ESP-JETA
ESP Journal of Engineering & Technology Advancements
ISSN: 2583-2646 / Volume 3 Issue 1 March 2023 / Page No: 229-236
Paper Id: JETA-V3I3P118 / Doi: 10.56472/25832646/JETA-V3I3P118

Original Article

PII De-Identification Techniques for Healthcare Data Warehouses

Narasimha Chaitanya Samineni
Vice President, Quality Assurance Supervisor.

Received Date: 28 February 2023 Revised Date: 25 March 2023 Accepted Date: 31 March 2023

Abstract: *Healthcare data warehouses consolidate electronic health records, claims, lab systems, and operational data to support analytics, quality improvement, and research. However, these warehouses often contain protected health information (PHI) and other personally identifiable information (PII) that can be linked across sources, making privacy risk higher than in single-system datasets. This article presents a practical, risk-based framework for de-identifying healthcare warehouse data while preserving analytic utility. The framework aligns regulatory expectations for de-identification, emphasizes realistic threat models, and maps common privacy techniques to healthcare warehouse patterns such as longitudinal patient linkage, cohort discovery, and machine learning feature stores. We describe a technique taxonomy spanning masking, suppression, pseudonymization and tokenization, generalization-based privacy models (k-anonymity family), distributional protections (t-closeness), and differential privacy for aggregate outputs. We then propose a layered warehouse architecture separating a restricted PHI vault from an analytics-ready de-identification zone, supported by strong access control, key management, and auditing. Finally, we provide utility and quality validation methods, operational governance controls, and an implementation blueprint to help organizations deploy de-identification as an engineered capability rather than a one-time data export step. [1][2][4][5][8]*

Keywords: *Healthcare Data Warehouse, PHI, PII, De-identification, Tokenization, k-anonymity, Differential Privacy, Re-identification Risk, Governance. [1][4][5][8]*

I. INTRODUCTION

Healthcare organizations increasingly rely on enterprise data warehouses to unify fragmented clinical and administrative data into consistent, queryable structures for reporting and secondary use. Common warehouse use cases include cohort identification for research, readmission and quality metrics, fraud and abuse analysis, clinical operations dashboards, and population health segmentation. These use cases often require longitudinal linkage across encounters, facilities, and time, which is exactly what makes privacy risk harder: linkage increases the uniqueness of patient trajectories even when obvious identifiers are removed. [5][10][11]

Regulatory and ethical expectations also push organizations to share data more broadly while reducing privacy risk. In the United States, HIPAA provides two recognized paths for de-identification, Safe Harbor and Expert Determination, and both approaches have practical implications for how warehouse data products should be engineered, validated, and released. In the European context, GDPR distinguishes between anonymized data and pseudonymized data, with pseudonymized data still treated as personal data if re-linkage is possible. These requirements create a common challenge: how to preserve analytic value while reducing re-identification risk to an acceptably low level for the intended sharing context. [1][2][3]

This paper contributes three items. First, it provides a technique taxonomy specifically grounded in healthcare data warehouse realities such as longitudinal linkage and multi-domain integration. Second, it proposes a layered warehouse architecture that separates a restricted PHI vault from a governed de-identification zone. Third, it supplies operational validation and governance guidance so de-identification can be repeatable, measurable, and auditable. [1][4][13][15]

II. HEALTHCARE DATA WAREHOUSES AND PII/PHI CHARACTERISTICS

A healthcare data warehouse typically ingests data from EHR systems, claims and billing platforms, lab and imaging systems, and sometimes external registries. Many organizations also standardize data into common models to support consistent analytics and reusable tooling, such as i2b2 for cohort and translational research workflows, or OMOP CDM for observational studies across heterogeneous sources. These warehouse patterns emphasize longitudinal patient identity resolution and repeated use across teams, which makes privacy engineering central rather than optional. [19][20]

Healthcare PII and PHI differ from consumer datasets in two critical ways. First, clinical facts can be highly sensitive (diagnoses, medications, procedures), and harm from disclosure can be substantial. Second, quasi-identifiers are plentiful: demographics, ZIP code, timestamps, provider location, encounter sequences, and rare conditions can uniquely identify patients when combined. Classic research shows that even datasets without direct identifiers can be vulnerable to linkage attacks when quasi-identifiers overlap with external data sources. [5][10][12]

Warehouse data also contains "longitudinal fingerprints," for example repeated visit patterns, sequences of lab tests, and time-stamped care events. These patterns improve analytics but can also enable re-identification by matching unique trajectories. The risk increases as the warehouse integrates more domains, which is why de-identification for warehouses must address linkability and composition, not only remove names or IDs. [10][11][12]

Table 1 : Common Phi Elements in Healthcare Warehouses and Recommended De-Identification Treatment

PHI/PII Element Category	Examples in Warehouse	Common Treatment Options	Notes for Warehouse Analytics	Reference(s)
Direct identifiers	Name, MRN, SSN, phone, email	Remove, tokenize, or store only in PHI vault	Keep link keys separated from analytics zone	[1][2][4]
Geographic detail	Street address, full ZIP	Generalize (3-digit ZIP), suppress, region bucketing	Geography often needed, prefer controlled generalization	[1][5][7]
Dates and timestamps	DOB, admission time, procedure date	Date shifting, age banding, suppress rare timestamps	High re-ID risk with exact timestamps in longitudinal data	[1][5][11]
Rare conditions or events	Uncommon diagnosis, rare procedure	Suppress, generalize codes, apply t-closeness	Risk rises sharply with rarity and external knowledge	[7][11][12]
Free-text fields	Notes, referrals, messages	NLP-based redaction, remove, or synthesize	Often contains hidden identifiers; high leakage risk	[1][13][15]
Device or account IDs	Portal IDs, device IDs	Tokenize, rotate tokens per context	Prevent cross-context linkage unless justified	[4][13][15]

III. REGULATORY AND GOVERNANCE REQUIREMENTS

Healthcare de-identification programs must satisfy not only legal requirements, but also institutional governance expectations for ethical use, auditability, and data minimization. In the United States, the HIPAA Privacy Rule recognizes two primary pathways for de-identification: Safe Harbor and Expert Determination [1][2]. Safe Harbor prescribes removal of specific identifiers and requires that the covered entity have no actual knowledge that remaining data could identify an individual [2]. While Safe Harbor is straightforward for static datasets, it can reduce warehouse utility because it frequently forces removal or coarse handling of dates, geographic details, and other attributes essential for longitudinal analysis and operational reporting [1][2].

Expert Determination is often more compatible with enterprise analytics because it permits tailored transformation strategies, provided a qualified expert concludes that the re-identification risk is "very small" in the intended context and documents the reasoning, methods, and results [1][2]. Practically, this means de-identification must be treated as a measurable engineering outcome. Organizations should define risk metrics (such as equivalence class sizes, uniqueness measures, and attribute disclosure checks) and store the evidence as part of release documentation and audit trails [1][13]. Expert Determination also encourages context-aware choices, such as controlled date shifting windows, geographic bucketing aligned to population density, and suppression rules for rare combinations that create uniqueness [1][5][7].

In the European context, GDPR introduces an important boundary between anonymization and pseudonymization [3]. Pseudonymization, such as replacing identifiers with tokens while retaining a re-linkage key, remains personal data under GDPR if re-identification is reasonably possible [3]. This is operationally significant for data warehouse design because tokenization and key separation are necessary safeguards, but they are not automatically equivalent to anonymization. Therefore, governance must specify which zones and outputs are "restricted," "pseudonymized," or "de-identified," and apply controls accordingly [3][4].

Modern programs also benefit from standardized terminology and technique classification. ISO/IEC 20889 provides structured language for describing de-identification methods and helps align technical controls with compliance narratives, which is particularly helpful when multiple teams and vendors participate in warehouse pipelines [4]. Beyond regulatory

requirements, most healthcare organizations must satisfy internal privacy boards, IRB expectations, and data access committees. These governance structures typically require: (1) purpose limitation, (2) minimum necessary data, (3) access approvals, (4) auditability, and (5) retention controls for extracts and downstream products [1][3][13]. A mature governance approach ties these requirements to data product tiers and ensures that every release has a documented policy basis, repeatable transformation configuration, and measurable privacy validation report [1][4][13].

IV. THREAT MODELS AND RE-IDENTIFICATION RISK IN WAREHOUSE SETTINGS

A healthcare data warehouse increases re-identification risk because it consolidates data across domains and time, enabling powerful linkage and filtering. Effective de-identification begins with a clear threat model that answers: who might attempt re-identification, what auxiliary information they might have, and what access or query capabilities they possess [10][12]. In warehouse settings, threats include external recipients of extracts, analysts with broad internal access, contractors, or malicious insiders. Even without malicious intent, analysts can inadvertently expose identity by building small cohorts, running differencing queries, or exporting high-granularity datasets for convenience [1][13].

Linkage attacks are a primary risk: an attacker links quasi-identifiers in a “de-identified” dataset (age, ZIP, dates, provider location, visit patterns) to an external dataset that contains identifiers [5][10]. Classic evidence shows that sparse, high-dimensional datasets can be vulnerable when the attacker has partial auxiliary information, especially when data includes timestamps or unique event combinations [10]. Inference attacks are also important: even if a patient is not uniquely identified, the dataset may reveal sensitive attributes with high confidence for a small group (for example, when a cohort slice has near-uniform diagnosis distribution) [6][7]. Finally, composition attacks matter for warehouses because releases and queries happen repeatedly. Multiple extracts or repeated interactive queries can combine to reveal information that each individual release would not expose [8][9][12].

Healthcare data is particularly vulnerable because it contains rare events and rare conditions. Uncommon diagnoses, rare medication combinations, and specialized procedures can function like fingerprints. When those fingerprints are combined with time and geography, uniqueness increases sharply, which is why date handling and geo granularity are critical design choices [5][7][11]. A systematic review of re-identification attacks on health data shows that the feasibility and success of re-identification attempts depend on context, methodology, and the presence of auxiliary information, reinforcing that de-identification must be evaluated as a risk management exercise rather than a checklist [11][13].

A practical approach is to define **risk tiers** aligned to sharing context. For example:

- Tier A: Internal analytics with strict controls. Use pseudonymization or tokenization with strong access boundaries, plus minimization of direct identifiers.
- Tier B: Internal research and broad internal sharing. Add generalization and suppression policies to reduce uniqueness, and enforce minimum cohort size rules.
- Tier C: External research release. Apply formal privacy models such as k-anonymity family constraints plus attribute disclosure controls (l-diversity, t-closeness), and include documented expert determination evidence.
- Tier D: Public reporting and interactive exploration. Prefer aggregate-only outputs protected using differential privacy or equivalent query controls to address composition risk. [1][3][5][7][8][9]

Risk evaluation should also consider “linkability across datasets.” Stable tokens enable longitudinal analysis, but also enable linkage across products if tokens are reused broadly. A warehouse program should specify token scope and rotation strategies so that linkability is enabled only where justified and approved. Key separation, contract controls, and monitoring are part of the threat model, not just the cryptography or masking method [3][4][13][14].

V. PII DE-IDENTIFICATION TECHNIQUES (TAXONOMY AND TRADEOFFS)

Healthcare warehouses typically need two properties that can conflict: stable linkage (to track patients longitudinally) and reduced identifiability (to protect privacy). No single technique solves all scenarios, so effective solutions combine methods, aligned to risk and utility goals. Foundational work on k-anonymity formalized the idea that each record should be indistinguishable from at least k−1 others on quasi-identifiers. Later work showed that k-anonymity alone can leak sensitive attributes via homogeneity and background knowledge, motivating l-diversity and t-closeness to reduce attribute disclosure. [5][6][7]

A. Direct Identifier Removal and Masking

The first layer removes or masks direct identifiers such as names, phone numbers, addresses, and medical record numbers. In warehouse practice, this is implemented as column-level policies enforced at ingestion and transformation layers. Masking alone is insufficient when quasi-identifiers remain highly unique, but it is a necessary baseline for most releases. [1][4]

B. Pseudonymization and Tokenization (Stable Linkage with Separation)

Pseudonymization replaces identifiers with pseudonyms, while tokenization typically replaces values with non-derivable tokens, storing the mapping in a separate system. In warehouses, tokenization is often used to preserve joins across fact tables without exposing direct identifiers. The key privacy requirement is separation: re-identification keys must be isolated in a restricted PHI vault, with strict access control and auditing. Under GDPR, pseudonymization remains within scope as personal data if re-linkage is possible, so governance must reflect that. [3][4][13]

C. Generalization and Suppression (K-Anonymity Family)

Generalization (age bands, ZIP truncation) and suppression (removing outliers or rare combinations) are common ways to achieve k-anonymity-like protections. In healthcare, timestamps and location granularity are frequent drivers of uniqueness; thus date shifting and controlled geographic bucketing are common under expert determination approaches. [1][5][15]

D. Attribute Disclosure Controls (L-Diversity, T-Closeness)

For sensitive attributes like diagnosis categories, l-diversity aims to ensure variety within each equivalence class, and t-closeness ensures class distributions are close to global distributions. These are relevant when analysts can filter to small groups where a sensitive attribute becomes nearly certain. t-closeness is particularly useful in healthcare where disease prevalence is skewed and "rare disease implies identity" is a realistic risk. [6][7]

E. Perturbation And Differential Privacy (DP)

For aggregated outputs (counts, rates, averages), differential privacy provides a formal guarantee that the presence or absence of a single individual has limited effect on the result. DP is especially useful for dashboards, cohort exploration tools, and public reporting where queries can be repeated and combined. Practical DP deployments require careful privacy budget accounting and sensitivity analysis, but they directly address composition risks that harm ad hoc query systems. [8][9]

F. Synthetic Data

Synthetic data generation can reduce privacy risk when done correctly, but naive synthesis can still leak information or preserve unique patterns. A risk-based approach treats synthetic data as a release option that still requires validation against disclosure risk metrics and utility metrics, consistent with statistical disclosure control principles. [15][16][17][18]

G. Format-Preserving Encryption (FPE) and Controlled Cryptographic Transforms

FPE can preserve data formats (for example, a token that "looks like" an ID) while encrypting values, useful for systems that require format constraints. In warehouse contexts, FPE can help integrate with downstream tools while keeping raw identifiers encrypted, but key management and access controls still determine overall privacy posture. [14]

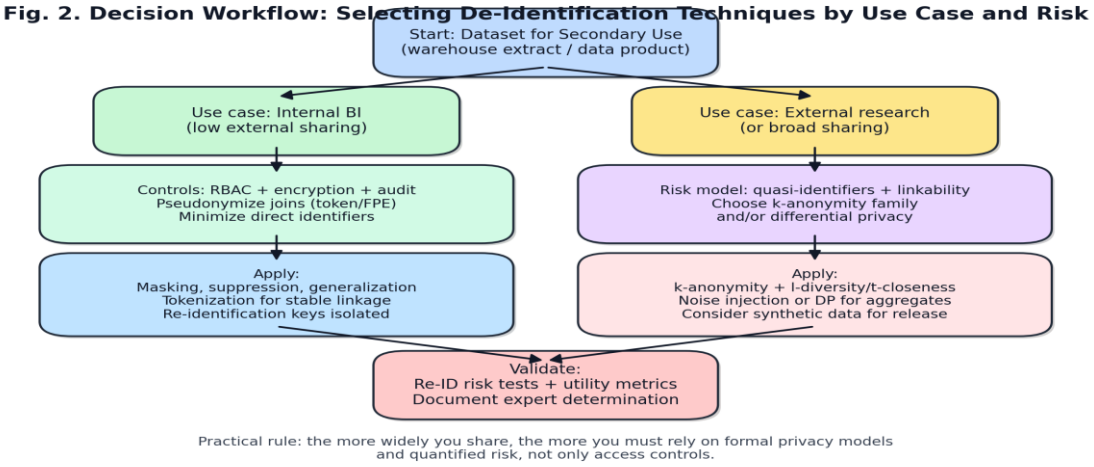

Figure 1 : Selecting De-Identification Techniques by Use Case and Risk

Table 2 : De-Identification Techniques for Healthcare Warehouses: Privacy, Utility, and Operational Fit

Technique	Primary Protection Goal	Utility Impact (Typical)	Warehouse Fit	Key Risks / Failure Modes	Reference(s)
Direct identifier removal	Remove obvious identifiers	Low to medium	Baseline for all tiers	Quasi-identifier linkage remains	[1][5]
Tokenization with key separation	Stable linkage without exposing IDs	Low	Excellent for longitudinal joins	Mapping compromise, over-broad access	[4][13]
Generalization + suppression	Reduce uniqueness (k-anonymity)	Medium	Strong for curated extracts	Poor handling of high-dimensional sparsity	[5][15]
l-diversity	Reduce attribute disclosure	Medium	Useful for diagnosis-driven slices	Can fail with skewness or similarity	[6]
t-closeness	Control distribution leakage	Medium to high	Strong for rare conditions	Utility loss if distributions heavily constrained	[7]
Differential privacy (aggregates)	Formal guarantee for query outputs	Low to medium	Great for dashboards and exploration	Budget exhaustion, mis-set sensitivity	[8][9]
Synthetic data	Reduce direct disclosure	Variable	Good for sharing and testing	Memorization, hidden leakage	[15][16][17][18]

VI. DE-IDENTIFICATION PIPELINE ARCHITECTURE FOR HEALTH DATA WAREHOUSES

A robust warehouse approach treats de-identification as a pipeline capability with zones, policies, and measurable controls. A practical pattern is a layered design: (1) ingestion and validation, (2) restricted raw zone, (3) PHI vault (re-identification keys and direct identifiers), (4) de-identification service (policy-driven transforms), and (5) an analytics-ready de-identification zone that is the default source for BI and ML. This design limits PHI exposure while preserving stable linkage where permitted. [4][13][19]

Key architectural principles include separation of duties and key separation. Token mapping tables and encryption keys must be physically and logically separated from the de-identified zone. Access should be role-based with approvals and strong logging. For "re-identification allowed" workflows (for example, care management), the PHI vault can support controlled re-linkage via audited services, rather than copying identifiers into analytic datasets. This reduces uncontrolled proliferation of PHI across marts and extracts. [1][4]

Composition risk must be addressed in repeated query systems. If the warehouse supports flexible cohort exploration, repeated "slicing" can reveal sensitive facts through differencing attacks. DP-based protections can be applied at the semantic layer for counts and rates, while row-level extracts can be gated behind expert determination, minimum cohort thresholds, and suppression policies. [8][9][10]

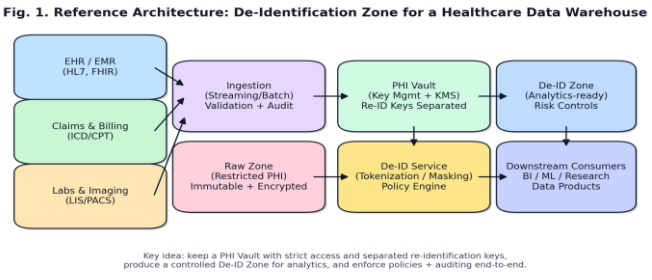

Figure 2 : De-Identification Zone for a HealthCare Data Warehouse

VII. DATA UTILITY, QUALITY, AND VALIDATION METRICS

De-identification is only successful if the resulting data remains fit for purpose. Utility validation should be use-case-specific. For BI reporting, key metrics include aggregate accuracy, trend preservation over time, and stability of dimension distributions (age bands, region buckets). For ML feature stores, metrics include feature distribution similarity, rank ordering stability, and model performance degradation relative to restricted baselines. For research extracts, cohort counts, incidence rates, and code coverage are often the primary utility targets. [15][19][20]

Privacy validation should quantify re-identification risk rather than rely on "identifiers removed" assumptions. Common metrics include k (minimum equivalence class size), uniqueness of quasi-identifier combinations, l-diversity levels for sensitive attributes, and distribution distance constraints for t-closeness. Risk testing should also include attack simulations aligned to likely auxiliary data, such as public demographics or known event timestamps. Evidence suggests that context and methodology strongly influence observed re-identification outcomes, reinforcing the need for documented, contextual validation. [5][6][7][11]

A practical validation package for each release tier includes: (1) privacy metrics and thresholds, (2) utility metrics for intended analytics, (3) documentation of transforms and parameters (date shift windows, generalization hierarchies), and (4) recipient environment controls. This directly supports expert determination style documentation and repeatability across releases. [1][13]

VIII. OPERATIONALIZATION: ACCESS CONTROL, AUDIT, AND LIFECYCLE MANAGEMENT

Operationalizing de-identification in a healthcare data warehouse requires treating privacy as a product capability with defined tiers, controls, and monitoring. In practice, de-identification fails not only due to weak transformations, but also due to operational leakage: uncontrolled exports, excessive internal permissions, unmanaged copies of extracts, and ambiguous policy ownership. Governance must define who can access which zone, under what conditions, and how compliance is measured over time [1][3][4][13].

A. Data Product Tiers and Permitted Uses

A practical model defines tiered data products, for example:

- Restricted PHI Zone: raw or minimally transformed PHI, accessed only by essential pipeline services and approved clinical operations teams.
- Pseudonymized Internal Zone: tokenized or pseudonymized data for internal analytics requiring longitudinal linkage, with strict access control and no direct identifiers.
- De-identified Research Zone: generalization and suppression applied, plus formal privacy constraints as needed, with documented release validations.
- Aggregate or Public Outputs: aggregate-only endpoints with protections such as differential privacy and strict small-cell suppression. [1][4][8][9][13]

Each tier should explicitly define allowed users, allowed query tools, export rules, and retention periods. This prevents "tier drift," where a dataset intended for restricted internal use slowly becomes shared broadly without revisiting risk assumptions [1][3][13].

B. Access Controls and Separation of Duties

Access controls should implement least privilege and separation of duties. Analysts with access to the de-identified zone should not automatically access token mappings or re-identification keys. Re-link operations, when needed, should be performed through audited services rather than by sharing mapping tables. Key management must be centralized and controlled, with restricted key access and rotation strategies [4][14]. Under GDPR reasoning, pseudonymization is a safeguard, and the key separation and access control practices are part of the privacy posture [3].

C. Auditing and Monitoring

Auditing must capture not only dataset access, but also risky behaviors such as repeated small cohort queries, excessive slicing, and frequent exports. Logs should include user, time, datasets accessed, columns touched, and export events. For interactive cohort tools, monitoring should detect query patterns consistent with differencing or reconstruction attempts. This is especially important because repeated queries can reveal sensitive information even when each single result appears safe [8][9][12].

D. Lifecycle Management for Extracts

Many privacy incidents occur after the warehouse release step. Extracts are often copied into email attachments, personal drives, unmanaged notebooks, or third-party tools. A mature program enforces retention limits, export approvals, and controlled storage locations for extracts. It should also require re-validation when: (1) new data sources are added, (2) quasi-identifiers change granularity, (3) recipient scope expands, or (4) releases become more frequent. This aligns with long-standing findings that anonymization assumptions can degrade under evolving auxiliary information and repeated releases [10][12][15].

IX. IMPLEMENTATION BLUEPRINT (REFERENCE WORKFLOW AND CONTROLS)

A practical implementation can be executed as an incremental program:

- Inventory and classify fields into direct identifiers, quasi-identifiers, sensitive attributes, and operational metadata. Map fields to de-identification requirements and tier policies. [1][4]
- Define release tiers and recipients (internal BI, internal ML, external research, public reporting) and pair each with technique bundles and validation thresholds. [1][3][15]
- Build the PHI vault and token service with hardened key management, strict access policies, and auditable re-link endpoints. [4][13][14]
- Implement policy-driven transforms in ETL/ELT, including generalization hierarchies (age bands, geo buckets), date shifting rules, and suppression logic for rare combos. [5][7][15]
- Add privacy and utility validation gates as pipeline checks, producing a release report with metrics and decisions. [1][11][13]
- Protect interactive analytics using minimum cohort thresholds and DP mechanisms for aggregate endpoints, preventing repeated-query leakage. [8][9]
- Operationalize governance with approval workflows, audit review, retention and re-validation cadence, and incident response playbooks. [1][4][12]

This blueprint aligns well with common healthcare warehousing patterns (i2b2 style cohort workflows and common data model transformations). It also supports repeatable de-identification as data products are refreshed daily or near real-time, rather than requiring manual privacy work for every extract. [19][20]

X. LIMITATIONS AND FUTURE DIRECTIONS

De-identification is not a one-size-fits-all solution, and no approach guarantees zero risk in all contexts. A core limitation is that privacy risk depends on external auxiliary information and the attacker's capabilities, both of which can evolve. Methods that appeared "safe" at release time may become less safe when new external datasets become available or when multiple extracts allow composition attacks [10][12][15]. This is particularly relevant in healthcare because public information, social data, and consumer datasets can indirectly provide linkage points.

Another limitation concerns the tradeoff between privacy and utility in high-dimensional clinical data. k-anonymity-based approaches often require suppression or heavy generalization when many quasi-identifiers exist, which can reduce utility for rare disease research, time-to-event analysis, and geographic health equity studies [5][7]. l-diversity and t-closeness improve attribute disclosure resistance but may increase information loss, especially when sensitive attributes are highly skewed, as is common in clinical datasets [6][7]. For row-level releases intended for sophisticated research, expert determination must therefore be tightly coupled with use-case-specific utility testing, not generic checks [1][13][15].

Differential privacy offers strong guarantees for aggregates, but it introduces practical constraints: privacy budget management, sensitivity estimation, and the need to design user experiences around noisy results. DP is most mature for dashboards and interactive cohort exploration outputs, but less directly applicable for detailed patient-level extracts required in certain research workflows [8][9]. Similarly, synthetic data is promising for sharing and testing, but it is not automatically safe. Poorly designed synthesis can memorize rare records or preserve unique trajectories that enable re-identification. Synthetic data must therefore be evaluated with disclosure risk testing and utility testing, consistent with statistical disclosure control principles [15][16][17][18].

Future directions likely include hybrid designs that combine strong operational controls with formal privacy protections where they are most needed. Examples include DP-protected cohort discovery layers that sit above a de-identified warehouse zone, advanced risk scoring that updates as new sources are integrated, and standardized automation for expert determination documentation so privacy validation becomes part of the data pipeline release process [1][8][9][13]. Another direction is

improving robust, domain-aware transformations for free text and unstructured fields, since narrative notes often contain hidden identifiers that can bypass traditional structured-field de-identification controls [1][13]. Over time, the most successful programs will likely be those that treat de-identification as continuous, measurable privacy engineering rather than as a one-time export step.

XI. CONCLUSION

Healthcare data warehouses create exceptional analytic value but also amplify privacy risk through multi-source linkage and longitudinal detail. Effective de-identification must therefore be engineered as a layered capability combining technique bundles, quantified validation, and strong governance. This paper presented a practical framework: a taxonomy of de-identification techniques aligned to warehouse realities, a zone-based architecture separating PHI vaults from analytics-ready de-identification zones, and operational practices for repeatable releases. By matching techniques to sharing scope and threat models, and by measuring both risk and utility, healthcare organizations can enable responsible secondary use while meeting regulatory and ethical expectations. [1][4][5][8][13]

XII. REFERENCES

[1] U.S. Department of Health and Human Services (HHS), Office for Civil Rights, "Guidance on De-identification of Protected Health Information," Nov. 26, 2012.

[2] U.S. Government Publishing Office, "45 CFR § 164.514: Standard: De-identification of protected health information (HIPAA Privacy Rule)," Electronic Code of Federal Regulations, accessed 2022.

[3] European Union, "Regulation (EU) 2016/679 (General Data Protection Regulation)," Official Journal of the European Union, 2016.

[4] ISO/IEC, "ISO/IEC 20889:2018 Privacy enhancing data de-identification terminology and classification of techniques," International Organization for Standardization, 2018.

[5] L. Sweeney, "k-anonymity: A model for protecting privacy," International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, vol. 10, no. 5, pp. 557–570, 2002.

[6] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "l-Diversity: Privacy beyond k-anonymity," ACM Transactions on Knowledge Discovery from Data, vol. 1, no. 1, 2007.

[7] N. Li, T. Li, and S. Venkatasubramanian, "t-Closeness: Privacy beyond k-anonymity and l-diversity," in Proc. IEEE International Conference on Data Engineering (ICDE), 2007.

[8] C. Dwork, "Differential privacy," in Proc. International Colloquium on Automata, Languages and Programming (ICALP), 2006.

[9] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in Proc. Theory of Cryptography Conference (TCC), 2006.

[10] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in Proc. IEEE Symposium on Security and Privacy, 2008.

[11] K. El Emam, E. Jonker, L. Arbuckle, and B. Malin, "A systematic review of re-identification attacks on health data," PLoS ONE, vol. 6, no. 12, e28071, 2011.

[12] P. Ohm, "Broken promises of privacy: Responding to the surprising failure of anonymization," UCLA Law Review, vol. 57, pp. 1701–1777, 2010.

[13] K. El Emam, Guide to the De-Identification of Personal Health Information. Boca Raton, FL, USA: Auerbach Publications, 2013.

[14] National Institute of Standards and Technology (NIST), "SP 800-38G: Recommendation for block cipher modes of operation: Methods for format-preserving encryption," 2016.

[15] B. C. M. Fung, K. Wang, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," ACM Computing Surveys, vol. 42, no. 4, 2010.

[16] L. Willenborg and T. de Waal, Elements of Statistical Disclosure Control. New York, NY, USA: Springer, 2001.

[17] D. B. Rubin, "Statistical disclosure limitation," Journal of Official Statistics, vol. 9, no. 2, pp. 461–468, 1993.

[18] T. Dalenius, "Towards a methodology for statistical disclosure control," Statistik Tidskrift, vol. 15, pp. 429–444, 1977.

[19] S. N. Murphy et al., "Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2)," Journal of the American Medical Informatics Association (JAMIA), 2010.

[20] E. A. Voss et al., "Feasibility and utility of applications of the common data model to multiple, disparate observational health databases," Journal of the American Medical Informatics Association (JAMIA), vol. 22, no. 3, pp. 553–564, 2015.