology data combined with patient-specific health markers. Supported by the described platform, outbound messages can be delivered by API-based communications, delivering messages (HL7 or FIHR) for the posting of descriptive reports (e.g., PDF-based) to the medical record or for the triggering of configured actions within the local EHR. Establishing these integrations requires close collaboration and partnership with both administrative and operational teams in addition to approvals from administrative leaders (e.g., Chief Medical Information Officer).

Deployment and program validation
Before institutional approval for the PULSE program launch across all clinic environments, inaugural feasibility testing of our program design was performed across four clinic locations at two hospitals, conducted under REB-approval as the Cardiovascular Imaging Registry of Calgary (CIROC). This service-focused pilot has successfully deployed self-directed electronic patient consent for multi-domain data abstraction and automated SDOH/PROM deployment across 36,979 encounters in 28,829 unique patients. Program enrollment rates of 88% for out-patients and 80% for in-patients have been achieved. Cumulative matched data resources from this pilot implementation have delivered objective value for exploring personalized cardiovascular care. For example, studies focused on phenomics-based prediction of cardiovascular outcomes, inclusive of heart failure hospitalization[31–33], sudden cardiac death[32], atrial fibrillation recurrence (following ablation)[34], and composite major adverse cardiovascular events[35] have been supported. Genotype-phenotype association[36] and phenotype discovery studies have also been facilitated in patients with active cancer[37], genetic

cardiomyopathy[38] and infiltrative cardiomyopathy[39]. The value of multi-domain data integration to support AI-CDS for ML-based prediction of heart failure hospitalization[40] and new-onset atrial fibrillation[41] have been trained and validated. Finally, unstructured data resources have been leveraged to support the development of novel image-processing pipelines[38,42–51].

Long-term sustainability
The PULSE program is designed with sustainability and adaptability in mind, ensuring alignment with evolving technologies and regulations. Its modular infrastructure allows seamless updates to components such as AI models, data pipelines, and cloud systems. A dedicated data governance team actively monitors changes in privacy laws and institutional policies to maintain compliance, while regular stakeholder engagement ensures relevance and alignment with community needs. To address model drift and evaluate equity and fairness, PULSE is incorporating protocols for periodic review and revalidation of AI models using updated data and patient feedback. Institutional investment ensures long-term operational support, while collaborations with technology vendors and research partners provide access to emerging tools and innovations. Finally, an embedded monitoring and evaluation framework continuously assesses the program's impact and guides iterative improvements, ensuring PULSE remains a dynamic and sustainable resource for AI-CDS development and implementation.

Pulse implementation – overview and timelines
A graphical guide to the implementation of the PULSE program is provided in Supplementary Figs. 1 and 2. These respective figures
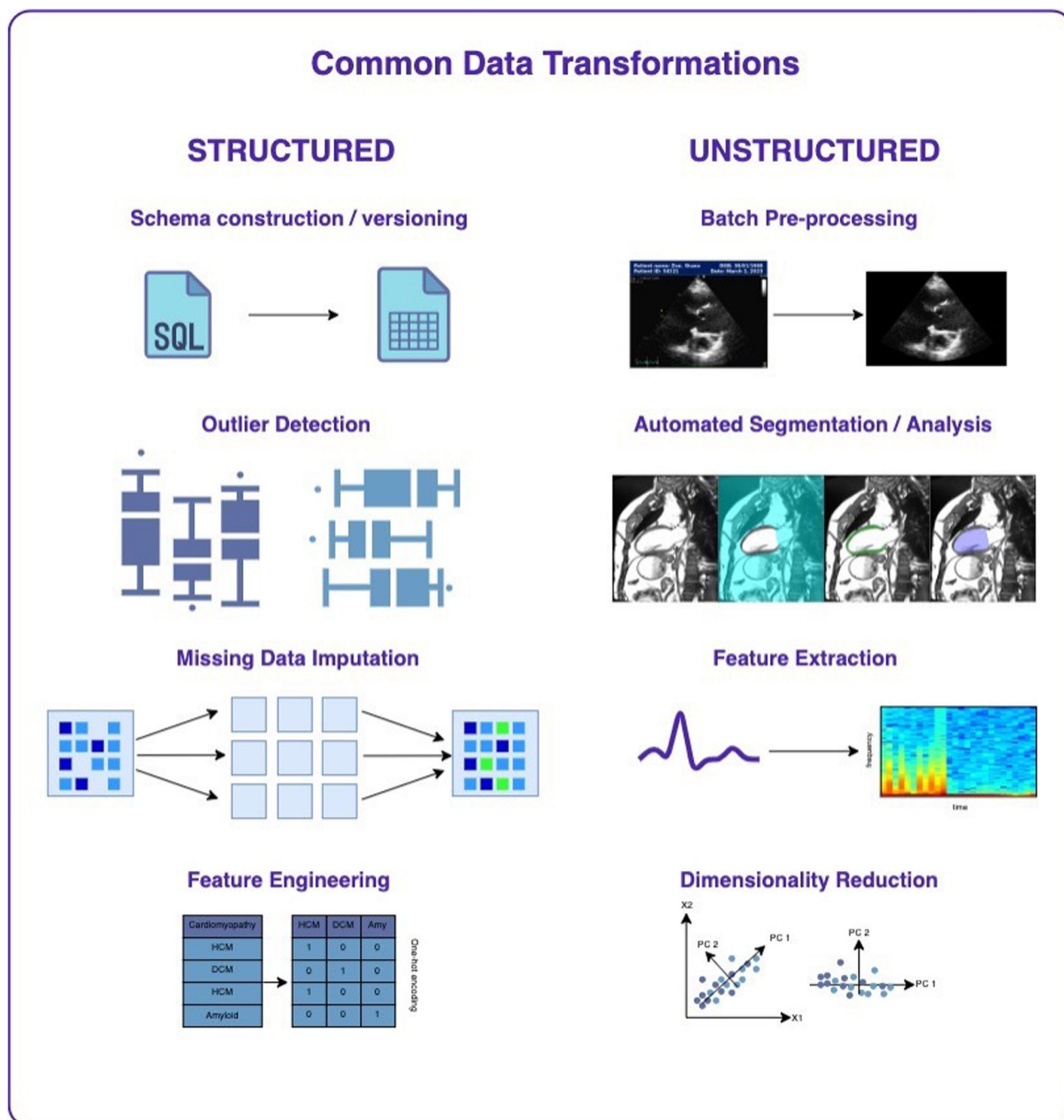
Fig. 6 | Common data transformations. Common data transformations applied to meet data product specifications for a publishable schema. All data transformations are documented and versioned, establishing each schema's data lineage, provided as part of each published schema's data dictionary. For structured data, this includes methodologies applied for managing outliers, missing data, and feature engineering. For unstructured data, this includes transformation applied to prepare raw data, select target data regions (e.g., segmentation), extract features, and performance of dimensionality reduction.

provide milestone completion targets for both a preparatory phase (inclusive of establishing a leadership team, hiring of core personnel, and development of core infrastructure and compute capabilities), and deployment phase (training, integration, shadow testing, and deployment of AI-CDS). Timelines were based on the achieved completion dates at our institute, however, will vary based on local resource availability. Additionally, a summary of core personnel required to support the PULSE program and estimated costs for implementation and maintenance are presented in Supplementary Table 1.

While additional resources developed for the PULSE program are not publicly available at this time, we are dedicated to fostering collaboration to support similar initiatives. Institutions or individuals interested in accessing specific PULSE resources, including data schema templates, data governance documentation, and training materials, are encouraged to contact us directly to explore potential data-sharing agreements.

Data availability

The data resources generated and analyzed during this study are not publicly available but can be shared upon reasonable request. Interested researchers are encouraged to contact the corresponding author to discuss potential access and usage conditions.

Code availability

The code supporting this study is not publicly available but can be shared upon reasonable request. Interested parties are encouraged to contact the corresponding author to discuss potential access and usage conditions.

References

1. Benjamin, E. J. et al. Heart Disease and Stroke Statistics—2019 Update: A Report From the American Heart Association. *Circulation* **139**, e56–e528 (2019).
2. Birger, M. et al. Spending on cardiovascular disease and cardiovascular risk factors in the United States: 1996 to 2016. *Circulation* **144**, 271–282 (2021).
3. Alowais, S. A. et al. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Med. Educ.* **23**, 689 (2023).
4. Davenport, T. & Kalakota, R. The potential for artificial intelligence in healthcare. *Future Healthc. J.* **6**, 94–98 (2019).
5. Dey, D. et al. Artificial intelligence in cardiovascular imaging. *J. Am. Coll. Cardiol.* **73**, 1317–1335 (2019).
6. Elias, P. et al. Artificial intelligence for cardiovascular care - Part 1: advances: JACC review topic of the week. *J. Am. Coll. Cardiol.* **83**, 2472–2486 (2024).
7. Jain, S. et al. Artificial intelligence in cardiovascular care — Part 2: applications: JACC review topic of the week. *J. Am. Coll. Cardiol.* **83**, 2487–2496 (2024).
8. Zhang, J. et al. Moving towards vertically integrated artificial intelligence development. *Npj Digit. Med.* **5**, 1–9 (2022).
9. Lowe, H. J., Ferris, T. A., Hernandez, P. M. & Weber, S. C. STRIDE – An integrated standards-based translational research informatics platform. *Amia. Annu. Symp. Proc.* **2009**, 391–395 (2009).
10. Callahan, A. et al. The Stanford Medicine data science ecosystem for clinical and translational research. *JAMIA Open* **6**, ooad054 (2023).
11. Putzier, M. et al. Implementation of cloud computing in the German healthcare system. *NPJ Digit. Med.* **7**, 12 (2024).
12. Hampf, C. et al. A survey on the current status and future perspective of informed consent management in the MIRACUM consortium of the German Medical Informatics Initiative. *Transl. Med. Commun.* **6**, 7 (2021).
13. Probst-Hensch, N. et al. Swiss Cohort & Biobank – The White Paper. *Public Health Rev.* **43**, 1605660 (2022).
14. Sudlow, C. et al. UK Biobank: an Open Access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Med.* **12**, e1001779 (2015).
15. Ahmed, M. I. et al. A systematic review of the barriers to the implementation of artificial intelligence in healthcare. *Cureus* **15**, e46454 (2023).
16. Rumsfeld, J. S. et al. Cardiovascular health: the importance of measuring patient-reported health status: a scientific statement from the American Heart Association. *Circulation* **127**, 2233–2249 (2013).
17. Sanchez-Gavilan, E. et al. Added value of patient-reported outcome measures (PROMs) after an acute stroke and early predictors of 90 days PROMs. *J. Patient-Rep. Outcomes* **6**, 66 (2022).
18. Verma, D., Bach, K. & Mork, P. J. Application of machine learning methods on patient reported outcome measurements for predicting outcomes: a literature review. *Informatics* **8**, 56 (2021).
19. Cachat-Rosset, G. & Klarsfeld, A. Diversity, equity, and inclusion in artificial intelligence: an evaluation of guidelines. *Appl. Artif. Intell.* **37**, 2176618 (2023).
20. Celi, L. A. et al. Sources of bias in artificial intelligence that perpetuate healthcare disparities—A global review. *PLOS Digit. Health* **1**, e0000022 (2022).
21. Ueda, D. et al. Fairness of artificial intelligence in healthcare: review and recommendations. *Jpn. J. Radiol.* **42**, 3–15 (2024).
22. Da Silva, M., Flood, C. M., Goldenberg, A. & Singh, D. Regulating the safety of health-related artificial intelligence. *Healthc. Policy* **17**, 63–77 (2022).
23. TerKonda, S. P. & Fish, E. M. Artificial intelligence viewed through the lens of state regulation. *Intell. -Based Med.* **7**, 100088 (2023).
24. Artificial Intelligence Act: deal on comprehensive rules for trustworthy AI | News | European Parliament. http://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai (2023).
25. Aiming for truth, fairness, and equity in your company's use of AI. *Federal Trade Commission* http://www.ftc.gov/business-guidance/blog/2021/04/aiming-truth-fairness-equity-your-companys-use-ai (2021).
26. U.S. Department of Health and Human Services. Health Insurance Portability and Accountability Act of 1996 (HIPAA). Public Law 104-191. 1996. https://www.hhs.gov/hipaa/.
27. Statutes of Canada. Personal Information Protection and Electronic Documents Act (PIPEDA). S.C. 2000, c. 5. https://laws-lois.justice.gc.ca/eng/acts/P-8.6/.
28. Dash, S., Shakyawar, S. K., Sharma, M. & Kaushik, S. Big data in healthcare: management, analysis and future prospects. *J. Big Data* **6**, 54 (2019).
29. Ziegler, E. et al. Open health imaging foundation viewer: an extensible open-source framework for building web-based imaging applications to support cancer research. *JCO Clin. Cancer Inform.* **4**, 336–345 (2020).
30. Spector-Bagdady, K. et al. Principles for health information collection, sharing, and use: a policy statement from the American Heart Association. *Circulation* **148**, 1061–1069 (2023).
31. Purmah, Y. et al. Right ventricular ejection fraction for the prediction of major adverse cardiovascular and heart failure-related events. *Circ. Cardiovasc. Imaging* **14**, e011337 (2021).
32. Mikami, Y. et al. Right ventricular insertion site fibrosis in a dilated cardiomyopathy referral population: phenotypic associations and value for the prediction of heart failure admission or death. *J. Cardiovasc. Magn. Reson.* **23**, 79 (2021).
33. Purmah, Y. et al. Mid-wall striae fibrosis predicts heart failure admission, composite heart failure events, and life-threatening arrhythmias in dilated cardiomyopathy. *Sci. Rep.* **12**, 1739 (2022).
34. Yakimenka, A. et al. Influence of sex-based differences in cardiac phenotype on atrial fibrillation recurrence in patients undergoing pulmonary vein isolation. *Front. Cardiovasc. Med.* **9**, 894592 (2022).
35. Alba, A. C. et al. Prognostic value of late gadolinium enhancement for the prediction of cardiovascular outcomes in dilated cardiomyopathy. *Circ. Cardiovasc. Imaging* **13**, e010105 (2020).
36. Kashyap, N. et al. Prognostic utility of cardiovascular magnetic resonance–based phenotyping in patients with muscular dystrophy. *J. Am. Heart Assoc.* **12**, e030229 (2023).
37. Labib, D. et al. Effect of active cancer on the cardiac phenotype: a cardiac magnetic resonance imaging-based study of myocardial tissue health and deformation in patients with chemotherapy-Naïve Cancer. *J. Am. Heart Assoc.* **10**, e019811 (2021).
38. Satriano, A. et al. 3-Dimensional regional and global strain abnormalities in hypertrophic cardiomyopathy. *Int. J. Cardiovasc. Imaging* **35**, 1913–1924 (2019).
39. Avitzur, N. et al. 3D myocardial deformation analysis from cine MRI as a marker of amyloid protein burden in cardiac amyloidosis: validation versus T1 mapping. *Int. J. Cardiovasc. Imaging* **34**, 1937–1946 (2018).
40. Cornhill, A. K. et al. Machine learning patient-specific prediction of heart failure hospitalization using cardiac MRI-Based phenotype and electronic health information. *Front. Cardiovasc. Med.* **9**, 890904 (2022).
41. Dykstra, S. et al. Machine learning prediction of atrial fibrillation in cardiovascular patients using cardiac magnetic resonance and

electronic health information. *Front. Cardiovasc. Med.* **9**, 998558 (2022).

42. Farrag, N. A., Lochbihler, A., White, J. A. & Ukwatta, E. Evaluation of fully automated myocardial segmentation techniques in native and contrast-enhanced T1-mapping cardiovascular magnetic resonance images using fully convolutional neural networks. *Med. Phys.* **48**, 215–226 (2021).

43. Satriano, A. et al. Neural-Network-Based Diagnosis Using 3-dimensional myocardial architecture and deformation: demonstration for the differentiation of hypertrophic cardiomyopathy. *Front. Cardiovasc. Med.* **7**, 584727 (2020).

44. Satriano, A. et al. Clinical feasibility and validation of 3D principal strain analysis from cine MRI: comparison to 2D strain by MRI and 3D speckle tracking echocardiography. *Int. J. Cardiovasc. Imaging* **33**, 1979–1992 (2017).

45. Satriano, A. et al. Three-dimensional thoracic aorta principal strain analysis from routine ECG-gated computerized tomography: feasibility in patients undergoing transcatheter aortic valve replacement. *BMC Cardiovasc. Disord.* **18**, 76 (2018).

46. Satriano, A. et al. Characterization of right ventricular deformation in pulmonary arterial hypertension using three-dimensional principal strain analysis. *J. Am. Soc. Echocardiogr.* **32**, 385–393 (2019).

47. Garcia, J. et al. Left atrial vortex size and velocity distributions by 4D flow MRI in patients with paroxysmal atrial fibrillation: Associations with age and CHA2 DS2 -VASc risk score. *J. Magn. Reson. Imaging JMRI* **51**, 871–884 (2020).

48. Hassanabad, A. F. et al. Pressure drop mapping using 4D flow MRI in patients with bicuspid aortic valve disease: a novel marker of valvular obstruction. *Magn. Reson. Imaging* **65**, 175–182 (2020).

49. Geeraert, P. et al. Bicuspid aortic valve disease is associated with abnormal wall shear stress, viscous energy loss, and pressure drop within the ascending thoracic aorta: A cross-sectional study. *Medicine* **100**, e26518 (2021).

50. Hudani, A. et al. 4D-Flow MRI characterization of pulmonary flow in repaired tetralogy of fallot. *Appl. Sci.* **13**, 2810 (2023).

51. Aliabadi, S. et al. Intra-cardiac pressure drop and flow distribution of bicuspid aortic valve disease in preserved ejection fraction. *Front. Cardiovasc. Med.* **9**, 903277 (2022).

Acknowledgements

Author contributions
S.D. performed data analysis, was involved in program design, infrastructure development, feasibility analysis, and manuscript authorship. J.A.W. was senior author, conceived and designed the infrastructure, and finalized manuscript contents. All other authors either assisted in program design (M.M., R.B., D.L., M.K., Y.F., J.F., J.B., B.L., S.D., M.G., and P.F.) or infrastructure development (M.M., R.B., D.L., and Y.F.). All authors reviewed and edited the final manuscript.

Competing interests
Dr. James White, Jackie Flewitt, and Steven Dykstra are shareholders in Cohesic Inc. (affiliations not related to the described work). Dr. James White has received research funding from Siemens Healthineers, Bristol Meyers Squibb, Pfizer Inc., and Circle Cardiovascular Inc. The remaining authors have nothing to disclose.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-025-01490-9.

Correspondence and requests for materials should be addressed to James A. White.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.