Original Research Article

DIGITAL HEALTH

DIGITAL HEALTH
Volume 11: 1–11
© The Author(s) 2025
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20552076251323110
journals.sagepub.com/home/dhj
Sage

Structuring and centralizing breast cancer real-world biomarker data from pathology reports through C-LAB® artificial intelligence platform

Florent Le Borgne[1,#], Camille Garnier[2,#], Camille Morisseau[1], Yanis Navarrete[2], Yanina Echeverria[2], Juan Mir[2], Jaume Calafell[2], Tanguy Perennec[3], Olivier Kerdraon[4], Jean-Sébastien Frenel[5,6], Judith Raimbourg[5,6], Mario Campone[5,6], Maria Fe Paz[2] and François Bocquet[1,7]

Abstract

Purpose: To evaluate the effectiveness of C-LAB®, an artificial intelligence (AI) platform, in extracting, structuring, and centralizing biomarker data from breast cancer pathology reports within the challenging, heterogeneous dataset of the Institut de Cancérologie de l'Ouest (ICO).

Methods: C-LAB® was deployed at the ICO to analyze HER2 and hormonal receptor data from breast cancer pathology reports. During the development phase, 292 anatomic pathology reports were used to design and refine the rule-based extraction algorithm through an iterative process of monitoring and adjustments. After finalizing the algorithm, it was applied to a total of 2323 anatomic pathology reports. To evaluate the platform's accuracy, performance metrics could only be calculated for a subset of these reports that were also available in the structured National Epidemiological Strategy and Medical Economics (ESME) database. Out of the 2323 pathology reports belonging to 487 patients analyzed by C-LAB®, 666 corresponded to 97 patients present in the ESME database. These reports were used as the gold standard for performance assessment, as ESME provides structured data against which the outputs of the C-LAB® algorithm could be compared.

Results: C-LAB® achieved over 80% agreement with human extractions (precision, recall, and F1-score) in structuring biomarker data from complex, unstructured pathology reports, despite dataset variability and optical character recognition errors. While the ESME database served as a benchmark, its reliance on single manual data entry without secondary review introduces potential inaccuracies, suggesting the observed performance reflects close alignment between human and algorithmic extractions rather than absolute accuracy. C-LAB® demonstrates significant potential to reduce manual workload, centralize data, and enable scalable, real-time reporting.

Conclusion: AI technologies like C-LAB® show significant potential in creating accessible and actionable digital factories from complex pathology data, aiding in the precision management of diseases such as breast cancer diagnostics and treatment.

[1]Data Factory & Analytics Department, Institut de Cancérologie de l'Ouest, Nantes-Angers, France
[2]Connect By Circular Lab, Madrid, Spain
[3]Department of Radiation Oncology, Institut de Cancérologie de l'Ouest, Nantes-Angers, France
[4]Department of Pathology, Institut de Cancérologie de l'Ouest, Nantes-Angers, France
[5]Oncology Department, Institut de Cancérologie de l'Ouest, Nantes-Angers, France
[6]Center for Research in Cancerology and Immunology Nantes-Angers, Nantes University and Angers University, Nantes-Angers, France
[7]Law and Social Change Laboratory, Faculty of Law and Political Sciences, Nantes University, Nantes, France

[#]These authors contributed equally.

Corresponding author:
Florent Le Borgne, Data Factory & Analytics Department, Institut de Cancérologie de l'Ouest, 44805 Nantes-Angers, France.
Email: florent.leborgne@ico.unicancer.fr

Keywords

Cancer disease, artificial intelligence general, machine learning general, oncology medicine, biomarker, genetics medicine

Submission date: 22 August 2024; Acceptance date: 3 February 2025

Introduction

In oncology, precision medicine is well established, applying treatments tailored to the unique characteristics of each patient's cancer.[1,2] This approach relies on a detailed tumor characterization to identify features, structures, and biomarkers that targeted therapies can act upon. The relevance of precision medicine has grown over the past 20 years, particularly in breast cancer, with notable impact from new antibody-drug conjugates targeting the HER2 protein.[3–5] Diagnosis depends on HER2 staining scores in tumor samples, ranging from 0 to 1+, 2+ (requiring further ISH analysis), or 3 + . This scoring initially led to two patient categories: HER2-negative (score 0, 1+, or 2 + with ISH negative) and HER2-positive (score 2 + with ISH positive or 3+), granting access to trastuzumab (HER2-targeted monoclonal antibody) treatment.[6] Recently, new antibody-drug conjugates like trastuzumab deruxtecan have shown benefits for a new intermediate category of breast cancer patients, termed HER2-Low (score 1 + or 2 + with ISH negative).[7,8] These advances highlight the crucial role of pathologists, who must assess and adapt to evolving HER2 testing nomenclature.

Anatomic and molecular pathology labs use specialized techniques to process tissue and liquid biopsy samples, revealing morphology, genetic mutations, and protein expression. The diagnostic data reported by pathologists is essential in precision medicine and is typically recorded in the narrative format of a pathology report. For both prospective patient monitoring and retrospective research, including real-world data and quality monitoring, there is growing interest in extracting and structuring information from these reports.[9,10] This demand has spurred research aimed at creating digital factories to centralize diagnostic data, yet a gap remains in making this data accessible and actionable.

A major challenge is the lack of standardized, structured reporting in pathology labs, where report formats and content can vary widely, often including free text (Figure 1). Reports may be fully structured, semi-structured, or highly unstructured, with inconsistencies across labs, pathologists, and timeframes. To address this, initiatives like the International Collaboration on Cancer Reporting (ICCR)[11] aim to provide a unified approach, and ICCR has recently collaborated with SNOMED International[12] to standardize terminology and coding.

However, adopting structured synoptic reporting has been challenging, as new guidelines are continually released, and the diversity of cases makes it difficult for pathologists to abandon free text.[13] Free text also allows flexibility for describing rare conditions, which could be lost with rigid structuring. As a complementary approach, some initiatives[14] and registries focus on post-structuring data from narrative reports. However, manual data extraction remains common, requiring extensive time and introducing potential for human error. Semi-automated tools exist[15–18] but generally lack the scalability to adapt to the variety of unstructured data in pathology reports, typically focusing on a single biomarker.

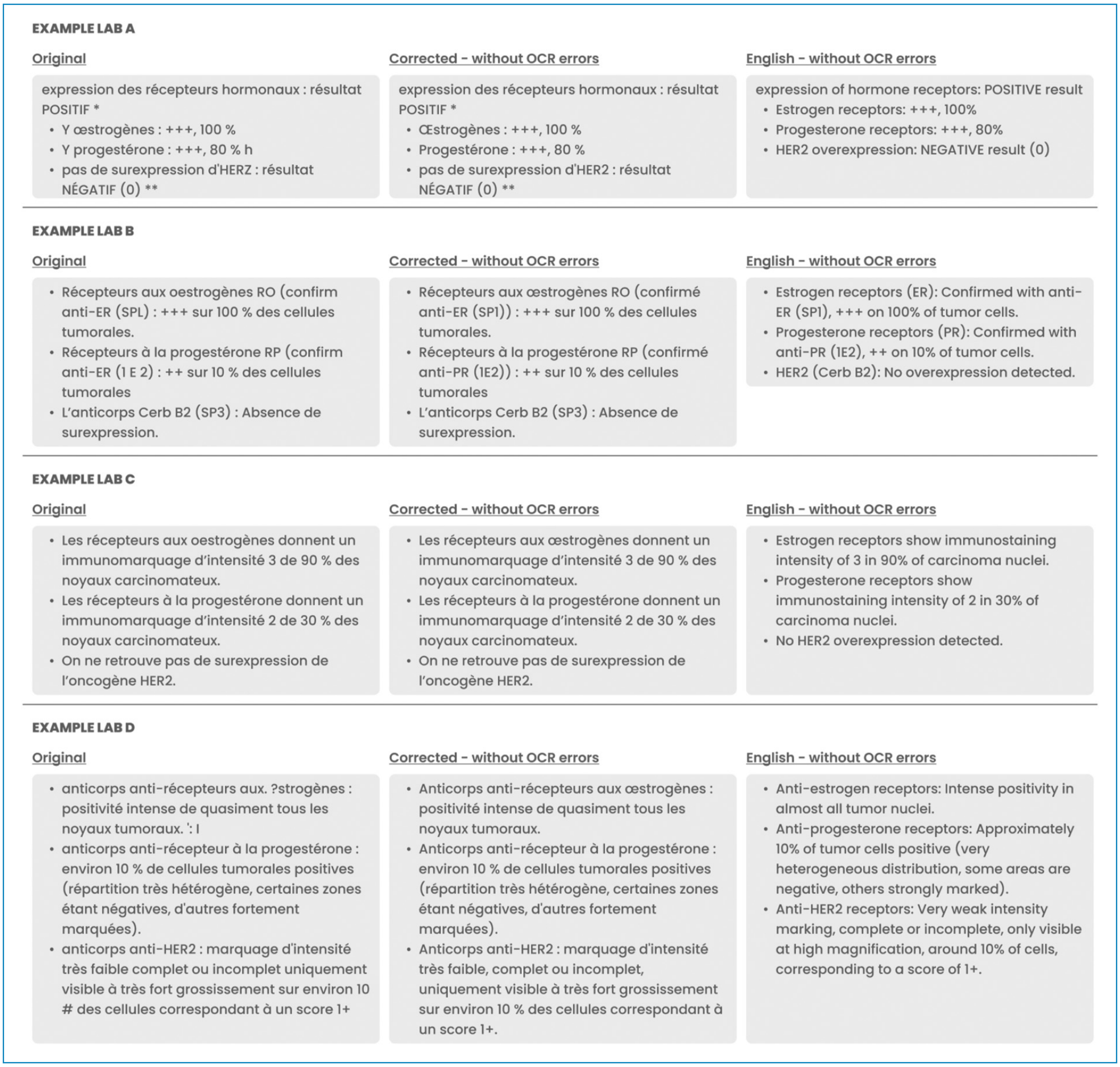
Recent advances in natural language processing (NLP), particularly deep learning and transformer models, have significantly improved information extraction from unstructured clinical data.[19] This study used a rule-based approach. It was chosen for interpretability and precision in handling heterogeneous datasets. Rule-based methods are proven effective for specific healthcare tasks, as seen in applications for extracting structured data from pathology and radiology reports.[20]

To address this infrastructure gap, Connect by Circular-Lab developed C-LAB®, an artificial intelligence (AI) platform that automates and centralizes pathology testing data. C-LAB® supports digitalization of any diagnostic data—structured or unstructured, from any indication—and provides labs with rapid, precise access to actionable digital data.

This study evaluated C-LAB®'s performance in extracting, structuring, and centralizing biomarker data from breast cancer pathology reports. C-LAB® was deployed at the Institut de Cancérologie de l'Ouest (ICO) and compared with the ESME database, a structured national database for cancer patient data from the Unicancer network in France, which served as the gold standard for this analysis. The study focused on biomarkers HER2 and hormonal receptors from 2323 anatomic pathology reports, corresponding to 487 breast cancer patients.

Materials and methods

This study was carried out at the ICO, a not-for-profit Comprehensive Cancer Center. ICO is located in the west of France (Pays de la Loire) and is part of the UNICANCER network of 18 French non-for-profit Comprehensive Cancer Centers. The study was conducted



EXAMPLE LAB A

Original

expression des récepteurs hormonaux : résultat POSITIF *

- Y œstrogènes : +++, 100 %
- Y progestérone : +++, 80 % h
- pas de surexpression d'HERZ : résultat NÉGATIF (0) **

Corrected - without OCR errors

expression des récepteurs hormonaux : résultat POSITIF *

- Œstrogènes : +++, 100 %
- Progestérone : +++, 80 %
- pas de surexpression d'HER2 : résultat NÉGATIF (0) **

English - without OCR errors

expression of hormone receptors: POSITIVE result

- Estrogen receptors: +++, 100%
- Progesterone receptors: +++, 80%
- HER2 overexpression: NEGATIVE result (0)

EXAMPLE LAB B

Original

- Récepteurs aux oestrogènes RO (confirm anti-ER (SPL) : +++ sur 100 % des cellules tumorales.
- Récepteurs à la progestérone RP (confirm anti-ER (1 E 2) : ++ sur 10 % des cellules tumorales
- L'anticorps Cerb B2 (SP3) : Absence de surexpression.

Corrected - without OCR errors

- Récepteurs aux œstrogènes RO (confirmé anti-ER (SP1)) : +++ sur 100 % des cellules tumorales.
- Récepteurs à la progestérone RP (confirmé anti-PR (1E2)) : ++ sur 10 % des cellules tumorales
- L'anticorps Cerb B2 (SP3) : Absence de surexpression.

English - without OCR errors

- Estrogen receptors (ER): Confirmed with anti-ER (SP1), +++ on 100% of tumor cells.
- Progesterone receptors (PR): Confirmed with anti-PR (1E2), ++ on 10% of tumor cells.
- HER2 (Cerb B2): No overexpression detected.

EXAMPLE LAB C

Original

- Les récepteurs aux oestrogènes donnent un immunomarquage d'intensité 3 de 90 % des noyaux carcinomateux.
- Les récepteurs à la progestérone donnent un immunomarquage d'intensité 2 de 30 % des noyaux carcinomateux.
- On ne retrouve pas de surexpression de l'oncogène HER2.

Corrected - without OCR errors

- Les récepteurs aux œstrogènes donnent un immunomarquage d'intensité 3 de 90 % des noyaux carcinomateux.
- Les récepteurs à la progestérone donnent un immunomarquage d'intensité 2 de 30 % des noyaux carcinomateux.
- On ne retrouve pas de surexpression de l'oncogène HER2.

English - without OCR errors

- Estrogen receptors show immunostaining intensity of 3 in 90% of carcinoma nuclei.
- Progesterone receptors show immunostaining intensity of 2 in 30% of carcinoma nuclei.
- No HER2 overexpression detected.

EXAMPLE LAB D

Original

- anticorps anti-récepteurs aux. ?strogènes : positivité intense de quasiment tous les noyaux tumoraux. ': I
- anticorps anti-récepteur à la progestérone : environ 10 % de cellules tumorales positives (répartition très hétérogène, certaines zones étant négatives, d'autres fortement marquées).
- anticorps anti-HER2 : marquage d'intensité très faible complet ou incomplet uniquement visible à très fort grossissement sur environ 10 # des cellules correspondant à un score 1+

Corrected - without OCR errors

- Anticorps anti-récepteurs aux œstrogènes : positivité intense de quasiment tous les noyaux tumoraux.
- Anticorps anti-récepteur à la progestérone : environ 10 % de cellules tumorales positives (répartition très hétérogène, certaines zones étant négatives, d'autres fortement marquées).
- Anticorps anti-HER2 : marquage d'intensité très faible, complet ou incomplet, uniquement visible à très fort grossissement sur environ 10 % des cellules correspondant à un score 1+.

English - without OCR errors

- Anti-estrogen receptors: Intense positivity in almost all tumor nuclei.
- Anti-progesterone receptors: Approximately 10% of tumor cells positive (very heterogeneous distribution, some areas are negative, others strongly marked).
- Anti-HER2 receptors: Very weak intensity marking, complete or incomplete, only visible at high magnification, around 10% of cells, corresponding to a score of 1+.

Figure 1. Narrative content and OCR errors in reports. This figure showcases examples of report formats and wording variations in breast biomarker results sourced from various laboratories, anonymized as Labs A, B, C, and D. These examples highlight the substantial diversity in structure and terminology used in pathology reports, particularly in biomarker data presentation, which poses significant challenges for standardized data description and analysis. The narrative data ICO accessed was processed through OCR technology, a tool designed to extract text based on character shape recognition. While OCR is effective for basic text extraction, it has inherent limitations in accurately distinguishing visually similar characters. For instance, the OCR software struggled to differentiate between "o" and "0," or "2" and "Z," resulting in transcription errors such as "HER2" being misinterpreted as "HERZ." Similarly, nuanced distinctions between accented characters, such as "à" and "a," were occasionally missed, leading to inaccuracies. Another frequent error involved symbols, where OCR occasionally rendered "%" as "#." The English translation is provided to improve clarity.

in 2022 based on retrospective reports of patients followed at ICO between January 2014 and December 2021.

ICO dataset

The ICO repository includes nearly 10 million clinical texts in French, with approximately 650,000 new reports each year, 5% of which are pathology reports. For this study, C-LAB® analyzed 2323 anatomic pathology reports from ICO's digital archive corresponding to 487 patients randomly selected from all breast cancer patients followed at ICO between January 2014 and December 2021. These reports span both ICO's internal records and cases centralized from numerous labs across western France, adding substantial variability in reporting styles and formats, which increases data extraction complexity. Figure 1 provides anonymized samples illustrating the narrative nature of the reports and optical character recognition (OCR) errors observed that made the reports even more complex. Figure 2 provides a Venn diagram of the data selection process. Among the 2323 anatomic pathology reports, 292 reports were used during the development phase to design and refine the rule-based extraction algorithm. Once finalized, the algorithm was applied to the entire set of 2323 reports. A subset of 666 reports belonged to 97 patients also included in the national ESME database; these served as our "gold standard" for comparative performance analysis.

ESME database

The ESME cohort, initiated in 2014 by Unicancer, aims to centralize real-world data on metastatic breast cancer through a comprehensive dataset collected across 18 French Comprehensive Cancer Centers. This academic initiative aggregates data from patient records, including demographics and treatment details, from electronic medical records. More detailed information on the ESME cohort is available in prior publications.[21,22] Among the ICO reports processed by C-LAB®, 666 correspond to cases also present in the ESME database, which is used as the gold standard for comparison in this study. However, as ESME relies on single manual data entry, there may be potential inaccuracies, suggesting that C-LAB®'s actual performance could be even higher than indicated by this benchmark. The ESME database represents a benchmark for evaluation, but it is not without potential errors due to its reliance on manual data entry without secondary review. Therefore, the comparison between C-LAB® and ESME reflects how closely the model aligns with human interpretations, not an absolute measure of correctness.

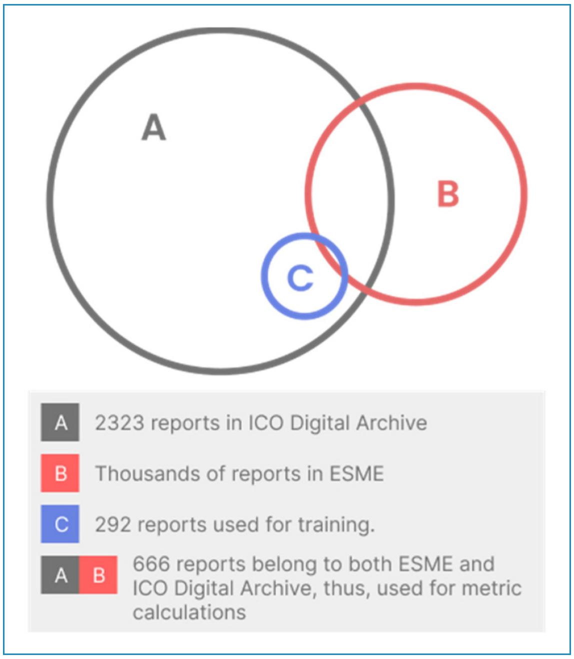

A
B
C
A 2323 reports in ICO Digital Archive
B Thousands of reports in ESME
C 292 reports used for training.
A B 666 reports belong to both ESME and ICO Digital Archive, thus, used for metric calculations

Figure 2. Venn diagram of study cohort and report selection.

C-LAB® platform and data processing

C-LAB® is an AI-driven platform developed by Connect by Circular-Lab to digitize, extract, centralize, and structure diagnostic and biomarker data from structured and unstructured pathology reports in real time. This automation enables prospective and retrospective data monitoring, supports real-world evidence generation, and facilitates drug efficacy assessments and post-market surveillance.

The platform uses NLP techniques, specifically rule-based algorithms, to anonymize, extract, and structure relevant data from pathology reports. C-LAB® accepts various file formats, including PDFs with text, Word, and Excel files, enabling flexibility across reporting styles.

Central to C-LAB®'s data extraction is a predefined ontology—a structured set of specific entities, or tags, tailored to each medical indication. In this study, C-LAB® extracted and normalized under specific tags multiple variables, further compared with the ESME database, as shown in Table 1. C-LAB® has not inferred or interpreted any data beyond what is explicitly stated in the pathologist's report. For instance, the Estrogen and Progesterone Immunohistochemistry (IHC) results are determined solely by the presence of the words "negative" or "positive" and do not consider staining percentage thresholds. For information, pathologists in France use the 10% staining threshold to conclude on positivity of estrogen receptors (ERs) or progesterone receptors (PRs), as opposed to the US practice where the 1% threshold is used.[23,24]

Connect by Circular-Lab has developed an annotation tool to streamline the training process, allowing a training set of reports to be annotated, which provides the basis for C-LAB®'s rule-based NLP system. The extraction algorithm is built on regex and token-based pattern matching using SpaCy's, an open-source Python library, Matcher module, which supports flexible recognition of clinical narrative variations. SpaCy's robust NLP capabilities, including tokenization and customizable entity recognition, allow C-LAB® to scale across different medical report formats with accuracy. SpaCy divides text into "tokens," which are individual units such as words, punctuation marks, or numbers, enabling a fine-grained analysis of the language structure. By creating rules with ordered sequences of tokens, the system can accurately capture entities across various text formats and clinical language variations. Token-based patterns allow flexible matching across phrasing variations, which is particularly useful in clinical narratives. Once the ontology is established, C-LAB® enables the creation of inter-tag relationships

Table 1. List of tags (ontology) extracted by C-LAB® on ICO anatomic pathology reports and compared with the ESME database.

Tag	Definition	Normalization
% Estrogen IHC	The percentage of cells in breast tissue sample that show positive staining for estrogen receptors using Immunohistochemistry (IHC).	#%
% Progesterone IHC	The percentage of cells in breast tissue sample that show positive staining for progesterone receptors using IHC.	#%
Estrogen IHC result	The overall result indicating whether estrogen receptors are present in the breast tissue sample based on IHC analysis.	Positive/negative
Her2 IHC result	The score indicating the level of HER2 protein expression in breast tissue sample using Immunohistochemistry (IHC).	0 / 1+ / 2+ / 3+
Patient number	A unique identifier assigned to a patient for tracking and record-keeping purposes.	#
Progesterone IHC result	The overall result indicating whether progesterone receptors are present in the breast tissue sample based on IHC analysis.	Positive/negative
Result ISH Her2	The result of the In Situ Hybridization (ISH) test for HER2 gene amplification in the breast tissue sample.	Amplified/non amplified
Sample date	Date of retrieval of sample material (when the sample is received in the lab)	Day/month/year

that capture clinical dependencies between tags, enhancing data standardization and minimizing interpretation bias. Although these relationships were not applied in this study, which focused solely on raw data extraction, they represent an important feature for enriching data connections in clinical practice.

We followed a sequential process to develop the rule-based algorithm, utilizing 292 reports to design and refine the extraction rules. This iterative approach involved careful monitoring and adjustments to ensure the model's robustness. The finalized model was then used in inference to extract data from the rest of the reports. To calculate the performance metrics we relied on an existing gold standard, the ESME database, which, despite its limitations, provided a sufficient benchmark to verify the model's performance.

Comparative analysis

In order to assess the quality of the data produced by the C-LAB® algorithm, we compared the predictions from the C-LAB® algorithm with the manually collected data present in the ESME database. The comparison focuses on four biomarkers: HER2, ISH, Estrogen, and Progesterone. For each biomarker, final results and staining percentage, when available, were evaluated independently. Note that only mention of tags combined with results were considered (i.e. tags without associated results are deleted). For each biomarker independently, the results found by the algorithm and those present in ESME were matched by date. As biomarker measurements are longitudinal data, to be able to compare the results between two databases it is necessary to first reconcile the measurements by patient and by date of measurement. Although the same reports used for manual entry were provided to the C-LAB® algorithm, the dates do not correspond exactly (entry error, taking into account the date of sampling or the date of result or the date of the report, etc.). Matching was done so that the results that can be matched with an exact date were matched, then among the remaining unmatched results those that can be matched with a difference of 1 day were matched, and so on up to a maximum accepted difference of 15 days. Each result present in ESME or found by the C-LAB® algorithm can only be matched once.

Matched biomarker results were considered true positives (TP) if the result found by C-LAB® was equal to the one present in ESME. If the result found by C-LAB® was different from the one present in ESME, C-LAB® result was counted as false positive (FP) and false negative (FN) as it needs to be counted as an error in both Precision and Recall. Unmatched C-LAB® results were considered FP and unmatched ESME results were considered FN. Since true negatives (TN) represent cases where the algorithm correctly did not tag something as an entity, these cases usually do not contribute directly to evaluating the algorithm's performance in detecting entities.

Classification	Definition
True Positives (TP)	C-LAB®'s result matches the result in ESME.
False Positives (FP)	C-LAB® identifies a result not in ESME. C-LAB®'s result differs from ESME (affecting Precision).
False Negatives (FN)	Result is in ESME but not identified by C-LAB®. C-LAB®'s result differs from ESME (affecting Recall).
True Negatives (TN)	Neither C-LAB® nor ESME identify a result (excluded from performance evaluation).

The performance of the C-LAB® algorithm was estimated in terms of Precision (or positive predictive value), Recall (or sensitivity) and F1 score for each biomarker as follows:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\ Score = 2 \times \frac{Precision\ \times Recall}{Precision + Recall}$$

The 95% confidence intervals (95% CI) were obtained by nonparametric bootstrap (1000 iterations).

Results

The results aggregated for all patients are presented in Table 2.

Among the 666 reports used as gold standard, associated with 98 metastatic breast cancer patients there are between 220 and 317 results per biomarker except for ISH where there are only 43 results. For example, for the HER2 biomarker there are between 0 and 7 results by patient with a median equal to three measures by patient. The results show a percentage of recall varying from 70% to 82% for all biomarkers tags analyzed, from 73% to 82% for precision and from 71% to 81% for F1 Score, revealing a significantly higher amount of true positive extractions than FP and FN ones. The results are close in terms of recall and precision for both ER and PR biomarkers whether in staining percentage or binary final results (positive, negative) with results close to 80% except for the PR biomarker in staining percentage with a precision equal to 73.3% (95% CI from 67.8% to 78.8%). The results for the HER2 biomarker (desired result in the form 0, 1+, 2+, 3+) were slightly lower with a recall of 74.6% (95% CI from 69.7% to 79.4%), a precision of 77.8% (95% CI from 72.9% to 82.9%) and an F1 score of 76.1% (95% CI from 72.1% to 79.9%). The ISH biomarker, which is only searched for patients with HER2 status equal to 2+, was associated with the lowest results (i.e. F1 score was equal to 71.4% (95% CI from 60.9% to 81.3%)). The matching temporal distances are presented in Figure 3.

Discussion

In precision medicine, especially in breast cancer, anatomic pathology testing has a central role to support the access of patients to the right treatment. Furthermore, the increasing development of antibody-drug conjugates which are often associated with a specific biomarker IHC test, reinforces even more the importance of pathology labs. Breast cancer illustrates this evolution well, with the arrival of the HER2-Low concept associated with the prescription

Table 2. Aggregated results from C-LAB®-ESME comparison for the biomarker of interest.

Biomarker	TP	FP (unmatched)	FN (unmatched)	Precision	Recall	F1 score
HER2 (0, 1+, 2+, 3+)	217	62 (53)	74 (65)	77.8 [72.9–82.9]	74.6 [69.7–79.4]	76.1 [72.1–79.9]
ISH (amplified, not amplified)	30	11 (11)	13 (13)	73.2 [60.6–86.1]	69.8 [56–84.1]	71.4 [60.9–81.3]
ER (positive, negative)	250	55 (50)	67 (62)	82.0 [77.9–85.9]	78.9 [73.7–83.9]	80.4 [76.6–84.2]
ER (number in %)	214	56 (37)	46 dfd(47)	79.3 [74.6–83.5]	82.3 [77.4–86.7]	80.8 [77.1–84.0]
PR (positive, negative)	249	59 (53)	58 (52)	80.8 [76.8–84.9]	81.1 [76.5–85.3]	81.0 [77.8–83.8]
PR (number in %)	181	66 (58)	39 (31)	73.3 [67.8–78.8]	82.3 [77.4–87.1]	77.5 [73.5–81.5]

For each biomarker (HER2 score, ISH result, ER and PR final results, and ER and PR staining percentage), true positive (TP), false positive (FP), and false negative (FN) have been calculated together with recall, precision and F1 score including 95% confidence intervals. In the FP and FN columns are indicated in parentheses the number of FP and FN due to unmatched C-LAB® results and unmatched ESME results, respectively. The other FP and FN are matched entities but with a different biomarker result.

of trastuzumab-deruxtecan, drawing new attention on the HER2 biomarker IHC test among the pathology community.[25]

In spite of the increasing digitalization of the pathology labs all over the world,[26,27] the last step of the testing workflow, corresponding to the reporting of the diagnostic method and results by the pathologist to the clinician, is associated with gaps in terms of digitalization, structuring, centralization, monitoring, actionability and archiving of reports data, as shown in Figure 4.

This gap in the workflow represents a challenge to access easily and timely both retrospective and prospective testing data which can impact testing quality optimization, patient outcome, and monitoring. Also, it potentially leads to delayed and/or non-appropriate treatment, unequal access of patients to diagnostics and treatment due to the lack of reactivity upon testing deviation, higher time and cost of research and clinical studies. Even in centers where reports are digitized, it often ends up with large data factories with several years of retrospective unstructured data that are barely accessible, often requiring manually transfer data of interest into excel tables or other manual and time-consuming solutions also associated with possible human errors during manual data transfer.

In this context, our comparison study between the AI-based C-LAB® digital platform and the analogic ESME database is essential to address the infrastructural gap in the digitalization, centralization, and structuration of diagnostic data to make digital factories more accessible and actionable. The key being here to enable a more automated and accurate extraction and structuring of pathology report's data toward enhanced real-time monitoring.

Our results show that the C-LAB® algorithm achieves recall (the ability to identify all relevant data) and precision (the ability to avoid FPs) rates close to 80% for extracting and structuring most hormonal receptor and HER2 status data. However, the performance for ISH was slightly lower, with an F1 score of 71.4% and precision of 73.2%. This reflects the inherent challenges of detecting less common data points, especially within unstructured and OCR-processed reports. The acceptable level of errors for an AI-based algorithm depends on its intended use. For tasks like improving the efficiency of manual data entry, high recall is crucial to ensure that as much relevant data as possible is captured, even if it means reviewing additional results due to lower precision. In contrast, for applications such as real-world evidence studies, where data must be highly accurate, the tolerance for errors is much lower. Balancing recall and precision during algorithm development is essential, as it affects both the completeness and the accuracy of the extracted data, and this tradeoff must align with the specific goals of the application. Note that, C-LAB® includes internal alerts to flag potential extraction errors, reinforcing quality control. These alerts ensure timely identification and correction of extraction errors, further enhancing the platform's reliability and accuracy.

However, this work has several limitations which should be noted. Firstly, we did not separate the data into training and validation sets, so we could not exclude overoptimistic results due to overfitting. At the time of C-LAB®'s application to ICO reports, Circular-Lab was not aware of the patients included in ESME, excluding an over selection of reports of patients included in ESME in the training set. Secondly, we used the ESME database as a reference (i.e. gold standard), although this is a cohort for which data have been entered manually by a single person without second person review. Some discrepancies may therefore be due to an error in the ESME database and not to the C-LAB® platform, particularly for FP results. An improvement to this work would be to have a gold standard independent of the learning set and with double entry and reviewing of inconsistencies in order to ensure that the differences could be attributed to AI errors. However, our current gold standard with its errors is to the disadvantage of our algorithm because it may underestimate the performance of C-LAB® platform. Thirdly, it is also important to mention that in this study we matched the measurements with a maximum deviation of 15 days, an arbitrary threshold which we consider acceptable. Note that, most of the matched measurements are matched on exact dates or with a difference of 1 day (Figure 3).

Our findings are consistent with those reported by Santos et al.,[19] which also highlighted the challenges of extracting information from unstructured pathology reports. In our experiment, F1 scores for entity extraction ranged from 71.4% to 81.0%, closely aligning with results from other studies in the field. For example, Mendoza-Urbano et al.,[28] which focused on oncology pathology reports, reported F1-scores ranging from 52.9% to 100%. Similarly, Yoon et al.[29] examining pre-anesthesia evaluation summaries, found F1-scores between 65.4% and 77.2%, while Wieneke et al.[17] achieved F1-scores of 80.0%, 92.0%, and 50.0% for extracting procedure, laterality, and result entities from pathology reports, respectively. These comparisons demonstrate that our rule-based algorithm performs comparably to other NLP approaches, including machine learning methods, despite the added complexities of unstructured data and OCR-related errors.

The reports' samples as shown in Figure 1 show how unstructured some reports can be, which reinforce the need for AI and demonstrate the power of C-LAB® to reach good extraction accuracy even with such unstructured reports. On top of it, odd characters such as omega or spaces into words or mis-wordings can arise in some reports due to OCRization processes.

While rule-based approaches offer extensive customization and handling of OCR-related issues but can be

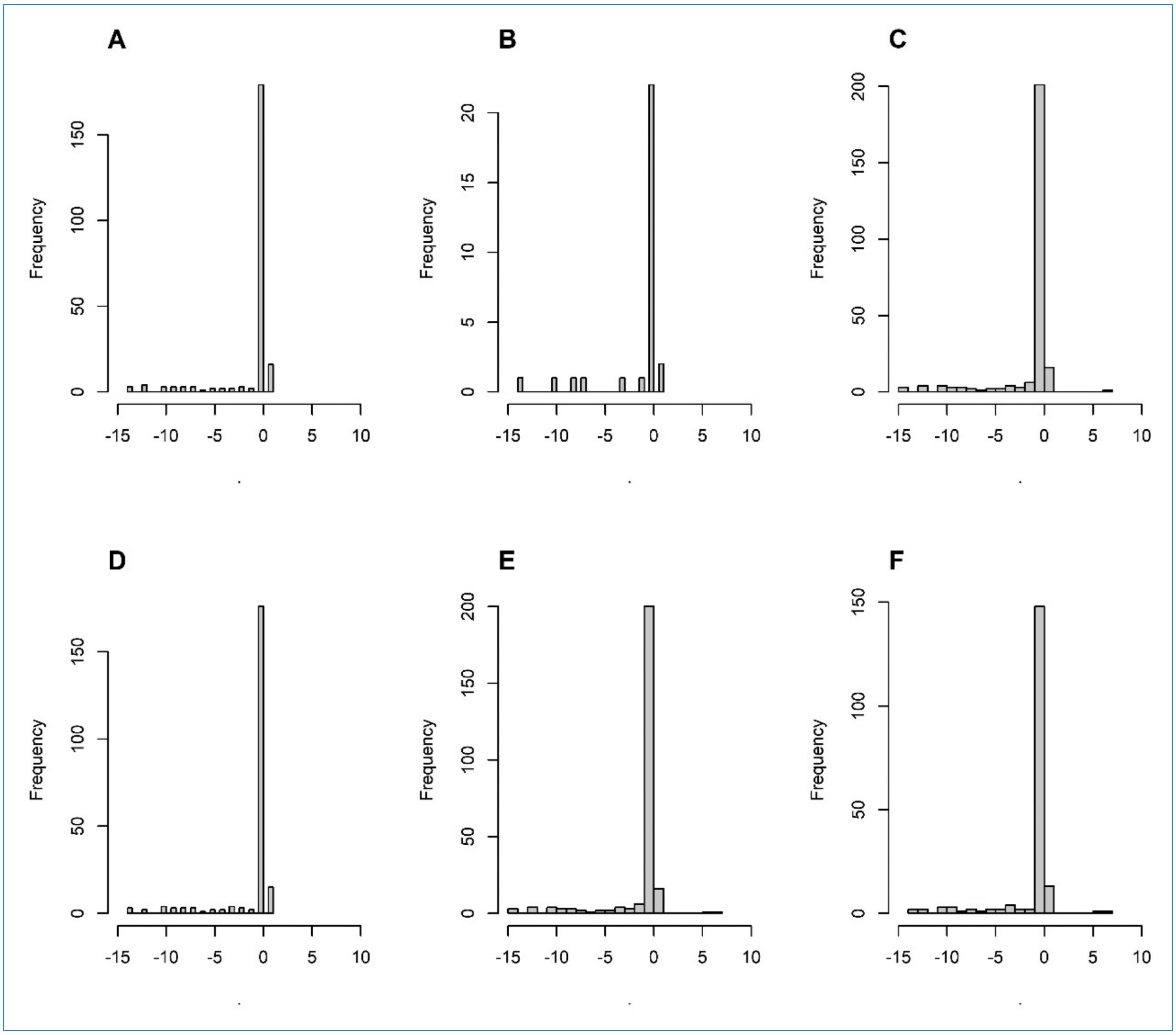


Figure 3. Distribution of matching temporal distances between C-LAB® results and ESME results. Positive distances mean that the date of the result found by C-LAB® is earlier than the date of the result present in ESME. A: HER2; B: ISH; C: ER (Positive, Negative), D: ER (percentage); E: PR (Positive, Negative), F: PR (percentage).

associated with manual efforts and be vulnerable to minor shifts in language,[19] deep learning methods, which can model intricate relationships between words and labels,[19] represent an interesting area of research for future development.

In the end, this study shows the promising capacity of NLP AI technologies such as rule-based systems in extracting and structuring complex testing data to create digital factories that are accessible and actionable and better connected to healthcare stakeholders. As a next step, it would be good to aggregate the results of the three biomarkers (ER, PR, HER2) as they are all three required to make therapeutic decisions.

In line with this, such a platform can be highly beneficial to retrieve easily and quickly reported information when a change occurs in the testing nomenclature. For example, in the context of breast cancer and the arising new HER2-Low tumor entity, having a tool that can track diagnostic scores precisely and in real time will be key to optimize controlled implementation of new testing recommendations and patient access to new therapies. C-LAB® can facilitate the management of retrospective and prospective studies by monitoring HER2-Low prevalence and even extrapolating potential HER2-Low patients from the reported staining and scoring data.

Conclusion

This study demonstrates the significant potential of the C-LAB® AI platform in automating the extraction, structuring, and centralization of complex biomarker data from

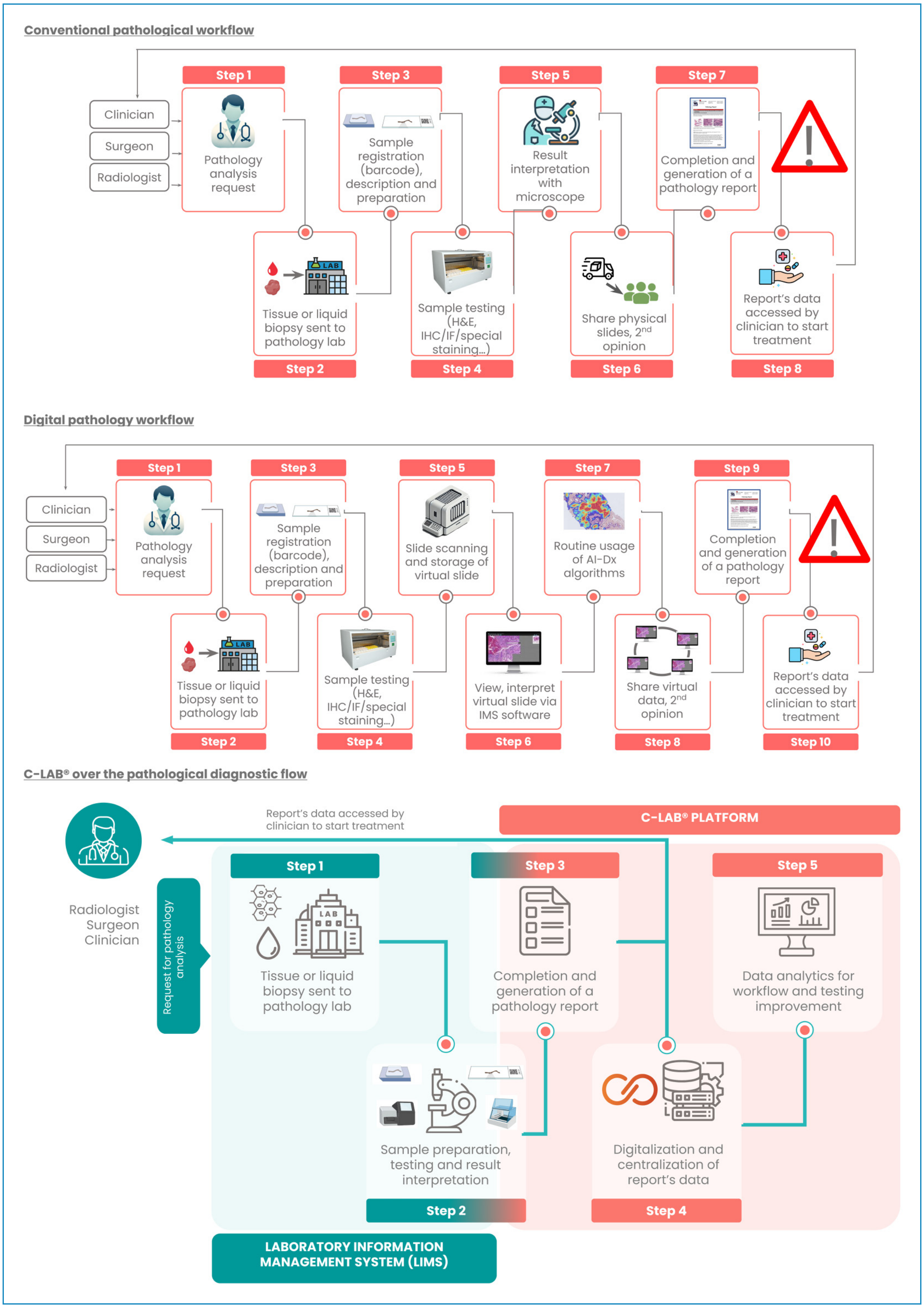

Conventional pathological workflow
Clinician
Surgeon
Radiologist
Step 1
Pathology analysis request
Step 2
Tissue or liquid biopsy sent to pathology lab
Step 3
Sample registration (barcode), description and preparation
Step 4
Sample testing (H&E, IHC/IF/special staining...)
Step 5
Result interpretation with microscope
Step 6
Share physical slides, 2nd opinion
Step 7
Completion and generation of a pathology report
Step 8
Report's data accessed by clinician to start treatment
Digital pathology workflow
Clinician
Surgeon
Radiologist
Step 1
Pathology analysis request
Step 2
Tissue or liquid biopsy sent to pathology lab
Step 3
Sample registration (barcode), description and preparation
Step 4
Sample testing (H&E, IHC/IF/special staining...)
Step 5
Slide scanning and storage of virtual slide
Step 6
View, interpret virtual slide via IMS software
Step 7
Routine usage of AI-Dx algorithms
Step 8
Share virtual data, 2nd opinion
Step 9
Completion and generation of a pathology report
Step 10
Report's data accessed by clinician to start treatment
C-LAB® over the pathological diagnostic flow
Report's data accessed by clinician to start treatment
C-LAB® PLATFORM
Radiologist
Surgeon
Clinician
Request for pathology analysis
Step 1
Tissue or liquid biopsy sent to pathology lab
Step 2
Sample preparation, testing and result interpretation
Step 3
Completion and generation of a pathology report
Step 4
Digitalization and centralization of report's data
Step 5
Data analytics for workflow and testing improvement
LABORATORY INFORMATION MANAGEMENT SYSTEM (LIMS)

Figure 4. Conventional versus digital pathology workflow and C-LAB® positioning. The current digitalization of the pathology workflow shows a gap in addressing the reporting and monitoring of testing data.

unstructured pathology reports in breast cancer patients. By achieving precision and recall rates close to 80% for key biomarkers such as hormonal receptors and HER2 status, C-LAB® proves to be a valuable tool in bridging the infrastructural gap in the digitalization of pathology data. The platform's ability to handle heterogeneous and unstructured reports, including those with OCR-induced errors, underscores its robustness and adaptability, particularly in the context of precision medicine.

By transforming narrative reports into a centralized digital format, this study demonstrates the platform's capability to support data accessibility and provide a foundation for harmonized reporting. It also demonstrates the ability to adapt efficiently to changes in testing nomenclature and guidelines, making it a promising tool in the evolving field of cancer diagnostics and treatment. For instance, in the context of the emerging HER2-Low classification in breast cancer, the platform could aid in identifying and monitoring patient populations that may benefit from new therapeutic options. While these findings illustrate the promise of this technology in facilitating more efficient data handling, further validation and analysis are necessary to accurately assess its performance and to fully assess its impact on patient outcomes and research initiatives.

In conclusion, AI-driven solutions like C-LAB® hold significant promise in enhancing the precision and efficiency of pathology data management. By overcoming the challenges associated with unstructured reporting and OCR limitations, such platforms could play a pivotal role in advancing personalized medicine and improving patient care outcomes in oncology.

Contributorship: FLB, CG, TP, OK, JSF, JR, MC, MFP, and FB were responsible for the conceptualization. CG and CM were in charge of the data curation. CG, MFP, and FB were responsible for funding acquisition. FLB, CG, and CM conducted investigation. FLB, CG, YN, YE, JM, JC, and MFP designed the methodology. CG, MFP, and FB were responsible for project administration. CG and MFP provided resources. CG, YN, YE, and JM designed the software. CG, MFP, and FB supervised all the research project. CG and MFP proposed visualization. FLB performed comparative analysis. FLB and CG wrote the initial draft of the article. All the authors reviewed and edited the article and approved the final manuscript.

Declaration of conflicting interests: The authors declared the following potential conflicts of interest with respect to the research, authorship, and/or publication of this article: FLB, CM, TP, OK, JSF, JR, MC and FB declare no conflict of interest. CG, YN, YE, JM, JC, MFP declare a potential conflict of interest as they were involved in both the development of the C-LAB® platform and the authorship of this study, as consultants or employees of Connect by Circular-Lab. However, every effort was made to ensure the integrity and objectivity of the research, including adherence to rigorous evaluation protocols. Note that none of these people had access to the ESME data and that the evaluation of the algorithm's performance was carried out by FLB, which has no connection with Connect by Circular-Lab.

Ethical approval: ICO Data Protection Officer included the study in the register of ongoing studies as requested by the Commission nationale de l'informatique et des libertés (National Commission for Data Protection and Liberties). All the patients included for the study have signed a generic research consent form. At any time, the patient can object to its participation in the study by contacting the ICO Data Protection Officer. A data processing sheet containing the essential information of the study and the data collected is public and available online for all patients at mesdonnees.unicancer.fr.

Funding: The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iD: Florent Le Borgne https://orcid.org/0000-0003-2361-1608

References

1. Jameson JL and Longo DL. Precision medicine–personalized, problematic, and promising. *N Engl J Med* 4 juin 2015; 372: 2229–2234.
2. Garrido P, Aldaz A, Vera R, et al. Proposal for the creation of a national strategy for precision medicine in cancer: a position statement of SEOM, SEAP, and SEFH. *Clin Transl Oncol* avr 2018; 20: 443–447.
3. Harbeck N and Gnant M. Breast cancer. *Lancet* 18 mars 2017; 389: 1134–1150.
4. Kunte S, Abraham J and Montero AJ. Novel HER2-targeted therapies for HER2-positive metastatic breast cancer. *Cancer* 1 oct 2020; 126: 4278–4288.
5. Gámez-Chiachio M, Sarrió D and Moreno-Bueno G. Novel therapies and strategies to overcome resistance to anti-HER2-targeted drugs. *Cancers (Basel)* 19 sept 2022; 14: 4543.
6. Hurvitz SA, Hegg R, Chung WP, et al. Trastuzumab deruxtecan versus trastuzumab emtansine in patients with HER2-positive metastatic breast cancer: updated results from DESTINY-Breast03, a randomised, open-label, phase 3 trial. *Lancet* 14 janv 2023; 401: 105–117.
7. Modi S, Jacot W, Yamashita T, et al. Trastuzumab deruxtecan in previously treated HER2-low advanced breast cancer. *N Engl J Med* 7 juill 2022; 387: 9–20.
8. Tarantino P, Viale G, Press MF, et al. ESMO Expert consensus statements (ECS) on the definition, diagnosis, and management of HER2-low breast cancer. *Ann Oncol* août 2023; 34: 645–659.
9. Lam H, Nguyen F, Wang X, et al. An accessible, efficient, and accurate natural language processing method for extracting diagnostic data from pathology reports. *J Pathol Inform* 2022; 13: 100154.
10. Odisho AY, Bridge M, Webb M, et al. Automating the capture of structured pathology data for prostate cancer

clinical care and research. *JCO Clin Cancer Inform* 17 juill 2019; 3: CCI.18.00084.

11. Ellis IO, Rakha EA, Tse GM, et al. An international unified approach to reporting and grading invasive breast cancer. an overview of the international collaboration on cancer reporting (ICCR) initiative. *Histopathology* janv 2023; 82: 189–197.
12. Hufstedler H, Roell Y, Peña A, et al. Navigating data standards in public health: a brief report from a data-standards meeting. *J Glob Health* 5 avr 2024; 14: 03024.
13. Sluijter CE, van Lonkhuijzen LRCW, van Slooten HJ, et al. The effects of implementing synoptic pathology reporting in cancer diagnosis: a systematic review. *Virchows Arch* juin 2016; 468: 639–649.
14. Lee J, Song HJ, Yoon E, et al. Automated extraction of biomarker information from pathology reports. *BMC Med Inform Decis Mak* 21 mai 2018; 18: 29.
15. Zheng S, Lu JJ, Appin C, et al. Support patient search on pathology reports with interactive online learning based data extraction. *J Pathol Inform* 1 janv 2015; 6: 51.
16. Buckley JM, Coopey SB, Sharko J, et al. The feasibility of using natural language processing to extract clinical information from breast pathology reports. *J Pathol Inform* 1 janv 2012; 3: 23.
17. Wieneke AE, Bowles EJA, Cronkite D, et al. Validation of natural language processing to extract breast cancer pathology procedures and results. *J Pathol Inform* 1 janv 2015; 6: 38.
18. Yala A, Barzilay R, Salama L, et al. Using machine learning to parse breast pathology reports. *Breast Cancer Res Treat* 1 janv 2017; 161: 203–211.
19. Santos T, Tariq A, Gichoya JW, et al. Automatic classification of cancer pathology reports: a systematic review. *J Pathol Inform* 2022 Jan 20; 13: 100003. PMID: 35242443; PMCID: PMC8860734.
20. Lam H, Nguyen F, Wang X, et al. An accessible, efficient, and accurate natural language processing method for extracting diagnostic data from pathology reports. *J Pathol Inform* 2022; 13: 100154.
21. Pérol D, Robain M, Arveux P, et al. The ongoing French metastatic breast cancer (MBC) cohort: the example-based methodology of the epidemiological strategy and medical economics (ESME). *BMJ Open* 21 févr 2019; 9: e023568.
22. Grinda T, Antoine A, Jacot W, et al. Evolution of overall survival and receipt of new therapies by subtype among 20 446 metastatic breast cancer patients in the 2008-2017 ESME cohort. *ESMO Open Juin* 2021; 6: 100114.
23. *Agence de la Biomédecine*. Mise à jour des recommandation concernant l'activité de prélèvement et de greffe d'organes et de tissus durant l'épidémie du coronavirus le SARS-CoV-2 [Internet]. 2020 [cité 2 juin 2020]. Disponible sur: https://www.agence-biomedecine.fr/Recommandation-concernant-l-activite-de-prelevement-et-de-greffe-d-organes-et-1314.
24. Allison KH, Hammond MEH, Dowsett M, et al. Estrogen and progesterone receptor testing in breast cancer: ASCO/CAP guideline update. *J Clin Oncol* 20 avr 2020; 38: 1346–1366.
25. Shirman Y, Lubovsky S and Shai A. HER2-Low Breast cancer: current landscape and future prospects. *Breast Cancer (Dove Med Press)* 2023; 15: 605–616.
26. Schwen LO, Kiehl TR, Carvalho R, et al. Digitization of pathology labs: a review of lessons learned. *Lab Invest* nov 2023; 103: 100244.
27. Grobholz R, Janowczyk A, Frei AL, et al. National digital pathology projects in Switzerland: a 2023 update. *Pathologie (Heidelb)* déc 2023; 44: 225–228.
28. Mendoza-Urbano DM, Garcia JF, Moreno JS, et al. Automated extraction of information from free text of Spanish oncology pathology reports. *Colomb Med (Cali)* 2023; 54: e2035300.
29. Yoon SB, Lee J, Lee HC, et al. Comparison of NLP machine learning models with human physicians for ASA physical Status classification. *NPJ Digit Med* 28 sept 2024; 7: 259.