

PyxGen Data Specification (v0.1)

1. Overview

PyxGen accepts structured CSV input files as the primary method for ingesting external healthcare data.

This specification defines the standard structure, formatting rules, and validation requirements for those files.

The objective is to ensure **consistent, predictable ingestion** of data originating from:

- Electronic Health Record (EHR) exports
- SQL database extracts
- Flat-file integrations or manual submissions

All input data must comply with the format and conventions described herein.

2. General CSV Rules

Rule	Description
File Encoding	UTF-8
Delimiter	Comma (,)
Header Row	Mandatory; first row in file
Date Format	ISO 8601 (YYYY-MM-DD)
Missing Values	Leave blank (no “NULL”)
Boolean Values	True/False or Y/N
File Naming Convention	<EntityName>.csv (e.g., Patient.csv)
File Size Limit (initial)	≤ 100MB per upload (configurable)

Each file represents one entity type (Patient, Encounter, Immunization, etc.). All files are related via key fields such as `patient_id`.

3. Core Entity Definitions

3.1 Patient.csv

Field	Required	Type	Example	Description
patient_id	✓	string	P12345	Unique patient identifier
first_name	✓	string	John	Given name
last_name	✓	string	Doe	Family name
dob	✓	date	1984-07-12	Date of birth
gender	✓	string	M	M/F/O/U
address		string	123 Main St	Street address
city		string	Seattle	
state		string	WA	
postal_code		string	98101	
phone		string	555-123-4567	
email		string	john.doe@email.com	

Primary Key: patient_id

FHIR Mapping: Patient resource

3.2 Encounter.csv

Field	Required	Type	Example	Description
encounter_id	✓	string	E98765	Unique encounter identifier
patient_id	✓	string	P12345	References Patient.patient_id
encounter_date	✓	date	2025-10-12	Encounter date/time

encounter_type	✓	string	outpatient	e.g., inpatient/outpatient/emergency
provider_id		string	DR001	
location		string	Mercy Hospital	

Primary Key: `encounter_id`

Foreign Key: `patient_id` → Patient.csv

FHIR Mapping: Encounter resource

3.3 Immunization.csv

Field	Required	Type	Example	Description
immunization_id	✓	string	IM001	Unique record ID
patient_id	✓	string	P12345	References Patient.patient_id
immunization_code	✓	string	207	CVX code or equivalent
immunization_name	✓	string	COVID-19 mRNA vaccine	
immunization_date	✓	date	2021-03-12	
lot_number		string	12345ABC	
manufacturer		string	Pfizer	
administering_provider		string	DR001	

Primary Key: `immunization_id`

Foreign Key: `patient_id` → Patient.csv

FHIR Mapping: Immunization resource

3.4 Observation.csv

Field	Required	Type	Example	Description
observation_id	✓	string	OBS001	Unique ID
patient_id	✓	string	P12345	References Patient.patient_id
observation_code	✓	string	8480-6	LOINC code
observation_name	✓	string	Systolic Blood Pressure	
observation_value	✓	string	120	Numeric or text result
observation_unit		string	mmHg	
observation_date	✓	date	2025-10-15	

Primary Key: observation_id

Foreign Key: patient_id → Patient.csv

FHIR Mapping: Observation resource

3.5 Provider.csv (Optional)

Field	Required	Type	Example	Description
provider_id	✓	string	DR001	Unique provider ID
provider_name	✓	string	Dr. Alice Smith	
npi		string	1234567890	National Provider Identifier
specialty		string	Family Medicine	

Primary Key: provider_id

FHIR Mapping: Practitioner resource

4. Data Validation Rules

When files are uploaded into PyxGen, the following validations occur automatically:

Validation	Description
Header Check	All required headers must be present.
Data Type Check	Each field must conform to its declared type (date, string, etc.).
Referential Integrity	Foreign key references must exist (e.g., Encounter.patient_id must exist in Patient.csv).
Unique ID Check	Primary keys must be unique per file.
Null Check	Required fields cannot be blank.
Date Validation	Dates must be valid and in ISO format.
Encoding Validation	File must be UTF-8 encoded.

If any rule fails, PyxGen will return a **Validation Report** in JSON format containing:

- Filename
 - Row number
 - Field name
 - Error type and description
-

5. Versioning

PyxGen maintains backward compatibility via versioned specifications:

- Current Version: **v0.1**
 - Next Version: **v0.2 (planned)** → Add support for Procedures, Medications, and Conditions
-

6. Future Extensions

Planned roadmap for later schema expansion:

- **Procedures.csv** → align with FHIR Procedure
- **MedicationStatement.csv** → track administered meds

- **Condition.csv** → diagnoses/problem list
 - **FHIR Export Mapping** → automated JSON Bundle generation
-

7. Example File Layout

/input_data/

```
|— Patient.csv  
|— Encounter.csv  
|— Immunization.csv  
|— Observation.csv  
|— Provider.csv
```

8. Notes for Developers

- All ingestion modules should reference this specification.
- Schema changes must update both the spec version and the validation engine.
- Always prefer **explicit over inferred** data typing.
- When possible, align coded fields (e.g., LOINC, CVX) to their FHIR equivalents.