original reports

Text Mining in Electronic Medical Records Enables Quick and Efficient Identification of Pregnancy Cases Occurring After Breast Cancer

Julie Labrosse, MD[1]; Thanh Lam, MD[2]; Clara Sebbag, MD[1]; Milena Benque, MD[1]; Ines Abdennebi, MD[1]; Hilde Merckelbagh, MD[3]; Marie Osdoit, MD[1]; Maël Priour[1]; Julien Guerin[1]; Thomas Balezeau[1]; Beatriz Grandal, MD[1]; Florence Coussy, MD[1]; Angélique Bobrie, MD[4]; Loïc Ferrer, PhD[5]; Enora Laas, MD[1]; Jean-Guillaume Feron, MD[1]; Fabien Reyal, MD, PhD[6]; and Anne-Sophie Hamy, MD, PhD[6]

abstract

PURPOSE To apply text mining (TM) technology on electronic medical records (EMRs) of patients with breast cancer (BC) to retrieve the occurrence of a pregnancy after BC diagnosis and compare its performance to manual curation.

MATERIALS AND METHODS The training cohort (Cohort A) comprised 344 patients with BC age ≤ 40 years old treated at Institut Curie between 2005 and 2007. Manual curation consisted in manually reviewing each EMR to retrieve pregnancies. TM consisted of first applying a keyword filter ("accouch*" or "enceinte," French terms for "deliver*" and "pregnant," respectively) to select a subset of EMRs, and, second, checking manually EMRs to confirm the pregnancy. Then, we applied our TM algorithm on an independent cohort of patients with BC treated between 2008 and 2012 (Cohort B).

RESULTS In Cohort A, 36 pregnancies were identified among 344 patients (10.5%; 2,829 person-years of EMR). Thirty were identified by manual review versus 35 by TM. TM resulted in a lower percentage of manual checking (26.7% *v* 100%, respectively) and substantial time gains (time to identify a pregnancy: 13 minutes for TM *v* 244 minutes for manual curation, respectively). Presence of any of the two TM filters showed excellent sensitivity (97%) and negative predictive value (100%). In Cohort B, 67 pregnancies were identified among 1,226 patients (5.5%; 7,349 person-years of EMR). Similarly, for Cohort B, TM spared 904 (73.7%) EMRs from manual review and quickly generated a cohort of 67 pregnancies after BC. Incidence rate of pregnancy after BC was 0.01 pregnancy per person-year of EMR in both cohorts.

CONCLUSION TM is highly efficient to quickly identify rare events and is a promising tool to improve rapidity, efficiency, and costs of medical research.

INTRODUCTION

Offering a major potential to improve safety, quality, and efficiency of health care, electronic medical records (EMRs) have become the standard for patient information storage. Containing various sources of clinical data (including demographics, diagnostic history, medications, laboratory test results, and vital signs), EMRs are used for information input, storage, display, retrieval, and sharing.[1] EMRs appear as a treasure trove for large-scale analysis of health data and are increasingly used for purposes of clinical and translational research.[2-4]

Although EMRs store a lot of valuable medical information, the complexity of data processing and analysis is increased by the lack of common structural frameworks. Data in EMRs are mostly present in the form of unstructured text containing errors (improper grammatical use, spelling, local dialects, and semantic ambiguities). So far, medical records were manually reviewed to retrieve data, resulting in considerable time, finances, and human efforts. Collecting follow-up data was particularly difficult, since these dynamic data evolve throughout time and have to be continuously updated to avoid being outdated. Hence, a large fraction of subjects lack follow-up available to ascertain whether they experienced the health outcome or adverse event of interest within a given time period.[5]

Text mining (TM) technology has emerged as a solution to accelerate data review by bridging the gap between free text and structured information representation. TM uses computational technologies, such as natural language processing, knowledge management, data mining, and machine learning, to efficiently identify patients, diseases, or specific terms.[6,7] Although its application has been proved successful in identifying information from biomedical and clinical sources,[2] so far, studies analyzing TM performance were performed on small sample sizes and suggested that detection errors, the lack of robust EMRs, and

ASSOCIATED CONTENT

Appendix

Author affiliations and support information (if applicable) appear at the end of this article.

Accepted on August 5, 2019 and published at ascopubs.org/journal/cci on October 18, 2019: DOI https://doi.org/10.1200/CCI.19.00031

CONTEXT

Key Objective

To apply text mining technology on electronic medical records to identify a rare event (presence of a pregnancy after breast cancer) and compare results obtained to those obtained by manual curation.

Knowledge Generated

Text mining outperformed manual curation. Text mining methods considerably reduced reviewer burden with a minimal loss of true cases and enabled substantial time gains compared with manual curation.

Relevance

Text mining technology offers the possibility to quickly browse and merge data. Text mining appears to be highly efficient to quickly generate cohorts of patients who experience a rare event, resulting in substantial cost savings. Filling defaults, smoothing noise, and using standardized, structured medical records will improve its performance. Given the promising impact of computational technologies, it opens a wide range of new perspectives for research purposes and future in silico discoveries.

likely missed information resulting from unrecorded data could limit TM technology's efficiency.[8,9]

Pregnancy following breast cancer (BC) is a rare event, occurring in 3% to 10% of patients with BC,[10] and is a topic of growing interest for research purposes. However, because it can occur from diagnosis until the end of reproductive life, retrieving the occurrence of a pregnancy after BC is a challenge that requires the full review of medical charts and subsequent cohort updates.

The aim of the current study was to investigate if TM technology could be an alternative to manual curation to retrieve the occurrence of pregnancies after BC diagnosis in a cohort of patients with BC.

MATERIALS AND METHODS

Patients and Tumors

We analyzed a cohort of 344 female patients with BC age ≤ 40 years treated at Institut Curie between January 1, 2005 and December 31, 2007 (Cohort A). We chose 40 years old as age limit given the very low expected pregnancy rates after BC reported after that age[11] and in accordance with the cutoff used in previous studies.[12-14] Furthermore, because of the important and age-increasing incidence of BC, extending the age cutoff would have brought very few additional pregnancies while considerably increasing the number of EMRs to review manually, which would have compromised study feasibility. The cohort included T0 to T3 tumors (TNM classification) without distant metastasis at diagnosis. T4 and/or patients with metastases at diagnosis and patients with a previous history of cancer were excluded. Information on clinical and tumor characteristics was retrieved prospectively from EMRs and registered as structured data in the institutional database. Approved by the Breast Cancer Study Group of Institut Curie, the study was conducted according to institutional and ethical rules concerning research on tissue specimens and patients. Tumor characterization and treatment protocol data are available in the Appendix.

Pregnancy After BC Diagnosis

Pregnancy after BC was defined by a pregnancy occurring from BC diagnosis until last follow-up date. Pregnancy cases occurring after June 2014 were censored (cutoff date of the analysis: May 31, 2014). Information on occurrence of a pregnancy after BC was retrieved in free full-text documents in either oncological, radiologic, or surgical consultations or clinical notes, including both in-house and external numerized notes. To retrieve this information, we compared two curation methods in a blinded manner.

Manual curation technique. First, each EMR was manually reviewed by a medical doctor (H.M.) from April to May 2014 to document the occurrence of a pregnancy following BC.

Extraction by TM. Extraction by TM technology consisted of a two-step process. First, we established a filter to select files containing specific string character patterns expected to have high specificity and to be present in EMRs in case of pregnancy. We chose the string character pattern "accouch*" to retrieve the following words: "accouchement," "accouché," "accoucher" (in English "delivery," "deliver," "delivered," respectively). In addition, we chose the string character pattern "enceinte" (in English "pregnant") to select all pregnant women, regardless of outcome (delivery, miscarriage, abortion). We did not retain the term "grossesse" (in English "pregnancy") because it engendered too much noise while referring in most cases to past obstetrical history rather than to pregnancies occurring after BC diagnosis (Appendix Table A1). These string character patterns were detected by TM in all documents except pdf image files. TM identified the number of patients for which "accouch*" and/or "enceinte" were present, the number of documents in which these patterns were found, and date and localization of concerned documents in EMRs.

Second, two independent researchers (C.S., M.B.) manually checked medical records selected after the first step from January 3-5, 2017 to confirm presence of a pregnancy after BC diagnosis. Only documents containing patterns

TABLE 1. Patient Characteristics

Characteristic	All (N = 344)	No Pregnancy (n = 308)	Pregnancy (n = 36)	*P*
Age, years, mean ± SD	36.4 ± 3.9	36.8 ± 3.9	32.3 ± 3.9	< .01
Body mass index, mean ± SD	22.9 ± 4.0	22.9 ± 4.2	22.8 ± 3.7	.82
< 20	82 (24.3)	75 (24.9)	7 (19.4)	.46
20-25	178 (52.8)	157 (52.2)	21 (58.3)	
26-30	50 (14.8)	43 (14.3)	7 (19.4)	
> 30	27 (8.0)	26 (8.6)	1 (2.8)	
Gravidity before BC diagnosis				< .01
Yes	79 (23.0)	63 (20.5)	16 (44.4)	
Parity before BC diagnosis				< .01
Yes	102 (29.7)	80 (26.0)	22 (61.1)	
Histology				
Nonspecific type	290 (84.3)	262 (85.1)	28 (77.8)	.33
Lobular	13 (3.8)	12 (3.9)	1 (2.8)	
Other	41 (11.9)	34 (11.0)	7 (19.4)	
Hormonal receptor status				
ER positive	218 (69.9)	199 (70.6)	19 (63.3)	.54
PR positive	121 (51.7)	111 (53.1)	10 (40.0)	.30
BC subtype				.52
Luminal	183 (59.8)	168 (60.6)	15 (51.7)	
TNBC	61 (19.9)	53 (19.1)	8 (27.6)	
HER2-positive	62 (20.3)	56 (20.2)	6 (20.7)	
T stage				.25
T0 or pTis	16 (4.7)	16 (7.9)	5 (20.0)	
T1	144 (42.1)	136 (67.0)	17 (68.0)	
T2	146 (42.7)	44 (21.7)	3 (12.0)	
T3	25 (7.3)	6 (3.0)	0 (0.0)	
N stage				.14
N0	260 (75.8)	178 (57.8)	25 (69.4)	
N1/N2/N3	83 (24.2)	87 (28.2)	10 (27.8)	
Surgery				.73
No	1 (0.3)	1 (0.3)	0 (0.0)	
Lumpectomy	181 (52.6)	160 (51.9)	21 (58.3)	
Mastectomy	162 (47.1)	147 (47.7)	15 (41.7)	
Neoadjuvant chemotherapy				
Yes	114 (33.1)	103 (33.4)	11 (30.6)	.87
Adjuvant chemotherapy				.96
No	87 (25.3)	77 (25.0)	10 (27.8)	
Anthracyclines only	73 (21.2)	65 (21.1)	8 (22.2)	
Anthracycline-taxane regimen	176 (51.2)	159 (51.6)	17 (47.2)	
Taxanes only	8 (2.3)	7 (2.3)	1 (2.8)	
Hormone therapy				.18
Yes	205 (59.6)	189 (61.4)	16 (44.4)	

NOTE. Data presented as No. (%) unless otherwise noted.

Abbreviations: BC, breast cancer; ER, estrogen receptor; PR, progesterone receptor; SD, standard deviation.

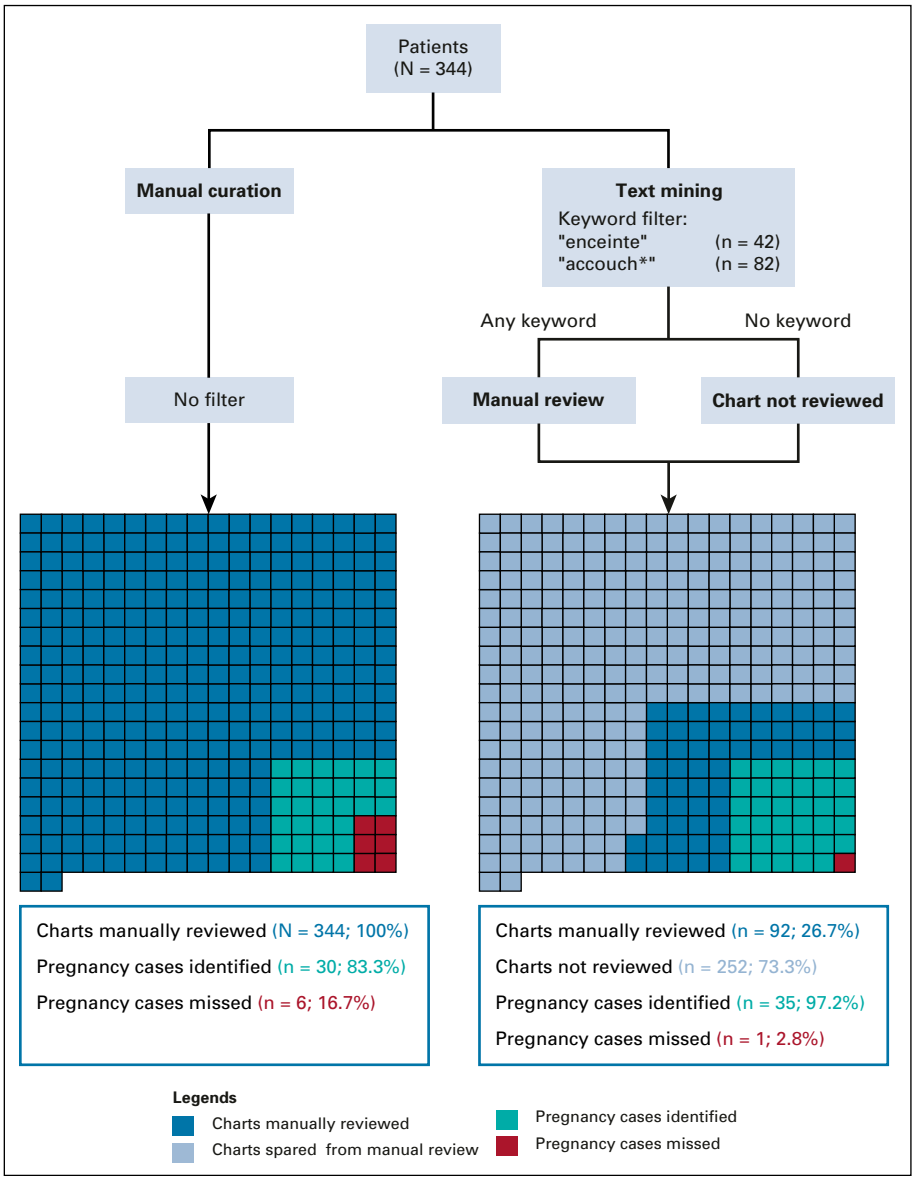

FIG 1. Flowchart for manual curation and text mining techniques.

"accouch*" or "enceinte" according to TM technology were checked.

After applying each technique on the cohort, each patient was classified in a binary manner as having or not having a pregnancy after BC.

Statistics

Person-years of EMR was defined by the sum of follow-up times (in years) for each individual. Incidence rate for pregnancy after BC was defined by the number of incident pregnancy cases after BC during follow-up divided by the number of person-years of EMR. Percentage of manual checking was defined as the number of EMRs checked divided by the total number of EMRs. Time to identify one pregnancy was calculated as the time necessary to identify all pregnancies with one technique (manual curation or TM) divided by the total number of pregnancies identified by that technique.

Our objective was to determine which method (TM or manual curation) was the most efficient to detect pregnancies occurring after BC in our cohort. The ability of each technique (TM or manual curation) to detect this event (pregnancy after BC) was evaluated considering the true number of events occurring in the cohort (ie, the true number of pregnancies occurring after BC in our cohort). Hence, we considered as our gold standard the number of true pregnancies, corresponding to pregnancies identified by the combination of the two methods. We compared the two techniques in terms of sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV).

True-positive (TP) cases were defined as patients with subsequent pregnancy who had been identified as having

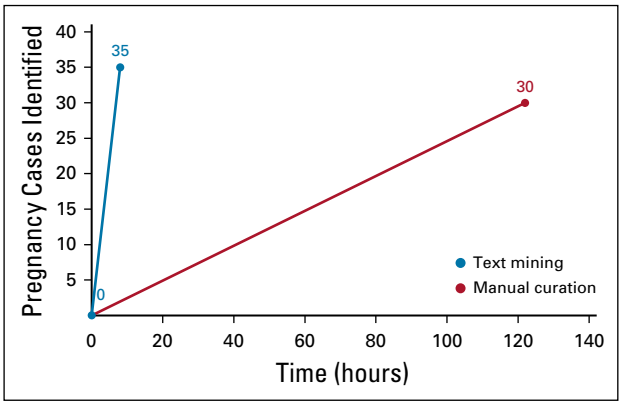

FIG 2. Time to identify pregnancy cases for manual curation and text mining techniques.

a pregnancy by the technique. False-positive (FP) cases were defined as patients without subsequent pregnancy who had been identified as having a pregnancy by the technique. True-negative (TN) cases were defined as patients without subsequent pregnancy who had not been identified as having a pregnancy by the technique. False-negative (FN) cases were defined as patients with subsequent pregnancy who had not been identified as having a pregnancy by the technique.

Sensitivity was defined as the probability of a technique to detect a pregnancy among women having a pregnancy, calculated as follows: sensitivity = TP/(TP + FN). Specificity was defined as the probability of a technique not to detect a pregnancy among women who did not have a pregnancy, as follows: specificity = TN/(TN + FP). PPV was defined as the probability to have a pregnancy when the technique detected a pregnancy, as follows: PPV = TP/(TP + FP). NPV was defined as the probability not to have a pregnancy when the technique did not detect a pregnancy, as follows: NPV = TN/(TN + FN).

Identification of Pregnancies in a Second Cohort of Patients With BC Using the TM Algorithm

To validate if TM could rapidly identify pregnancies in an independent cohort, we applied our TM algorithm on a larger cohort of patients with BC. The second data set (Cohort B) contained female patients age ≤ 40 years and treated at Institut Curie for BC from January 1, 2008 to December 31, 2012. After applying TM technology, medical records were manually checked by two independent researchers (J.L., I.A.) to confirm the presence of a pregnancy after BC diagnosis.

RESULTS

Training Cohort (cohort A)

Baseline characteristics and identification of pregnancy cases. Our cohort comprised 344 patients with BC age ≤ 40 years. The 344 records were complete. Median follow-up was 137 months (range, 6.3-143.3 months). Patient, tumors, and treatment characteristics are listed in Table 1 and were compared between patients who were pregnant after BC diagnosis and those who were not. Patients having experienced a pregnancy after BC diagnosis were significantly younger compared with those who had not (mean ± SD, 32.3 ± 3.9 *v* 36.8 ± 3.9, respectively; $P < .01$). A significantly greater proportion of patients having experienced a pregnancy after BC diagnosis had already been pregnant and had already given birth before cancer diagnosis (gravity, 44.4% *v* 20.5%, respectively; $P < .01$; parity, 61.1% *v* 26.0%, respectively; $P < .01$).

Using the manual technique, all 344 EMRs were reviewed (Fig 1). Percentage of manual checking was 100%, and total time needed to manually check all EMRs was 122 person-hours. At completion, 30 patients (8.7%) were identified as having experienced pregnancy after BC diagnosis. Time to identify a pregnancy was 244 minutes.

Using TM technique, filters "accouch*" and "enceinte" were applied on the 344 EMRs (Fig 1). In 252 EMRs (73.3%), none of these string patterns was identified. These EMRs were not further checked. In 92 EMRs (26.7%), either "accouch*" (n = 82), "enceinte" (n = 42), or both "accouch*" and "enceinte" (n = 32) were identified. These EMRs were further selected for manual review. Percentage of manual checking was 26.7%. Total time needed to manually check EMRs was 8 person-hours. At completion,

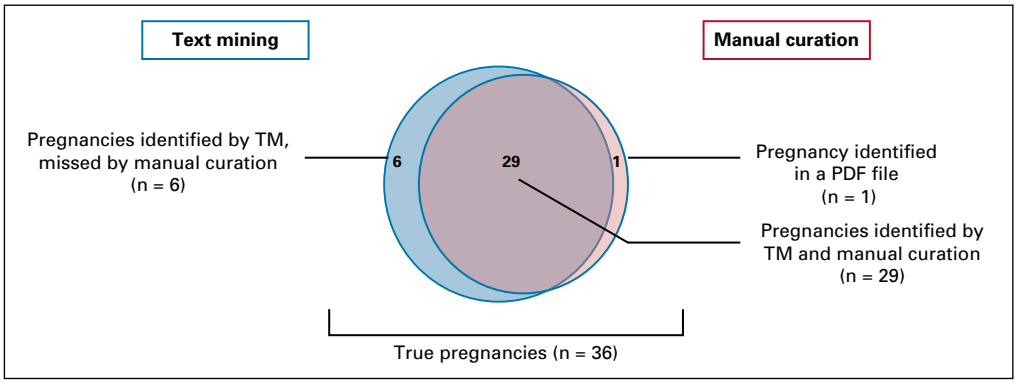

FIG 3. Distribution of pregnancies identified by manual curation and text mining (TM) techniques.

TABLE 2. Contingency Table for Manual Curation and Text Mining

Contingence	Pregnancy	No Pregnancy	Total
Manual reading only			
Pregnancy detection in EMR	30	0	30
No pregnancy detection in EMR	6	308	314
Total	36	308	344
Keyword "enceinte" only			
Pregnancy detection in EMR	24	18	42
No pregnancy detection in EMR	12	290	302
Total	36	308	344
Keyword "accouch*" only			
Pregnancy detection in EMR	32	50	82
No pregnancy detection in EMR	4	258	262
Total	36	308	344
Keyword "enceinte" or "accouch*"			
Pregnancy detection in EMR	35	57	92
No pregnancy detection in EMR	1	251	252
Total	36	308	344
Keyword "enceinte" and "accouch*"			
Pregnancy detection in EMR	23	9	32
No pregnancy detection in EMR	13	299	312
Total	36	308	344

Abbreviation: ERM, electronic medical record.

35 patients (35/344, 10.2%) were identified as having experienced pregnancy after BC diagnosis. Time to identify a pregnancy was 13 minutes.

Results obtained are detailed in Figure 1 and corresponding time in Figure 2. Altogether, TM technology substantially reduced the percentage of manual checking compared to manual review alone (26.7% *v* 100%, respectively) and resulted in substantial time and efficacy gains. Mean time to review one (true) pregnancy was found to be 17.8 times shorter using the TM technique (35 pregnancies identified in 8 person-hours) than using the full manual approach (30 in 122 person-hours).

Comparison of results obtained by manual curation versus TM technology. Combining both methods, we identified 36 pregnancies occurring after BC diagnosis (true number of pregnancies: n = 36 out of 344 patients, representing 10.5%; 2,829 person-years of EMR). The incidence rate of pregnancy after BC diagnosis was 0.01 pregnancies per person-year of EMR (36/2,829).

Manual curation identified 30 pregnancies and missed six pregnancies. Among them, 29 were common and also identified by TM. The pregnancy identified by manual curation only was reported in a pdf image file and was therefore not recognized by TM (Fig 3). The six pregnancies that manual curation did not detect were missed despite clear indication in free full text. No wording ambiguity was observed. We identified no specific feature that could explain why these pregnancy cases had been missed. Hence, these six missed pregnancies missed by manual curation were imputed to human error.

TM identified 35 pregnancies, 29 of them also being identified by manual curation. The six cases identified by TM only had been missed by manual curation despite clear indication in free full text (Fig 3). Hence, TM identified 35 pregnancies and missed one pregnancy.

Next, we investigated the relative performance of keywords used alone or in combination to identify pregnancies. The contingency table for each method is detailed in Table 2 and their accuracy in Figure 4.

Identification of any keyword "enceinte" or "accouch*" maximized the NPV of pregnancy (251/252 = 100%; thus decreasing the proportion of false-negative cases: 1/252 = 0%) while keeping an acceptable PPV (35/92 = 38%). Performances of the different strategies are described in Appendix Figure A1.

Manual curation showed 83% sensitivity (30/36) and 98% NPV (308/314). Sensitivity was highest (97%) when using keyword combination "enceinte" or "accouch*" (sensitivity: "enceinte" alone, 67%; "accouch*" alone, 89%; "enceinte" and "accouch*," 64%; manual reading, 83%).

By construction, specificity was perfect (100%) using manual curation. Using TM methods, specificity was very high (97%) using presence of both keywords and was satisfying using any keyword (81%).

Altogether, our results suggest that TM using the combination of any of the keywords "enceinte" or "accouch*" is associated with excellent NPV, meaning that this technique could be used to identify pregnancy, with a very low number of missed pregnancies. However, the 38% PPV rate highlights that TM alone is not sufficient to identify pregnancies and that complementary manual reading is required to filter noise generated by FP EMRs.

Identification of Pregnancies in a Second Cohort of Patients With BC Using the TM Algorithm (cohort B)

We selected an independent cohort of 1,226 patients treated between 2008 and 2012. All records were complete. Follow-up of Cohort B was significantly lower than Cohort A (mean follow-up, Cohort A, 110.0 months [108.0-113.0 months] *v* Cohort B, 73.4 months [71.8-75.3 months], respectively; $P < .001$). A total of 322 EMRs (26.3%) were selected by TM method after applying filters "accouch*" or "enceinte." TM eliminated 904 files (73.7%). All 322 selected records were then manually reviewed in 28 person-hours. At completion, 5.5% of the patients (n = 67 out of 1,226, representing 7,349 person-years of EMR) were identified as having experienced pregnancy after BC diagnosis. The incidence rate of pregnancy after BC diagnosis was 0.01 pregnancies per person-year of EMR (67/7,349).

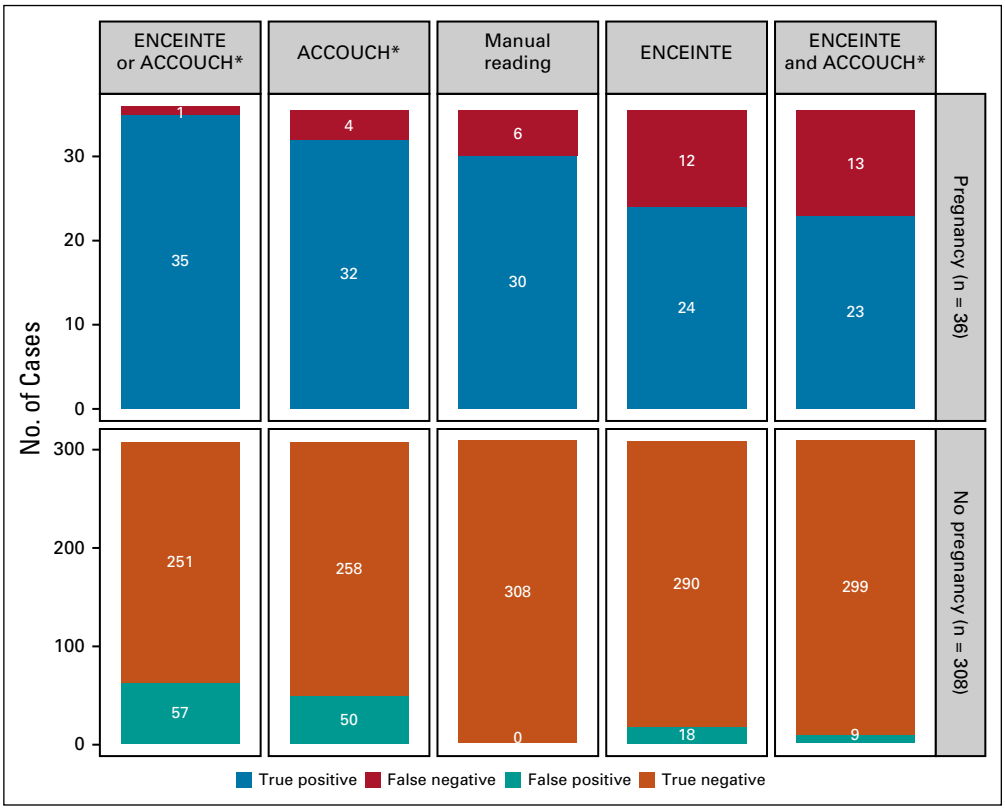

FIG 4. Accuracy of pregnancy cases identified by manual reading and text mining filters.

DISCUSSION

Our study demonstrates that TM approach is more efficient than manual curation to identify pregnancy cases occurring after BC diagnosis. TM considerably reduced the number of files to be reviewed, sparing 73.3% EMRs from manual review in Cohort A, and 73.7% in Cohort B. Moreover, while focusing manual review on a very small subset of the cohort, efficient targeting of documents containing keywords enabled us to decrease time spent without missing EMRs of interest. As a whole, in our study, TM outperformed manual curation by enabling substantial time, human resources, and financial gains.

Our results are consistent with previous studies highlighting the performance of TM technology. Alex et al[15] suggested that optimal TM techniques via natural language processing may be up to 34% faster than manual curation. O'Mara-Eves et al[16] showed in a systematic review of 44 studies that TM reduced workload, decreased number of items to be screened, and improved workflow efficiency. In 2012, Denny et al[8] showed that TM outperformed manual curation to identify patients having colorectal cancer testing, with only a modest number of FPs. Small et al[2] recently used TM to identify patients with trileaflet aortic stenosis and coronary artery disease on 282,569 echocardiography reports (81,164 patients) and 27,205 cardiac catheterization reports (14,567 patients). TM was superior to billing codes to identify patients, with a PPV of 95% compared with 53% by billing codes for trileaflet aortic stenosis and a PPV of 97% compared with 86% for coronary artery disease, respectively.

Second, TM appears to be highly efficient to quickly generate cohorts of patients who experience a rare event, resulting in substantial cost savings. Applying TM on a second, larger cohort of patients with BC enabled us to quickly generate one of the largest cohorts from a single institution of patients having experienced pregnancy after BC. We designed our detection method to minimize the number of missed pregnancies. We found an incidence of pregnancies after BC of 5.5%, which is similar to the rate described in the literature.[17]

Furthermore, TM is notably efficient, because the occurrence of rare events can be mentioned very briefly in EMRs and is thus easily missed by manual reading. Our results confirm that TM curation reduced reviewer burden with a minimal loss of true cases.

Our study has several limitations. Although the incidence of pregnancy after BC observed in our study was similar between the two cohorts, we cannot rule out the fact that we may have missed pregnancies among unverified files. On the basis of results of Cohort A in which TM identified 35 pregnancies out of 36, we can approximate that two were likely missed in Cohort B. However, the incidence of pregnancy after BC observed in our study is consistent with

the literature. Although our TM method was not able to process scanned reports and pdf files, quality of results does not seem significantly affected, as only one pregnancy of 36 was missed because of this issue. Optical Character Recognition softwares[18-20] converting scanned files into editable output format are expected to solve such problems in the near future. It is also possible that TM missed miscarriages and pregnancies without favorable outcomes. Nevertheless, TM missed fewer pregnancy cases than manual curation and was efficient in identifying pregnancies carried out successfully.

We must highlight that TM is dependent on the availability, suitability, adaptability, interoperability, and comparative accuracy of current TM resources.[21,22] Diverse, incomplete, and redundant data, as well as discordances between medical language and natural language processing, may affect mining results.[4,9,23,24] As such, performance of TM may be affected by the variability and quality of data contained in EMRs. An external validation of our TM method is warranted to confirm its efficiency on different types of EMRs coming from other centers. Furthermore, filters "accouch*" or "enceinte" were chosen to minimize missed true cases, but the combination of any keywords increased FP cases. Hence, manual checking of EMRs selected by TM is still required.

To conclude, the possibility to quickly browse and merge data is promising, and filling defaults, smoothing noise, and using standardized and structured medical records will improve performance. A huge amount of medical data will be released in the near future by the French National Health Data hub,[25] connecting medical sources and EMRs with medico-administrative sources and reimbursement. Given the promising impact of computational technologies, it opens a wide range of new perspectives for research purposes and future in silico discoveries.

AFFILIATIONS

[1]Institut Curie, Paris, France
[2]Geneva University Hospitals, Geneva, Switzerland
[3]Port-Royal Maternity Unit, Paris, France
[4]Cancerology Institute, Montpellier, France
[5]Institut Curie, U900, Hôpital René Huguenin, Saint-Cloud, France
[6]Paris 5 Research University, INSERM U932, Institut Curie, Paris, France

CORRESPONDING AUTHOR

Fabien Reyal, MD, PhD, Institut Curie, 26 rue d'Ulm, 75005 Paris, France; e-mail: fabien.reyal@curie.fr.

EQUAL CONTRIBUTION

F.R. and A.S.H. contributed equally to this work.

AUTHOR CONTRIBUTIONS

Conception and design: Julie Labrosse, Clara Sebbag, Milena Benque, Hilde Merckelbagh, Thomas Balezeau, Florence Coussy, Jean-Guillaume Feron, Fabien Reyal, Anne-Sophie Hamy

Administrative support: Anne-Sophie Hamy

Provision of study material or patients: Angélique Bobrie, Enora Laas, Anne-Sophie Hamy

Collection and assembly of data: Julie Labrosse, Thanh Lam, Clara Sebbag, Ines Abdennebi, Hilde Merckelbagh, Marie Osdoit, Thomas Balezeau, Beatriz Grandal, Angélique Bobrie, Jean-Guillaume Feron, Fabien Reyal, Anne-Sophie Hamy

Data analysis and interpretation: Julie Labrosse, Clara Sebbag, Ines Abdennebi, Maël Priour, Julien Guerin, Thomas Balezeau, Angélique Bobrie, Loïc Ferrer, Enora Laas, Jean-Guillaume Feron, Fabien Reyal, Anne-Sophie Hamy

Manuscript writing: All authors

Final approval of manuscript: All authors

Accountable for all aspects of the work: All authors

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated unless otherwise noted. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/cci/author-center.

Open Payments is a public database containing information reported by companies about payments made to US-licensed physicians (Open Payments).

No potential conflicts of interest were reported.

REFERENCES

1. van Velthoven MH, Mastellos N, Majeed A, et al: Feasibility of extracting data from electronic medical records for research: An international comparative study. BMC Med Inform Decis Mak 16:90, 2016
2. Small AM, Kiss DH, Zlatsin Y, et al: Text mining applied to electronic cardiovascular procedure reports to identify patients with trileaflet aortic stenosis and coronary artery disease. J Biomed Inform 72:77-84, 2017
3. Chen ES, Sarkar IN: Mining the electronic health record for disease knowledge. Methods Mol Biol 1159:269-286, 2014
4. Sun W, Cai Z, Li Y, et al: Data processing and text mining technologies on electronic medical records: A review. J Healthc Eng 2018:4302425, 2018
5. Vock DM, Wolfson J, Bandyopadhyay S, et al: Adapting machine learning techniques to censored time-to-event health record data: A general-purpose approach using inverse probability of censoring weighting. J Biomed Inform 61:119-131, 2016
6. Spasić I, Livsey J, Keane JA, et al: Text mining of cancer-related information: Review of current status and future directions. Int J Med Inform 83:605-623, 2014
7. Cohen R, Elhadad M, Elhadad N: Redundancy in electronic health record corpora: Analysis, impact on text mining performance and mitigation strategies. BMC Bioinformatics 14:10, 2013
8. Denny JC, Choma NN, Peterson JF, et al: Natural language processing improves identification of colorectal cancer testing in the electronic medical record. Med Decis Making 32:188-197, 2012

9. Meystre S, Haug PJ: Natural language processing to extract medical problems from electronic clinical documents: Performance evaluation. J Biomed Inform 39:589-599, 2006
10. Ives A, Saunders C, Bulsara M, et al: Pregnancy after breast cancer: Population based study. BMJ 334:194, 2007
11. Iqbal J, Amir E, Rochon PA, et al: Association of the timing of pregnancy with survival in women with breast cancer. JAMA Oncol 3:659-665, 2017
12. Ruddy KJ, Gelber SI, Tamimi RM, et al: Prospective study of fertility concerns and preservation strategies in young women with breast cancer. J Clin Oncol 32:1151-1156, 2014
13. Lambertini M, Martel S, Campbell C, et al: Pregnancies during and after trastuzumab and/or lapatinib in patients with human epidermal growth factor receptor 2-positive early breast cancer: Analysis from the NeoALTTO (BIG 1-06) and ALTTO (BIG 2-06) trials. Cancer 125:307-316, 2019
14. McCray DKS, Simpson AB, Flyckt R, et al: Fertility in women of reproductive age after breast cancer treatment: Practice patterns and outcomes. Ann Surg Oncol 23:3175-3181, 2016 [Erratum: Ann Surg Oncol 23:1063, 2016]
15. Alex B, Grover C, Haddow B, et al: Assisted curation: Does text mining really help? Pac Symp Biocomput 2008:556–567, 2008
16. O'Mara-Eves A, Thomas J, McNaught J, et al: Using text mining for study identification in systematic reviews: A systematic review of current approaches. Syst Rev 4:5, 2015 [Erratum: Syst Rev 4:59, 2015]
17. de Bree E, Makrigiannakis A, Askoxylakis J, et al: Pregnancy after breast cancer. A comprehensive review. J Surg Oncol 101:534-542, 2010
18. Rasmussen LV, Peissig PL, McCarty CA, et al: Development of an optical character recognition pipeline for handwritten form fields from an electronic health record. J Am Med Inform Assoc 19:e90-e95, 2012
19. Hawker CD, McCarthy W, Cleveland D, et al: Invention and validation of an automated camera system that uses optical character recognition to identify patient name mislabeled samples. Clin Chem 60:463-470, 2014
20. Lee YH, Park EH, Suh JS.: Simple and efficient method for region of interest value extraction from picture archiving and communication system viewer with optical character recognition software and macro program. Acad Radiol 22:113-116, 2015
21. Przybyła P, Shardlow M, Aubin S, et al: Text mining resources for the life sciences. Database 2016:baw145, 2016
22. Van Auken K, Jaffery J, Chan J, et al: Semi-automated curation of protein subcellular localization: A text mining-based approach to Gene Ontology (GO) Cellular Component curation. BMC Bioinformatics 10:228, 2009
23. Doan S, Conway M, Phuong TM, et al: Natural language processing in biomedicine: A unified system architecture overview. Methods Mol Biol 1168:275-294, 2014
24. Yim WW, Yetisgen M, Harris WP, et al: Natural language processing in oncology: A review. JAMA Oncol 2:797-804, 2016
25. Ministère des Solidarités et de la Santé: Rapport Health Data Hub, mission de préfiguration. https://solidarites-sante.gouv.fr/ministere/documentation-et-publications-officielles/rapports/sante/article/rapport-health-data-hub-mission-de-prefiguration

APPENDIX

Materials and Methods: Treatment Protocol

Histologic grade was described according to the Elston-Ellis modification of the Scarff-Bloom-Richardson grading system (Elston CW, et al: Histopathology 19:403-410, 1991).

Hormone receptor expression was analyzed by immunohistochemistry. Tumors were considered positive for estrogen receptor (ER) or progesterone receptor (PR) if 10% of carcinomatous cells displayed positive staining, as recommended by European guidelines (Balaton AJ, et al: Annales de Pathologie 19:336, 2008). *HER2* status was determined according to ASCO recommendations (Wolff AC, et al: J Clin Oncol 31:3997-4013, 2013). On the basis of immunohistochemistry surrogates, pathologic breast cancer subtypes were defined as follows: tumors positive for either ER or PR and negative for *HER2* were classified as luminal; tumors positive for *HER2* were considered *HER2*-positive BC; tumors negative for ER, PR, and *HER2* were considered triple-negative BC (TNBC). Pathological complete response was defined as the absence of residual invasive cancer cells in the breast and axillary lymph nodes (ypT0/is + / ypN0).

Patients were treated according to national guidelines. For patients treated by neoadjuvant chemotherapy, regimens changed over time (anthracycline-based regimen or sequential anthracycline-taxane regimen), with trastuzumab used in an adjuvant and/or neoadjuvant setting for *HER2*-positive tumors since the middle of the past decade. Trastuzumab treatments changed over time because of a change of marketing authorization during the study period. Surgery consisted of breast-conserving surgery or mastectomy. In case of neoadjuvant chemotherapy, surgery was performed 4 to 6 weeks after. Patients received adjuvant radiotherapy according to national guidelines. Indications of radiotherapy were: lumpectomy, radical mastectomy in case of initial T3 or T4 tumors, all patients with involved axillary lymph nodes, and patients with high-risk node-negative breast cancer.

For patients concerned, adjuvant chemotherapy was administered during radiotherapy to patients who had not reached pathologic response and/or with nodal involvement after neoadjuvant chemotherapy. Adjuvant hormone therapy (tamoxifen, aromatase inhibitor, or gonadotropin releasing hormone agonists) was prescribed when indicated to patients with luminal disease (sequentially to radiotherapy and adjuvant chemotherapy when administered). Patient follow-up after treatment was every 3 months during the first 2 years, then every 6 months during 3 years, and once a year starting from the 5th year. Follow-up consisted of clinical examination associated with mammography and mammary ultrasound once a year.

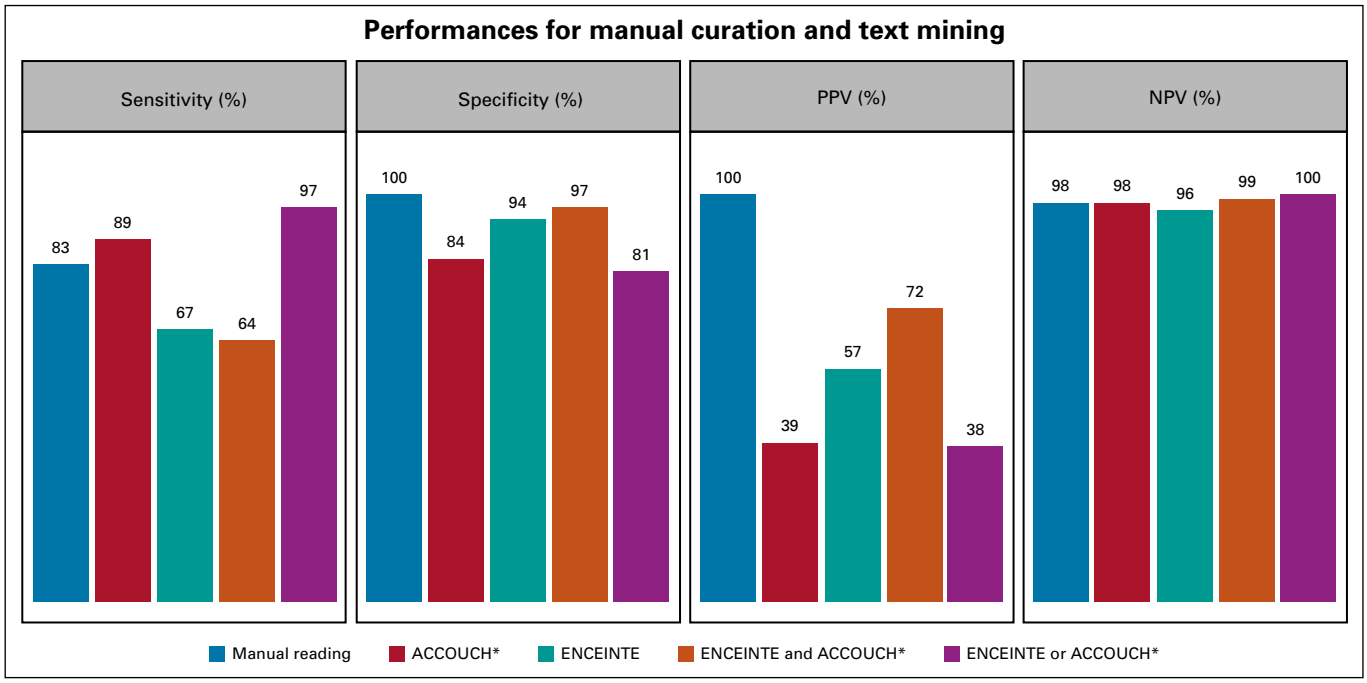

FIG A1. Performances for manual curation and text mining. NPV, negative predictive value; PPV, positive predictive value.

TABLE A1. Results of Text Mining Using Additional Keywords

	"acouch*", "acco*ch*" "en*einte"	"avort*"	"fausse couche"	"interruption"	"IVG"	"IMG"	"grossesse"
	(n = 1)	(n = 13)	(n = 0)	(n = 91)	(n = 125)	(n = 13)	(n = 337)
Referring to patient medical or obstetrical past history	1 (100)	13 (100)	0 (0.0)	13 (14.3)	124 (99.2)	2 (15.4)	232 (68.9)
Referring to the current breast cancer, without subsequent pregnancy case	0 (0.0)	0 (0.0)	0 (0.0)	70 (76.9)	0 (0.0)	0 (0.0)	80 (23.7)
Referring to the current breast cancer, with subsequent pregnancy case	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	1 (0.8)	0 (0.0)	25 (7.4)
Other	0 (0.0)	0 (0.0)	0 (0.0)	8 (8.8)	0 (0.0)	11 (84.6)	

NOTE. Data are presented as No. (%).

Abbreviations: IMG, interruption medicale de grossesse (pregnancy interruption for medical reasons); IVG, interruption volontaire de grossesse (pregnancy interruption for personal reasons).

Files also containing keyword "enceinte" and/or "accouch".

