Development of a Natural Language Processing Model for deriving breast cancer quality indicators : A cross-sectional, multicenter study

Etienne Guével, Sonia Priou, Rémi Flicoteaux, Guillaume Lamé, Romain Bey, Xavier Tannier, Ariel Cohen, Gilles Chatellier, Christel Daniel, Christophe Tournigand, et al.

Guével E, Priou S, Flicoteaux R, Lamé G, Bey R, Tannier X, et al. Development of a Natural Language Processing Model for deriving breast cancer quality indicators : A cross-sectional, multicenter study. Revue d'Épidémiologie et de Santé Publique. 2023;71(6):102189. https://doi.org/10.1016/j.respe.2023.102189

Développement d'un modèle de traitement automatique du langage pour calculer des indicateurs qualité du cancer du sein : une étude transversale multicentrique

Development of a Natural Language Processing Model for Deriving Breast Cancer Quality Indicators: A cross-sectional, Multicenter Study

Etienne Guével (1), Sonia Priou (2), Rémi Flicoteaux (3), Guillaume Lamé (2), Romain Bey (1), Xavier Tannier (4), Ariel Cohen (1), Gilles Chatellier (5), Christel Daniel (1), Christophe Tournigand (6), Emmanuelle Kempf (4,6), pour le groupe Cancer AP-HP, une initiative du CRAB*

1. Assistance Publique – Hôpitaux de Paris, Innovation and Data, IT Department, 75012 Paris, France
2. Université Paris-Saclay, CentraleSupélec, Laboratoire Génie Industriel, 91192 Gif-sur-Yvette, France
3. Assistance Publique – Hôpitaux de Paris, Department of medical information, 75012 Paris, France
4. Université Sorbonne Paris Nord, Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances pour la e-Santé, LIMICS, 75006 Paris, France
5. Université Paris CIté, Department of medical informatics, Assistance Publique Hôpitaux de Paris, Centre-Université de Paris (APHP-CUP), 75015 Paris, France
6. Université Paris Est Créteil, Assistance Publique – Hôpitaux de Paris, Department of medical oncology, Henri Mondor and Albert Chenevier University Hospital, 94000 Créteil, France

* CRAB: Cancer Research Application on Big Data

Auteur de correspondance :
Emmanuelle Kempf
Department d'oncologie médicale,
Groupe hospitalo-universitaire Henri Mondor and Albert Chenevier,
Assistance Publique – Hôpitaux de Paris,
1 rue Gustave Eiffel
94000 Créteil, France
Emmanuelle.kempf@aphp.fr
P +1 4981 4531 / F +1 4981 2576

Contribution des auteurs :

- Conception : EK, EG, SP, RF, GL, XT, GC
- Recueil des données : EG, SP, AC
- Analyse des données : EK, EG, SP, RF, GL, AC, GC
- Interprétation des résultats : EK, SP, RF, GL, AC, GC, CD
- Draft : EG, SP, GL
- Relecture : EK, EG, SP, RF, GL, RB, XT, AC, GC, CD
- Validation de la version finale à publier : EK, EG, SP, RF, GL, RB, XT, AC, GC, CD, CT
- Assume la responsabilité : EK, EG, GC

Conflits d'intérêt :
Les auteurs n'ont aucun conflit d'intérêt à déclarer.

Guével E, Priou S, Flicoteaux R, Lamé G, Bey R, Tannier X, et al. Development of a Natural Language Processing Model for deriving breast cancer quality indicators : A cross-sectional, multicenter study. Revue d'Épidémiologie et de Santé Publique. 2023;71(6):102189. https://doi.org/10.1016/j.respe.2023.102189

Guével E, Priou S, Flicoteaux R, Lamé G, Bey R, Tannier X, et al. Development of a Natural Language Processing Model for deriving breast cancer quality indicators : A cross-sectional, multicenter study. Revue d'Épidémiologie et de Santé Publique. 2023;71(6):102189. https://doi.org/10.1016/j.respe.2023.102189

RESUME

Objectifs

Les données médico-administratives ne suffisent pas à automatiser le calcul des indicateurs de qualité et de sécurité des soins (IQSS). L'objectif de notre étude de faisabilité est d'analyser 1/ la disponibilité des sources de données ; 2/ la disponibilité de chaque variable élémentaire par indicateur, et 3/ d'appliquer des algorithmes de traitement du langage naturel pour extraire automatiquement ces informations.

Méthode

Nous avons réalisé une étude de faisabilité observationnelle transversale multicentrique sur l'entrepôt de données cliniques de l'Assistance Publique – Hôpitaux de Paris (AP-HP). Nous avons étudié la prise en charge des patients atteints de cancer du sein traités à l'AP-HP entre janvier 2016 et juin 2021, et les indicateurs publiés par l'European Society of Breast Cancer Specialist, à partir des données administratives du Programme de Médicalisation du Système d'Information (PMSI) et des comptes-rendus d'anatomopathologie. Pour chaque indicateur, nous avons calculé le nombre (%) de patients pour lesquels toutes les sources de données nécessaires étaient disponibles, et le nombre (%) de patients pour lesquels toutes les variables élémentaires étaient disponibles dans les sources, et pour lesquels l'IQSS associé était calculable. Pour extraire des données utiles des comptes rendus textuels, nous avons développé et validé des algorithmes dédiés basés sur des règles, dont les mesures de performance ont été évaluées par rappel, précision et score f1.

Résultats

Des 5 785 patientes diagnostiquées d'un cancer du sein (60,9 ans, IQR [50,0-71,9]), 5 147 (89,0 %) avaient des actes liés au cancer enregistrés dans le PMSI, et 3 732 (72,5 %) avaient au moins une chirurgie. Des 34 indicateurs cibles, 9 étaient calculables avec le PMSI seul, et 6 autres le devenaient en utilisant les données présentes dans les comptes-rendus d'anatomopathologie. Dix variables élémentaires étaient nécessaires au calcul des 6 indicateurs combinant Programme de Médicalisation du Système d'Information et comptes-rendus d'anatomopathologie. Les comptes-rendus nécessaires étaient disponibles pour 58,8% à 94,6% des patients, suivant les indicateurs. Les algorithmes d'extraction textuelle avaient une exactitude moyenne de 76,5 % (min-max [32,7 %-93,3 %]), une précision moyenne de 77,7 % [10,0 %-97,4 %] et une sensibilité moyenne de 71,6 % [2,8 % à 100,0 %]. Une fois ces algorithmes appliqués, les variables nécessaires au calcul des indicateurs étaient possibles à extraire pour 2% à 88% des patients, suivant les indicateurs.

Discussion

La disponibilité des comptes-rendus dans l'entrepôt de données, celle des variables élémentaires au sein des comptes rendus, et la performance des algorithmes d'extraction limite la population pour laquelle les indicateurs sont calculables.

Conclusions

Le calcul automatisé d'indicateurs qualité à partir des dossiers patients informatisés est une perspective qui se heurte à de nombreux freins pratiques.

Guével E, Priou S, Flicoteaux R, Lamé G, Bey R, Tannier X, et al. Development of a Natural Language Processing Model for deriving breast cancer quality indicators : A cross-sectional, multicenter study. Revue d'Épidémiologie et de Santé Publique. 2023;71(6):102189. https://doi.org/10.1016/j.respe.2023.102189

Termes MeSH: Indicateurs de qualité, soins de santé ; traitement du langage naturel ; traitement électronique de données

Guével E, Priou S, Flicoteaux R, Lamé G, Bey R, Tannier X, et al. Development of a Natural Language Processing Model for deriving breast cancer quality indicators : A cross-sectional, multicenter study. Revue d'Épidémiologie et de Santé Publique. 2023;71(6):102189. https://doi.org/10.1016/j.respe.2023.102189

ABSTRACT

Objectives

Medico-administrative data are promising to automate the calculation of Healthcare Quality and Safety Indicators. Nevertheless, not all relevant indicators can be calculated with this data alone. Our feasibility study objective is to analyze 1/ the availability of data sources; 2/ the availability of each indicator elementary variables, and 3/ to apply natural language processing to automatically retrieve such information.

Method

We performed a multicenter cross-sectional observational feasibility study on the clinical data warehouse of Assistance Publique – Hôpitaux de Paris (AP-HP). We studied the management of breast cancer patients treated at AP-HP between January 2019 and June2021, and the quality indicators published by the European Society of Breast Cancer Specialist, using claims data from the *Programme de Médicalisation du Système d'Information* (PMSI) and pathology reports. For each indicator, we calculated the number (%) of patients for whom all necessary data sources were available, and the number (%) of patients for whom all elementary variables were available in the sources, and for whom the related HQSI was computable. To extract useful data from the free text reports, we developed and validated dedicated rule-based algorithms, whose performance metrics were assessed with recall, precision, and f1-score.

Results

Out of 5,785 female patients diagnosed with a breast cancer (60.9 years, IQR [50.0-71.9]), 5,147 (89.0%) had procedures related to breast cancer recorded in the PMSI, and 3,732 (72.5%) had at least one surgery. Out of the 34 key indicators, 9 could be calculated with the PMSI alone, and 6 others became so using the data from pathology reports. Ten elementary variables were needed to calculate the 6 indicators combining the PMSI and pathology reports. The necessary sources were available for 58.8% to 94.6% of patients, depending on the indicators.
The extraction algorithms developed had an average accuracy of 76.5% (min-max [32.7%-93.3%]), an average precision of 77.7% [10.0%-97.4%] and an average sensitivity of 71.6% [2.8% to 100.0%]. Once these algorithms applied, the variables needed to calculate the indicators were extracted for 2% to 88% of patients, depending on the indicators.

Discussion

The availability of medical reports in the electronic health records, of the elementary variables within the reports, and the performance of the extraction algorithms limit the population for which the indicators can be calculated.

Conclusions

The automated calculation of quality indicators from electronic health records is a prospect that comes up against many practical obstacles.

MeSH terms: Quality Indicators, Health Care; Natural Language Processing; Electronic Data Processing

Guével E, Priou S, Flicoteaux R, Lamé G, Bey R, Tannier X, et al. Development of a Natural Language Processing Model for deriving breast cancer quality indicators : A cross-sectional, multicenter study. Revue d'Épidémiologie et de Santé Publique. 2023;71(6):102189. https://doi.org/10.1016/j.respe.2023.102189

INTRODUCTION

Healthcare Quality and Safety Indicators (HQSI) contribute to the management of care facilities over time and the comparison of facilities of the same type. HQSI measurement can improve care, reduce practice heterogeneity and reduce costs (1). However, the production of HQSI often requires time-consuming manual data entry, so that the assessment of the quality and safety of care is often based on *ad hoc* campaigns and therefore covers only a very small sample of the practices being evaluated. Faced with this problem, the French 2018-22 national health strategy set the ambition to "develop result, vigilance and alert indicators for the three sectors of the care offer" whose collection "will have to be automated without overwork for professionals" (2). In this context, we are interested in the feasibility of the automatic collection of HQSI for the management of breast cancer, the most common cancer in women.

Different studies have attempted to automatically calculate HQSI for breast cancer from structured medico-administrative data. Among the indicators of interest to the medical community, only a small part is calculable from medico-administrative data: 9 indicators out of 46 in a first study, 9 out of 367 in a more recent study (3,4). The set of indicators developed by the French National Cancer Institute (INCa) to assess the quality of breast cancer care also follows this strategy of exploiting structured data: the HQSI selection process explicitly excluded indicators that could not be calculated from structured data from medico-administrative databases (5). While this choice is understandable because it allows a calculation at the national level, management sciences have largely shown that the choice of measured performance indicators affects the behavior of the actors who are subject to them (6). In the case of breast cancer, indicators that cannot be calculated from medico-administrative data cover important dimensions of the care process and are for some requested in accreditation standards such as that of the OECI (Organisation of European Cancer Institutes - oeci.eu). It is therefore important to try to cover the relevant indicators as best as possible, and not just those that are easily obtained. This requires going beyond medico-administrative data and using a richer range of data to obtain indicators.

The development of hospital Electronic Health Records (EHR) presents an opportunity to go further (7,8). However, one of the main barriers to the use of EHR for HQSI calculation is the level of data quality of EHRs. Thus, in a recent American study on the quality of care in oncology, the average availability rate of the variables needed to calculate HQSI was only 23% in EHR (9). Out of 19 quality indicators chosen, only two were calculable for more than 1% of the patients studied. In addition, most of the information in EHR is contained in the free text of the reports. The exploitation of EHR on a large scale therefore presupposes the ability to automatically extract data from free text, using natural language processing (NLP) algorithms.

We present the first results of a feasibility study of automated HQSI calculation for breast cancer, using hospital EHR data: we analyzed 1/ the availability of data sources; 2/ the availability of each indicator elementary variables, and 3/ we applied natural language processing to automatically retrieve such information. We focused on the HQSI offered by the European Society of Breast Cancer Specialists (EUSOMA), and on the hospital EHR data contained in the clinical data warehouse (CDW) of the Assistance Publique – Hôpitaux de Paris (AP-HP) Teaching hospital (10). In this pilot, we focused on pathology reports and claim data solely.

METHODS

We conducted a multicenter cross-sectional study on the hospital EHR data available in the AP-HP CDW. The AP-HP CDW contains the data of 11.4 million patients, collected during their care in all 38 AP-HP university hospitals. In this study, we used data from the *Programme de Médicalisation des Systèmes d'Information* (PMSI, the national hospital claims database) and data contained in the Orbis® EHR (demographic data, administrative visits, and clinical reports). The PMSI contains structured data including the coding of diagnoses according to the International Classification of Diseases 10th revision (ICD10) and the coding of medical procedures performed according to the *Classification Commune des Actes Medicaux* (CCAM).

The constitution of the CDW of the AP-HP was authorized by the CNIL (Commission Nationale de l'Informatique et des Libertés) on January 19, 2017 (authorization n ° 19800120). The research work presented here was approved by the Scientific and Ethical Committee of AP-HP (IRB00011591) on May 15, 2020 (authorization CSE_20-0055_COVONCO-AP). The working database was extracted on December 5, 2022. We reported the study results according to the RECORD statement.

1. Population identification

We included adult female patients newly referred to AP-HP for breast cancer between January 1, 2019, and June 30, 2021, and with a breast cancer resection at AP-HP. We focused the study on this period because previous data had lower completeness. Patients with multiple cancers were excluded.
Breast cancer hospitalizations were identified by the presence of an ICD10 code C50 (invasive cancer) or D05 (*in situ* tumor) in primary or related diagnosis in the PMSI. A breast cancer hospitalization was considered a *new cancer* if no breast cancer hospitalization for the patient was identified in the 18 months prior to the start date of hospitalization. This hospitalization was then considered the baseline hospitalization for the patient.
Hospitalizations related to another type of cancer were identified in the PMSI by an ICD10 cancer code (see Appendix Table 1). Patients with hospitalization for another type of cancer between June 2017 and December 2022 were excluded.
Identification of anticancer treatments and their related dates were performed by PMSI ICD10 codes for chemotherapy (Z511), radiotherapy (Z510) and palliative care (Z515) and CCAM codes for surgery (see Appendix Table 2). For ICD-10 codes, the dates of treatment occurrence were defined as the first day of the related patient visit.

2. HQSI calculation methods

EUSOMA proposes 34 HQSI (10,11). The titles of the HQSI were translated into French (Appendix Table 3). Each HQSI was broken down into elementary variables necessary for its calculation, available in the PMSI data, in a free text pathology report, or in another data source. We classified HQSI into three groups according to the data needed to calculate them: PMSI only; PMSI and pathology reports; Other. In this pilot, we focused on indicators that could be evaluable with PMSI and pathology reports.
For each indicator, we calculated the number (%) of patients for whom all necessary data sources were available (Table 1). For example, the calculation of indicator 2 "*Ratio of mild to malignant diagnoses"* required both PSMI and pathology reports sources to be available. Then, we calculated the number (%) of patients for whom all elementary variables were available in the sources, and for whom the related HQSI was computable. For example, the calculation of indicator 2 required the "*Malignancy of the sample studied*" variable to be available in the related operative pathology report.

3. Extraction of pathology reports elementary variables

Availability of pathology reports
The date of the pathology report was identified by extracting the date from the text of the report via regular expression (Appendix Table 4).

Three types of pathology reports were identified:

- Diagnostic pathology reports corresponded to pathology reports dated before the administration of the first treatment (surgery or chemotherapy) for the patient.
- Preoperative pathology reports corresponded to pathology reports dated at least 3 days before the date of surgery.
- Postoperative pathology reports were pathology reports dated within 3 days before or after the date of the first surgery.

We assessed the availability of different types of pathology reports in the AP-HP CDW by calculating the number of patients for whom a pathology report was available in the CDW in relation to the total number of patients in the relevant population.

Availability of elementary variables in pathology reports

The elementary variables mentioned in the pathology reports were annotated by an engineer (EG) with a medical oncologist (EK) (development set, 259 pathology reports randomly sampled among those edited before June 2021), or only by a medical oncologist (test set, 48 pathology reports randomly sampled among those edited after June 2021 for patients operated for primary breast cancer). We assessed the availability of elementary variables in pathology reports by calculating the percentage of pathology reports containing them in the test set.

Development of an algorithm for automatic extraction of elementary variables by NLP

We developed an NLP algorithm for extracting the elementary variables available in pathology reports based on regular expressions (Appendix Table 4). The regular expressions were developed using only the 259 pathology reports of the development set. The developed NLP algorithm was based on components from the EDS-NLP v0.7.4 software library (12).

We calculated the performance of the elementary variable extraction algorithm on the test set. For each elementary variable, we measured accuracy (ratio of true predicted positive cases on all predicted positive cases, i.e., positive predictive value), recall (ratio of true predicted positive cases on all positive cases, i.e., sensitivity), and f1-score (harmonic mean of accuracy and recall). We assessed the weighted average values, defined as the average of the calculated values of the metric for each of the classes, weighted by the number of elements in the class.

The algorithm returned, for each document, a value for each elementary variable. If several values of the same elementary variable were found for a document, the value indicating the worst prognosis was selected. If the elementary variable was not detected in the text, the return value was null.

RESULTS

1. Population characteristics

Among the 5,785 patients newly referred to AP-HP between January 2019 and June 2021 for a breast cancer (60.9 years IQR [50.0-71.9]), 3,575 were operated at AP-HP. 3532 patients (97.8%) had an invasive cancer and 43 (1.2%) had an *in situ* tumor.

2. Identification of computable HQSI

The 34 HQSI proposed by EUSOMA were broken down into 41 elementary variables (Appendix Table 5), of which 12 were available in the PMSI data, 13 were available in the pathology reports, and 16 were available in other data source. From these variables, 9/34 HQSI (26%) were calculable only with PMSI and 15/34 HQSI (47%) were calculable with PMSI and pathology reports. To calculate the 6 additional HQSI using pathology reports, 10 elementary variables had to be extracted from the pathology reports (Table 1): the type of pathological technique used, the malignancy of the sample studied, the estrogen receptor status, the HER2 status, the tumor grade, the histological type of the tumor, the pTNM score, the size of the tumor, the presence of vascular embolisms and the distance to the resection margins.

3. Availability of elementary variables in pathology reports

Availability of pathology reports

The 6 HQSI calculable using PMSI and pathology reports were only for patients for whom a diagnostic, preoperative or postoperative pathology report was available, depending on the case (Table 1). A diagnostic pathology report was available in the AP-HP CDW for 3,426 (91.8%) patients, a preoperative pathology report for 2176 (58.3%) patients, and a postoperative pathology report for 3,515 (94.2%) patients (Table 2). The 6 HQSI were calculable for patients with all the data necessary to calculate the elementary variables of HQSI (Table 1).

Availability of elementary variables in pathology reports

The availability of elementary variables in pathology report ranged from 75% (36/48 papers) for distance to margins to 100% (48/48 papers) for histological type and tumor malignancy (Table 3).

4. Performance and results of NLP algorithms

The performance of the algorithms developed to extract elementary variables from pathological reports is listed in Table 3. The algorithms had a weighted average accuracy of 77.1%, ranging from 14.1% to 100%, and a weighted average sensitivity of 76.2%, ranging from 27% to 100% for each elementary variable.

5. HQSI computability

The calculation of each HQSI depends on the combination of the availability of pathology report (diagnostic, preoperative or operative depending on the case), the presence of elementary variables in the pathology report, and the performance of the NLP algorithm for extracting elementary variables. The HQSI 2 reporting the ratio of mild to malignant diagnoses was calculable for 88% of the target population; HQSI 3b for 49.9% of targeted patients, HQSI 4a only for 17.6% of patients operated on with invasive cancer; HQSI 4b for 2.3% of patients operated with *in-situ* tumor; HQSI 11d was calculable for 46.5% of *in situ* tumor and HQSI 13a was calculable for 44.9% of patients with invasive cancer (Table 1).

DISCUSSION

This study presents the opportunities and limitations of using hospital HER data to automate HQSI calculation. More than a quarter of the EUSOMA indicators (9/34) were theoretically automatically computable from PMSI data, and almost half (16/34) using information extracted from the pathology reports in addition to the PMSI data. Nevertheless, for these indicators combining PMSI and data from reports, the proportion of patients for whom the indicator can be produced automatically varies widely, from 2.3% to 88%. These results are better than those obtained by Schorer *et al.*, who could only calculate two indicators for more than 1% of their population (9). Nevertheless, these results remain unsatisfactory and do not allow a transition to production.

The manual calculation of HSQI requires small sample sizes and may lack generalizability (13). The automatability of the production of an indicator relies on the secondary use of patient EHR, whose feasibility depends on the quality and the availability of data of interest (14). A recent simulation study performed on the APHP CDW related to breast cancer patient pathways showed a substantial loss of key information due to the complexity of data flows (15). Another study addressed the need of a constant assessment of EHR data quality life cycle (16). In 2023, the French *Haute Autorité de Santé* stakeholder required efforts in data quality and documentation to be made when developing a CDW (17). When it comes to structured data, this availability is calculated directly. When it comes to data extracted from free text documents by NLP, the availability of data depends on the availability of the relevant textual document, the presence of the information of interest within the textual document, and the performance of the algorithm for extracting this data (18)(19). The exploratory work carried out highlights that pathology reports are regularly missing, especially pre-operative pathology reports (in more than 40% of cases), most often because the biopsy did not take place at AP-HP. In this case, there may be a pathology report in the CDW, but this pathology report is in a scanned document format and is not usable to date.
When the reports are available, the data sought must still be present. This is not always the case, and the availability of some variables drops rapidly, around 49% for grade or 54% for pTNM score. Whether it is the availability of documents or the information in these documents, it is impossible to say whether the missing data are random or whether bias is present.
Finally, indicators based on data extracted using NLP depend on the performance of extraction algorithms which are so far mixed and vary greatly depending on the variable (20). The algorithms used, based on rules, do not require any computing power, but take a long time to develop and validate. The performance of the algorithms would probably be improved by increasing the size of the annotated set for development and validation, but the annotation work is long and tedious, and requires a significant investment in medical time (21). Large language models are an interesting perspective for free text information retrieval (22).
Our conclusions are in line with those of a similar experiment, conducted in France in 2015: the automatic calculation of indicators appears feasible, but depends on the quality and availability of data, and requires significant resources both medical and IT to develop methods (23).
Faced with this situation, it would first be necessary to conduct data quality campaigns to improve the availability of sources – in our case, pre- and post-operative pathology reports– but also the completeness and quality of the data of interest within these sources. This implies a significant effort with random results. Another approach is to multiply the sources, also including the reports of

multidisciplinary meetings, consultations, etc., and to cross the data from these different sources. This could increase the availability of data but possibly highlight inter-source inconsistencies. Nevertheless, this requires the development and validation of algorithms. In any case, a lot of work seems necessary before we can deploy indicators calculated automatically from unstructured data.

This work was based on data from the AP-HP CDW, a database of 11 million patient records. It has benefited from the experience accumulated during several projects on the quality of care in oncology. Nevertheless, the choice of breast cancer, a large part of whose diagnosis and management can be done outside the hospital, may have darkened the results, compared to what could perhaps have been observed on less common cancers treated in hospital. Besides, the availability of data sources relies on the complex infrastructure of the AP-HP CDW which can generate substantial loss of key information (15). Moreover, the availability of elementary variables was estimated based on NLP algorithms whose performance metrics could be improved.

Guével E, Priou S, Flicoteaux R, Lamé G, Bey R, Tannier X, et al. Development of a Natural Language Processing Model for deriving breast cancer quality indicators : A cross-sectional, multicenter study. Revue d'Épidémiologie et de Santé Publique. 2023;71(6):102189. https://doi.org/10.1016/j.respe.2023.102189

CONCLUSION

The automated calculation of HQSI from EHRs is a promising prospect, but it faces many practical obstacles: availability of sources, availability of information in these sources, and resources to develop algorithms for extracting this data.

REFERENCES

1. Laronga C, Gray JE, Siegel EM, Lee JH, Fulp WJ, Fletcher M, et al. Florida initiative for quality cancer care: Improvements in breast cancer quality indicators during a 3-year interval. J Am Coll Surg. 2014;219(4):638-645.e1.
2. Ministère des Solidarités et de la Santé. Stratégie nationale de santé 2018-2022. 2022;1–53.
3. Andreano A, Anghinoni E, Autelitano M, Bellini A, Bersani M, Bizzoco S, et al. Indicators based on registers and administrative data for breast cancer: routine evaluation of oncologic care pathway can be implemented. J Eval Clin Pract [Internet]. 2016 Feb 1 [cited 2022 Sep 9];22(1):62–70. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/jep.12436
4. Guarneri V, Pronzato P, Bertetto O, Roila F, Amunni G, Bortolami A, et al. Use of electronic administrative databases to measure quality indicators of breast cancer care: Experience of five regional oncology networks in Italy. J Oncol Pract. 2020 Feb 1;16(2):81.
5. Houzard S, Courtois E, Le Bihan Benjamin C, Erbault M, Arnould L, Barranger E, et al. Monitoring breast cancer care quality at national and local level using the French National Cancer Cohort. Clin Breast Cancer. 2022 May 21;
6. Bevan G, Hood C. What's measured is what matters: targets and gaminf in the English public health care system. Public Adm [Internet]. 2006;84(3):517–38. Available from: https://doi.org/10.1111/j.1467-9299.2006.00600.x
7. Amster A, Jentzsch J, Pasupuleti H, Subramanian KG. Completeness, accuracy, and computability of National Quality Forum-specified eMeasures. J Am Med Informatics Assoc. 2015;22(2):409–16.
8. Ahmad FS, Rasmussen L V., Persell SD, Richardson JE, Liss DT, Kenly P, et al. Challenges to electronic clinical quality measurement using third-party platforms in primary care practices: The healthy hearts in the heartland experience. JAMIA Open. 2019;2(4):423–8.
9. Schorer AE, Moldwin R, Koskimaki J, Bernstam E V, Venepalli NK, Miller RS, et al. Chasm Between Cancer Quality Measures and Electronic Health Record Data Quality. JCO Clin Cancer Informatics [Internet]. 2022;(6):e2100128. Available from: https://doi.org/10.1200/CCI.21.00128
10. Biganzoli L, Marotti L, Hart CD, Cataliotti L, Cutuli B, Kühn T, et al. Quality indicators in breast cancer care: An update from the EUSOMA working group. Eur J Cancer. 2017 Nov;86:59–81.
11. van Dam PA, Tomatis M, Marotti L, Heil J, Wilson R, Rosselli Del Turco M, et al. The effect of EUSOMA certification on quality of breast cancer care. Eur J Surg Oncol J Eur Soc Surg Oncol Br Assoc Surg Oncol. 2015 Oct;41(10):1423–9.
12. Dura B, Wajsburt P, Petit-Jean T, Cohen A, Jean C, Bey R. EDS-NLP: efficient information extraction from French clinical notes (v0.7.4). Zenodo [Internet]. 2022. Available from: https://doi.org/10.5281/zenodo.7428752
13. Couralet M, Leleu H, Capuano F, Marcotte L, Nitenberg G, Sicotte C, et al. Method for developing national quality indicators based on manual data extraction from medical records. BMJ Qual Saf. 2013;22(2):155–62.
14. van Velthoven MH, Mastellos N, Majeed A, O'Donoghue J, Car J. Feasibility of extracting data from electronic medical records for research: an international comparative study. BMC Med Inform Decis Mak. 2016 Jul;16:90.
15. Priou S, Lame G, Jankovic M, Chatellier G, Bey R, Tournigand C, et al. Why Are Data Missing in Clinical Data Warehouses? A Simulation Study of How Data Are Processed (and Can Be Lost). Stud Health Technol Inform. 2023 May;302:202–6.
16. Chelico JD, Wilcox AB, Vawdrey DK, Kuperman GJ. Designing a Clinical Data Warehouse Architecture to Support Quality Improvement Initiatives Northwell Health , Manhasset , NY ; 2 University of Washington , Seattle , WA ; NewYork Presbyterian Hospital , New York , NY. AMIA Annu Symp Proc. 2016;2016:381–90.
17. Doutreligne M, Degremont A, Jachiet P-A, Lamer A, Tannier X. Good practices for clinical data

warehouse implementation: A case study in France. PLOS Digit Heal. 2023 Jul;2(7):e0000298.

18. Lindvall C, Lilley EJ, Zupanc SN, Chien I, Udelsman B V., Walling A, et al. Natural Language Processing to Assess End-of-Life Quality Indicators in Cancer Patients Receiving Palliative Surgery. J Palliat Med [Internet]. 2019 Feb 1 [cited 2020 Oct 7];22(2):183–7. Available from: https://pubmed.ncbi.nlm.nih.gov/30328764/
19. Bae JH, Han HW, Yang SY, Song G, Sa S, Chung GE, et al. Natural Language Processing for Assessing Quality Indicators in Free-Text Colonoscopy and Pathology Reports: Development and Usability Study. JMIR Med informatics. 2022 Apr;10(4):e35257.
20. Warner JL, Levy MA, Neuss MN, Warner JL, Levy MA, Neuss MN. Feasibility and accuracy of extracting cancer stage information from narrative electronic health record data. J Oncol Pract. 2016 Feb 1;12(2):e169–79.
21. Neves M, Ševa J. An extensive review of tools for manual annotation of documents. Brief Bioinform. 2021 Jan;22(1):146–63.
22. Jiang K, Mujtaba MM, Bernard GR. Large Language Model as Unsupervised Health Information Retriever. Stud Health Technol Inform. 2023 May;302:833–4.
23. Ficheur G, Schaffar A, Caron A, Balcaen T, Beuscart J-B, Chazard E. Elderly Surgical Patients: Automated Computation of Healthcare Quality Indicators by Data Reuse of EHR. Stud Health Technol Inform. 2016;221:92–6.
24. Kempf E, Priou S, Lamé G, Daniel C, Bellamine A, Sommacale D, et al. Impact of two waves of Sars-Cov2 outbreak on the number, clinical presentation, care trajectories and survival of patients newly referred for a colorectal cancer: A French multicentric cohort study from a large group of University hospitals. Int J Cancer [Internet]. 2021 Jan 17 [cited 2022 Jan 20];accepted(September 2021):1–10. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/ijc.33928

Guével E, Priou S, Flicoteaux R, Lamé G, Bey R, Tannier X, et al. Development of a Natural Language Processing Model for deriving breast cancer quality indicators : A cross-sectional, multicenter study. Revue d'Épidémiologie et de Santé Publique. 2023;71(6):102189. https://doi.org/10.1016/j.respe.2023.102189

Table legends and legends

Table 1. Indicators (name shortened, correspondence to the EUSOMA indicator in Annex Table 3) of quality and safety of the care considered, and elementary variables necessary for their calculation

HQSI	Population of interest	Number of patients	Elementary variables needed to calculate the HQSI	Data sources needed to calculate elementary variables	Number (%) of patients for whom all necessary data sources were available	Number (%) of patients for whom all elementary variables were available in the sources, and for whom the HQSI was computable
2. Ratio of mild to malignant diagnoses	Operated patients	3575	Date of primary surgery	PMSI	3381 (94,6 %)	3147 (88,0 %)
			Malignancy of the sample studied	Operative pathology report		
3.b Preoperative PCR diagnosis rate	Operated patients	3575	Date of primary surgery	PMSI	2101 (58,8 %)	1781 (49,8 %)
			Malignancy of the sample studied	Pre-operative pathology report		
			Type of pathological technique used	Pre-operative pathology report		
4.a Invasive component prognostic factor assessment rate (12 prognostic factors)*	Invasive cancer for operated patients	3532	chemotherapy date	PMSI	2709 (76,7 %)	622 (17,6 %)
			type of cancer	PMSI		
			date of surgery of the primary	PMSI		
			grade of tumour	Diagnostic pathology report		
			estrogen receptor status	Diagnostic pathology report		

HQSI	Population of interest	Number of patients	Elementary variables needed to calculate the HQSI	Data sources needed to calculate elementary variables	Number (%) of patients for whom all necessary data sources were available	Number (%) of patients for whom all elementary variables were available in the sources, and for whom the HQSI was computable
			HER2 status	Diagnostic pathology report		
			histological type	Diagnostic pathology report		
			pTNM score	Operative pathology report		
			vascular embolisms	Operative pathology report		
			distance to margins	Operative pathology report		
				Operative pathology report		
			tumour size	Operative pathology report		
4.b Assessment rate of prognostic factors of the non-invasive component (5 prognostic factors)	Cancer *in situ* for operated patients	43	date of surgery of the primary	PMSI	39 (90,7 %)	1 (2,3 %)
			type of cancer	PMSI		
			grade of tumour	Diagnostic pathology report		
			histological type	Diagnostic pathology report		
			estrogen receptor status	Diagnostic pathology report		

HQSI	Population of interest	Number of patients	Elementary variables needed to calculate the HQSI	Data sources needed to calculate elementary variables	Number (%) of patients for whom all necessary data sources were available	Number (%) of patients for whom all elementary variables were available in the sources, and for whom the HQSI was computable
			distance to margins	Operative pathology report		
			tumour size	Operative pathology report		
11.d Conservative surgery rates for patients with small in situ cancer	Cancer *in situ* for operated patients	43	type of cancer	PMSI	39 (90,7 %)	20 (46,5 %)
			date of surgery of the primary	PMSI		
			type of primary surgery	PMSI		
			tumour size	Operative pathology report		
13.a Adjuvant chemotherapy rate in patients with invasive breast cancer, T > 1 cm or N+ and ER+	Invasive cancer for operated patients	3532	date of surgery of the primary	PMSI	3342 (94,6 %)	1588 (44,9 %)
			chemotherapy date	PMSI		
			estrogen receptor status	Operative pathology report		
			tumour size	Operative pathology report		
			pTNM score	Operative pathology report		

Abbreviations: HQSI, indicators of quality and safety of care; PMSI, Programme de médicalisation des systems d'information

Guével E, Priou S, Flicoteaux R, Lamé G, Bey R, Tannier X, et al. Development of a Natural Language Processing Model for deriving breast cancer quality indicators : A cross-sectional, multicenter study. Revue d'Épidémiologie et de Santé Publique. 2023;71(6):102189. https://doi.org/10.1016/j.respe.2023.102189

Table 2. Availability of pathology reports

Data source	Number (%) of patients with the data source		
	Invasive cancer N = 3532	*In-situ* tumor N = 43	Total N = 3575
Diagnostic pathology report	3253 (92,1)	39 (90,7)	3292 (92,1)
Pre-operative pathology report	2084 (59,0)	17 (39,5)	2101 (58,8)
Post-operative pathology report	3342 (94,6)	39 (90,7)	3381 (94,6)
PMSI	3532 (100)	43 (100)	3575 (100)

Abbreviations: PMSI, Programme de médicalisation des systems d'information

Table 3. Performance on the validation set (N=48) of natural language processing algorithms.

Elementary variable	Availability in validation set documents	Weighted* accuracy	Weighted* recall	F1-weighted* score
estrogen receptor status	41 (85,4%)	92,9	91,7	92,1
HER2 status	40 (83,3%)	88,2	87,5	87,5
rank	45 (93,8%)	92,2	82,5	85,6
histological type	48 (100%)	81,5	85,4	83,4
Malignancy of the sample studied	48 (100%)	100	100	100
pTNM score	45 (93,8%)	91,9	93,3	92,1
tumour size	47 (97,9%)	49,0	50,0	47,0
vascular embolisms	42 (87,5%)	86,0	75,0	77,8
distance to margins	36 (75,0%)	14,1	27,1	15,1

** The weighted average values were defined as the average of the calculated values of the metric for each of the classes, weighted by the number of elements in the class.*

Table 4. Extraction of elementary variables from pathology reports

Basic variable	pathology report used for the calculation of the elementary variable	Number (%) of patients for whom the elementary variable is extracted by NLP from the pathology report of interest		
		Invasive cancer (N=3532)	Cancer *in situ* (N = 43)	Total (N = 3575)
Estrogen receptor status	Diagnostic pathology report	2341 (66,3)	4 (9,3)	2345 (65,6)
	Post-operative pathology report	2223 (62,9	4 (9,3)	2227 (62,3)
HER2 status	Diagnostic pathology report	2353 (66,6)	4 (9,3)	2357 (65,9)
rank	Diagnostic pathology report	1764 (49,9)	1 (2,33)	1765 (49,4)
histological type	Diagnostic pathology report	2653 (75,1)	36 (83,7)	2688 (75,2)
Malignancy of the sample studied	Pre-operative pathology report	1878 (53,2)	16 (37,2)	1894 (53,0)
	Post-operative pathology report	3111 (88,1)	34 (79,1)	3147 (88,0)
pTNM score	Post-operative pathology report	1935 (54,8)	3 (7,0)	1938 (54,2)
tumour size	Post-operative pathology report	2567 (72,7)	20 (46,5)	2587 (72,4)
vascular embolisms	Post-operative pathology report	2548 (72,1)	9 (20,9)	2557 (71,5)
distance to margins	Post-operative pathology report	1523 (43,1)	17 (39,5)	1540 (43,1)
Type of pathological technique used	Pre-operative pathology report	1808 (51,2)	16 (37,2)	1824 (51,0)

Abbreviations: NLP, natural language processing

Guével E, Priou S, Flicoteaux R, Lamé G, Bey R, Tannier X, et al. Development of a Natural Language Processing Model for deriving breast cancer quality indicators : A cross-sectional, multicenter study. Revue d'Épidémiologie et de Santé Publique. 2023;71(6):102189. https://doi.org/10.1016/j.respe.2023.102189

Additional Table 1: ICD10 codes used to identify cancer-related hospitalizations

Type of Cancer	ICD-10 code
Anus	C21
Bile ducts	C23 C24 D01.5 D37.6
Bladder	C66 C67 C68 D09.0 D09.1 D41.2 D41.3 D41.4 D41.7 D41.9
Intestine	C17 D01.4 D37.2
Uterus	C53 D06
Central nervous system	C70 C71 C72.0 C72.2 C72.3 C72.8 C72.9 D42 D43.0 D43.1 D43.2 D43.4 D43.7 D43.9
Colon	C18 C19 D01.0 D01.1 D37.3 D37.4
UPC	C76 C80 C97 D09.7 D09.9 D48.7 D48.9 D48.3
Endometrium	C54 C55 D07.0 D39.0
Eye	C69 D09.2
Gastric	C16 D00.2 D37.1
Head & neck	C0 C10 C11 C12 C13 C14 C30 C31 C32 D00.0 D02.0 D37.0 D38.0
Hodgkin lymphoma	C81
Kidney	C6.4 C6.5 D41.0 D41.1
Leukaemia	C91 C92 C93 C94.0 C94.1 C94.2 C94.3 C94.4 C94.5 C94.7 C95
Liver	C22
Lung	C33 C34 D02.1 D02.2 D38.1
Melanoma	C43 D03
Mesothelioma	C45.0 C45.1 C45.2 C45.7 C45.9
Myeloma	C90
Non-Hodgkin lymphoma	C82 C83 C84 C85 C86
Oesophagus	C15 D00.1
Osteosarcoma	C40 C41 D48.0
Other digestive	C26 C48 D01.7 D01.9 D37.7 D37.9 D48.4
Other endocrine	C74 C75 D09.3 D44.1 D44.2 D44.3 D44.4 D44.5 D44.6 D44.7 D44.8 D444.0 D444.8
Other gynaecological	C51 C52 C57 C58 D07.1 D07.3 D39.2 D39.7 D39.9
Other hematological malignancies	C88 C96 C94.6 D45 D46 D47
Other lung	C37 C38 C39 D02.3 D02.4 D38.2 D38.3 D38.4 D38.5 D38.6
Other urothelial	C60 C63 D07.4 D07.6 D40.7 D40.9
Ovary	C56 D39.1
Pancreas	C25
Peripheral nervous system	C47 C72.1 C72.4 C72.5 D43.3 D48.2
Prostate	C61 D07.5 D40.0
Rectum	C20 D01.2 D37.5
Other skin	C44 D04 D48.5
Tissue	C46 C49 D48.1
Testicle	C62
Thyroid	C73 D44.0

Guével E, Priou S, Flicoteaux R, Lamé G, Bey R, Tannier X, et al. Development of a Natural Language Processing Model for deriving breast cancer quality indicators : A cross-sectional, multicenter study. Revue d'Épidémiologie et de Santé Publique. 2023;71(6):102189. https://doi.org/10.1016/j.respe.2023.102189

Supplementary Table 2: CCAM codes for primary breast cancer resection surgery

Type of surgery	ACPC codes
Lumpectomy with lymph node dissection	QEFA001, QEFA008
Lumpectomy without lymph node dissection	QEFA004, QEFA016, QEFA017, QEFA018
Mastectomy with lymph node dissection	QEfA003, QEFA005, QEFA010, QEFA020
Mastectomy without lymph node dissection	QEFA007, QEFA012, QEFA013, QEFA015

Guével E, Priou S, Flicoteaux R, Lamé G, Bey R, Tannier X, et al. Development of a Natural Language Processing Model for deriving breast cancer quality indicators : A cross-sectional, multicenter study. Revue d'Épidémiologie et de Santé Publique. 2023;71(6):102189. https://doi.org/10.1016/j.respe.2023.102189

Supplementary table 3. Translation of EUSOMA indicators into French

Indicator number	Version française	English version
2.	Précision des procédures de diagnostique (ratio des cancers Bénins/Malins pour les patientes opérées)	Ratio of benign to malignant diagnoses based on definitive pathology report (surgery only, non-operative biopsies excluded)
3.b	Réalisation du diagnostic préopératoire histologiquement/cytologiquement des patientes	Proportion of women with breast cancer (invasive or in situ) who had a preoperative histologically or cytologically confirmed malignant diagnosis (B5 or C5)
4.a	Complétude de la collecte des paramètres pronostiques/prédictifs pour les patientes avec un cancer invasif	Proportion of invasive cancers cases for which the following prognostic/predictive parameters have been recorded: histological type), grading, ER, HER-2/neu For patients receiving primary systemic treatment, characterisation on core biopsy prior to therapy is mandatory. For patients receiving primary surgery characterisation may be performed on the surgical specimen only In addition to the above parameters, the following parameters must be recorded after surgery: Pathological stage, size in mm for the invasive component, peritumoral vascular invasion, distance to nearest radial margin
B	Complétude de la collecte des paramètres pronostiques/prédictifs pour les patientes avec un cancer in situ	Proportion of non-invasive cancer cases for which the following prognostic/predictive parameters have been recorded: Grading, dominant histological pattern, size in mm, distance to nearest radial margin, ER
11.a	Evitement du sur-diagnostic des ganglions sentinels pour les patientes avec un cancer invasif	Proportion of patients with invasive cancer and clinically negative axilla who underwent sentinel lymph-node biopsy only (excluding patients who received primary systemic treatment)
11.d	Chirurgies conservatrices effectuées pour les patientes avec un cancer in situ de petite taille (moins de 2 cm)	Proportion of patients with non-invasive breast cancer not greater than 2 cm who underwent BCT
13.a	Traitement chimiothérapeutique adapté pour les patientes ayant un cancer du sein invasif et RE-	Proportion of patients with ER- (T>1cm or Node +) invasive carcinoma who received adjuvant chemotherapy

Guével E, Priou S, Flicoteaux R, Lamé G, Bey R, Tannier X, et al. Development of a Natural Language Processing Model for deriving breast cancer quality indicators : A cross-sectional, multicenter study. Revue d'Épidémiologie et de Santé Publique. 2023;71(6):102189. https://doi.org/10.1016/j.respe.2023.102189

Supplementary Table 4. Lists of regular expressions used for entity retrieval

	Term	Regular expressions
pathology report Dates	date	r"([0-9]{2}\/[0-9]{2}\/[0-9]{4})"
	levy	r"(pr[ée]lev[ée]\sle|date\sd.\spr[ée]l[èe]vement)"
	reception	r"(r?e[cç]u\sle|date\sde\sr.ception)"
Section Titles	indication	r"indications?"
	History of the disease	r"histoire de la maladie", r"histoire de la maladie - explorations", r"histoire de la maladie actuelle", r"histoire du poids", r"histoire recente", r"histoire recente de la maladie", r"rappel clinique", r"resume", r"resume clinique", r"resume clinique - histoire de la maladie", r"antecedents et histoire de la maladie", r"renseignements? cliniques?"
	immunohistochemistry	r".{0,15}immuno-?histochimi(e|que).{0,80}"
	conclusion	r"au total", r"conclusion", r"conclusion de sortie", r"synthese medicale / conclusion", r"synthese", r"synthese medicale", r"synthese medicale/conclusion", r"conclusion medicale", r"examen histologique.{1,4}conclusion"
Entities	RE	r"r[eé]cepteur.{1,10}[(oe)œ]strog[eè]ne", r"\bre\b", r"\bro\b"
	RP	r"r[ée]cepteur.{1,10}progest[ée]rone", r"\brp\b"
	HER2	r"c-?erb.{0,3}2", r"her\s?2"
	Other	r"t[ée]moin", r"\bra\b|r[ée]cepteur.{1,8}androg[èe]ne", r"seuil|statut her2 (positif|équivoque|négatif)", r"tumeur (non)?proliférant", r"composante", r"(mono|poly)somie", r"e-cadh[eé]rine", r"\bpdl1\b", r"(facteur|indice).{1,10}prolif[ée]ration", r"ki.{0,2}67"
	rank	r"[Ee]lston (& |et |and)?[Ee]llis", r"\b[Ee]{2}\b" + r"(?s).(\([^\d\(\)]*[0-3].{0,2}[\+,][^\d\(\)]*[0-3].{0,2}[\+,][^\d\(\)]*[0-3]\))"

Guével E, Priou S, Flicoteaux R, Lamé G, Bey R, Tannier X, et al. Development of a Natural Language Processing Model for deriving breast cancer quality indicators : A cross-sectional, multicenter study. Revue d'Épidémiologie et de Santé Publique. 2023;71(6):102189. https://doi.org/10.1016/j.respe.2023.102189

Supplementary Table 5. Elementary variables needed to calculate EUSOMA HQSI

Basic variable	Type of report	HQSI EUSOMA Number
Malignancy of the operating room	pathology report	2.
tumour size	pathology report	4a., 4b., 11c., 11d., 13a., 13b.
distance to margins	pathology report	4a., 4b.
vascular embolisms	pathology report	4a.
pTNM score	pathology report	13a., 10b., 10c., 4a., 13b.
Preoperative malignancy	pathology report	3b.
Type of pathological technique used	pathology report	3b.
histological type	pathology report	4a., 4b.
HER2 status	pathology report	4a., 13b., 13c.
estrogen receptor (ER) status	pathology report	4a., 4b., 12.
grade EE	pathology report	4a., 4b.
progesterone receptor (pathology report) status	pathology report	12., 4a.
cTNM score	Other	11a.
Date of taking trastuzumab	Other	13b., 13c.
RCP date	Other	8.
Date of diagnosis	Other	5.
Physical exam date	Other	1.
type of endocrine therapy	Other	12.
BRCA1/2	Other	11c.
stage of cancer	Other	14a., 14b.
asymptomatic status of patients	Other	15a.
IBC or locally advanced unresectable ER carcinoma	Other	13d.
Nurse consultation date	Other	16a., 16b.
areas targeted by radiation therapy	Other	10c., 10b.
Presence of a Data Manager in the Breast Cancer Center	Other	17.
Post-processing data collection	Other	15b.
Clinical Assessment Date	Other	15a.
Baseline staging date	Other	14a., 14b.
referral for genetic counselling	Other	7.

Guével E, Priou S, Flicoteaux R, Lamé G, Bey R, Tannier X, et al. Development of a Natural Language Processing Model for deriving breast cancer quality indicators : A cross-sectional, multicenter study. Revue d'Épidémiologie et de Santé Publique. 2023;71(6):102189. https://doi.org/10.1016/j.respe.2023.102189

Figure legends

NA

Guével E, Priou S, Flicoteaux R, Lamé G, Bey R, Tannier X, et al. Development of a Natural Language Processing Model for deriving breast cancer quality indicators : A cross-sectional, multicenter study. Revue d'Épidémiologie et de Santé Publique. 2023;71(6):102189. https://doi.org/10.1016/j.respe.2023.102189

APPENDICES

ADDITIONAL METHODS

Development of NLP algorithms for the extraction of elementary variables from pathological reports

In this section the methods used to extract each of the elementary variables are described.

Immunohistochemical data (ER status, RP status, HER2 status)

Immunohistochemistry data are extracted from pathology reports as follows:

1. Search for the section describing immunohistochemistry results using regular expressions (regex). For this several types of sections are indicated each with a list of regex corresponding to their possible titles. The text between the title of a section and the next title is assigned to the first title. In our case, each text was divided into the following sections: indication, history of the disease, immunohistochemistry, and conclusion (regex available in supplementary table 4.)
2. The entities sought, as well as other entities that can be mentioned in immunohistochemical analyses are matched by regex and the text between two entities is assigned to the first entity (lists of entities and regex available in supplementary table 4.)
3. We search in each of these texts
    a. Score mentions
    b. Mentions of %
    c. Intensity mentions
    d. Marking information

    These mentions are associated with the entity of the sentence
4. The value of the entity is normalized from these references
    a. RE/pathology report status
        i. If indicated + or - then the value is taken
        ii. If a percentage is indicated and is greater than or equal to 10, the score deduced is +, if less than 10 then -
        iii. Otherwise no score
    b. HER2 status
        i. If a score (0, 1, 2 or 3) is indicated then it is taken
        ii. If percentage, marking and intensity the score deduced is:
            1. Score 3: Strong full membrane labeling > 10%
            2. score 2: strong full membrane labeling <= 10%; Moderate > 10% complete membrane labeling or moderate to high > 10% incomplete membrane labeling
            3. Score 1: Low > 10% low complete or incomplete membrane labeling
            4. Score 0: No labeling or moderate or low membrane labeling <= 10%
        iii. Otherwise no score

5. If there are several values for the same entity then the most serious is taken
    a. - for RE/RP status
    b. lowest for HER2 status

Rank

The grade of the tumor is obtained by looking for the structure of the Elston-Ellis histoprognostic grade (regex in supplementary table 4.) which is: (architecture: x1 + nucleus: x2+ mitotic activity: x3) with x1, x2 and x3 between 1 and 3.
Normalization is made from the value G = x1 + x2 + x3

- If G >= 8 then the tumor is grade 3
- if 8 > G >= 6 then the tumor is grade 2
- if G < 6 then the tumour is grade 1

If there are several grades for a patient, the highest is kept.

Histological type / malignancy of the tumor

The histological type of the tumor and the malignancy/benignity of the tumor are information obtained from the ADICAP codes. ADICAP codes are extracted with the eds.adicap pipeline from the edsnlp python library (12).
The information contained in these codes was retrieved according to the established thesaurus.

The ADICAP codes verifying these conditions have been selected:

- The organ of the code is the breast
- In the case where one is interested in the operating room, the method of sampling is operating room with complete excision of the organ

Malignancy was deduced as follows:

- If the ADICAP code is that of a tumor pathology or a particular pathology of the organs, the 6th character has been selected. If this character is between 0 and 3 then the code indicates a benign character, otherwise the code indicates a malignant character
- If the ADICAP code is that of a non-tumor pathology then it indicates a benign character
- If a patient has multiple ADICAP codes, then the malignant character is retained above the benign character

The histological type of the tumor is inferred from ADICAP codes indicating tumor pathology (and --GSA5B2 codes that indicate intragalactophoric adenocarcinoma in situ). If a patient has several ADICAP codes indicating different histological types then the codes indicating invasive tumors are kept above those indicating in situ tumors, and the codes indicating rarer pathologies (papillary carcinoma, mucinous) are kept above those indicating more common pathologies (ductal, lobular).

pTNM score

The pTNM score was extracted from patients' postoperative pathological reports using the edsnlp python library. Such algorithms were developed and validated in other observational cancer studies performed on the AP-HP CDW (24).
If several TNM scores are found for the same patient, then the most severe score is kept: M the highest, then N the highest, then T the highest.

Tumour size / Vascular embolus / Distance to margins

6. Search for the section describing immunohistochemistry results using regular expressions (regex). For this several types of sections are indicated each with a list of regex corresponding to their possible titles. The text between the title of a section and the next title is assigned to the first title. In our case, each text was divided into the following sections: indication, history of the disease, immunohistochemistry, and conclusion (regex available in supplementary table 4.)