

Natural language processing solution (REALLI) to evaluate the limitations of the French national health insurance database (SNDS): example of HER2⁺ metastatic breast cancer

AUTHORS

A. Groenez¹, B. Lebas², G. Rejasse², C.Roux², P.A. Squara¹, M Jouve²

¹Pfizer, Paris, France

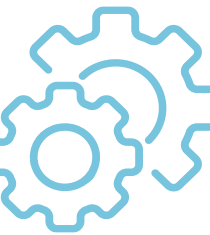
²Sancare, Paris, France



INTRODUCTION AND OBJECTIVES

The French National Health Insurance Database (SNDS) is one of the largest claims databases in Europe, covering the entire French population and patient care pathways. Despite its strengths, evidence generation using SNDS is limited by the lack of detailed ICD-10 coding and potential delays in coding within the hospital claims system (PMSI).

This study aims to evaluate the limitations of SNDS for identifying specific subpopulations, such as HER2+ metastatic breast cancer (BC) patients, by leveraging a natural language processing (NLP) solution (REALLI) that integrates both structured and unstructured data from electronic health records (EHRs).



METHODOLOGY

Data Source : Hospital claims data (PMSI) and Electronic Health Records (HER) data from 3 general hospitals and 1 university hospital were extracted.

Hospital claims data (PMSI)		EHR
<div><div>Discharge summaries</div><ul style="list-style-type: none">•Hospital diagnosis (ICD-10)•Stays & patient pathway•Medical procedures performed•Treatments on top of DRG•DRG</div>	<div><div>Administrative data</div><ul style="list-style-type: none">•Age•Gender•Date of birth•Date of stay•Area of hospitalization</div>	<div><div>All medical content</div><ul style="list-style-type: none">•Medical and nursing reports, Clinician and Multidisciplinary consultation meeting notes (text)•Lab test results (Structured or unstructured)•Drug prescription (structured or unstructured)</div>

Period : Between January 1st, 2017, and December 31st, 2021.

Population : Adult women hospitalized for breast Cancer were identified through ICD10 diagnosis code

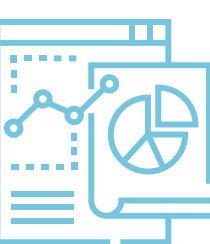
Algorithms Compared (Table 1) :

- PMSI Algorithm: Relies exclusively on structured PMSI data (ICD-10 codes, ATC/UCD codes).
- REALLI Algorithm: Integrates PMSI data and unstructured data extracted from EHRs using NLP

Table 1: Comparison of information retrieved using PMSI vs REALLI

	PMSI algorithm	REALLI algorithm
Hospital's PMSI coding (ICD 10 code)	✓	✓
Administrative data	✓	✓
On top of DRG medication file (ATC and UCD codes)	✓	✓
Textual elements (unstructured data from medical reports and notes)	✗	✓
Lab test results	✗	✓
Content of the in-hospital administration/prescription	✗	✓

Analysis: Comparison of metastatic and HER2+ metastatic BC patient counts between REALLI and PMSI algorithms. Assessment of coding delays for metastatic diagnosis dates between sources.



RESULTS

> METASTATIC AND HER2+ STATUS

Between January 2017 and December 2021, 3,214 BC patients were included.

Table 2 : Repartition of the number of BC patients identified using PMSI-VS REALLI algorithms according to metastatic and HER2⁺ status

	PMSI algorithm	REALLI algorithm
Metastatic status	913	1045
HER2 ⁺ status	48	67

The metastatic status of 1045 and 913 patients was identified by the REALLI and PMSI algorithms, respectively. Thirteen percent of metastatic BC patients were not identified in the PMSI (**Table 2**).

Among the 749 metastatic BC patients who were not managed in a university hospital, HER2⁺ status of 67 and 48 patients was identified by the REALLI and PMSI algorithms, respectively (**Table 2**). Twenty-eight percent of HER2⁺ metastatic BC patients were not identified in the PMSI.

> CODING TIMELINE OF METASTASES

Among the 671 patients with a metastatic diagnosis date in both REALLI and PMSI, 82 (12%) had a coding delay, including 31 (5%) with a delay of more than 12 months (**Figure 1**).

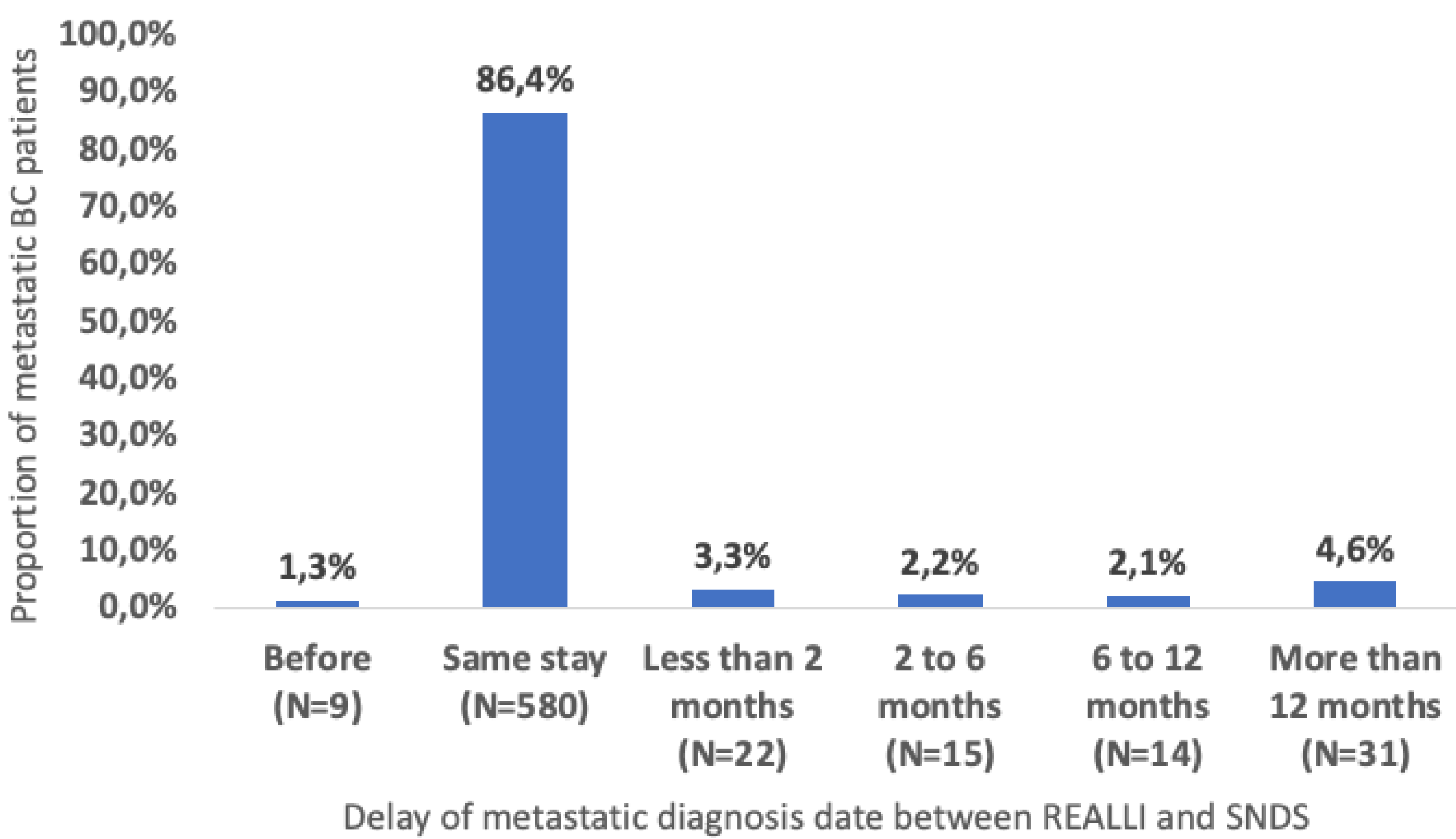


Figure 1 : Repartition of the proportion of metastatic BC patients with a delay in metastatic diagnosis date between REALLI and PMSI



CONCLUSION

- > Leveraging NLP to analyze EHRs significantly improves the identification of HER2+ metastatic breast cancer patients in the SNDS database, highlighting underrepresentation and coding delays in PMSI.
- > Further studies involving various medical centers are needed to validate the results, considering variability in coding practices.



DOWNLOAD THE
DIGITAL VERSION

Scan this QR code to download a copy of the poster. Please note that reproducing copies of this poster and additional content via the Quick Response (QR) code is prohibited without permission from both ISPOR and the authors.

