TATEEDA

AI data strategy for healthcare organizations. How hospitals can prepare data for AI

For hospital CIOs, data teams, informatics leaders, and digital health stakeholders

Inc. 5000 | San Diego Business Journal Fastest 100 Growing Private Companies | Clutch ★★★★★ 10 REVIEWS

What this paper covers

This white paper focuses on the part many AI projects skip: **data readiness.**

It covers four practical areas:

- data quality
- data architecture
- data governance
- anonymization and de-identification

> **Main point:** if the data is fragmented, inconsistent, or hard to access safely, AI projects slow down or fail.

Why hospitals struggle with AI data

Common issue

Data is split across systems

Important fields are incomplete

Records are inconsistent

Notes are hard to use safely

No clear ownership

What it looks like in practice

EHR, billing, labs, notes, devices, portals

Missing meds, allergies, timestamps, discharge details

Same concept stored in different ways

Useful detail mixed with PHI

Nobody owns field definitions or fixes

Result

Teams cannot build one reliable view

Weak models and weak reporting

Extra cleanup work

Privacy and access problems

Slow decisions

7220 Trade Street, Suite 103
San Diego, CA 92121

What “AI-ready data” actually means

AI-ready data is not “perfect data.”

It is data that is good enough for **a specific use case**, is **easy to locate**, is **structured consistently**, and can be used **without creating privacy risk**.

Example

For readmission prediction, medication history, discharge details, recent labs, and prior encounters matter far more than dozens of unused demographic or marketing fields.

7220 Trade Street, Suite 103
San Diego, CA 92121

Start with the right data quality checks

Quality check	Why it matters
Duplicate patient records	The same person may appear twice in training data
Missing allergy data	Clinical risk and poor care-support outputs
Missing medication stop/start dates	Hard to understand treatment history
Lab result without units	A model can misread the number
Bad encounter timestamps	Patient timeline becomes unreliable
Local codes not mapped	Data cannot be compared across systems

Rule: do not clean everything first.
Clean the fields your first AI use case depends on.

7220 Trade Street, Suite 103
San Diego, CA 92121

The first 10 metrics to track

Metric	Target owner
Duplicate patient rate	Registration / identity team
Medication completeness	Clinical informatics
Allergy completeness	Clinical informatics
Lab unit completeness	Lab / data engineering
Timestamp validity	Data engineering
Code mapping coverage	Terminology / integration team
Missing discharge disposition	Care operations
Encounter type accuracy	EHR / analytics
Note type classification rate	Informatics / NLP team
Failed data loads	Platform team

Simple operating model: review these weekly and assign one owner per metric.

7220 Trade Street, Suite 103
San Diego, CA 92121

A simple data architecture for AI

Layer	What goes there	Main rule
Raw	Source extracts as received	Do not manually edit
Clean	Standardized, linked, corrected records	Fix format and obvious errors
Trusted / curated	Approved datasets for reporting and AI	Add definitions and owners
Feature layer	Model-ready tables	Version everything

*FHIR bulk export exists for sharing **large FHIR datasets** between systems such as EHRs, data warehouses, and other clinical or administrative platforms.*

7220 Trade Street, Suite 103
San Diego, CA 92121

What a hospital lake or lakehouse should do

A central data platform should help the team do five things:

Need	Practical outcome
Bring data together	One place for EHR, labs, claims, notes, and device feeds
Preserve raw source data	Teams can trace where values came from
Create trusted datasets	Reporting and AI use the same approved source
Control access	Sensitive data is not exposed too broadly
Reuse work	Features built once can support multiple projects

> **Example:** A no-show model, staffing forecast, and patient outreach tool may all reuse the same appointment-history table.

7220 Trade Street, Suite 103
San Diego, CA 92121

Governance: who makes decisions

Role	Main responsibility
Data owner	Approves meaning, usage, and access
Data steward	Maintains definitions, mappings, and issue logs
Data engineer / platform team	Builds pipelines, quality checks, and access controls
Analytics / AI team	Defines what data is needed for the use case
Security / compliance	Reviews privacy, access, and release conditions

Use governance to answer simple questions:

- What does this field mean?
- Who can use it?
- How is it checked?
- What happens when the source changes?

NIST's AI Risk Management Framework is designed to help organizations manage AI risk as an ongoing program, and its core functions are Govern, Map, Measure, and Manage.

7220 Trade Street, Suite 103
San Diego, CA 92121

Standards that reduce future rework

Standard	Why it helps
USCDI	Defines core health data elements to exchange and use
US Core	Gives common FHIR profiles for representing those elements
OMOP CDM	Helps standardize longitudinal data for research and large-scale analytics

*ASTP/ONC's current platform shows **USCDI v6**, published in July 2025. HL7's current published **US Core** version is **8.0.0**. OHDSI describes **OMOP CDM** as an open common data standard for observational data, with standardized vocabularies for consistent analytics across sources.*

7220 Trade Street, Suite 103
San Diego, CA 92121

A practical standards strategy

Do not try to standardize everything at once.

Good approach: standardize the data that moves between teams, vendors, and use cases first.

High Priority
Map core patient, encounter, lab, medication, and allergy data

High Priority
Use shared definitions for dates, units, and status fields

Medium Priority
Align new integrations to FHIR / US Core patterns

Medium Priority
Map longitudinal analytics data to OMOP when cross-source consistency matters

Low Priority
Rebuild stable internal systems just for the sake of theory

7220 Trade Street, Suite 103
San Diego, CA 92121

How to structure records for AI

Data type	Best structure for AI
Patient demographics	Stable profile table
Encounters	One row per encounter with type and timing
Labs	Value + unit + reference range + timestamp
Medications	Drug, dose, start date, stop date, status
Allergies	Substance, reaction, severity, status
Notes	Separate controlled text pipeline
Appointments	Scheduled, canceled, missed, completed
Device data	Timestamped event stream or daily summary

Main principle: build a clean **patient timeline.**

7220 Trade Street, Suite 103
San Diego, CA 92121

Example: from messy chart data to model-ready data

Messy input

- one medication appears twice
- a lab result has no unit
- discharge date is present in one system but missing in another
- note text contains social-risk details but also PHI

Model-ready output

- one medication history record with start/stop logic
- lab result standardized to one unit
- one approved discharge timestamp
- note facts extracted into a controlled dataset with proper access rules

This is the kind of work that makes AI outputs more useful.

7220 Trade Street, Suite 103
San Diego, CA 92121

Notes and unstructured text need a separate workflow

• Problem with notes	• Safer approach
PHI is embedded in free text	Restrict access and de-identify where required
Different departments write differently	Tag note type and section
Notes include copied content	Track provenance and update timing
Extracted facts may be wrong	Validate against structured data when possible

Example: A discharge note may contain follow-up risks, caregiver issues, or housing concerns that never appear in coded fields. That makes notes valuable — but only if they are handled carefully.

7220 Trade Street, Suite 103
San Diego, CA 92121

De-identification: two main options

HHS guidance explains two HIPAA-recognized methods for de-identification:
Safe Harbor and **Expert Determination**.

Method	Best when	Tradeoff
Safe Harbor	You want a clear rules-based approach	May remove too much detail
Expert Determination	You need more useful data retained	Requires expert review and documentation

HHS also makes clear that de-identification reduces risk, but does not create a magical zero-risk dataset.

7220 Trade Street, Suite 103
San Diego, CA 92121

What Safe Harbor means in practice

• Usually removed or generalized	• Why it matters
Names and direct identifiers	Obvious identity risk
Detailed geography	Small-population re-identification risk
Full dates in many cases	Timelines can reveal identity
Contact details	Direct exposure risk
Free-text identifiers	Hardest part to clean reliably

Practical point:
Safe Harbor is easier to explain to leadership, but it can make some analytics and AI use cases less useful if too much detail is removed.

7220 Trade Street, Suite 103
San Diego, CA 92121

What Expert Determination means in practice

Requirement	What the team should ask for
Qualified expert	Who made the judgment and why they are qualified
Risk method	How re-identification risk was assessed
Assumptions	What external data or recipients were considered
Documentation	Written methods and final conclusion
Re-review triggers	When the assessment must be updated

This option is often better when the hospital needs more analytical value from the data, but it must be documented well.

7220 Trade Street, Suite 103
San Diego, CA 92121

A 90-day plan

Days 1–30

Pick 1–2 AI use cases, list required fields, define 10 quality metrics

Days 31–60

Build raw + clean + trusted layers, assign owners, set access rules

Days 61–90

Create first model-ready dataset, review privacy approach, launch weekly quality reporting

Good first use cases

- Readmission support
- No-show prediction
- Staffing demand forecasting
- Prior authorization support

7220 Trade Street, Suite 103
San Diego, CA 92121

What success looks like

By 90 days

- Core datasets have owners
- Quality metrics are reviewed weekly
- Access rules are documented
- One AI-ready dataset is live

By 6 months

- Trusted datasets are reused across projects
- Fewer one-off extracts and spreadsheet workflows
- Faster launch of new analytics and AI pilots
- Better auditability and lower rework

This option is often better when the hospital needs more analytical value from the data, but it must be documented well.

7220 Trade Street, Suite 103
San Diego, CA 92121

Final takeaway

Healthcare organizations do not need to "solve all data problems" before using AI.

They do need to:

1. Choose a real use case
2. Clean the data that matters for it
3. Centralize it in a controlled platform
4. Define ownership and access
5. Structure it for timelines and reuse
6. Choose a de-identification method on purpose

That is what turns scattered hospital data into something AI teams can safely use.

tateeda.com

7220 Trade Street, Suite 103
San Diego, CA 92121

Need help assessing your data readiness for AI?

Healthcare organizations do not need to “solve all data problems” before using AI.

TATEEDA helps healthcare organizations:

- Review data quality gaps
- Design practical lake / lakehouse flows
- Structure records for AI use cases
- Define governance and privacy controls

A strong AI program usually starts with better data decisions, not bigger models.

tateeda.com

7220 Trade Street, Suite 103
San Diego, CA 92121