Information-Theoretic Patient Record Matching in Medical Databases: A Discriminative Power and Feature Analysis Using MIMIC-IV

Vitalijs Teze[a], Erika Nazaruka[b] and Dmirtijs Bliznuks[c]
Institute of Applied Computer Systems, Riga Technical University, 10 Zunda Embankment, Riga, Latvia

Keywords: Patient Record Matching, Information Theory, Feature Extraction, MIMIC-IV, Critical Care, Shannon Entropy, Feature Stability, Patient Identification, Healthcare Data Integration, Temporal Pattern Analysis.

Abstract: This paper presents an information-theoretic framework to evaluate feature discriminative power and stability for patient record matching. We analyse the discriminative power and temporal stability of features through Shannon entropy, evaluating their effectiveness for patient identification without unique identifiers. Our framework categorizes features into demographics/administrative ($D(F)$=12247.56 bits), ICU care patterns ($D(F)$=266.40 bits), and clinical records ($D(F)$=12.10 bits), achieving a combined discriminative power of 12526.06 bits. This significantly exceeds the theoretical minimum threshold ($\log_2(N) \approx 16$ bits) for our population of 65,366 patients. The framework employs hierarchical feature weighting based on information content and stability coefficients, revealing that temporal patterns and service transitions contain higher discriminative power than traditional demographic identifiers. We demonstrate that effective matching requires balancing feature stability against information content while maintaining computational efficiency. The framework provides a foundation for implementing reliable patient matching systems, though further validation across diverse healthcare environments is needed.

1 INTRODUCTION

Healthcare data is recognized as a cornerstone for improving patient outcomes, optimizing resource utilization, and advancing medical research. In critical care settings, the ability to accurately link patient records across fragmented datasets is essential for ensuring continuity of care and conducting robust retrospective analyses (Duggal et al., 2015; Kho et al., 2015). In a fragmented healthcare system, such as in the United States, where patients may receive care at multiple institutions, individual-level data is often scattered across disparate systems (Ong et al., 2020). Accurately linking patient records is crucial for creating a comprehensive view of a patient's medical history and enabling data analysis across a wider range of inquiries pertinent to research, clinical, and commissioning arenas (Clark et al., 2019).

However, achieving reliable record matching is a persistent challenge in the absence of unique patient identifiers (Duggal et al., 2015; Kho et al., 2015; Ong et al., 2020). Without a nationwide unique patient identifier, accurately matching multiple records for the same patient from disparate sources is challenging, particularly in large and complex datasets (Duggal et al., 2015). The reliance on basic matching methods using existing records often results in inaccurate patient identification (Fernandes & O'Connor, 2015).

Patient record matching typically relies on identifiers such as social security numbers or medical record numbers (Duggal et al., 2015; Godlove & Ball, 2015). When these are unavailable, alternative approaches must leverage features such as demographics, clinical events, and temporal patterns (Evans et al., 2016; Nie & Roantree, 2019). These approaches may include deterministic and probabilistic matching algorithms. Deterministic

[a] https://orcid.org/0009-0001-2165-4789
[b] https://orcid.org/0000-0002-1731-989X
[c] https://orcid.org/0000-0003-4252-9220

Teze, V., Nazaruka, E. and Bliznuks, D.
Information-Theoretic Patient Record Matching in Medical Databases: A Discriminative Power and Feature Analysis Using MIMIC-IV.
DOI: 10.5220/0013475400003928
In *Proceedings of the 20th International Conference on Evaluation of Novel Approaches to Software Engineering (ENASE 2025)*, pages 280-291
ISBN: 978-989-758-742-9; ISSN: 2184-4895

matching relies on exact matches of key identifiers, while probabilistic matching uses algorithms to calculate the likelihood of a match based on the similarity of data elements (Blecker et al., 2016; Godlove & Ball, 2015; McCoy et al., 2013; Riplinger et al., 2020).

Despite the availability of sophisticated methods, existing research often lacks a systematic approach to identifying the minimal set of features necessary for accurate matching, particularly in critical care contexts where data heterogeneity and missing values are prevalent. The reliance on a limited number of identifiers and inconsistencies in data quality often lead to suboptimal matching rates. Standardizing demographic data elements, such as telephone numbers, dates of birth, and addresses, can improve matching algorithm accuracy (Godlove & Ball, 2015; Riplinger et al., 2020).

This paper addresses this gap by presenting a probabilistic matching framework tailored for the Medical Information Mart for Intensive Care (MIMIC)-IV database. Our framework uses Shannon entropy to quantify feature utility, distinct from deployable matching algorithms. Our primary focus is to determine the minimal set of features required for reliable patient matching and to evaluate the performance of our approach under realistic clinical scenarios. By doing so, we aim to provide a scalable and practical solution for healthcare data integration in critical care settings. Evaluating the effectiveness of referential matching software, which augments patient data with information from external sources, and exploring big data analytics approaches like fuzzy matching algorithms and MapReduce techniques can potentially enhance matching rates and improve clinical decision-making.

This study addresses two questions: (1) Which features provide the greatest discriminative power for patient matching in MIMIC-IV? (2) How do stability and information content trade off in feature selection?

2 RELATED WORK

The primary challenge lies in accurately linking records from disparate sources that pertain to the same individual without compromising patient privacy. Inaccurate matching can lead to medical errors, compromised patient safety, billing mistakes, and flawed research outcomes (Fernandes & O'Connor, 2015; Godlove & Ball, 2015; Just et al., 2016; Riplinger et al., 2020; Zech et al., 2016).

2.1 Matching Techniques

Unique Patient Identifiers (UPIs): Some countries, such as Singapore, Canada, and Australia, have implemented national healthcare identifiers to facilitate patient matching. However, these identifiers often face limitations in cross-border information sharing and incorporating data from non-traditional sources like social care settings (Fernandes & O'Connor, 2015). In the United States, the lack of a nationwide unique patient identifier poses a significant challenge to accurately matching records (Duggal et al., 2015; Godlove & Ball, 2015).

Algorithmic Approaches: These methods utilize demographic data, such as name, date of birth, social security number, and address, to match patient records. Algorithms range in complexity from basic deterministic matching, requiring exact matches on specific identifiers, to sophisticated probabilistic matching techniques that employ statistical models and threshold limits. However, even advanced algorithms fall short of achieving a 100% match rate (Fernandes & O'Connor, 2015; Riplinger et al., 2020).

Referential Matching Software. This approach enhances algorithmic matching by utilizing third-party databases containing verified patient information. This supplementary data can help resolve ambiguities and improve match rates (Riplinger et al., 2020).

Hybrid Models. Recognizing the limitations of individual approaches, researchers have proposed hybrid models that combine different techniques. For example, combining algorithmic matching with referential matching software can potentially enhance accuracy. Other examples include combining structured and unstructured data. Big data analytics techniques like fuzzy matching algorithms and MapReduce have also been proposed for handling large datasets (Blecker et al., 2016; Duggal et al., 2015; Riplinger et al., 2020).

Privacy-Preserving Record Linkage (PRL). With growing concerns about patient privacy, researchers are actively developing techniques that enable record linkage without disclosing sensitive patient identifiers. These techniques often involve masking or encrypting identifiers before performing matching operations. Examples of such techniques include using Bloom filters and one-way hashing algorithms to protect patient privacy while enabling record linkage (Godlove & Ball, 2015; Sehili et al., 2015; Toth et al., 2014; Vatsalan et al., 2017).

Machine Learning and Deep Learning Methods: Recent advancements in machine learning have

significantly improved patient identification and matching in healthcare by leveraging both structured and unstructured data. Integrating data types such as demographics, clinical notes, and diagnostic codes allows these algorithms to capture complex patterns and nuances that traditional rule-based methods often miss. Deep learning models have shown remarkable potential in enhancing accuracy and inclusivity, especially when applied to diverse patient populations and challenging clinical scenarios. The approach might overcome limitations of rule-based systems, although it is constrained by dataset diversity (Blecker et al., 2016; Gehrmann et al., 2018; Hua et al., 2023).

2.2 Gap Analysis

Despite the breadth of existing research on patient matching — ranging from algorithmic approaches to referential matching and hybrid models — none of the works discussed above explicitly apply Information Theory (Shannon, 1948) in evaluating or optimizing patient record linkage. Scholars often focus on improving matching accuracy via algorithmic refinements (deterministic, probabilistic, referential), but do not frame the problem in terms of entropy or minimal information requirements.

In applying an information-theoretic viewpoint, we aim to systematically quantify how much information each feature (or combination of features) carries by computing their entropy, assessing joint and conditional entropies, and comparing the cumulative information gained to $\log_2(N)$ — the theoretical threshold for uniquely identifying a single patient within a population of size N. We will demonstrate how this approach can be applied to the MIMIC-IV database, leveraging its set of attributes and patient cohort to evaluate the minimal set of features required for both reliable and efficient patient matching.

3 METHODOLOGY

3.1 Data Description

We utilized the MIMIC-IV database (v3.1), which contains de-identified electronic health record data of patients admitted to the Beth Israel Deaconess Medical Center in Boston, MA. The dataset spans from 2008 to 2022 (Goldberger et al., 2000; A. Johnson et al., 2024; A. E. W. Johnson et al., 2023). Our analysis focuses on Intensive Care Unit (ICU) stays, resulting in a subset of 65366 unique patient admissions.

There are 31 tables from 2 schemas (hosp, icu) of the MIMIC-IV database that we have access to. Several auxiliary tables in MIMIC-IV contain only identifier mappings without additional attributes that could contribute to patient matching. For example, the caregiver and provider tables consist solely of unique identifiers (caregiver_id and provider_id respectively) without any supplementary information about the healthcare providers themselves. While these tables are essential for maintaining referential integrity within the database, they do not provide discriminative features for patient identity resolution and are therefore excluded from our analysis.

The MIMIC-IV database structure for patient record matching can be organized into four main categories:

1. Core Patient Data

- Patient demographics (patients table): Contains fundamental identifiers (subject_id), demographics, and mortality data.
- Hospital encounters (admissions, transfers, icustays tables): Track patient movement through hospital units using hadm_id and stay_id as linking keys.
- Clinical assignments (services table): Documents care team responsibilities independent of physical location.
- Outpatient measurements (omr table): Contains longitudinal measurements like vital signs and anthropometrics.

2. Clinical Events and Orders

- ICU documentation (chartevents): Main repository (313M+ rows) for vital signs, labs, and clinical measurements.
- Temporal events (datetimeevents, ingredientevents, inputevents, outputevents): Track time-based clinical activities and patient I/O.
- Procedures and diagnoses: Captured through multiple coding systems (ICD-9/10, CPT) in diagnoses_icd, procedures_icd, drgcodes, and hcpcsevents tables.
- Provider orders (poe, poe_detail): Comprehensive order tracking using a flexible EAV model.

3. Laboratory and Diagnostic Data

- Laboratory measurements (labevents): Specimen-linked test results with 98% hospital stay coverage.

- Microbiology cultures (microbiologyevents): Hierarchical culture results including negative findings and antibiotic sensitivities.

4. Medications and Prescriptions

- Administration records (emar, emar_detail): Barcode-verified medication delivery data post-2011.
- Pharmacy management (pharmacy, prescriptions): Detailed medication orders with standardized identifiers (GSN, NDC).

All tables are interconnected through key identifiers (subject_id, hadm_id, stay_id) and supported by reference tables (d_icd_diagnoses, d_icd_procedures, d_labitems, d_items, d_hcpcs) that provide standardized definitions and classifications.

3.2 Feature Extraction Framework

In patient record matching without unique identifiers, categorical variables play a critical role in establishing identity linkage. However, the reliability and discriminative power of these variables depends heavily on their consistency across multiple encounters. We evaluated categorical features across the MIMIC-IV database to assess their suitability for patient matching.

First, we identified potentially useful categorical variables across major database tables, focusing on features that could contribute to patient identification.
Demographic Identifiers: Gender, race, language, and marital status from patients and admissions tables; Insurance type and admission location; Admission type (including AMBULATORY OBSERVATION, DIRECT EMER., ELECTIVE, etc.); Anchor year grouping formatted as specific year ranges (e.g., 2008-2010).
Clinical Service Patterns: Care unit transitions; Clinical services (e.g., CMED, CSURG, DENT); Event types for transfers (ed, admit, transfer, discharge); Admission and transfer patterns.
Clinical Categorizations: Laboratory test priorities (routine/stat) and flags; Medication routes and types (MAIN, BASE, ADDITIVE); Order types and status (Active/Inactive); Procedure status (Paused, FinishedRunning, Stopped); Administration types from medication records; Specimen types from microbiology data.

For each categorical variable, we analyzed the following metrics:

- The number of subjects with multiple distinct values.
- The percentage of subjects showing variation across encounters.
- The total number of unique values in each category.
- The distribution of values across the population.

This analysis serves multiple purposes:

- Information Content Assessment: Variables with high consistency across encounters but good variation across the population provide strong discriminative power.
- Data Quality Evaluation: Unexpected variations in supposedly stable characteristics (e.g., multiple recorded genders) may indicate data quality issues.
- Feature Selection: Guides the selection of reliable categorical features for the matching algorithm.
- Entropy Calculation: Informs the theoretical information content available from each categorical variable.

In addition to categorical variables, temporal patterns provide important discriminative information for patient matching. We analyzed temporal features across multiple dimensions:
Admission Patterns: Time intervals between hospital admissions; Emergency department registration to admission delays; Length of stay distributions; Season and time-of-day admission patterns.
Care Transitions: Service-to-service transfer intervals; ICU transfer timing sequences; Department-to-department movement patterns.
Treatment Timelines: Laboratory test ordering patterns; Medication administration sequences (from emar); Procedure scheduling patterns.
Documentation Patterns: Time deltas between chart time and store time across various events (lab results, medications, procedures); Order-to-administration intervals.

For each temporal feature, we computed:

- Intra-patient timing consistency.
- Inter-patient timing variations.
- Cyclic pattern detection (daily, weekly, seasonal).
- Sequential pattern stability.

The temporal analysis provides:

- Pattern Recognition: Identification of characteristic temporal signatures in patient care sequences.

- Timing Fingerprints: Development of patient-specific temporal patterns that persist across encounters.
- Quality Control: Detection of temporal anomalies that might indicate record matching errors.
- Information Gain: Quantification of additional discriminative power when temporal features are combined with categorical variables.

This temporal dimension adds context to our matching framework, particularly for distinguishing between patients with similar categorical profiles but distinct care patterns.

3.3 Feature Analysis Framework

Our analysis framework employs information theory to evaluate the discriminative power and reliability of features for patient matching. This approach consists of three main components:

3.3.1 Information Content Analysis

For each feature f, we compute:

Shannon entropy (1), where $p(x)$ is the probability of value x.

$$H(f) = -\sum p(x)\log_2 p(x) \quad (1)$$

Conditional entropy (2) for feature f given subject s.

$$H(f|s) = -\sum_s p(s)\sum_x p(x|s)\log_2 p\,(x|s) \quad (2)$$

Mutual information (3) between feature and subject identity.

$$I(f;s) = H(f) - H(f|s) \quad (3)$$

Null rate penalty factor (4).

$$\alpha(f) = 1 - \frac{\text{null\_count}}{\text{total\_records}} \quad (4)$$

The effective information content $I_E(f)$ is calculated as the product of mutual information and the null rate penalty factor (5).

$$I_E(f) = I(f;s) \times \alpha(f) \quad (5)$$

3.3.2 Temporal Stability Assessment

For features that vary over time, we evaluate:

Intra-patient variance (6) computed across multiple encounters. Where N is the number of patients, n_i is the number of encounters for patient i, f_{ij} is the feature value for patient i at encounter j, $\overline{f_i}$ (7) is the mean feature value for patient i.

$$\sigma^2_{intra}(f) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{n_i - 1}\sum_{j=1}^{n_i}\left(f_{i,j} - \overline{f_i}\right)^2 \quad (6)$$

$$\overline{f_i} = \frac{1}{n_i}\sum_{j=1}^{n_i} f_{i,j} \quad (7)$$

Inter-patient variance (8) between different patients. Where N is the number of patients, $\overline{f_i}$ (7) is the mean feature value for patient i, $\bar{f}$ (9) is the grand mean across all patients.

$$\sigma^2_{inter}(f) = \frac{1}{N-1}\sum_{i=1}^{N}\left(\overline{f_i} - \bar{f}\right)^2 \quad (8)$$

$$\bar{f} = \frac{1}{N}\sum_{i=1}^{N}\overline{f_i} \quad (9)$$

Stability coefficient (10).

$$S(f) = \frac{\sigma^2_{intra}(f)}{\sigma^2_{inter}(f)} \quad (10)$$

Time-decay factor (11) modelling feature stability over temporal gaps. Where Δt is the time difference between measurements, α_f (12) is the decay constant specific to feature f, $t_{1/2}(f)$ is the half-life period for feature f, $s(f)$ (13) is the feature stability score, $I(condition)$ is the indicator function (1 if true, 0 if false), N is the number of patients, n_i is the number of encounters for patient i, $f_{i,j}$ is the feature value for patient i at encounter j.

The half-life $t_{1/2}(f)$ can be empirically determined for each feature type. For example:

- Demographics (gender, race): Very long half-life (years).
- Insurance status: Medium half-life (months).
- Clinical measurements: Short half-life (days/weeks).

$$\lambda(\mathrm{f}, \Delta\mathrm{t}) = \mathrm{e}^{-\alpha_\mathrm{f}\Delta\mathrm{t}} \times \mathrm{S(f)} \quad (11)$$

$$\alpha_f = -\frac{\ln(0.5)}{t_{1/2}(f)} \quad (12)$$

$$s(f) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{n_i-1}\frac{I(f_{i,j} = f_{i,j+1})}{n_i - 1} \quad (13)$$

3.3.3 Hierarchical Feature Weighting

Two stability thresholds are defined for feature classification:

- $\text{threshold}_1 = 0.3$: Upper bound for high stability features
- $\text{threshold}_2 = 0.7$: Upper bound for moderate stability features

Thresholds (0.3, 0.7) were set at the 25th and 75th percentiles of stability coefficients across MIMIC-IV features, balancing high-stability primaries (e.g., gender) and variable tertiaries (e.g., ICU duration).

Features are classified into tiers based on their information content and stability:

Primary Features (Tier 1, $\beta_{\text{tier}} = 1$):

- High stability. $S(f) <= \text{threshold}_1$
- Low entropy $H(f) <= 2$
- Minimal null rate ($\alpha(f) > 0.99$)

Secondary Features (Tier 2, $\beta_{\text{tier}} = 0.5$):

- Moderate stability. $\text{threshold}_1 < S(f) <= \text{threshold}_2$
- Medium entropy $2 < H(f) <= 12$
- Acceptable null rate. $\alpha(f) > 0.95$

3. Tertiary Features (Tier 3, $\beta_{\text{tier}} = 0.25$):

- Variable stability.
- High entropy.
- Used for disambiguation.

The final feature weight is computed as (14), where β_{tier} is a tier-specific base weight.

$$w(f) = \beta_{\text{tier}} \times I_E(f) \times \frac{1}{S(f)} \quad (14)$$

3.3.4 Combined Feature Space

The total discriminative power D of a feature set F is evaluated against the theoretical minimum required information content (15) for $f \in F$.

$$D(F) = \sum_{f \in F} w(f) \times I_E(f) \quad (15)$$

This must satisfy (16) where N is the total patient population size.

$$D(F) \geq \log_2(N) \quad (16)$$

4 RESULTS AND DISCUSSION

Our analysis of the MIMIC-IV database (v3.1) examined the information content and discriminative power of various features for patient matching. The study encompassed core hospital data, admission records, intensive care unit information, and outpatient measurements from the Online Medical Record (OMR) system.

The analysed features fall into six distinct categories: Demographics, Admission Patterns, Clinical Services, ICU Stays, Care Transitions, and Outpatient Measurements. Initial analysis covered 27 distinct features across these categories, with results indicating that temporal patterns and service transitions contain significantly higher information content than demographic data alone.

The information theoretic analysis showed that:

- Only Online Medical Record sequential measurements exceeded the theoretical minimum information threshold ($\log_2(N) \approx 16$ bits for our population of N=65,366 patients).
- A cluster of high-information features (service transitions, ICU stays, care transitions) provided between 12-16 bits of information, followed by a significant gap to the next feature at 6.107 bits.
- Clinical time-based features consistently showed higher discriminative power than static demographic data, with traditional identifiers providing less than 2 bits of information.

4.1 Categorical Feature Analysis

Based on the framework defined in section 3.3.1, we analyzed each feature's information content through four metrics:

- Shannon entropy ($H(f)$),
- Conditional entropy ($H(f|s)$),
- Mutual information ($I(f;s)$),
- The null rate penalty factor ($\alpha(f)$).

The effective information content $I_E(f)$ was then calculated as the product of mutual information and the null rate penalty factor.

Table 1 presents the information theoretic metrics for the categorical features analyzed, including Shannon entropy ($H(f)$), conditional entropy

Table 1: Information theoretic metrics for the main categorical features.

Category	Feature	H(f)	H(f|s)	I(f;s)	I_E(f)
Demographics	gender	0.998	0.000	0.998	0.998
Demographics	anchor_age	6.107	0.000	6.107	6.107
Demographics	anchor_year_group	2.283	0.000	2.283	2.283
Admission Patterns	race	2.361	0.100	2.262	2.262
Admission Patterns	language	0.765	0.000	0.765	0.765
Admission Patterns	marital_status	1.680	0.077	1.602	1.602
Admission Patterns	insurance	1.649	0.168	1.481	1.481
Admission Patterns	admission_type	2.644	0.959	1.685	1.685
Admission Patterns	admission_location	2.112	0.742	1.370	1.370
Admission Patterns	discharge_location	2.114	0.648	1.466	1.466
Care Transitions	careunit	3.599	2.046	1.554	1.554
Care Transitions	eventtype	1.980	1.653	0.327	0.327
Care Transitions	unit_stay_duration	15.720	3.319	12.401	12.401
Care Transitions	transfer_patterns	18.274	3.930	14.344	14.344
Clinical Services	curr_service	2.840	0.732	2.108	2.108
Clinical Services	prev_service	2.962	0.546	2.416	2.416
Clinical Services	service_transition_timing	18.440	2.514	15.926	15.926
ICU Stays	first_careunit	2.908	0.399	2.509	2.509
ICU Stays	last_careunit	2.908	0.399	2.509	2.509
ICU Stays	los	16.317	0.797	15.520	15.520
ICU Stays	icu_duration	16.317	0.797	15.520	15.520
ICU Stays	unit_transitions	2.908	0.399	2.509	2.509
ICU Stays	readmission_intervals	14.824	1.234	13.590	13.590
Online Medical Record	result_name	1.958	1.822	0.135	0.135
Online Medical Record	measurement_patterns	3.473	2.459	1.014	1.014
Online Medical Record	value_patterns	11.417	5.993	5.423	5.423
Online Medical Record	sequential_measurements	17.563	0.001	17.562	17.562

Note: H(f) = Shannon entropy, H(f|s) = Conditional entropy, I(f;s) = Mutual information, I_E(f) = Effective information content. All values in bits.

($H(f|s)$), and mutual information ($I(f;s)$). All features exhibited complete data availability (null rate penalty factor $\alpha(f) = 1.0$), indicating no missing values in the examined MIMIC-IV subset—or an oversight in detecting them—thus making effective information content ($I_E(f)$) equivalent to mutual information ($I(f;s)$).

The analysis revealed distinct tiers of feature informativeness:

High Information Content (>12 bits):

- Online Medical Record sequential measurements (17.562 bits).
- Clinical service transition timing (15.926 bits).
- ICU length of stay and duration (15.520 bits each) - notably, while 'los' was provided directly in MIMIC-IV, 'duration' was calculated as the difference between ‘intime’ and ‘outtime’; their identical information content validates data consistency.
- Care transition patterns (14.344 bits).
- Readmission intervals (13.590 bits).
- Unit stay duration (12.401 bits).

Medium Information Content (2-6 bits):

- Anchor age (6.107 bits).
- Online Medical Record value patterns (5.423 bits).
- ICU unit features (2.509 bits).
- Service assignments (2.108-2.416 bits).

Low Information Content (<2 bits):

- Traditional demographics (gender: 0.998 bits, language: 0.765 bits)
- Basic admission data (1.370-1.685 bits)
- Event types (0.327 bits)
- Result names (0.135 bits)

Table 2: Temporal stability analysis results.

Category	Feature	$\sigma^2_{intra}(f)$	$\sigma^2_{inter}(f)$	S(f)	λ(f,Δt)	s(f)
Demographics	gender	0.000	0.249	0.000	0.000	1.000
Demographics	anchor_age	1.248E-33	0.076836395	1.624E-32	1.620E-32	1.000
Demographics	anchor_year_group	0.000	2.069	0.000	0.000	1.000
Admission Patterns	race	1.225E+36	9.108E+36	0.134	0.134	0.940
Admission Patterns	language	36900.595	2.924E+36	1.262E-32	1.262E-32	1.000
Admission Patterns	marital_status	1.171E+36	1.196E+37	0.098	0.096	0.965
Admission Patterns	insurance	6.667E+36	4.354E+37	0.153	0.150	0.915
Admission Patterns	admission_type	1.379E+37	1.420E+37	0.971	0.586	0.339
Admission Patterns	admission_location	2.402E+37	2.511E+37	0.957	0.438	0.478
Admission Patterns	discharge_location	1.186E+37	1.637E+37	0.725	0.319	0.551
Care Transitions	careunit	1.607E+37	4.733E+36	3.396	2.744	0.230
Care Transitions	eventtype	1.702E+37	5.677E+36	2.999	2.297	0.255
Care Transitions	unit_stay_duration	5239.687	1237.562	4.234	3.246	0.000
Care Transitions	transfer_patterns	101250993.455	79868037.776	1.268	1.009	1.668E-05
Clinical Services	curr_service	1.195E+37	1.446E+37	0.826	0.464	0.462
Clinical Services	prev_service	1.830E+37	1.497E+37	1.222	0.848	0.151
Clinical Services	service_transition_timing	258013226.523	277993281.699	0.928	0.493	0.000
ICU Stays	first_careunit	2.161E+37	3.489E+37	0.619	0.277	0.459
ICU Stays	last_careunit	2.161E+37	3.489E+37	0.619	0.278	0.459
ICU Stays	los	30.752	21.453	1.433	0.728	0.000
ICU Stays	icu_duration	17713.190	12356.699	1.433	0.728	0.000
ICU Stays	unit_transitions	9.853E+36	1.383E+37	0.712	0.318	0.459
ICU Stays	readmission_intervals	212203693.193	239175517.099	0.887	0.564	0.000
Online Medical Record	result_name	3.037E+37	6.653E+36	4.565	3.680	0.166
Online Medical Record	measurement_patterns	12397959.817	3136807.982	3.952	2.979	0.443
Online Medical Record	value_patterns	2.848E+37	5.460E+36	5.216	4.205	0.014
Online Medical Record	sequential_measurements	9.985E+35	1.047E+35	9.541	7.188	0.957

Note: $\sigma^2_{intra}(f)$= Intra-patient Variance, $\sigma^2_{inter}(f)$ = Inter-patient Variance, S(f) = Stability Coefficient, λ(f,Δt)= Time Decay, s(f) = feature stability score.

Notably, conditional entropy values revealed that temporal and sequential features (like Online Medical Record measurements and service transitions) retained significant information content even after accounting for patient identity, indicating their value for disambiguation.

4.2 Temporal Feature Analysis

Analysis of temporal stability metrics across feature categories revealed distinct patterns in feature reliability and degradation over time. Table 2 presents the complete temporal analysis results.

Key findings include:

Demographics and Persistent Features (Stability Score ($s(f)$)> 0.90):

- Basic demographic features showed perfect stability (gender: 1.0, language: 1.0)
- Race and marital status maintained high stability (0.94 and 0.96 respectively)
- Insurance information showed good stability (0.92)

Clinical Service Data ($s(f)$ 0.40-0.50):

- First/last careunit: 0.46 stability
- Current service: 0.46 stability
- Admission location: 0.48 stability
- Service transitions showed moderate stability with temporal variability

Dynamic Care Features ($s(f)$ 0.20-0.35):
- Care unit events: 0.23 stability
- Event type patterns: 0.26 stability
- Previous service: 0.15 stability
- Result names and value patterns showed lower stability

Highly Variable Features ($s(f) < 0.001$):
- Length of stay
- Unit stay duration
- Transfer patterns: 0.00002 stability
- Readmission intervals: very low stability

The temporal analysis revealed several key insights:
- Demographic and administrative features maintain high stability across encounters but offer limited discriminative power
- Service-based features provide moderate stability with better discriminative potential
- Care transition patterns, while less stable, contain rich information content for temporally proximate encounters
- Dynamic clinical features show low stability but high information content, suggesting their utility for short-term matching

4.3 Hierarchical Feature Weighting

Our hierarchical feature weighting analysis revealed distinctive patterns in the relationships between information content, stability, and overall feature utility for patient matching. Using the three-tier classification system, features were weighted according to their stability scores and information content.

Primary Features (βtier = 1.0) demonstrated high stability but varied in information content. Four features qualified for this tier:
- Gender (weight: 997.97, $I_E(f)$: 1.00 bits)
- Language (weight: 764.93, $I_E(f)$: 0.76 bits)
- Marital status (weight: 16.36, $I_E(f)$: 1.60 bits)
- Insurance (weight: 9.67, $I_E(f)$: 1.48 bits)

Secondary Features (βtier = 0.5) comprised care unit identifiers. These features showed moderate stability with information content around 2.5 bits each:
- First careunit
- Last careunit

Tertiary Features (βtier = 0.25) formed the largest group with 21 features, including notably:
- Anchor age (highest overall weight: 1526.67, $I_E(f)$: 6.11 bits)
- Service transition timing (weight: 4.29, $I_E(f)$: 15.93 bits)
- Transfer patterns (weight: 2.83, $I_E(f)$: 14.34 bits)
- Readmission intervals (weight: 3.83, $I_E(f)$: 13.59 bits)

The final weights revealed several important insights:
- **Stability-Information Trade-off**: While some features like service transition timing and transfer patterns contained high information content (>14 bits), their lower stability scores resulted in reduced final weights. Conversely, demographically stable features like gender and language achieved higher weights despite lower information content due to their high stability.
- **Anchor Age Anomaly**: Despite being classified as a tertiary feature, anchor age achieved the highest overall weight (1526.67) due to its unique combination of moderate information content and computational stability characteristics.
- **Care Unit Features**: First and last careunit assignments maintained moderate weights through balanced stability and information content, positioning them as reliable secondary matching criteria.
- **Dynamic Features**: Highly variable features like length of stay and unit stay duration, despite high information content, received lower weights due to their temporal instability, aligning with their expected variability across patient encounters.

This weighting scheme effectively balanced the trade-off between feature stability and information content, prioritizing features that maintain consistent discriminative power across patient encounters while appropriately discounting unstable or low-information features.

4.4 Combined Feature Space

Analysis of the feature groups reveals distinct patterns in discriminative power and information content. The three major feature groups demonstrate varying levels of effectiveness for patient matching:

Demographics and Admission Features ($D(F)$ = 12247.56 bits) that combines the following features: 'Anchor Age', 'Ggender', 'Language',

'Anchor Year Group', 'Marital Status', and 'Insurance':

- Provides the highest discriminative power despite relatively low information content (13.24 bits)
- Dominated by anchor_age and gender contributions
- Achieves 96% of total discriminative power across all features
- High stability characteristics enable reliable long-term matching

ICU Care Patterns ($D(F)$ = 266.40 bits) combining 'Los', 'ICU Duration', 'Last Careunit', 'First Careunit', 'Readmission Intervals', 'Transfer Patterns', 'Service Transition Timing', 'Unit Transitions', 'Unit Stay Duration':

- Highest information content (94.83 bits)
- Moderate discriminative power driven by temporal patterns
- Service transitions and readmission intervals provide key disambiguation capabilities
- Effective for matching temporally proximate encounters

Clinical Records ($D(F)$ = 12.10 bits) consisting of 'Current Service', 'Previous Service', 'Sequential Measurements', 'Value Patterns', 'Measurement Patterns', 'Result Name':

- Moderate information content (28.66 bits)
- Limited discriminative power
- Most suitable for secondary verification
- Value in combination with other feature groups for edge cases

The total discriminative power across all groups (12526.06 bits) substantially exceeds the theoretical minimum threshold of $\log_2(N) \approx 16$ bits for our population, with demographic features providing the primary matching power and ICU patterns offering important secondary discrimination.

Comparison with Existing Approaches

4.5 Comparison with Existing Approaches

Unlike probabilistic methods, which weight features empirically (e.g., Ong et al., 2020), our framework quantifies information content using entropy, highlighting temporal features' dominance (e.g., 15.52 bits for ICU duration vs. 0.998 bits for gender). Ong et al.'s hybrid approach, combining deterministic and probabilistic record linkage, outperformed independent methods by identifying 18%-24% more correct pairs in congenital heart disease surveillance, leveraging PII like names and dates. However, their reliance on exact matches contrasts with our entropy-based feature stability analysis, which prioritizes temporal patterns over static identifiers. While Ong et al. addressed data quality via harmonization, our approach assumes complete data (e.g., $\alpha(f) = 1.0$), necessitating future validation for missingness. Direct accuracy comparisons remain future work, but Ong's findings underscore the potential of hybrid strategies, suggesting our framework could integrate temporal stability thresholds to enhance matching robustness across diverse datasets.

4.6 Limitations and Future Work

Our current approach to temporal anchoring of demographic features through admission times, while functional, introduces potential selection bias by excluding patients without admission records. Future work should explore alternative temporal reference points. It should also develop methods to incorporate outpatient encounters for more comprehensive patient matching.

A notable limitation is that our feature combinations often provide discriminative power significantly exceeding the theoretical minimum of $\log_2(N)$ bits - in some cases by several orders of magnitude. While this redundancy provides robustness against missing data, it may indicate computational inefficiency and potential overfitting to institutional patterns. Future research should investigate optimal feature selection methods that balance discriminative power with computational efficiency while maintaining matching accuracy.

The assumption of demographic stability across encounters needs careful examination, particularly for long-term longitudinal studies where characteristics like insurance status, marital status, and language preferences may evolve. Research into dynamic feature weighting that adapts to temporal distance could enhance matching accuracy.

Missing value patterns significantly impact join quality between administrative and clinical tables, potentially skewing stability metrics. Development of robust imputation methods specifically designed for temporal healthcare data could address this limitation and improve feature extraction reliability.

Our stability thresholds, while empirically derived from MIMIC-IV data, require validation across diverse healthcare settings and populations. Multi-institutional studies would help establish generalizable parameters for feature classification and weighting.

Privacy considerations currently limit cross-institutional validation of our matching approach. Future work should incorporate privacy-preserving computation methods that enable collaborative model validation without compromising patient confidentiality.

The current implementation overlooks potential feature interactions by treating each characteristic independently. Development of composite features that capture relationships between administrative, clinical, and temporal patterns could enhance discriminative power.

Real-time feature extraction and matching present computational challenges not addressed in our retrospective analysis. Research into efficient algorithms and optimization techniques would facilitate clinical deployment of our information-theoretic approach.

Specialty-specific matching requirements and varying documentation practices across clinical domains warrant investigation. Adaptive frameworks that account for department-specific feature stability and information content could improve matching accuracy in specialized care settings.

Future work could explore hybridizing our entropy-based framework with deterministic methods, as Ong et al. suggest, to address missing data and validate across diverse healthcare systems.

5 CONCLUSIONS

The information-theoretic analysis successfully established a framework for patient matching in critical care settings, revealing three complementary feature groups: demographics/administrative ($D(F)$ =12247.56 bits), ICU care patterns ($D(F)$ =266.40 bits), and clinical records ($D(F)$ =12.10 bits). While the combined discriminative power (12526.06 bits) substantially exceeds the theoretical minimum threshold ($\log_2(N) \approx 16$ bits), this significant redundancy presents both advantages and challenges.

The excess discriminative power provides robustness against missing data and institutional variability. However, it suggests potential computational inefficiencies and possible overfitting to institution-specific patterns. Future implementations should focus on optimizing feature selection to maintain matching accuracy while reducing computational overhead.

The research demonstrates that effective patient matching requires balancing:

- Feature stability vs. information content
- Computational efficiency vs. redundancy
- Institutional generalizability vs. local pattern optimization

This framework provides a foundation for implementing reliable patient matching systems, though further validation across diverse healthcare environments and optimization of feature selection methods is needed.

REFERENCES

Blecker, S., Katz, S. D., Horwitz, L. I., Kuperman, G., Park, H., Gold, A., & Sontag, D. (2016). Comparison of Approaches for Heart Failure Case Identification From Electronic Health Record Data. JAMA Cardiology, 1(9), 1014. https://doi.org/10.1001/jamacardio.2016.3236

Clark, S. J., Halter, M., Porter, A., Smith, H. C., Brand, M., Fothergill, R., Lindridge, J., McTigue, M., & Snooks, H. (2019). Using deterministic record linkage to link ambulance and emergency department data: Is it possible without patient identifiers?: A case study from the UK. International Journal of Population Data Science, 4(1). https://doi.org/10.23889/ijpds.v4i1.1104

Duggal, R., Khatri, S. K., & Shukla, B. (2015). Improving patient matching: Single patient view for Clinical Decision Support using Big Data analytics. 2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), 1–6. https://doi.org/10.1109/ICRITO.2015.7359269

Evans, R. S., Benuzillo, J., Horne, B. D., Lloyd, J. F., Bradshaw, A., Budge, D., Rasmusson, K. D., Roberts, C., Buckway, J., Geer, N., Garrett, T., & Lappé, D. L. (2016). Automated identification and predictive tools to help identify high-risk heart failure patients: Pilot evaluation. Journal of the American Medical Informatics Association, 23(5), 872–878. https://doi.org/10.1093/jamia/ocv197

Fernandes, L., & O'Connor, M. (2015). Accurate Patient Identification—A Global Challenge. Perspectives in Health Information Management, International, 1–7.

Gehrmann, S., Dernoncourt, F., Li, Y., Carlson, E. T., Wu, J. T., Welt, J., Foote, J., Moseley, E. T., Grant, D. W., Tyler, P. D., & Celi, L. A. (2018). Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. PLOS ONE, 13(2), e0192360. https://doi.org/10.1371/journal.pone.0192360

Godlove, T., & Ball, A. W. (2015). Patient Matching within a Health Information Exchange. https://pmc.ncbi.nlm.nih.gov/articles/PMC4696093/

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet:

Components of a new research resource for complex physiologic signals. Circulation, 101(23), e215–e220.

Hua, Y., Wang, L., Nguyen, V., Rieu-Werden, M., McDowell, A., Bates, D. W., Foer, D., & Zhou, L. (2023). A Deep Learning Approach for Transgender and Gender Diverse Patient Identification in Electronic Health Records. Journal of Biomedical Informatics, 147, 104507.

Johnson, A., Bulgarelli, L., Pollard, T., Gow, B., Moody, B., Horng, S., Celi, L. A., & Mark, R. (2024). MIMIC-IV (Version 3.1) [Dataset]. PhysioNet. https://doi.org/10.13026/KPB9-MT58

Johnson, A. E. W., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., Lehman, L. H., Celi, L. A., & Mark, R. G. (2023). MIMIC-IV, a freely accessible electronic health record dataset. Scientific Data, 10(1), 1. https://doi.org/10.1038/s41597-022-01899-x

Just, B. H., Marc, D., Munns, M., & Sandefer, R. (2016). Why patient matching is a challenge: Research on master patient index (MPI) data discrepancies in key identifying fields. Perspectives in Health Information Management, 13(Spring).

Kho, A. N., Cashy, J. P., Jackson, K. L., Pah, A. R., Goel, S., Boehnke, J., Humphries, J. E., Kominers, S. D., Hota, B. N., Sims, S. A., Malin, B. A., French, D. D., Walunas, T. L., Meltzer, D. O., Kaleba, E. O., Jones, R. C., & Galanter, W. L. (2015). Design and implementation of a privacy preserving electronic health record linkage tool in Chicago. Journal of the American Medical Informatics Association, 22(5), 1072–1080. https://doi.org/10.1093/jamia/ocv038

McCoy, A. B., Wright, A., Kahn, M. G., Shapiro, J. S., Bernstam, E. V., & Sittig, D. F. (2013). Matching identifiers in electronic health records: Implications for duplicate records and patient safety. BMJ Quality & Safety, 22(3), 219–224. https://doi.org/10.1136/bmjqs-2012-001419

Nie, D., & Roantree, M. (2019). Detecting Multi-Relationship Links in Sparse Datasets: Proceedings of the 21st International Conference on Enterprise Information Systems, 149–157. https://doi.org/10.5220/0007696901490157

Ong, T. C., Duca, L. M., Kahn, M. G., & Crume, T. L. (2020). A hybrid approach to record linkage using a combination of deterministic and probabilistic methodology. Journal of the American Medical Informatics Association, 27(4), 505–513. https://doi.org/10.1093/jamia/ocz232

Riplinger, L., Piera-Jiménez, J., & Dooling, J. P. (2020). Patient Identification Techniques – Approaches, Implications, and Findings. Yearbook of Medical Informatics, 29(01), 081–086. https://doi.org/10.1055/s-0040-1701984

Sehili, Z., Kolb, L., Borgs, C., Schnell, R., & Rahm, E. (2015). Privacy preserving record linkage with PPJoin.

Shannon, C. E. (1948). A mathematical theory of communication. The Bell System Technical Journal, 27(4), 623–656. https://doi.org/10.1002/j.1538-7305.1948.tb00917.x

Toth, C., Durham, E., Kantarcioglu, M., Xue, Y., & Malin, B. (2014). SOEMPI: a secure open enterprise master patient index software toolkit for private record linkage. AMIA Annual Symposium Proceedings, 2014, 1105.

Vatsalan, D., Christen, P., & Rahm, E. (2017). Scalable Multi-Database Privacy-Preserving Record Linkage using Counting Bloom Filters (No. arXiv:1701.01232). arXiv. https://doi.org/10.48550/arXiv.1701.01232

Zech, J., Husk, G., Moore, T., & Shapiro, J. (2016). Measuring the Degree of Unmatched Patient Records in a Health Information Exchange Using Exact Matching. Applied Clinical Informatics, 07(02), 330–340. https://doi.org/10.4338/ACI-2015-11-RA-0158