From Silos to a Treasure Island

A challenge for a nation-wide health data repository in Japan

Shinji Kobayashi
Specially-appointed lecturer, Gifu university
Chair, Open-source working group of IMIA

My bio

Shinji Kobayashi, MD, PhD

- Physician and medical informatics researcher
- Engaged in software programming since the 1980s
- Researching EHR interoperability and clinical data standards since 2007
- Contributor to MML, the Dolphin project, and Sennen-Karte in Japan
- Strong believer in open standards, open source software, and clinical narratives.

MML (Medical Markup Language) and The Dolphin project 1995-2010

Short history

- 1995: MML (Medical Markup Language) project initiated (volunteer-based)
- 2001: Dolphin project launched (regional deployment)
- 2003-2006: Dolphin expanded to Tokyo and Kyoto
- 2015: Sennen-Karte project started
- 2019: Secondary use of clinical data initiated

Before the dawn of EHRs (early 1990s)

- Electronic health records were not yet available.
- Most clinical records were paper-based
- Digitization was discussed, but not yet implemented in Japan.
- **Simply digitizing paper charts was not enough**

The impact of the Internet boom (late 1990s)

- The Internet changed how we thought about information.
- Information sharing became natural and expected
- Data no longer stayed within a single organization
- **Healthcare could no longer remain isolated**

Regional Health Information Sharing

- The goal was not digitization, but **continuity of care**
- Continuity required sharing clinical information within a region
- Regional health information networks were envisioned
- **MML and the Dolphin project were designed to enable this**

MML (Medical Markup Language)

- A domestic clinical data standard in Japan
- Designed to enable regional health information sharing
- Structured representation of clinical information
- Developed by a voluntary, multi-institutional community

Covered Clinical Information

- Patient information
- Health Insurance
- Diagnostic records
- Basic consultation information
  - Allergy, Blood type, Infection
- First consultation
  - Family history, age, birth
- Progress notes
- Surgical operation
- Clinical summary
- Laboratory test results
- Report
- Referral

MML Version 3

Dolphin project: regional deployment

- Regional health information exchange infrastructure
- Based on the MML standard
- Deployed in real clinical settings
- Supported as a national project under the Japanese Millennium Initiative (2001-2003)

Geographical deployment

What Dolphin achieved

- Established a regional health information exchange infrastructure
- Enabled standardized sharing of clinical data based on MML
- Operated in real-world clinical settings across multiple regions
- Included early attempts to interconnect regional networks (“Super Dolphin”)

What Dolphin could not solve

- Large-scale secondary use of accumulated clinical data
- Sustainable nation-wide integration beyond interconnected regions
- Governance and social frameworks for long-term data reuse

Sennen-Karte Project

An EHR for the Next Millennium

Why we needed Sen-nen Karte

- Fragmented privacy rules and opt-in consent made large-scale secondary use difficult
- Meaningful data reuse required nation-wide coverage, not isolated regions
- Value could not emerge from partial or local datasets alone
- New technologies (e.g., cloud computing) required a different system architecture

Overview of Sen-nen Karte

- Technical foundations
  - ISO 13606 / openEHR as internal logical model
    - Interoperability beyond individual standards (e.g., HL7, MML)
  - Hadoop-based architecture for large-scale storage and search
- Social and institutional foundations
  - Legal frameworks advanced by the government
    - controlled secondary use beyond opt-in only
  - Life Data Initiative (LDI) established as an institutional home for data reuse

Modernization of MML to Version 4

- Existing Dolphin sites needed continuity
  - migration cost had to be kept minimal
- MML had to be preserved for ongoing operations
- At the same time, MML needed modernization
  - easier implementation for new sites
- Internal logical model redesigned using openEHR / ISO 13606

MML Modules

Six additional modules were introduced in response to real clinical demands.

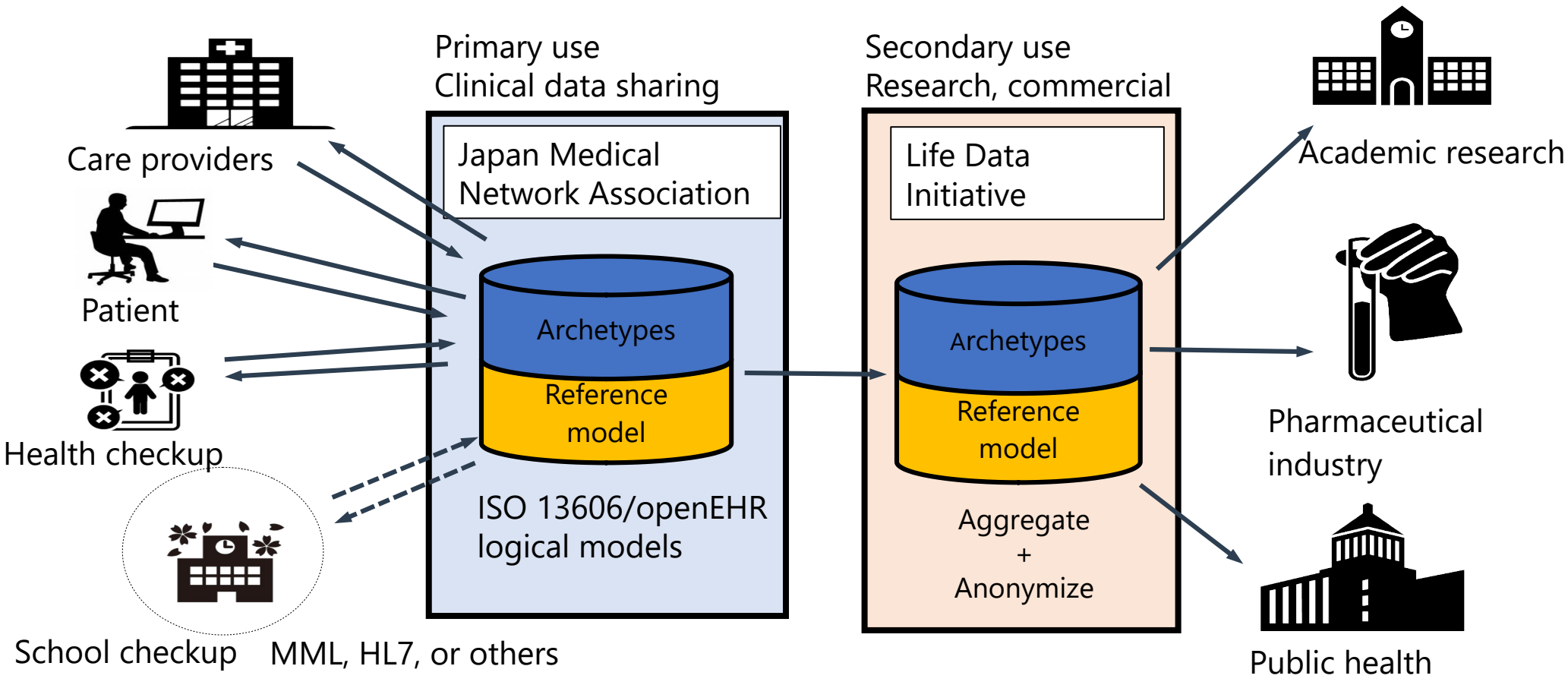
- Patient information
- Health Insurance
- Diagnostic records
- Basic consultation information
  - Allergy, Blood type, Infection
- First consultation
  - Family history, age, birth
- Progress notes
- Surgical operation
- Clinical summary

- Laboratory test results
- Report
- Referral
- Vital sign
- Flowsheet
- Prescription
- Injection
- Hemodialysis
- Procedure

Key Legislation

- Revised Act on the Protection of Personal Information (2015)
  - Redefined personal and sensitive data
  - Introduced new rules for secondary use of health data
- Act on Medical Infostructure for the Next Generation (2017)
  - Established governance frameworks for health data reuse
  - Enabled controlled secondary use through certified anonymization

Sennen-Karte “Millennial Electronic Health Record”



What Sen-nen Karte achieved

Scale of Sen-nen Karte (as of 2025)

- Primary use
  - ~10 million registered patients
  - Continuous use in daily clinical practice
- Secondary use
  - ~3.1 million patients available for secondary use
  - Governed, approved, and ethically controlled reuse
- In practice
  - Continuous data collection since 2016
  - Dozens of approved clinical research projects in progress

From regional pilots to nation-wide deployment

- Dolphin era

- Sennen-Karte era

Each dot represents a connected healthcare facility

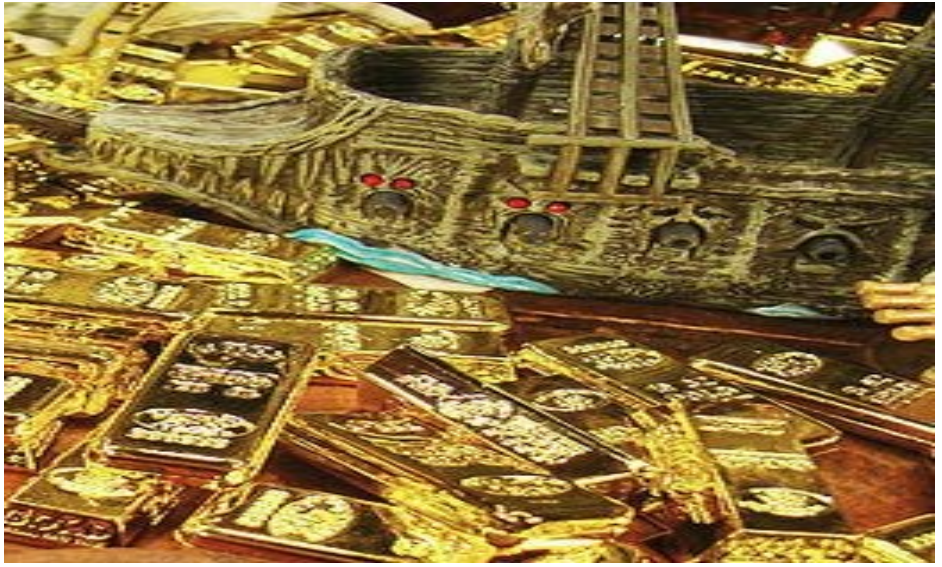
Major data instances accumulated

Contents	Number of instances (million)
Diagnoses	480
Progress notes	**450**
Insurance	330
Injection	195
Laboratory test results	180
Vital signs	170
Prescription	144
Patient demographics	118
Flowsheets	85

Current challenges

- Quality and consistency depend on human input
- Clinical narratives are hard to query and aggregate
  - Machine readability remains limited for free text
- Some hospitals send documents as PDF or images.
  - Value extraction requires significant manual effort

Junk pile or Treasure Island



Why clinical narratives were hard to reuse

- Written for humans, not for machines
- Context-rich but structure-poor
- Essential for clinical reasoning
- Ignored or underused in large-scale analysis

Large Language Models Changed the Game

2019年12月 初診。胸部X線で異常影を認めた。

胸腔穿刺により肺腺癌と診断。

EGFR L858R変異陽性、Stage IVB。

Osimertinibによる治療を開始。


```
{
  "encounter": {
    "date": "2019-12-*",
    "type": "initial_visit"
  },
  "diagnosis": {
    "primary": {
      "name": "lung adenocarcinoma",
      "method": "thoracentesis",
      "certainty": "confirmed"
    },
    "stage": {
      "system": "TNM",
      "value": "IVB"
    }
  },
  "genomic_profile": {
    "driver_mutations": [
      {
        "gene": "EGFR",
        "variant": "L858R",
        "status": "positive"
      }
    ]
  },
  "treatment": {
    "line": 1,
    "regimen": {
      "drug": "osimertinib",
      "class": "EGFR-TKI"
    },
    "intent": "systemic_therapy"
  },
  "source": {
    "original_format": "clinical_progress_note",
    "language": "ja",
    "extraction_method": "LLM-based semantic extraction"
  }
}
```

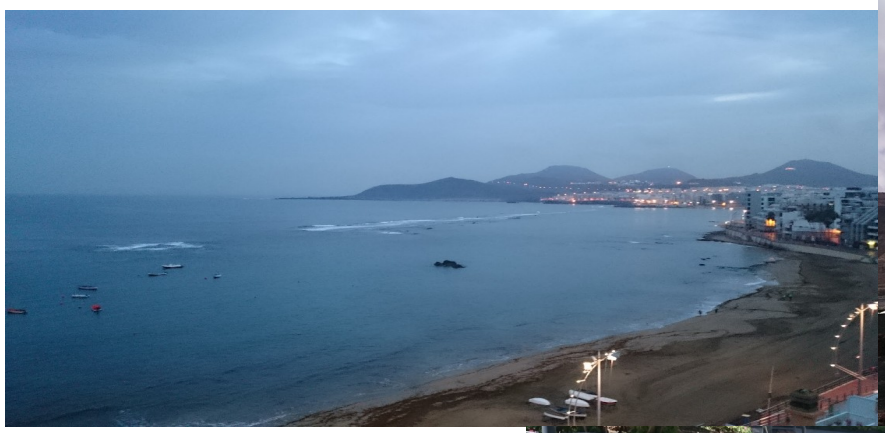

- Actual clinical narratives are much longer.

Human validation is required. This JSON represents extracted meaning, not ground truth.

Current challenges (after LLM)

- Output quality still depends on input quality
- Human validation remains necessary
- Bias and variability in clinical narratives persist
- Governance and accountability for secondary use are essential

Treasure Island!



La salud es un derecho humano no negociable.

Cada vida, cada una, es un tesoro.

Cada dato de salud es también un tesoro.

Health is a non-negotiable human right.

Every single life is a treasure.

Health data is also a treasure.

健康は交渉の余地がない基本的人権である。

一つひとつの生命は宝である。

医療データもまた宝である。