Lecture 7:

Clinical Data and Electronic Health Records

- Digital health transformation
- EHR adoption rates
- Data-driven medicine

Introduction to Biomedical Datascience

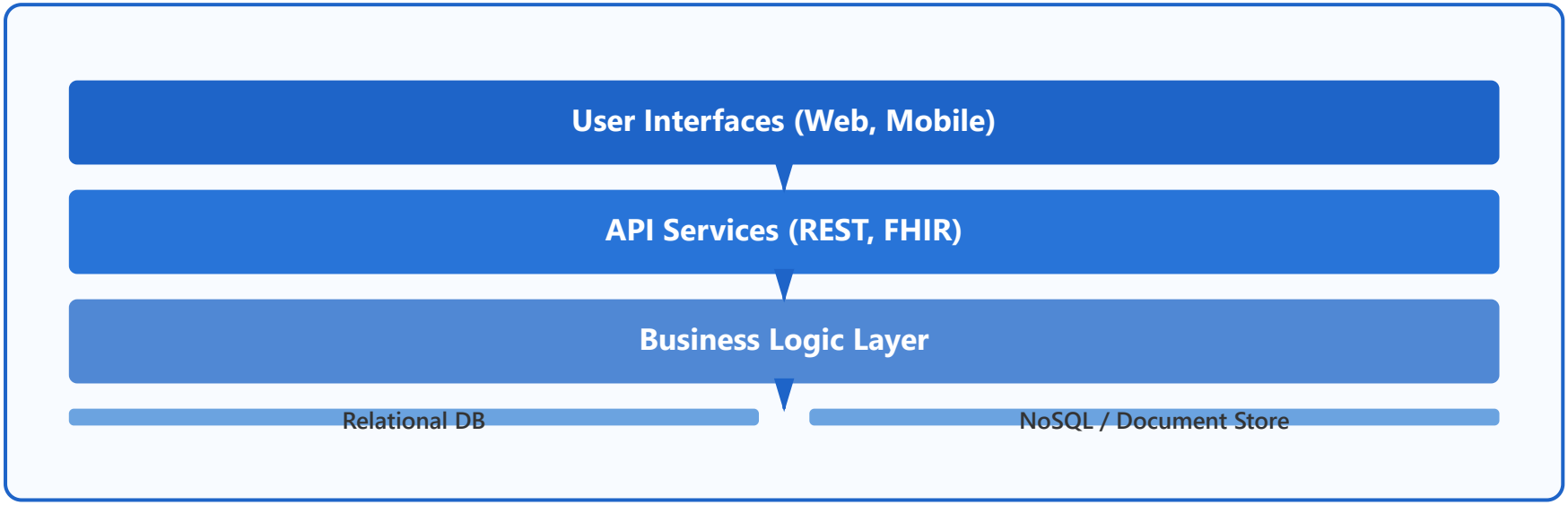
Lecture Contents

Part 1/3:

EHR Systems

- System components
- Data models
- Interoperability
- Security requirements

EHR Architecture



Database Design

- Relational databases (PostgreSQL, MySQL)
- NoSQL for unstructured data
- Data normalization strategies
- Indexing for performance

Application Layers

- Presentation layer (UI/UX)
- Business logic layer
- Data access layer
- Microservices architecture

💻 User Interfaces

- Web-based portals
- Mobile applications
- Clinical workflow integration
- Responsive design patterns

🔌 API Services

- RESTful APIs
- FHIR endpoints
- Authentication & authorization
- Rate limiting & monitoring

☁ Cloud Deployment

Modern EHRs leverage cloud infrastructure (AWS, Azure, GCP) for scalability, disaster recovery, and compliance with healthcare regulations (HIPAA, GDPR)

Data Types in EHR

Demographics

- Patient name, DOB, gender
- Address, contact information
- Insurance details
- Emergency contacts

Diagnoses/Procedures

- ICD-10 coded diagnoses
- CPT procedure codes
- Problem lists
- Surgical history

Medications

- Current medications
- Prescription history
- Allergies & adverse reactions
- Dosage and frequency

Laboratory Results

- Blood tests, imaging
- Pathology reports
- Vital signs
- LOINC coded values

Clinical Notes

- Progress notes
- Consultation reports

- Discharge summaries
- Nursing documentation

Structured vs Unstructured Data

Structured Data

ID	Diagnosis	Value
001	E11.9	140
002	I10	145/90
003	J45.909	Normal

DB

Unstructured Data

Patient presents with chest pain...

History of hypertension and diabetes

Physical exam shows...

NLP

Structured Data

- Predefined fields & formats
- Easily queryable
- Standardized codes (ICD, LOINC)
- Direct database storage
- Machine-readable

Unstructured Data

- Free text clinical notes
- Medical images (X-ray, MRI)
- Scanned documents
- Voice recordings
- Requires NLP for extraction

🔀 Hybrid Documents

Many clinical documents combine structured fields (dates, vital signs) with unstructured narratives (clinical impressions)

HL7 and FHIR Standards

HL7 v2 Messages

- Pipe-delimited format
- ADT, ORM, ORU message types
- Widely adopted legacy standard
- Complex parsing required

FHIR Resources

- JSON/XML formats
- Patient, Observation, Medication
- Modern web-based standard
- Easy to implement

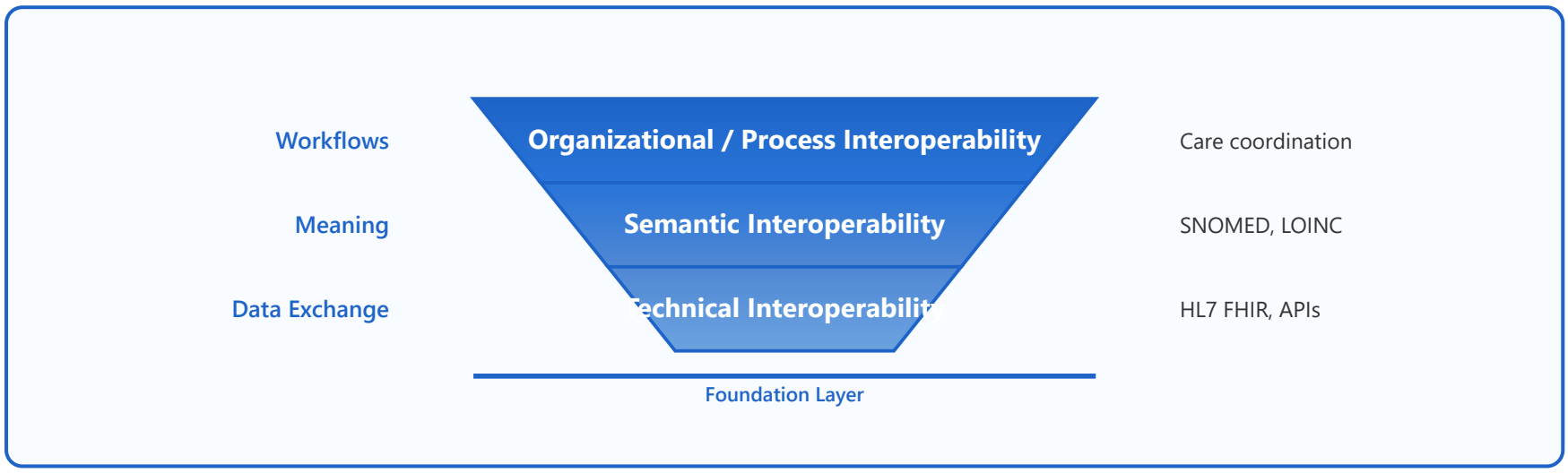
RESTful APIs

- HTTP GET, POST, PUT, DELETE
- Resource-based URLs
- OAuth 2.0 authentication
- SMART on FHIR apps

Implementation Guides

- US Core profiles
- Argonaut specifications
- Country-specific extensions
- Validation tools

Interoperability



Technical Standards

- HL7 FHIR
- Direct messaging
- APIs and web services
- Transport protocols (HTTPS, SFTP)

Semantic Standards

- Common terminologies (SNOMED, LOINC)
- Value set harmonization
- Concept mapping
- Unified Code Management

Process Interoperability

- Clinical workflows

Health Information Exchange (HIE)

- Regional/national networks

- Care coordination protocols
- Consent management
- Data governance policies

- Query-based vs push
- Patient matching algorithms
- Blockchain potential for trust

Data Warehousing for EHR

ETL Processes

- Extract from operational systems
- Transform & clean data
- Load into warehouse
- Incremental updates

Data Marts

- Disease-specific repositories
- Quality improvement data
- Research cohorts
- Departmental analytics

Star Schema

- Fact tables (encounters, labs)

Real-time vs Batch

- Batch: overnight processing

- Dimension tables (patient, time)
- Optimized for queries
- Aggregate calculations

- Real-time: streaming analytics
- Near real-time: micro-batching
- Trade-offs in complexity

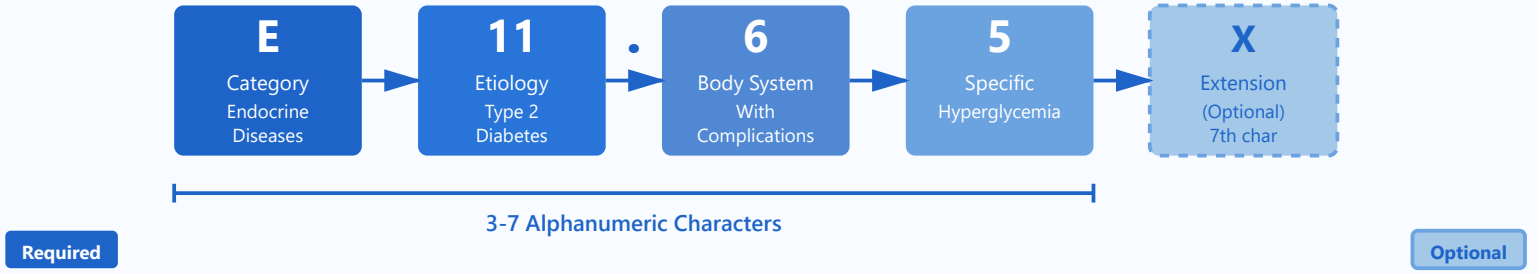
Part 2/3: Clinical Coding

Clinical Coding

- Terminology systems
- Ontology relationships
- Mapping challenges
- Use cases

ICD-10 Coding

ICD-10 Code Structure Example: E11.65



Common Categories

- A00-B99: Infectious diseases
- C00-D49: Neoplasms
- E00-E89: Endocrine, metabolic
- I00-I99: Circulatory system

Example Codes

- E11.9 - Type 2 diabetes
- I10 - Essential hypertension
- J45.909 - Asthma, unspecified
- M79.3 - Myalgia

ICD-10-PCS

- 7-character procedure codes

Coding Guidelines

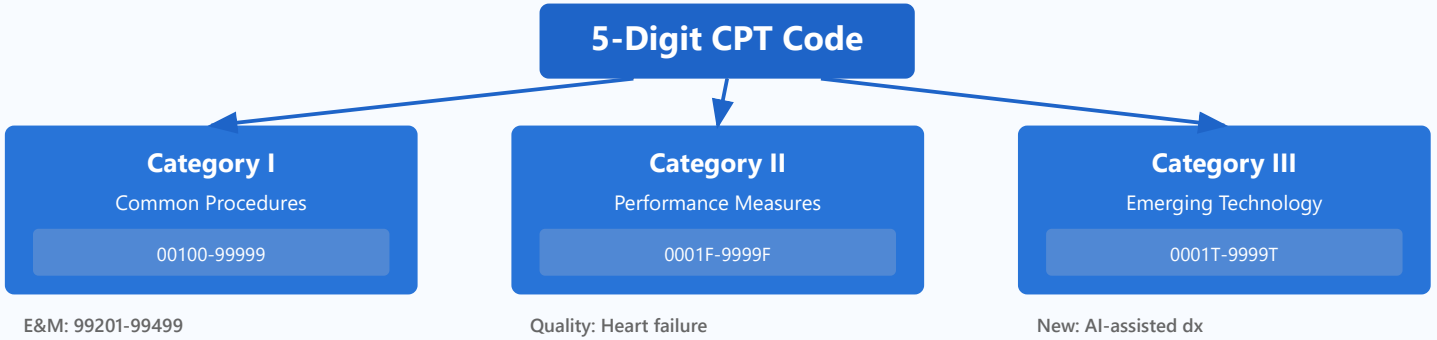
- Code to highest specificity

- Inpatient procedures only
- Section-Body-Root-Approach
- More specific than CPT

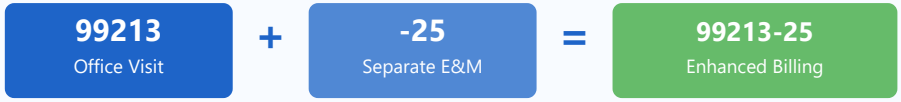
- Principal vs secondary diagnoses
- Excludes1 vs Excludes2 notes
- Use additional code notes

CPT Codes

CPT Code Categories & Structure



CPT Code with Modifiers Example



📊 Code Sections

- 00100-01999: Anesthesia
- 10004-69990: Surgery
- 70010-79999: Radiology

🏥 E&M Codes

- 99201-99215: Office visits
- 99217-99226: Hospital observation
- 99241-99255: Consultations

- 80047-89398: Laboratory
- 90281-99607: Medicine

- Level based on complexity

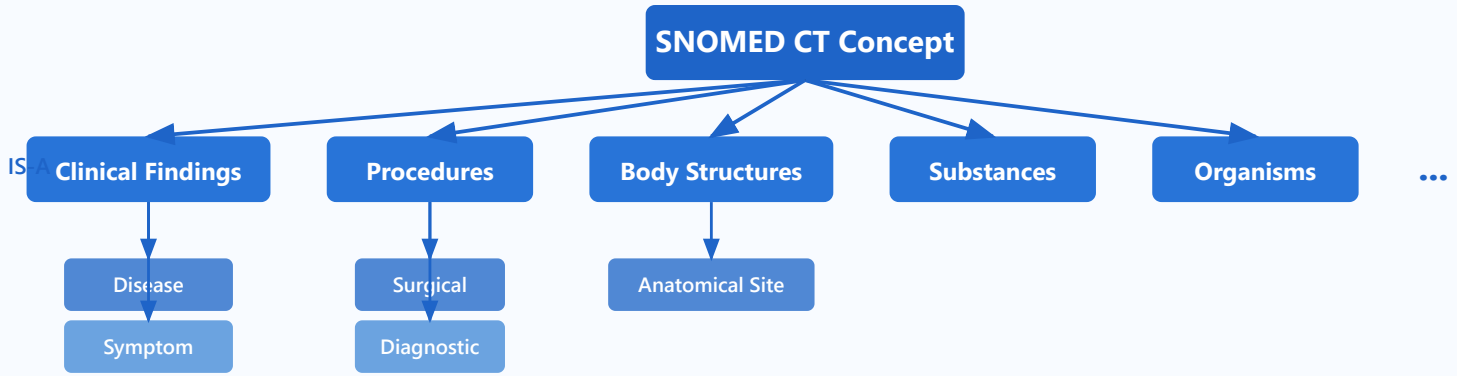
Common Modifiers

- -25: Significant separate E&M
- -59: Distinct procedural service
- -76: Repeat procedure
- -50: Bilateral procedure

RVU Values

- Work RVU: Physician effort
- Practice expense RVU
- Malpractice RVU
- Total RVU × Conversion = Payment

SNOMED CT



Concept Model

- Concepts, descriptions, relationships
- Unique concept IDs (SCTID)
- Fully specified names
- Synonyms and translations

Hierarchies

- Clinical findings
- Procedures
- Body structures
- Substances
- IS-A relationships

Relationships

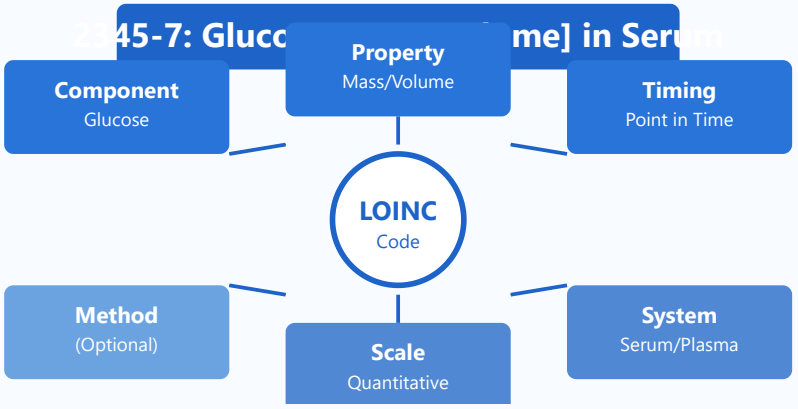
Post-coordination

- Finding site
- Associated morphology
- Causative agent
- Procedure site
- Compositional grammar

- Combine multiple concepts
- Express complex clinical meanings
- Example: 'Fracture of left femur'
- International adoption by 40+ countries

LOINC for Lab Tests

LOINC Six-Part Structure



Common LOINC Examples

2160-0 Creatinine	718-7 Hemoglobin	2951-2 Sodium	4544-3 Hematocrit

🧪 Test Categories

- Chemistry tests
- Hematology & coagulation
- Microbiology cultures

📊 Panel Organization

- Basic Metabolic Panel (BMP)
- Complete Blood Count (CBC)
- Comprehensive Metabolic Panel

- Serology & immunology
- Molecular pathology

- Lipid panel
- Liver function tests

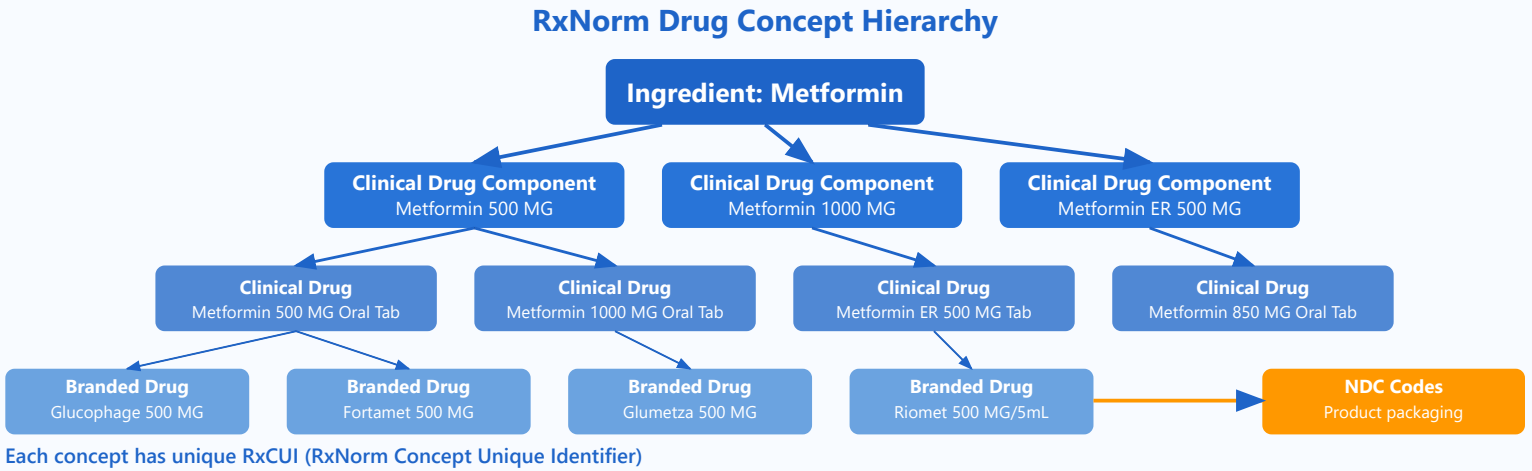
⚡ LOINC Properties

- MCnc: Mass concentration
- NCnc: Number concentration
- Prid: Presence/Identity
- Titr: Titer (dilution)
- Arb: Arbitrary units

🔬 UCUM Units

- mg/dL, mmol/L (chemistry)
- 10*3/uL (cell counts)
- IU/L (enzymes)
- Standardized unit conversion
- Reference range mapping

RxNorm for Medications



💊 Drug Concepts

- Normalized drug names
- Ingredient + strength + dose form
- Unique RxNorm CUI
- Links to other vocabularies

📊 Hierarchy Levels

- Ingredient (base substance)
- Precise ingredient (salt form)
- Clinical drug (generic)
- Branded drug (trade name)
- Drug pack (multi-ingredient)

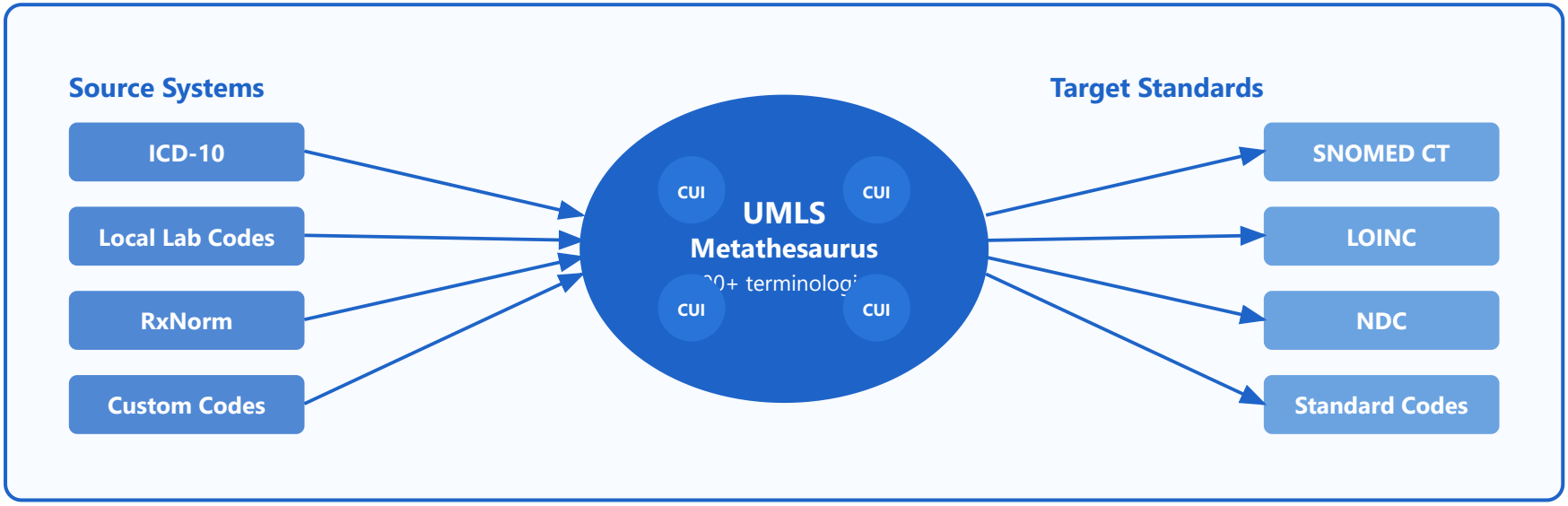
Dose Forms

- Oral tablet, capsule
- Injectable solution
- Topical cream, ointment
- Inhalation powder
- Transdermal patch

Integration

- NDC code mapping
- Drug-drug interactions
- Generic substitution
- Allergy checking
- Formulary management

Ontology Mapping



Crosswalk Creation

- ICD-10 to SNOMED CT
- LOINC to local lab codes
- RxNorm to NDC
- Manual and automated approaches

Automated Mapping

- String similarity algorithms
- Lexical matching
- Machine learning classifiers
- Natural language processing

✓ Validation Methods

UMLS Metathesaurus

- Expert review
- Dual coding
- Inter-rater reliability
- Continuous quality improvement

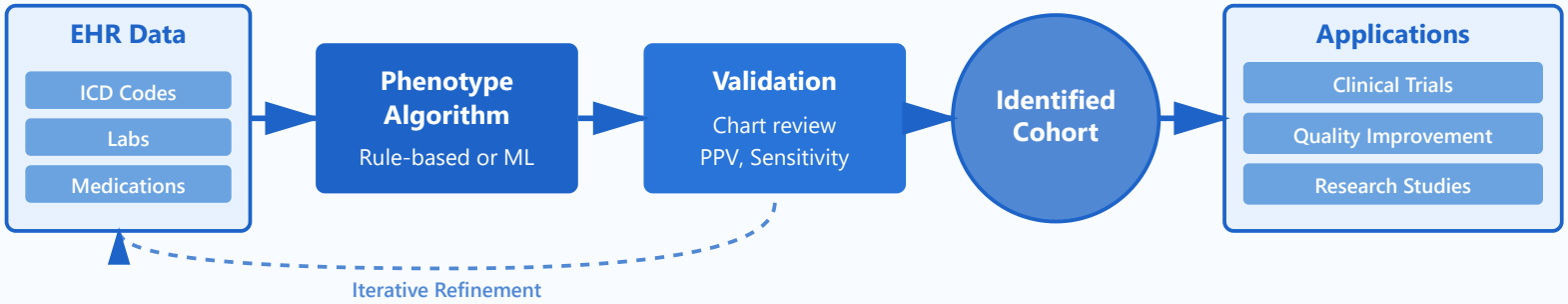
- Unified Medical Language System
- Integrates 200+ terminologies
- Concept Unique Identifiers (CUI)
- Relationship mappings across systems

Part 3/3: Data Analytics

Clinical Coding

- Phenotyping algorithms
- Predictive models
- Quality measures
- Population health

Clinical Phenotyping



📋 Computable Phenotypes

- Standardized disease definitions
- ICD codes + labs + meds
- Temporal logic criteria
- Inclusion/exclusion rules

📐 Rule-Based Methods

- Boolean logic (AND, OR, NOT)
- Diagnosis code combinations
- Lab value thresholds
- Medication orders

Machine Learning Approaches

- Supervised classification

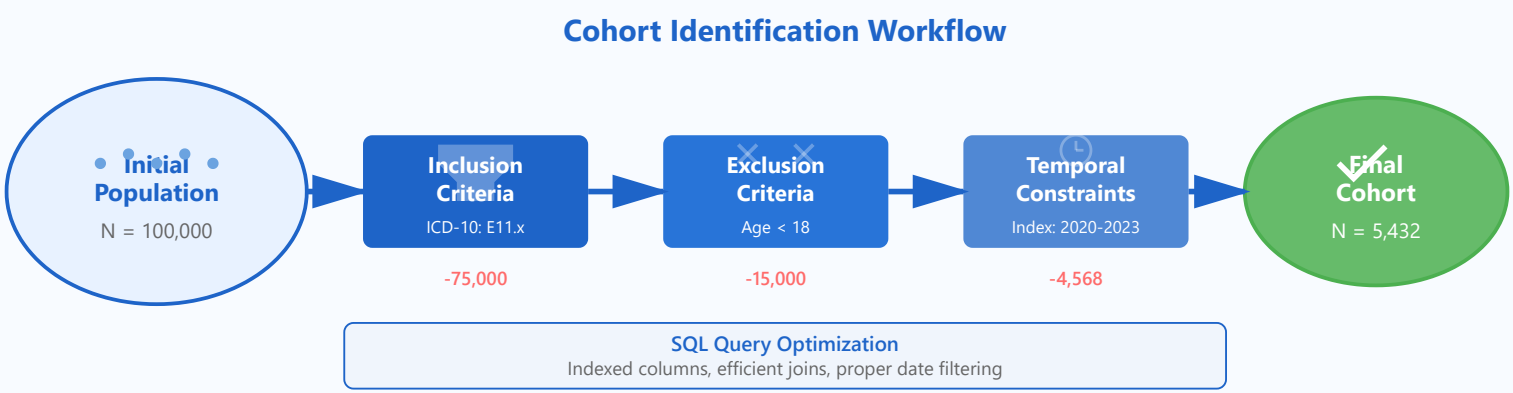
✓ Validation Strategies

- Chart review (gold standard)

- Feature engineering from EHR
- Random forests, deep learning
- Semi-supervised learning

- PPV, NPV, sensitivity, specificity
- Cross-institutional validation
- Phenotype libraries (PheKB, eMERGE)

Cohort Identification



✅ Inclusion Criteria

- Age range (18-65 years)
- Primary diagnosis codes
- Minimum encounter count
- Medication exposures
- Lab value thresholds

❌ Exclusion Criteria

- Competing diagnoses
- Prior treatments
- Missing key data
- Insufficient follow-up
- Pregnancy or nursing

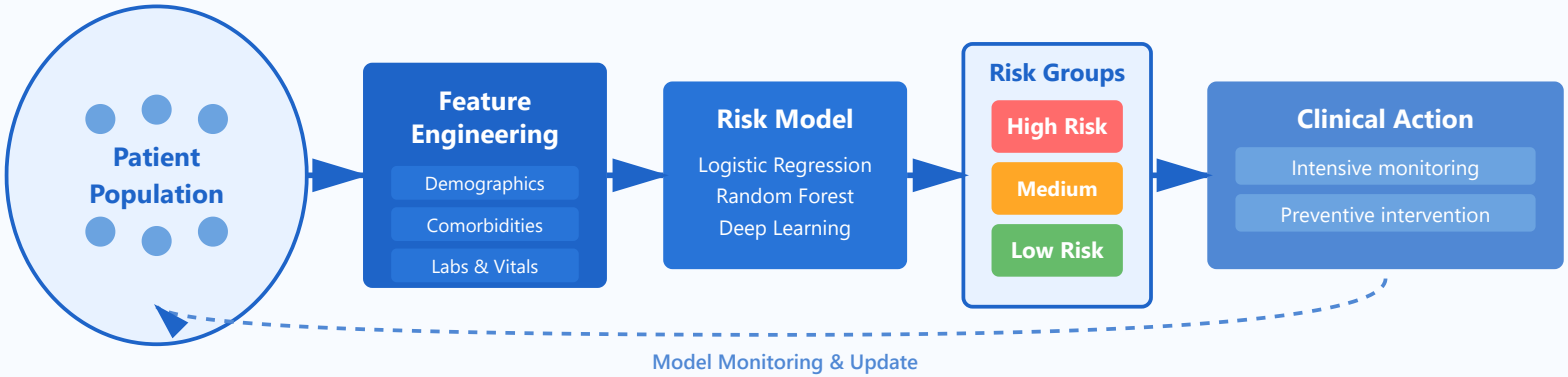
⏰ Temporal Logic

🔧 Implementation Tools

- Index date definition
- Washout periods (180 days)
- Follow-up windows
- Event sequence ordering
- Censoring rules

- OHDSI ATLAS interface
- SQL query builders
- Cohort validation metrics
- Attrition diagrams
- Sample size calculations

Risk Stratification



Clinical Risk Scores

- CHADS-VASc (stroke risk)
- MELD (liver disease)
- GRACE (cardiac events)
- Point-based scoring systems

Model Development

- Logistic regression
- Cox proportional hazards
- Gradient boosting machines
- Neural networks

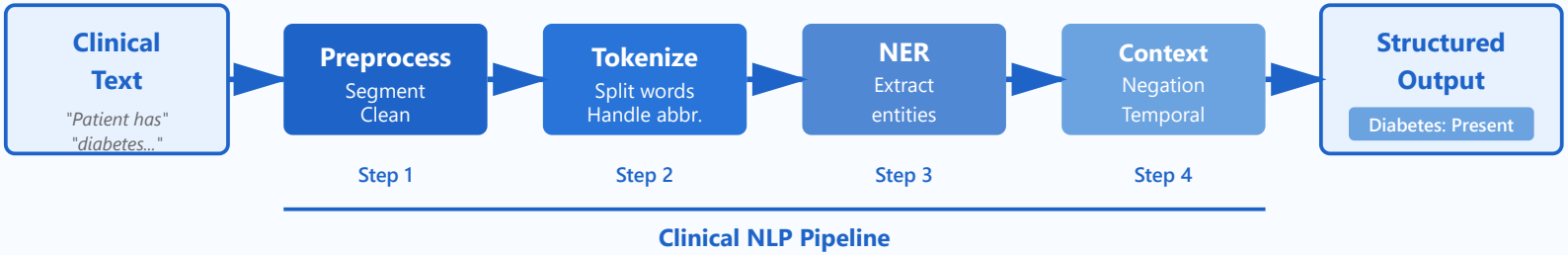
Feature Engineering

Calibration & Implementation

- Aggregating encounter data
- Temporal patterns
- Medication burden scores
- Comorbidity indices (Charlson, Elixhauser)

- Calibration plots
- Decision curve analysis
- Integration into EHR alerts
- Continuous model monitoring

Clinical NLP Basics



Text Preprocessing

- Sentence segmentation
- Lowercasing, punctuation removal
- Handling abbreviations
- PHI removal

Tokenization

- Word-level tokens
- Subword tokenization (BPE)
- Clinical-specific tokenizers
- Handling medical jargon

Named Entity Recognition

- Diseases, symptoms

Negation & Section Detection

- NegEx, ConText algorithms

- Medications, dosages
- Anatomical sites
- Procedures

- Identifying negated findings
- Section headers (HPI, ROS, A&P)
- Temporal expressions

Named Entity Recognition (NER)

Medical Entities

- Diseases & conditions
- Drugs & treatments
- Signs & symptoms
- Lab tests & values
- Anatomical structures

Rule-Based Systems

- Dictionary lookups (UMLS)
- Regular expressions
- Pattern matching
- Fast but limited coverage

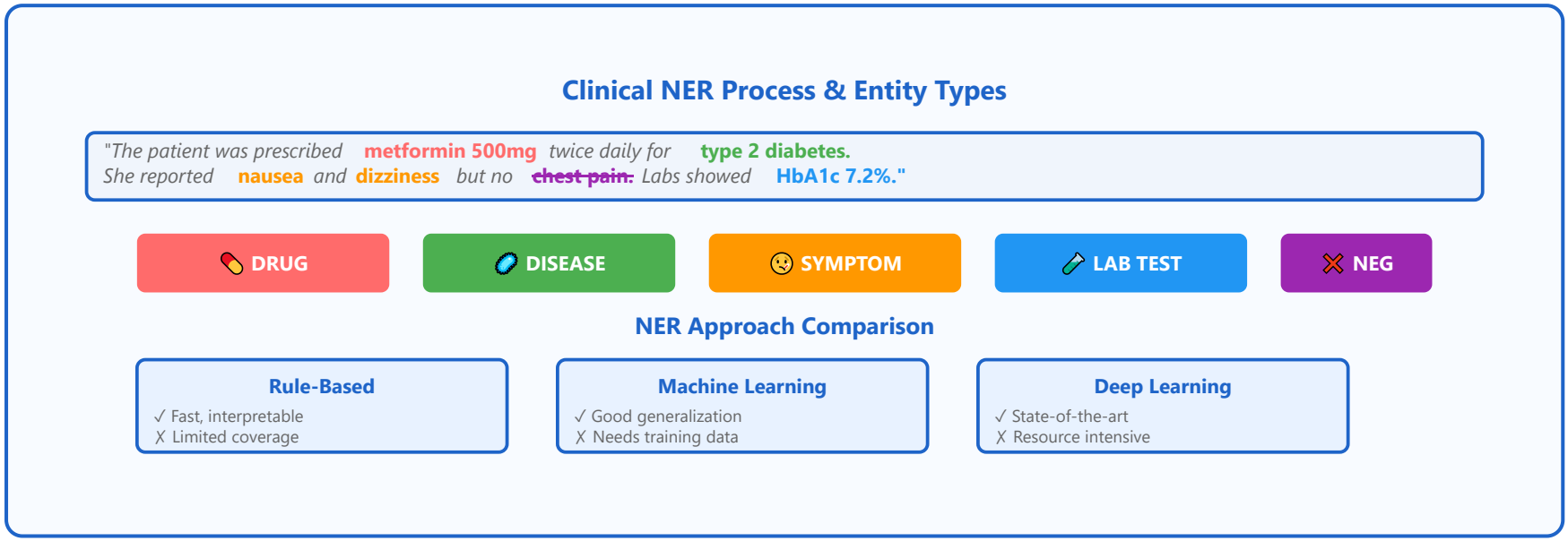
Machine Learning Models

- CRF, SVM models
- LSTM, BiLSTM
- Contextual embeddings
- Better generalization

Deep Learning & Hybrid

- BioBERT, ClinicalBERT
- Transfer learning
- Combining rules + ML
- State-of-the-art performance

Named Entity Recognition (NER)



Medical Entities

- Diseases & conditions
- Medications & dosages
- Signs & symptoms
- Procedures & surgeries
- Anatomical structures
- Lab tests & values

Rule-Based NER

- Dictionary lookup (UMLS)
- Regular expressions
- Pattern matching rules
- Fast execution
- Limited to known terms

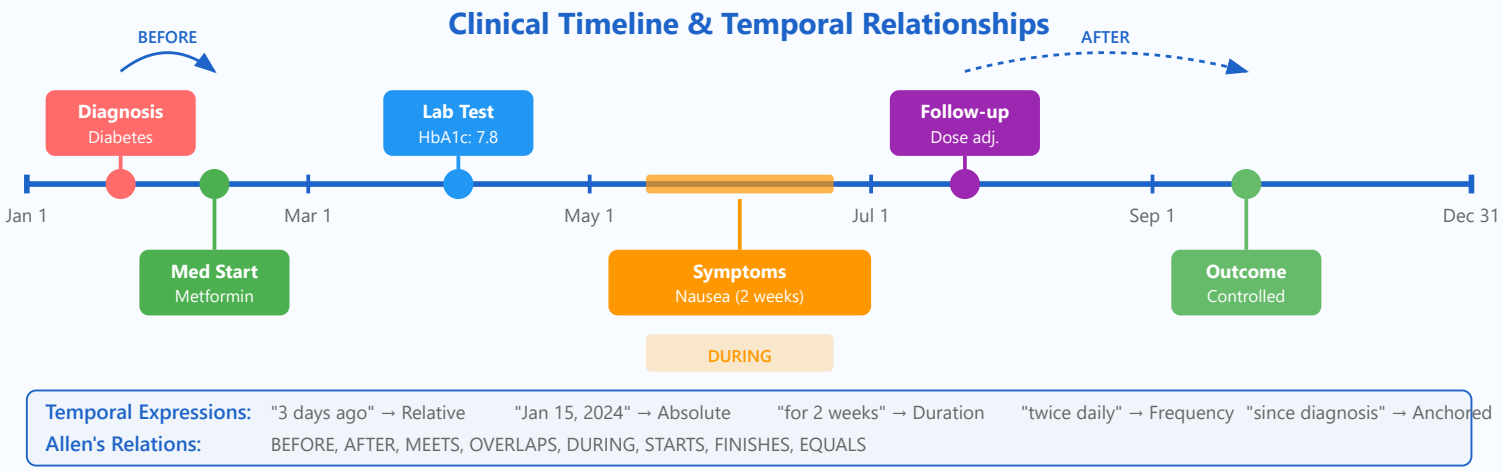
ML/DL Models

- CRF (Conditional Random Fields)
- BiLSTM-CRF networks
- BERT-based models
- BioBERT, ClinicalBERT
- Transfer learning approach

Performance Metrics

- Precision: correct predictions
- Recall: coverage of entities
- F1 score: harmonic mean
- Entity-level vs token-level
- Cross-validation testing

Temporal Reasoning



Time Expressions

- Absolute dates (Jan 1, 2024)
- Relative dates (3 days ago)
- Durations (for 2 weeks)
- Frequencies (twice daily)
- Anchored times (since surgery)

Event Ordering

- BEFORE / AFTER relations
- OVERLAPS / DURING
- STARTS / FINISHES
- Medication timelines
- Symptom progression

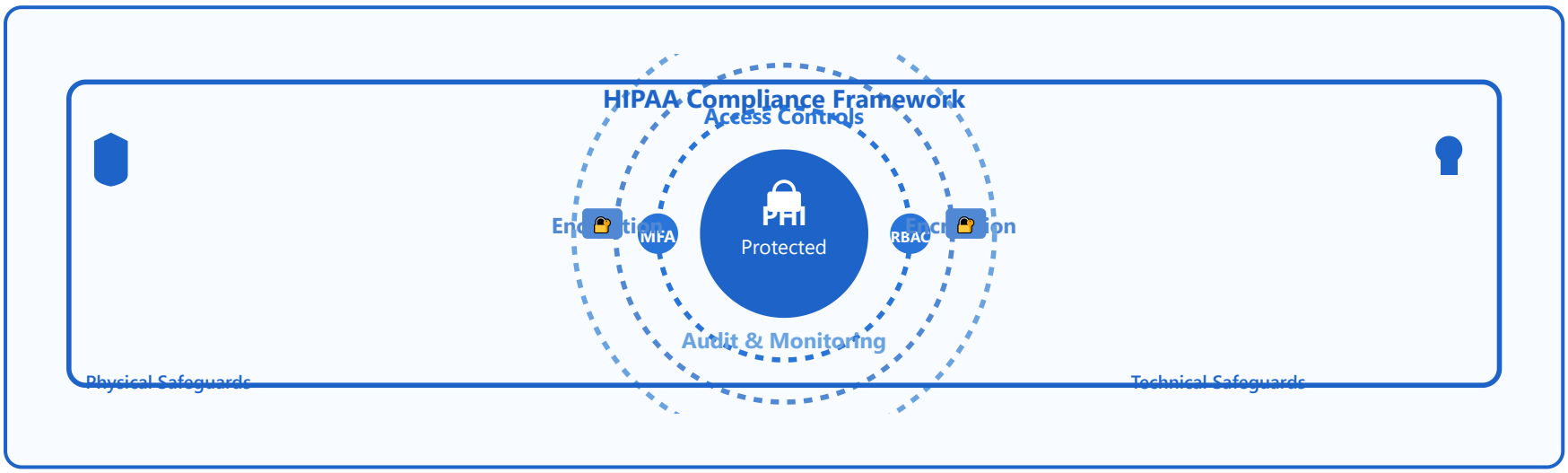
📊 Timeline Construction

- Patient journey visualization
- Multi-source data fusion
- Conflict resolution rules
- Uncertainty handling
- Missing data imputation

🎯 Clinical Applications

- Disease progression tracking
- Treatment response timing
- Adverse event detection
- Readmission prediction
- Longitudinal outcomes

Privacy and HIPAA



PHI Definition

- 18 identifiers under HIPAA
- Names, addresses, dates
- Medical record numbers
- Biometric identifiers

Minimum Necessary Rule

- Access only what's needed
- Role-based permissions
- Need-to-know principle
- Limit data sharing

Access Controls

- User authentication (MFA)

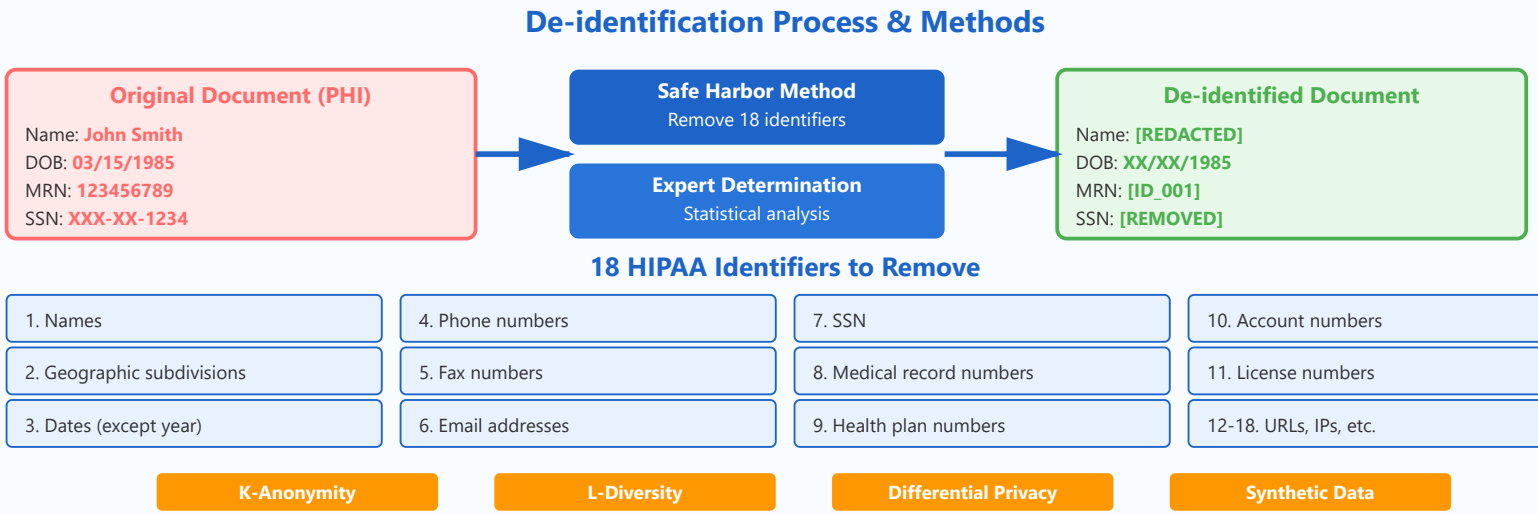
Breach Notification

- Report within 60 days

- Authorization levels
- Audit logs
- Encryption at rest & in transit

- Notify affected individuals
- Inform HHS if >500 patients
- Penalties for non-compliance

De-identification



Safe Harbor Method

- Remove all 18 identifiers
- Dates → year only
- Ages >89 → grouped as 90+
- Geographic: first 3 digits ZIP
- No statistical expertise needed

Expert Determination

- Statistical risk assessment
- Very small re-ID risk
- Retains more data utility
- Requires certified expert
- Document methodology

Automated Tools

- NER for PHI detection
- Philter, Scrubber, BoB
- Date shifting algorithms
- Validation required
- Hybrid human-AI review

Privacy Techniques

- K-anonymity (grouping)
- L-diversity (variety)
- Differential privacy (noise)
- Synthetic data (GANs)
- Utility vs privacy tradeoff

Real-World Evidence (RWE)

Randomized Controlled Trials

- Gold standard for efficacy
- Strict inclusion criteria
- Controlled environment
- Expensive & time-consuming
- Limited generalizability

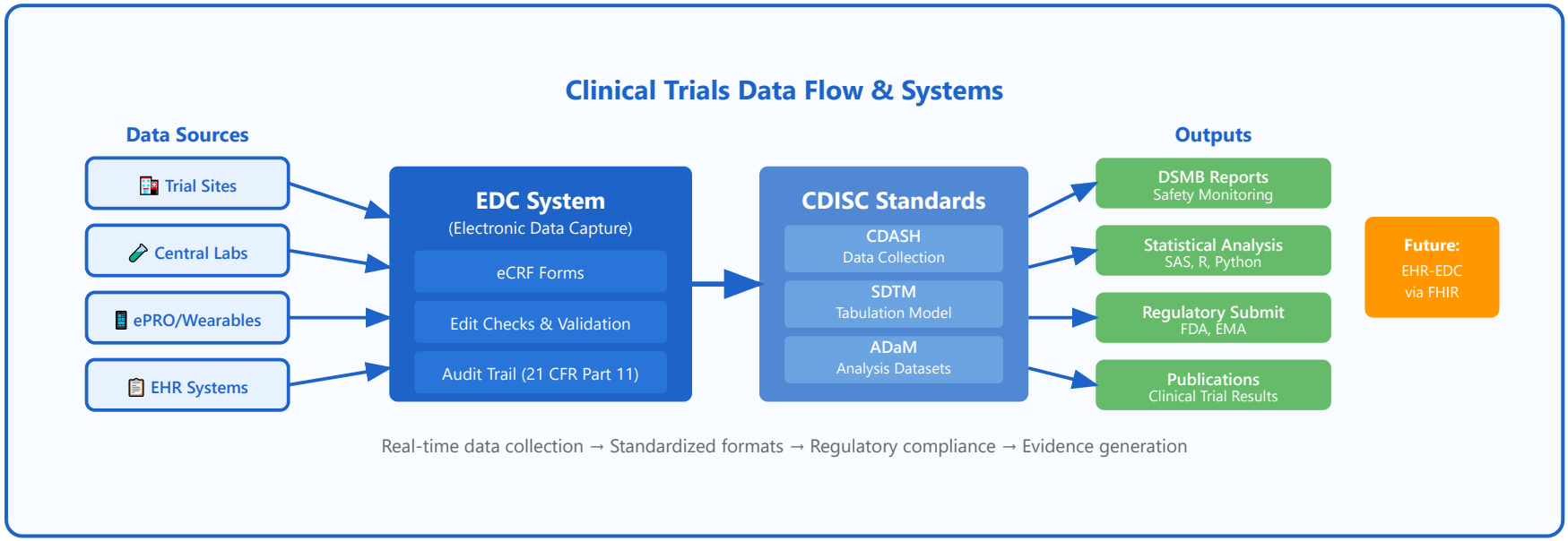
Real-World Evidence

- Effectiveness in practice
- Diverse patient populations
- Natural clinical settings
- Lower cost, faster
- Confounding & bias challenges

Regulatory Acceptance

FDA increasingly accepts RWE for drug approvals, label expansions, and post-market surveillance. Key: rigorous study design and bias mitigation.

Clinical Trials Data



💻 EDC Systems

- Electronic Data Capture
- eCRFs (case report forms)
- Real-time data validation
- Query management
- 21 CFR Part 11 compliance

📊 CDISC Standards

- CDASH: Collection standards
- SDTM: Tabulation model
- ADaM: Analysis datasets
- Define-XML: Metadata
- Regulatory submissions

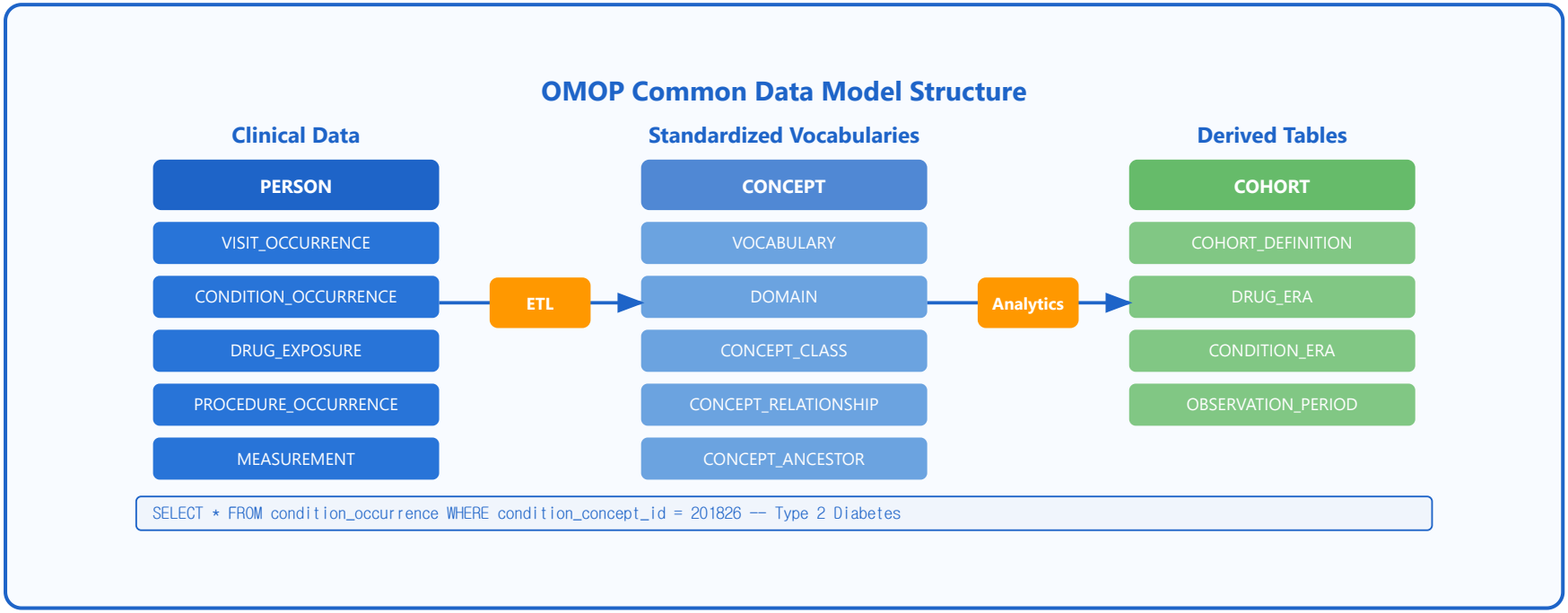
📈 Data Monitoring

- DSMB: Safety monitoring
- Interim analyses
- Stopping rules
- Adverse event tracking
- Risk-based monitoring

🖇️ EHR Integration

- Direct data transfer
- Automated eligibility screening
- FHIR for interoperability
- Reduces duplicate entry
- Real-world data linkage

Hands-on: OMOP CDM



📊 Data Model

- Standardized table structure
- Person-centric design
- Temporal relationships
- Standard vocabularies
- Source value preservation

🔄 ETL Process

- Source to standard mapping
- Concept ID assignment
- Date standardization
- White Rabbit profiling
- Rabbit-in-a-Hat mapping

🛠️ OHDSI Tools

- ATLAS: Cohort builder
- ACHILLES: Data quality
- HADES: R packages
- PLP: Prediction models
- CohortMethod: Causal inference

📈 Example Analysis

- Define T2DM cohort
- Characterize demographics
- Compare treatments
- Predict outcomes
- Network studies

Hands-on: Clinical NLP with Python

Clinical NLP Pipeline Example

```
# Install: pip install scispacy import spacy import scispacy from scispacy.linking import EntityLinker # Load clinical model nlp = spacy.load("en_core_sci_md") nlp.add_pipe("scispacy_linker", config={"resolve_abbreviations": True}) # Process clinical text text = "Patient prescribed metformin 500mg for T2DM. HbA1c was 7.2%" doc = nlp(text) # Extract entities for ent in doc.ents: print(f" {ent.text} → {ent.label_}") if ent._.kb_ents: cui = ent._.kb_ents[0][0] # UMLS CUI print(f" UMLS: {cui}")
```

```
# BioBERT for NER using Transformers from transformers import AutoTokenizer, AutoModelForTokenClassification import torch # Load BioBERT model tokenizer = AutoTokenizer.from_pretrained("dmis-lab/biobert-v1.1") model = AutoModelForTokenClassification.from_pretrained( "dmis-lab/biobert-v1.1" ) # Tokenize and predict inputs = tokenizer(text, return_tensors="pt") outputs = model(**inputs) predictions = torch.argmax(outputs.logits, dim=2)
```

NLP Pipeline Flow

scispaCy

- pip install scispacy
- Biomedical NER models
- UMLS entity linking

MedCAT

- Medical Concept Annotation
- Unsupervised learning
- Active learning interface

- Abbreviation detection
- Negation with NegEx

- Context detection
- SNOMED/UMLS linking

BioBERT/ClinicalBERT

- Pretrained on PubMed/MIMIC
- Fine-tuning for NER
- Relation extraction
- Question answering
- Hugging Face integration

Evaluation Metrics

- Precision: correct/predicted
- Recall: correct/actual
- F1 score: harmonic mean
- Entity vs token level
- Cross-validation

Thank You!
Future Directions in Clinical Informatics

AI in Healthcare

- Diagnostic assistance
- Drug discovery
- Personalized treatment
- Ambient clinical documentation

Precision Medicine

- Genomics integration
- Multi-omics data
- Pharmacogenomics
- Targeted therapies

Policy & Ethics

- Algorithmic bias
- Health equity
- Data governance
- International collaboration

Career Opportunities

- Clinical data scientist
- Health informatics researcher
- Bioinformatics engineer
- Healthcare AI developer