Adaptation and Validation of a Bidirectional Matching Algorithm to Enrich Hospital Data with Information from the French National Death Registry: A Case Study in Patients with Lung or Prostate Cancer (MO-REAL)

RWD167

AUTHORS

Stéphane Sanchez[1]
stephane.sanchez@hcs-sante.fr

Elodie Lehmann[2]
elodie.lehmann@alirahealth.com

Hélène Corneloup Savignol[3]
hcorneloup@ch-tarbes-lourdes.fr

Awa Baradji[4]
awa.baradji@ch-mdm.fr

Jean-François Ricci[5]
jf.ricci@alirahealth.com

Robin Nogues[6]
robin.nogues@sancare.fr

Charlotte Waegeneire[6]
charlotte.waegeneire@sancare.fr

Barbara Lebas[6]
barbara.lebas@sancare.fr

[1]CH Troyes, France, [2]Alira Health, Paris, France, [3]CH Tarbes, France, [4]CH Mont-de-Marsan, France, [5]Alira Health, Basel, Switzerland, [6]Sancare, Paris, France

RATIONALE AND OBJECTIVES

In-hospital Electronic Medical Records (EMRs) provide rich, detailed digitized form of a patient's chart data, but their usefulness to assess survival outcomes is limited by the inability to capture out-of-hospital death events, which undermines accurate mortality assessments. This project aims at exploring a new approach to overcome this limitation by validating the performance of the MatchID library (1) in a hospital setting to automatically link EMR data to the National French Death Registry (NFDR), enabling a more complete and accurate identification of patients who died outside of the hospital setting in France.

METHODOLOGY

Data sources: Two data sources were leveraged across three hospitals: EMRs from each hospital and the NFDR, maintained by the French National Institute of Statistics and Economic Studies (INSEE) (2). The NFDR provides data from the national register of deceased French citizens, with monthly updates on deaths occurring both in France and abroad since 1970.

Study period: January 2022 to December 2023.

Study population: Eligible adult patients were identified by searching hospital-native EMRs for all persons hospitalized at least once between 01 January 2022 and 31 December 2023, with a diagnosis of lung and/or prostate cancer (ICD-10-CM C34.X or C61.X) and for whom death was recorded in the hospital EMR, either as a discharge code "9" in the PMSI data or identified through text string searches.

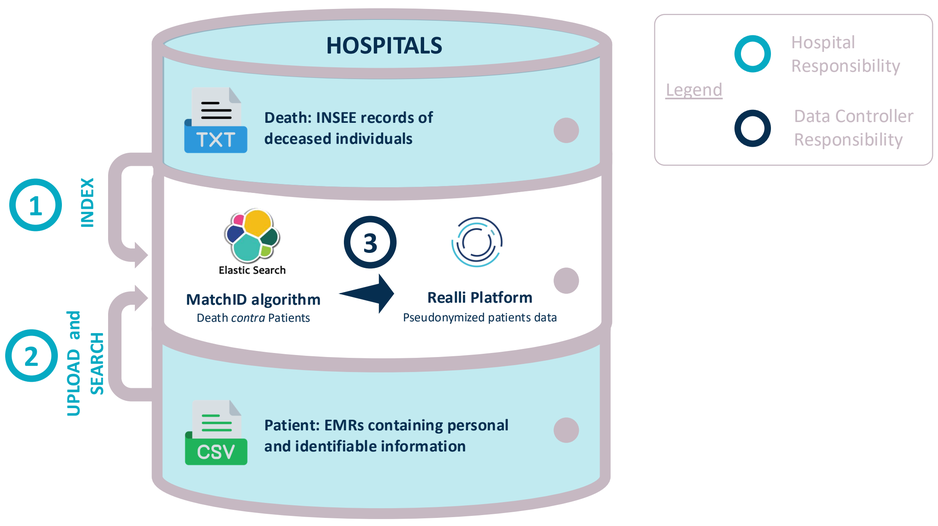

Figure 1: Description of the matching Process within Hospital Systems

Algorithm: Hospital EMRs were linked to the NFDR using the MatchID library (https://deces.MatchID.io/), via a three-step process:

1. Categorizing and organizing death-related data
2. Identifying patients' records from the CSV using the index
3. Scoring

The scoring system evaluates each identity component separately such as name, first name and birthdate. Text fields underwent normalization and tokenization before being compared between EMRs and NFDR using the Levenshtein distance, with penalties applied for discrepancies. Scores were multiplied, and a power coefficient was assigned based on the number of matching parameters. Fewer queried fields resulted in greater penalties for discrepancies. If certain data fields were missing, a smaller penalty (ranging from 50% to 100%, depending on the field) was applied.

Matching results were accessed through the "Realli solution," a machine learning-powered platform, providing near real-time in-depth extraction and analyses of full content structured and unstructured hospital EMRs.

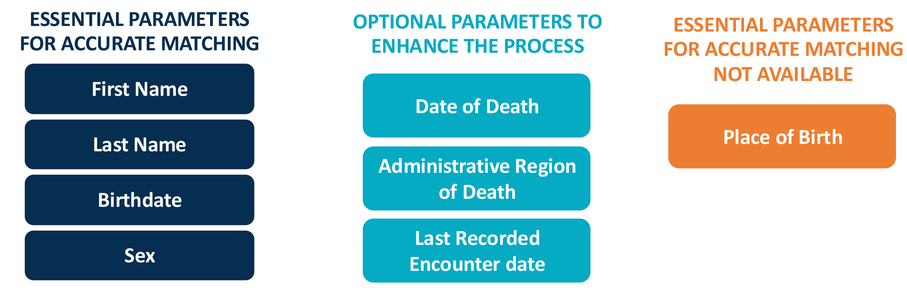

Figure 2: Matching Parameters

RESULTS

> Patients Classification

Between January 2022 and December 2023, among the **557 patients who met the inclusion criteria**, 140 (25.1%) were diagnosed with prostate cancer, 412 (74.0%) with lung cancer, and 5 (0.9%) with both cancers. Overall, 69.5% of the patients were at a metastatic stage, with more than half (57%) with metastatic lung cancer at the time of death.

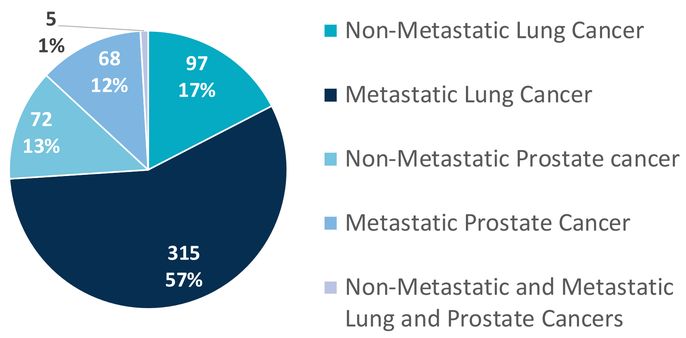

Figure 3: Distribution of Patients by Cancer Type and Metastatic Status

> Patients Demographics

Among the 557 patients, the mean age was higher in the non-metastatic group for lung cancer, whereas the metastatic group for prostate cancer had a higher mean age. In the lung cancer group, 32.9% were active smokers and 27.8% had quit smoking.

	Overall	Patients with Prostate Cancer		Patients with Lung Cancer	
		Non-metastatic	Metastatic	Non-metastatic	Metastatic
Number of patients	557	72	68	97	315
Mean age (SD)	72.75 (11.34)	85.36 (7.17)	77.67 (8.93)	72.46 (11.37)	68.81 (10.08)
Female (%)	123 (22,1%)	0%	0%	24 (4.3%)	99 (17.8%)
Active smoker (%)	209 (37,5%)	24 (4.3%)		183 (32,9%)	
Smoking cessation (%)	181 (32.5%)	24 (4.3%)		155 (27.8%)	

Note: Socio-demographic characteristics are not presented for the category of patients with both prostate and lung cancer due to the low number of patients in this group.

Table 1 – Patients Socio-demographic Characteristics

> Matching Performance Between Hospital EMRs and INSEE Database

Using all available personal information, nearly 100% (**554 out of 557; 99.64%**) of patients with a death record in the hospital system were identified in the INSEE database, with matching rates ranging from **94.4% for patients with prostate cancer** to **100% for patients with lung cancer**. Differences by cancer type were minimal (98.57% for prostate cancer; 99.76% for lung cancer), and the mapping algorithm showed **consistent performance across the three hospitals**, with matching rates between 98.18% and 100%.

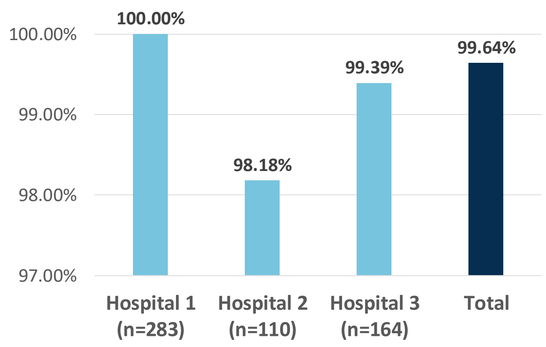

Figure 4: % of Deceased Patients in the Hospital Retrieved in the NFDR, by hospital

A **manual quality control** was conducted on a sample of 20% of patients to ensure the accuracy of the matching process. **All reviewed matches between hospital EMRs and the INSEE database were confirmed**. The algorithm demonstrated a **recall score of 100%** and **a precision of 100%,** resulting in an **F1 score of 1.**

> Sensitivity Analysis - Matching Performance when Removing NFDR Optional Parameters

The sensitivity analysis aimed at testing the algorithm's matching performance by creating a scenario in the hospital setting where there is no recorded death status in the EMR for a given patient. Hence, in our sample, we ignored available information on date and place of death to evaluate the sub-algorithm's performance. Using only partial information, the sub-algorithm's performance still closely approximated that of the complete algorithm, achieving a **matching rate of 99.46% (553 out of 557).**

A manual quality control review was conducted on a 20% sample of patients to verify the sub-algorithm's reliability. **All matches were confirmed as accurate**, with scores aligned to those from the complete algorithm's matching process.

In a confirmatory analysis of all deaths recorded in the three study hospitals in 2022 and 2023, regardless of the patient's diagnosis, 98% of death events (5,183 out of 5,289) were successfully matched to the NFDR, using either full or partial information from the NFDR.

CONCLUSION

This project validates the matching algorithm between hospital records and the mortality registry in France, showing high consistency across hospitals and patient diagnoses. By addressing the underreporting of hospital mortality in long-term studies, the project confirms the accuracy of the public mortality database and for long-term cohort studies in diseases/interventions with low inpatient mortality.

DOWNLOAD THE DIGITAL VERSION

Scan this QR code to download a copy of the poster.

Sources: *(1)* https://deces.matchid.io/about#algorithme-de-rapprochement-d-identites
(2) https://www.insee.fr/en/outil-interactif/5543645/conditions