A Platform for Regulatory-Grade Evidence Generation from Routine Electronic Health Record Data Capture

Claire A. Margolis, MS[1], Chhaya Shadra, MS[1], Alex Koshta, PhD[1], Cuau Leyva[1], Neha Seth, MS, BDS[1], Solmaz Bagherpour, PhD[1], Rakesh Puthalath[1], Helene Fevrier, MPH[1], Amanda Lien, ND[1], Rishabh Jain, PhD[1], Hariesh Rajasekar, MS[1], Theodore Leng, MD[1], Matthew Roe, MD, MHS[1]

[1]Verana Health, San Francisco, California, US

Background & Objectives

The role of real-world evidence (RWE) generation to inform clinical research, care delivery, and regulatory decision-making is rapidly evolving. Despite expanding uses of RWE, robust platforms and frameworks enabling and enforcing the cornerstones of regulatory-grade real-world data (RWD) — completeness, accuracy, traceability, generalizability, timeliness, scalability, and security — are yet to exist in the industry at scale.[1,2] Current RWD capture and storage systems were built for observational research and do not inherently support data of the quality required for regulatory use cases, raising the necessity for adapting solutions to fit future evidence generation needs.

The need for these platforms extends to the real-world data found in electronic health records (EHRs), which offer a uniquely rich clinical perspective on patient health yet require meaningful governance at scale due to challenges that arise from custom implementations and the utilization of free text notes. Verana Health partners with leading medical associations to transform electronic health record data found in clinical data registries into real-world evidence. This includes the American Academy of Ophthalmology IRIS® Registry (Intelligent Research in Sight), the nation's first comprehensive eye disease clinical database. As of September 2020, 349 million patient visits from 60 million unique patients across 60 EHRs exist in the database.[3]

Given this, our objective was to develop a unified technology platform for RWE generation based on routine EHR data capture with the goal of ensuring the fundamental qualities of robust, regulatory-grade data.

Information Model Design

We designed a fit-for-purpose information model to support prospective and observational research using patient EHR data that also accommodates linked data sources such as claims and images. To enable interoperability, the model incorporates elements of and builds upon existing industry standards, including the PCORnet Common Data Model and the OMOP Common Data Model, which are designed specifically to facilitate observational research.[4,5] Our information model extends these models to emphasize clinical completeness; it was built with direct specialty-specific clinical input and validated against HL7 FHIR resources.[6] For example, while both the PCORnet and OMOP models include observation objects, our model contains an expanded observation object including fields specifically designed to accommodate the types of observations collected during the clinical workflow in a specialty setting, as well as a child table to store modifiers and metadata about each observation and accommodate a one-to-many relationship. Emphasis is placed on ensuring traceability to the source EHR, one of the core requirements of regulatory uses. Our schemas are certified de-identified by expert determination, ensuring that there is no risk of re-identification when analyzing patient-level data.

Content Management & Record Harmonization

We developed a content management infrastructure to enable data quality improvement, organizational efficiency, and industry interoperability. This includes a stable terminology framework and unified approach to record harmonization and mapping. Records storing coded clinical data are algorithmically normalized to controlled terminologies to facilitate consistent downstream analysis as well as ensure compliance of string fields with de-identification specifications. The set of supported terminologies per domain is as follows:

Domain	Terminology or Ontology
Conditions	ICD-9-CM or ICD-10-CM or SNOMED-CT
Medications	RxNorm
Procedures	CPT or HCPCS or ICD-9-PCS or ICD-10-PCS
Observations	LOINC
Gender	HL7 Administrative Gender
Race	CDC Race Category
Ethnicity	HL7 Ethnicity

Table 1. Terminologies supported per domain by the Verana harmonization infrastructure

Across all eligible records within the American Academy of Ophthalmology's IRIS Registry, our harmonization platform was able to algorithmically harmonize 98% of patient records within the diagnosis domain, 86% of procedure records, and 94% of medication records, meaning that a majority of raw records within the data set were able to be assigned a code and description from a supported controlled terminology.

References

1. Miksad, RA and Abernethy, AP. Harnessing the Power of Real-World Evidence (RWE): A Checklist to Ensure Regulatory-Grade Data Quality. Clin Pharm & Ther 2017;103:202-205.
2. U.S. Department of Health and Human Services. Food and Drug Administration. Use of Electronic Health Record Data in Clinical Investigations, 2018. Retrieved from https://www.fda.gov/media/97567/download.
3. IRIS Registry. American Academy of Ophthalmology, 2020, www.aao.org/iris-registry
4. Observational Health Data Sciences and Informatics. OHDSI/CommonDataModel. GitHub, 2018, github.com/OHDSI/CommonDataModel/wiki.
5. PCORnet Data-Driven Common Model. The National Patient-Centered Clinical Research Network, 2019, pcornet.org/data-driven-common-model/.
6. "Welcome to FHIR®." Index - FHIR v4.0.1, hl7.org/fhir/.
7. "International Classification of Diseases (ICD) Information Sheet." *World Health Organization*, World Health Organization, 11 Oct. 2019, www.who.int/classifications/icd/factsheet/en/.

Concept Ontology & Application

To move beyond record- and code-level analyses and enable facile cohort creation with clinically-intuitive concepts, we developed a custom concept ontology for disease areas relevant for clinical and research use cases within our specialty areas. The ontology was created based on the way in which clinicians think about and categorize disease, in contrast to existing ontologies like the International Classification of Disease chapters, which were built for administrative and billing uses.[7] The ontology is structured graphically, with clinical concepts as nodes and relationships between concepts as edges. The current implementation is tree-based, with single root nodes designated for each disease area and children, or leaf nodes, having exactly one parent node. The design and infrastructure are extensible, such that future expansions may adopt different graphical structures. Each concept is defined in clinical terms as well as with logic wherein it may be mapped onto the IRIS Registry EHR data as well as any additional linked data sources. Concept definitions are applied at scale within our platform to discover all occurrences of each concept across all patients. This comprehensive set of information is used to efficiently build patient cohorts of interest and enforce definitional consistency across all use cases.

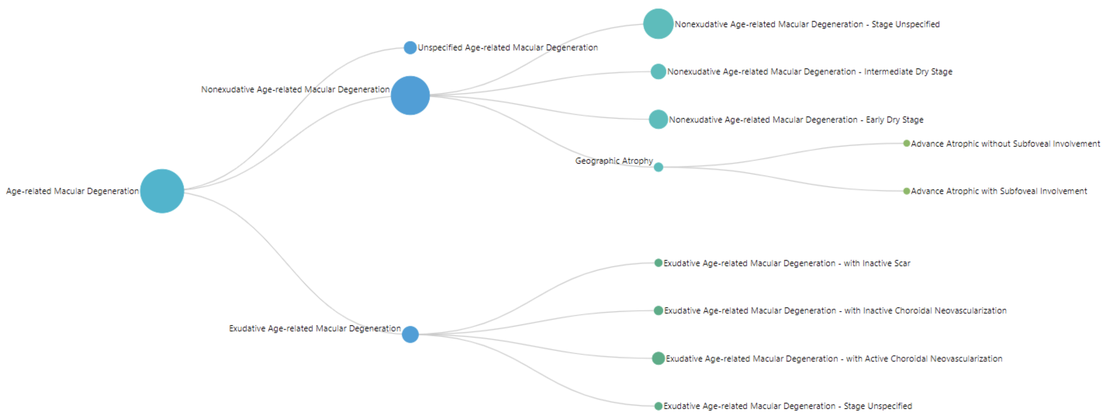

Fig. 1. Verana disease ontology for Age-Related Macular Degeneration (AMD)

Validation

A data governance framework is applied to the platform to ensure data integrity and continuously monitor platform function and validity. This framework includes master data management with source data provenance and metadata storage, technical and legal data security and access controls, robust data de-identification procedures, content management, and an end-to-end data quality framework shown below implemented on a scalable infrastructure across the entire platform, enabling automated, ongoing, actionable, quantified assessment of data quality for each dimension of regulatory-grade data. In this way, we are able to conduct ongoing validation of both the platform and embedded algorithm functionality as well as the datasets being curated by the platform themselves.

Cornerstones of Data Quality	Does the Data...	Technical	Clinical	Scientific
Completeness	*encompass the entire clinical picture?*	Field completeness is assessed among fields where data is expected. (e.g. each diagnosis must have a documented date)	Data completeness in clinical context exists (e.g., intraocular pressure is expected to be documented for patients with a diagnosis of Glaucoma)	Factors (e.g., confounders) have been considered in study design & analyses
Accuracy	*accurately reflect patient chart/reality?*	Data conforms to expected data types & constraints	EHR effectively captures patient journey & provider patterns	Results are within range of scientific acceptability
Traceability	*contain provenance back to source?*	Data elements & transformations are clear and auditable during ingestion and curation	Study specifies a clear, auditable patient cohort	Study design, methods, and analysis are clear and transparent
Consistency	*maintain integrity across structures, time, releases?*	Data are represented in a consistent data model, under congruent architecture & format	Cohort-specific trends & rates are tracked across time	Data is validated against published studies & external sources
Generalizability	*represent a minimally-biased sample?*	Data elements are harmonized to industry standards	Biases have been assessed & accounted for in clinical interpretation	External comparisons are used to identify and adjust for biases
Timeliness	*reflect recent practice patterns?*	Data is refreshed at appropriate frequency	Current practice patterns, treatments are incorporated	Data timeframe is relevant to current study

Table 2. Foundational concepts of Verana's data quality framework

Conclusions

We have created a unified technology platform for RWE generation based on routine EHR data capture with the goal of ensuring the fundamental qualities of robust, regulatory-grade data. This platform enables RWE generation with the potential to inform clinical research and future regulatory decision-making for use cases such as post-marketing surveillance studies, adverse event detection, and label expansion. Additionally, our platform design and approach proves that a single, extensible platform infused with clinical context can be used to deeply curate multiple specialty areas and can serve as an example to guide development of systems to support the use of data from clinical registries in disease areas beyond ophthalmology.