



Consolidating Fragmented Hospital Records Into a Dimensional Warehouse for a Regional Health Network: A Proposed Design Procedure

Pilar Petrosyan^{1*}, Larissa Mahlangu², Davit Moreno³

¹Department of Computer Science, Skagerrak Institute of Computer Science, Philippines

²Department of Computer Science, Danube Institute of Computer Science, Japan

³Department of Computer Science, Siberian Institute for Advanced Computer Science, Ethiopia

Article Info

Volume 1, Issue 1, 2026

doi: [10.62715/ijdwdm.2026.v1.i1.pp15-20](https://doi.org/10.62715/ijdwdm.2026.v1.i1.pp15-20)

Received: 13 January 2026

Accepted: 10 March 2026

Published: 22 May 2026

Abstract

Regional health networks are usually assembled from hospitals and clinics that each bought their own record systems, so that one patient's history lies scattered across several databases under different identifiers and code sets. Reporting on quality, capacity and population health across the network requires those records to be brought together, and a dimensional data warehouse remains the most widely understood structure for the purpose. This article proposes a seven-step procedure for building one. The steps are: agree governance and take an inventory of the sources; choose the clinical and administrative processes to model and declare the grain of each fact table; build conformed dimensions for patient, provider, facility, date and coded clinical concepts; resolve patient identity across sites with probabilistic record linkage; load the data through a staging area with history-preserving dimension handling; test data quality against explicit rules; and release the warehouse under role-based access and de-identification controls. Each step is described with the decisions it requires and the artefact it should produce. The procedure is assembled from the published literature on dimensional modelling, record linkage and health data quality. It is a proposal, it has not been applied or validated in a real network here, and the article sets out what an evaluation would need to establish.

Keywords: *Clinical data warehouse, Dimensional modelling, Record linkage, Master patient index, Data quality, Health information integration*

© 2026 Pilar Petrosyan et al. This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

1. Introduction

A regional health network seldom starts with a shared information system. It grows by affiliation and merger, and each member arrives with its own electronic health record, laboratory system, billing system and patient index. A person treated at two hospitals and a clinic in the same network may be known by three identifiers, have diagnoses recorded in different versions of a classification, and have laboratory results labelled with local test names. Clinicians cope by asking the patient. Planners and quality officers

* Corresponding author: Pilar Petrosyan, Department of Computer Science, Skagerrak Institute of Computer Science, Philippines. E-mail: pilar.petrosyan@skagerrak.edu

© 2026 Pilar Petrosyan et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

cannot, and questions such as how many patients were readmitted anywhere in the network within a month of discharge go unanswered or are answered by hand.

The secondary use of routinely collected health data for quality measurement, planning and research has long been recognised as valuable and as dependent on sound governance (Safran et al., 2007). A data warehouse is the usual technical response: a separate database, loaded periodically from the operational systems and organised for analysis (Chaudhuri and Dayal, 1997). Within warehousing, the dimensional approach arranges data as fact tables of measured events surrounded by descriptive dimensions, a form that analysts read easily and that query tools handle well (Kimball and Ross, 2013).

This article proposes a procedure for consolidating fragmented hospital records into such a warehouse. It is written for the informatics team of a network of modest size that cannot replace its members' source systems. The article describes a design and the reasoning behind it. It does not report an implementation, and no network, patient or dataset is described.

2. Background

The procedure combines four bodies of established work, and each is summarised briefly here so that the later steps can refer to it.

2.1. Dimensional and normalised designs

Two design traditions dominate warehousing. Inmon (2005) advocates an integrated, normalised enterprise warehouse from which subject-specific marts are derived. Kimball and Ross (2013) build the warehouse directly as a set of dimensional models tied together by conformed dimensions, meaning dimensions such as patient or date that carry the same keys and meaning wherever they are used. The dimensional route delivers usable results one business process at a time, which suits a network that must show value early. Its discipline lies in declaring the grain of each fact table, that is, exactly what one row represents, before anything else is designed.

2.2. Clinical data models and exchange standards

Health informatics has produced its own analytical models. The i2b2 platform stores observations in a single fact table keyed to patient, encounter, concept and time (Murphy et al., 2010), and the common data model evaluated by Overhage et al. (2012) standardises both table structure and vocabularies so that the same analysis can run at many sites. These models are optimised for research cohorts. A network that also needs operational reporting on beds, theatres and waiting times will want process-specific fact tables as well, though it should borrow their insistence on standard vocabularies. For getting data out of source systems, interfaces based on the HL7 FHIR standard increasingly supplement traditional message feeds and database extracts (Mandel et al., 2016).

2.3. Record linkage

Deciding whether two records refer to the same person is the oldest problem in health data integration. The probabilistic theory of Fellegi and Sunter (1969) compares pairs of records field by field, weights each agreement or disagreement by how much more likely it is among true matches than among non-matches, and classifies pairs as links, non-links or possible links for manual review. Modern practice adds blocking to limit the number of comparisons, approximate string comparison for names, and systematic evaluation of linkage quality (Christen, 2012). Where identifiers may not be exchanged in clear text, linkage can be performed on encoded values (Schnell et al., 2009).

2.4. Data quality and privacy

Data quality is a matter of fitness for the consumer's purpose and has several dimensions beyond accuracy (Wang and Strong, 1996). For electronic health records, Weiskopf and Weng (2013) identified completeness, correctness, concordance, plausibility and currency as the dimensions assessed in the literature. The cleaning problems met during integration, such as conflicting values, differing representations and duplicates, were classified by Rahm and Do (2000). Privacy law frames everything. In the United States the HIPAA Privacy Rule governs the use and disclosure of protected health

information and defines routes to de-identification ([US Department of Health and Human Services, 2002](#)). In the European Union the General Data Protection Regulation treats health data as a special category ([European Parliament and Council, 2016](#)).

3. The Proposed Procedure

The procedure has seven steps. Table 1 lists each with its central decision and the artefact it should leave behind, and Figure 1 shows the order in which they run. The order matters less than the dependencies: identity resolution needs the patient dimension to exist, and quality testing feeds corrections back to the owners of the source systems.

Table 1: Steps of the proposed consolidation procedure, with the central decision and output of each		
Step	Central decision	Artefact produced
1. Governance and inventory	Who may authorise use of each source, and what each contains	Data-sharing agreement; source catalogue
2. Processes and grain	Which events are modelled and what one fact row means	Bus matrix; grain statements
3. Conformed dimensions	Which shared descriptors and code systems the network adopts	Dimension specifications; terminology maps
4. Patient identity	Which record pairs are accepted as the same person	Master patient index; linkage report
5. Staged loading	How history, corrections and late records are handled	Load specifications; audit columns
6. Quality testing	Which rules a load must pass before release	Rule catalogue; quality report per load
7. Controlled release	Who sees identified, coded or aggregate data	Access policy; de-identified views

3.1. Step 1: Agree governance and take an inventory

Technical work should not begin until the member organisations have agreed in writing what the warehouse is for, who is accountable for it and under what legal basis each source will contribute. A steering group with clinical, information governance and technical members should own those decisions. The team then catalogues every candidate source: the system, its custodian, the tables or messages available, the identifiers used, the code systems, the period covered and the extraction route. The catalogue usually reveals that several sources hold overlapping versions of the same facts, and one must be named as authoritative for each.

3.2. Step 2: Choose the processes and declare the grain

A dimensional warehouse is organised around processes that generate measurable events. For a hospital network a reasonable first set is inpatient and outpatient encounters, diagnoses, procedures, laboratory results and medication orders. For each the team writes one sentence stating the grain, for example one row per diagnosis recorded against an encounter, or one row per reported result of a single laboratory test. A bus matrix, with processes as rows and dimensions as columns, records which dimensions each fact table uses and exposes where conformity will be needed. Starting with encounters is advisable, since most other facts hang from an encounter.

3.3. Step 3: Build the conformed dimensions

Patient, provider, facility and date dimensions are shared by nearly every fact table and must be built once. Each receives a surrogate key generated by the warehouse, so that source identifiers can change without breaking history. Coded clinical concepts need the most care. The network should choose a target code system for each domain, such as the International Classification of Diseases for diagnoses and

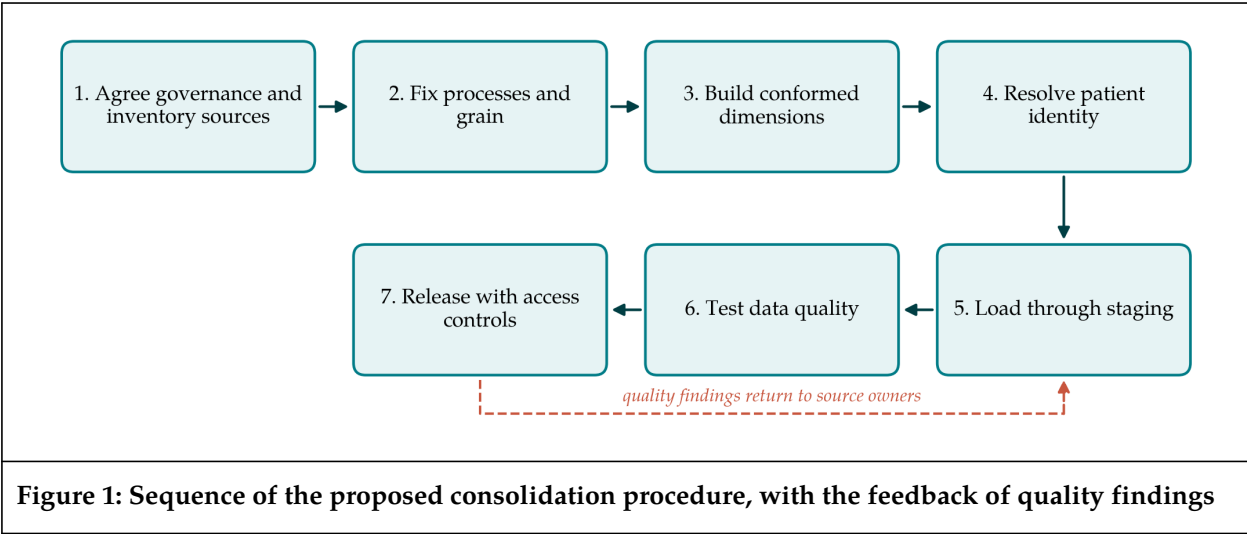
LOINC for laboratory tests, and maintain explicit maps from every local code to the target. The local code is kept beside the mapped one in the dimension, so that an analyst can always see what was originally recorded and a mapping error can be traced.

3.4. Step 4: Resolve patient identity

This step decides whether the warehouse can answer network-wide questions at all. Records from each site's patient index are standardised, with names parsed, dates validated and addresses normalised, and are then compared within blocks defined by fields such as date of birth or a phonetic name code. Pairs are scored in the Fellegi-Sunter manner. Two thresholds divide them into automatic links, automatic non-links and a middle band for clerical review. The thresholds reflect a judgement about which error is worse: a false link merges two people's histories, while a missed link merely leaves a history split. For analytical use a conservative upper threshold is the safer default. The result is a master patient index that maps every source identifier to one warehouse patient key, with the match score and method retained for audit.

3.5. Step 5: Load through a staging area

Extracts land unchanged in a staging area, where they are profiled, cleaned and mapped before any warehouse table is touched; the design choices available for such pipelines are surveyed by [Vassiliadis \(2009\)](#). Three behaviours should be specified in advance. Dimension attributes that change, such as a patient's address or a provider's department, are handled as slowly changing dimensions, with a new row and validity dates when history matters for analysis. Corrections issued by a source replace the earlier fact while an audit column records the load in which each version arrived. Facts that arrive before their dimension member, a laboratory result for a patient not yet registered in the index for instance, are attached to a placeholder member and repaired on a later load.



3.6. Step 6: Test data quality

Each load is tested against a catalogue of rules before it is released. The dimensions of [Weiskopf and Weng \(2013\)](#) give the catalogue its structure. Completeness rules check that expected fields and expected volumes are present for every site and period. Plausibility rules check ranges and sequences, such as a discharge that precedes its admission. Concordance rules compare related elements, such as a procedure against the patient's recorded sex where that is clinically determinate. Currency rules check the delay between an event and its arrival. Failures are reported to the source custodian for correction at origin, because a fix applied only in the warehouse will be undone by the next extract.

3.7. Step 7: Release under access and de-identification controls

Access is granted by role. Staff engaged in direct care or operational management may need identified data for their own facility; analysts generally do not. For them the warehouse should present views in

which direct identifiers are removed and dates and geography are coarsened. The HIPAA Privacy Rule offers two routes to de-identification, expert determination and the removal of eighteen listed types of identifier, and the second is the simpler to implement as a view. Removal of identifiers does not by itself prevent re-identification from combinations of ordinary attributes, the risk that the k -anonymity model was devised to measure (Sweeney, 2002), so small cells in released aggregates should be suppressed. Every query against identified data is logged.

4. Fitting the Procedure to a Regional Network

Three features of regional networks shaped the proposal. The first is uneven capability: a teaching hospital may offer a modern interface to its record system while a small clinic can supply only a nightly file. The staging design accepts both, and the source catalogue records the difference so that expectations of timeliness are set honestly. The second is limited staff. The procedure therefore delivers one process at a time, and a network that completes encounters and diagnoses has a useful warehouse before laboratory data are attempted. The third is divided ownership. Members remain separate legal entities, and the governance step exists because technical integration without agreed authority tends to stall at the first disputed request for data.

5. Limitations and What Remains to Be Shown

The procedure is a proposal built from published methods, and it has not been applied or validated here. Several things would need to be shown before it could be recommended with confidence. Linkage quality is the first: the rates of false and missed links achieved with the demographic fields that a given network actually holds, measured against a manually reviewed sample. The effort required for terminology mapping is the second, since it is commonly underestimated and depends on how idiosyncratic local codes are. The third is whether the quality rules detect the errors that matter to users, which can only be judged by comparing warehouse figures with trusted local reports.

The design also has known boundaries. A dimensional model represents coded, structured events well and free-text notes, images and waveforms poorly; those need other stores. Nightly loading suits retrospective reporting and not bedside decision support. And a warehouse inherits the documentation habits of its sources: a diagnosis that was never coded cannot be recovered by any amount of integration. Figures drawn from the warehouse describe recorded care, which is not always the same as care delivered.

6. Conclusion

Fragmented hospital records can be consolidated without replacing the systems that hold them, provided that identity, vocabulary and grain are settled deliberately. The seven steps proposed here place governance before engineering, treat patient linkage as the step on which network-wide analysis depends, and make data quality testing and controlled release part of every load. The procedure draws on methods that are individually well established. Whether their combination works at acceptable cost in a particular network is an empirical matter that this article leaves to those who apply it.

Conflict of interest

To be declared by the author(s) before publication.

Funding

To be declared by the author(s) before publication.

References

- Chaudhuri, S. and Dayal, U. (1997). [An overview of data warehousing and OLAP technology](#). *ACM SIGMOD Record*, 26(1): 65-74. <https://doi.org/10.1145/248603.248616>
- Christen, P. (2012). [Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and](#)

- Duplicate Detection**. Springer, Berlin.
- European Parliament and Council (2016). **Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation)**. *Official Journal of the European Union*, L 119: 1-88.
- Fellegi, I.P. and Sunter, A.B. (1969). **A theory for record linkage**. *Journal of the American Statistical Association*, 64(328): 1183-1210. <https://doi.org/10.1080/01621459.1969.10501049>
- Inmon, W.H. (2005). **Building the Data Warehouse**, 4th edn. Wiley, Indianapolis.
- Kimball, R. and Ross, M. (2013). **The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling**, 3rd edn. Wiley, Indianapolis.
- Mandel, J.C., Kreda, D.A., Mandl, K.D., Kohane, I.S. and Ramoni, R.B. (2016). **SMART on FHIR: A standards-based, interoperable apps platform for electronic health records**. *Journal of the American Medical Informatics Association*, 23(5): 899-908. <https://doi.org/10.1093/jamia/ocv189>
- Murphy, S.N., Weber, G., Mendis, M., Gainer, V., Chueh, H.C., Churchill, S. and Kohane, I. (2010). **Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2)**. *Journal of the American Medical Informatics Association*, 17(2): 124-130. <https://doi.org/10.1136/jamia.2009.000893>
- Overhage, J.M., Ryan, P.B., Reich, C.G., Hartzema, A.G. and Stang, P.E. (2012). **Validation of a common data model for active safety surveillance research**. *Journal of the American Medical Informatics Association*, 19(1): 54-60. <https://doi.org/10.1136/amiajnl-2011-000376>
- Rahm, E. and Do, H.H. (2000). **Data cleaning: Problems and current approaches**. *IEEE Data Engineering Bulletin*, 23(4): 3-13.
- Safran, C., Bloomrosen, M., Hammond, W.E., Labkoff, S., Markel-Fox, S., Tang, P.C. and Detmer, D.E. (2007). **Toward a national framework for the secondary use of health data: An American Medical Informatics Association white paper**. *Journal of the American Medical Informatics Association*, 14(1): 1-9. <https://doi.org/10.1197/jamia.M2273>
- Schnell, R., Bachteler, T. and Reiher, J. (2009). **Privacy-preserving record linkage using Bloom filters**. *BMC Medical Informatics and Decision Making*, 9: 41. <https://doi.org/10.1186/1472-6947-9-41>
- Sweeney, L. (2002). **k-anonymity: A model for protecting privacy**. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5): 557-570. <https://doi.org/10.1142/S0218488502001648>
- US Department of Health and Human Services (2002). **Standards for privacy of individually identifiable health information: Final rule**. 45 CFR Parts 160 and 164, Federal Register 67(157).
- Vassiliadis, P. (2009). **A survey of extract-transform-load technology**. *International Journal of Data Warehousing and Mining*, 5(3): 1-27. <https://doi.org/10.4018/jdwm.2009070101>
- Wang, R.Y. and Strong, D.M. (1996). **Beyond accuracy: What data quality means to data consumers**. *Journal of Management Information Systems*, 12(4): 5-33.
- Weiskopf, N.G. and Weng, C. (2013). **Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research**. *Journal of the American Medical Informatics Association*, 20(1): 144-151. <https://doi.org/10.1136/amiajnl-2011-000681>