Supplement to Kempf 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. Accepted Author Manuscript. https://doi.org/10.1002/ijc.33928

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

Emmanuelle Kempf, MD, MSc (1), Sonia Priou, MSc (2), Guillaume Lamé, PhD (3), Christel Daniel, MD, PhD (2,8), Ali Bellamine, MD, MSc (2), Daniele Sommacale, MD, PhD (4), Yazid Belkacemi, MD, PhD (5), Romain Bey, PhD (2), Gilles Galula, MD (6), Namik Taright, MD, PhD (7), Xavier Tannier, PhD (8), Bastien Rance, PhD (9), Rémi Flicoteaux, MD, PhD (7), François Hemery, MD, MSc (10), Etienne Audureau, MD, PhD (11), Gilles Chatellier, MD, PhD (9), Christophe Tournigand, MD, PhD (12) on behalf of the Assistance Publique – Hôpitaux de Paris Cancer Group.

Table of contents

Supplement to Kempf 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. Accepted Author Manuscript. https://doi.org/10.1002/ijc.33928

- Supplementary Materials & Methods

Origin of the data and variables

The study was conducted using the data of the AP-HP Clinical Data Warehouse (CDW) integrating administrative and health data from more than 11 million patients routinely collected during patient admission, consultation, or hospitalization with one of the 39 AP-HP hospitals. Such data mainly come from the AP-HP Electronic Healthcare Record (EHR) (ORBIS) being deployed across all hospitals and also from hundreds of other legacy clinical applications. The deployment of the AP-HP EHR started in 2014 for some hospitals and is still on-going for others. In order to have at least two years of medical history within the patients' EHRs, only the hospitals that installed this software before January 2016 were taken into account.

CDW storage solutions are based on a hybrid architecture combining relational databases (PostgreSQL) and NoSQL solutions offering the computing capacity adapted to Big Data analysis (clinical documents, medical images, and so on): computing cluster and high-tech components such as graphics processing unit (GPU) processors allowing parallelized mathematical operations to accelerate data processing using automatic learning algorithms. The security of the CDW system is ensured at the hardware, software, and organizational levels. Indeed, an authorization matrix compliant with data access and exploitation rules, as well as solutions ensuring data confidentiality, have been put in place. In particular, data pseudonymization is implemented on both structured and unstructured data (textual document and medical images) allowing their use in data research (in compliance with the CNIL, French data protection regulatory agency (and specially the MR004 reference methodology) and GDPR. Clinical data is interoperable - standardized using international terminologies (ICD-10, ICD-O, LOINC, ATC, etc.) and accessible to IRB-authorized users through a Jupyter portal (Python, R, Scala...).

The following variables were considered:

- Demographic data and data on hospital admission, transfer and discharge (healthcare trajectories)
- Medical history, coded according to 10th edition of the International Classification of disease (ICD-10)
- Medical procedures performed during hospitalization, coded according to the French Common Classification of Medical Procedures (CCAM, 11th edition)
- Clinical documents, especially anatomic pathology reports and medical imaging reports
- Outcomes (one-year survival rate, tumor stage)

Data transformation and data quality

Before their integration in the CDW, source data are transformed, pseudonymized and standardized using international terminologies (ICD-10, CIM-O, LOINC, ATC, etc.). Data quality checks are performed during the ETL (Extract-Transfer-Load) phase and also during the study (data cleaning).

The completeness of clinical reports was analysed. To estimate the availability of CT scans, pathology reports and MDM reports, we estimated the rate of visits with available reports. Concerning the pathology reports, we estimated the percentage of primary colon resection visits linked to at least

one pathology report (Supp Figure 4). This rate is stable from January 2018 to December 2020. Regarding CT scans, we estimated the percentage of patients with a colorectal cancer having at least one CT scan (Supp Figure 5). This percentage does not depend on the date of diagnosis of the patient's cancer. Concerning MDM reports, we estimated the percentage of patients with a colorectal cancer having at least one MDM report (Supp Figure 6).

Description of the development and validation of the NLP techniques used to extract tumor stage of CRC

Tumor stage of the CRC cases has been extracted from patients' clinical documents using NLP techniques. We report the development and validation of the algorithms according to the Minimum information about clinical artificial intelligence modeling (MI-CLAIM) statement.

The pathological tumor stage (pTNM) (according to the pTNM AJCC 8th edition) for cases with upfront resection of a primary localized colon cancer has been automatically extracted from the first related postoperative pathology report within the EHR of each patient using a regular expression algorithm.

We identified the first related postoperative pathology report within the EHR of each patient, based on the structured date of the medical file edition. Among them, 200 were annotated by a cancer specialist physician. The annotator was asked to extract the pTNM tumor stage when mentioned. The annotated dataset was randomly split along documents into a training set (50%) and a test set (50%).

We developed the following regular expression (Re Python library) algorithm on the training set solely:

```
([ycpP]{1,2}\s?(T([01234x]|is)[abcdx]?)[,\s]{0,2}[ycp]{0,2}\s?(N[xo01234\+][abcdx]?)*\s?(M[o01]?[\+x]?)?)|((T([01234x]|is)[abcdx]?)[,\s]{0,2}[ycp]{0,2}\s?(N[xo01234\+][abcdx]?)\s?(M[o01]?[\+x]?)?)
```

When incomplete information was extracted, the results were classified as a negative classification outcome.

On the training and the test sets, the following performance metrics were evaluated:

- sensitivity (recall),
- predictive positive value (precision),
- F1 score (harmonic mean between sensitivity and predictive positive value).

The **metastatic status of CRC cases** at initial presentation has been automatically extracted from the report of CT-scans performed between 90 days before and 45 days after the CRC diagnosis date using another regular expression algorithm.

CT-scans identification

The CT-scans were identified with a dedicated algorithm based on regular expression rules.

CT-scan examinations were identified and then defined as likely related to a baseline evaluation according to the procedure report's title as follow:

- CT-scan:
    - Title containing one of the following French expressions concerning MRI and CT-scan: SCANNER OR TDM OR TOMODENSITOM[EÉ]TRIE
- Examination likely related to a baseline evaluation:
    - Title containing the following French expression concerning the localization: TAP|TH?ORACO[-|]?ABDO|TH?ORACI[^\s]* *(?:ET)? *ABDO|
    - Procedure performed between 90 days before and 45 days after the CRC diagnosis date
- Metastatic status

We extracted the metastatic status of CRC cases at initial presentation from the available imaging text reports within the EHR of each patient using machine learning algorithms with two sequential steps.

To that aim, we identified the CT-scans within the imaging text reports available in the CDW according to the above-described algorithm. We selected the first hospital CT-scan performed between 90 days before and 45 days after the CRC diagnosis date. We then compared two methods of binary classification to extract staging CT-scans from non-staging CT-scans. Both methods are based on machine learning:

1) a random forest algorithm based on the frequency of words from the medical history and the conclusion sections of the text reports,
2) a convolutional neural network (CNN) using word2vec word embeddings pretrained on the CDW HER [49].

The metastatic status of CRC cases at initial presentation was extracted using the following regular expression algorithm on the training set solely:

(m[ée]tasta(se|tique)s?)|(diss[ée]min[ée]e?s?)|(carcinose)|(((allure|l[ée]sion|localisation|progression)s?\s)(suspecte?s?)?.{0,30}(secondaire)s?)|(l(a|â)ch(é|e|er)\sde\sballons?)|(l[ée]sions\s(non\s)?cibles)|(rupture.{1,20}corticale)|(envahissement.{0,15}parties\smolles)|((l[i,y]se).{1,20}os)|ost[eé]ol[i,y]|rupture.{1,20}corticale|envahissement.{1,20}parties\smolles|ost[eé]ocondensa.{1,20}(suspect|secondaire|[ée]volutive)|(l[ée]sion|anomalie|image).{1,20}os.{1,30}(suspect|secondaire|[ée]volutive)|os.{1,30}(l[ée]sion|anomalie|image).{1,20}(suspect|secondaire|[ée]volutive)|(l[ée]sion|anomalie|image).{1,20}l[i,y]tique|(l[ée]sion|anomalie|image).{1,20}condensant.{1,20}(suspect|secondaire|[ée]volutive)|fracture.{1,30}(suspect|secondaire|[ée]volutive)|((l[ée]sion|anomalie|image|nodule).{1,80}(secondaire))|((l[ée]sion|anomalie|image|nodule)s.{1,40}suspec?ts?)

Overall, 436 CT-scan text reports were manually annotated by a cancer specialist physician. The annotator was asked to classify the report as staging or non staging CT-scan, and to specify the metastatic status of the related tumor. This annotated dataset was randomly split into a training set (70%) and a test set (30%).

For both algorithms, on the training and the test sets, the following performance metrics were evaluated:

- sensitivity (recall),
- predictive positive value (precision),

F1 score (harmonic mean between sensitivity and predictive positive value).

Supplementary Results

pTNM tumor stage

The algorithm of pTNM identification resulted in a sensitivity and a positive predictive value of 98% and 96% on the training set. The F1 score* of the related algorithm reached 95% and 97% on the training set and the test set, respectively.

$$^{*}\mathrm{F}_1 = \frac{2\,(\,\text{Sensitivity} * \text{Positive Predictive Value})}{\text{Sensitivity} + \text{Positive Predictive Value}}$$

Metastatic status

The identification algorithm of staging CT-scans among the initial imaging reports resulted in a sensitivity and a positive predictive value of 86% and 78% on the training set.

The staging identification CNN algorithm reached a sensitivity of 68% and a positive predictive value of 83% on the test set. CNN performances are generally expected to outreach traditional statistical method outcomes. This is not the case in our study, which could be due to the relatively low number of annotated data. Older methods remain competitive in domains where annotated data are difficult to obtain [50].

The algorithm identified the metastatic status with a sensitivity and a positive predictive value of 78% and 86% on the train set. For the classification of text reports, the F1 score of the random forest reached 82% and 73% on the training set and the test set, respectively.

For the assessment of the tumor metastatic status, the F1 score of the regular expression algorithm reached 82% and 75% on the training set and on the test set, respectively.

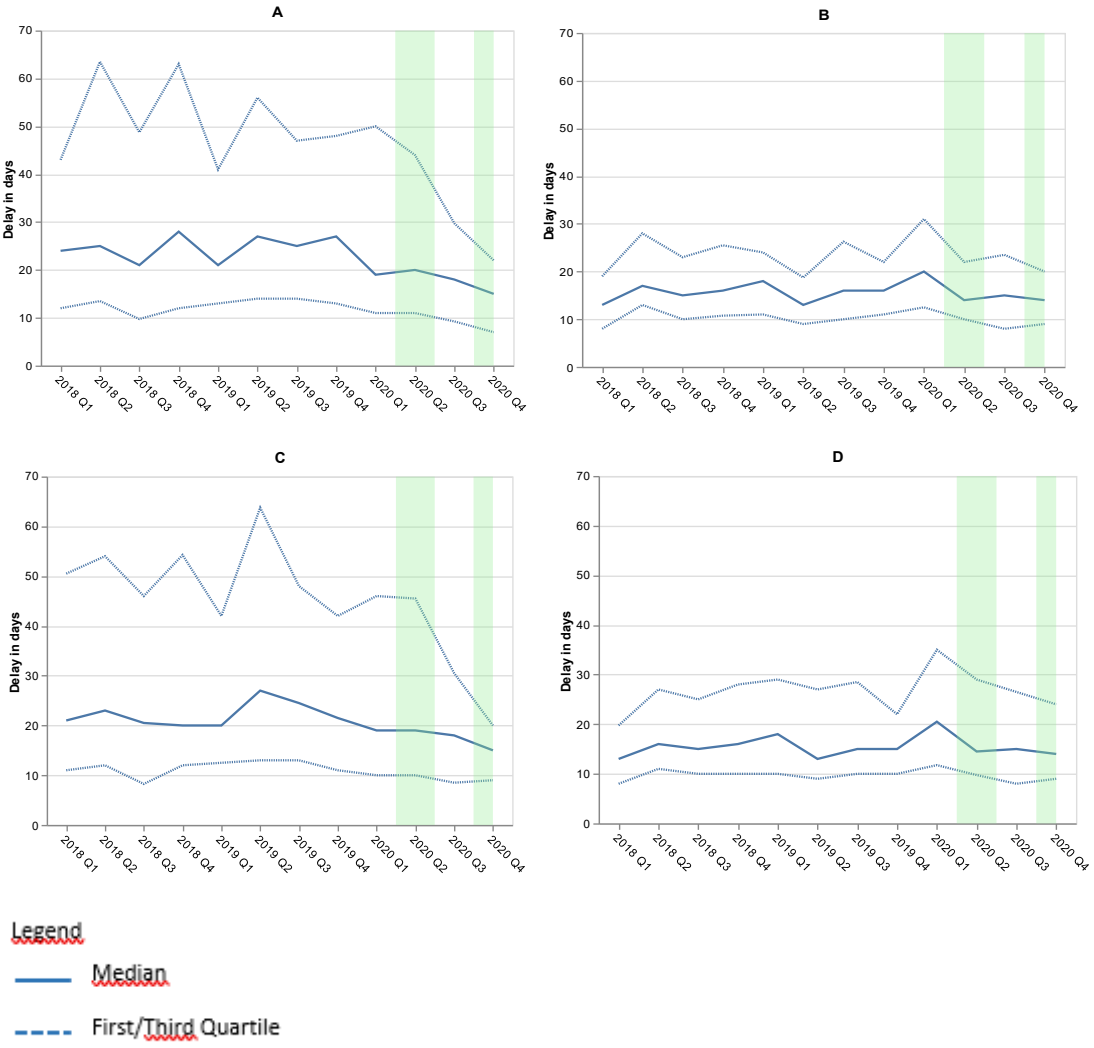
Interpretability

No interpretability is necessary for the algorithms based on regular expressions. Several examination techniques were used to understand the identification algorithm of staging CT-scans. First, SHapley Additive exPlanations (SHAP) summary plot was used to analyse the importance of each word in the classification of documents (Lundberg SM. A unified approach to interpreting model predictions. 2017). The words with the highest weight to classify a document as staging or not staging were "*bilan" "extension" "recherche" "reevaluation"* and "*masse"* (Supp Figure 7). Then, a sensitivity analysis was performed using SHAP force plots. These plots enable us to understand how words contributed to the model's prediction for a specific document. We analysed the plots for the two most confident and correctly classified documents and the two most confident and incorrectly classified documents. Finally, SHAP dependence plots were generated for all 15 top words to understand if the weight of a word was constant or if it depended on the value of other words (Supp Figure 8). Each words observed had a linear pattern.

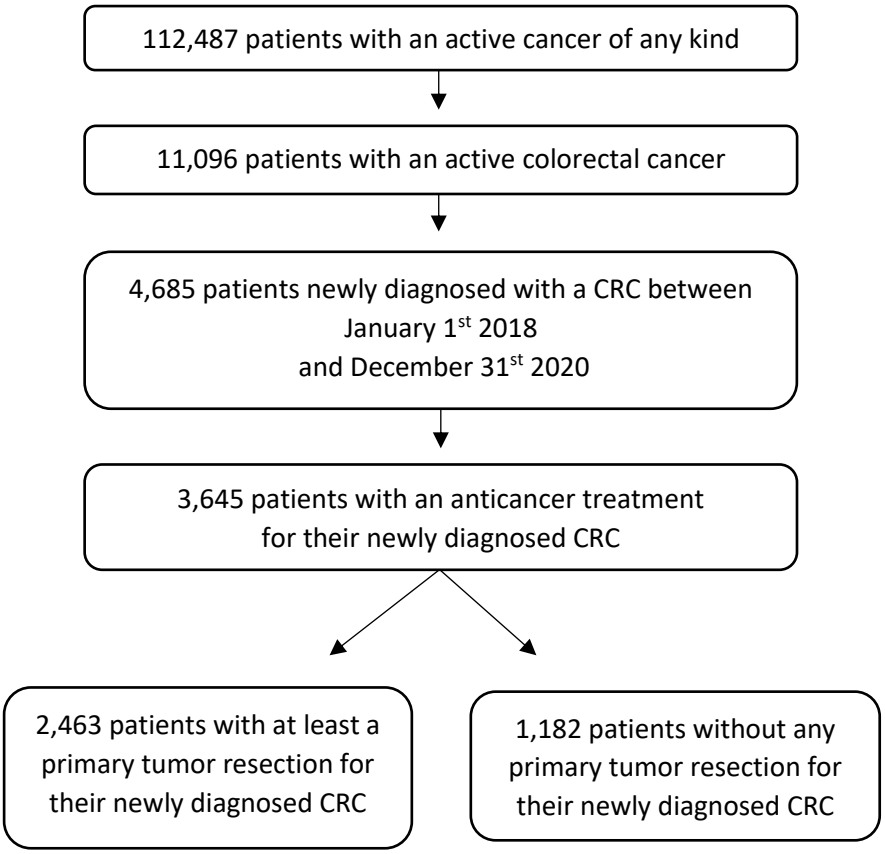
Reproducibility

After the approval by the AP-HP CDW IRB, the code used to select, and analyse the study data is in open access the algorithm is shared (AP-HP Github, Zenodo, License open source (BSD 3-clause) will be available for any reader of the manuscript.

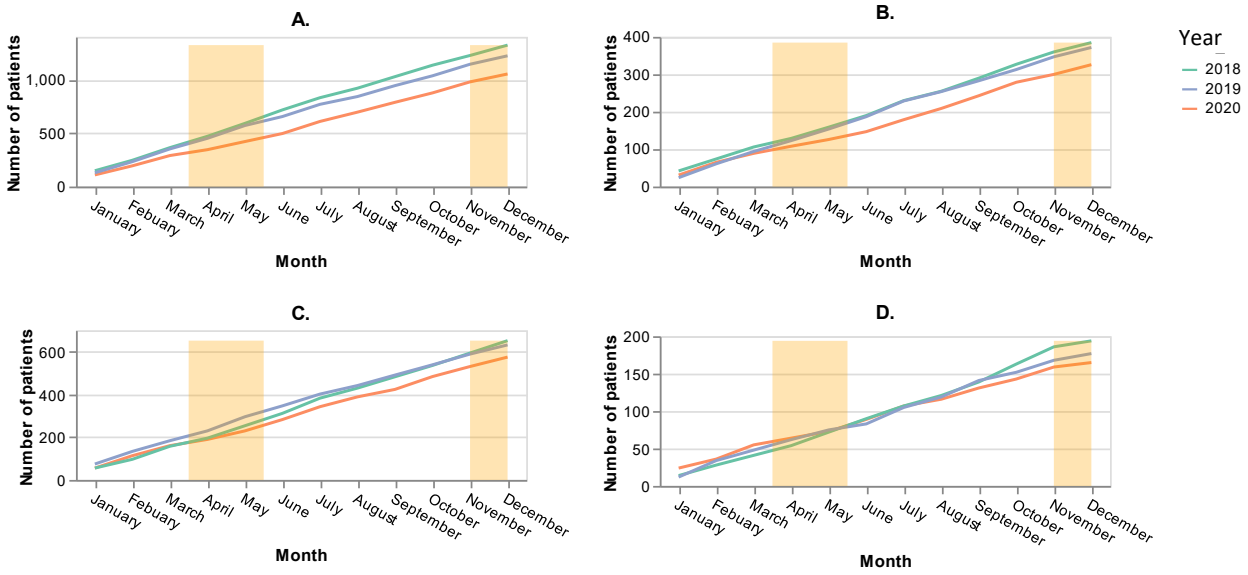
Supplementary Figures



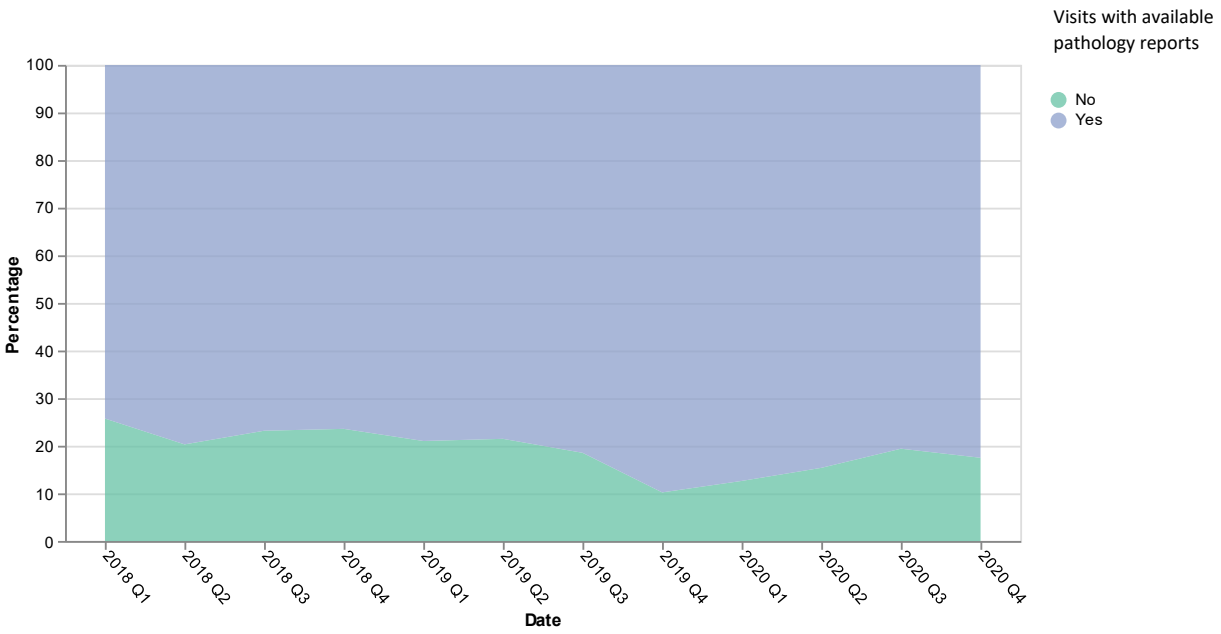
Supplementary Figure 1. Evolution of the median delay between the 1st multidisciplinary meeting (MDM) and the 1st therapeutic procedure for patients operated from a colorectal cancer at the APHP hospital between 2018 and 2020: patients having an MDM before the 1st anticancer treatment (Supp Fig. 1A) and patients having an anticancer treatment before the 1st MDM (Supp Fig. 1B). Supp Fig.1C and Supp Fig. 1D refer to the population of patients undergoing any kind of anticancer therapy.



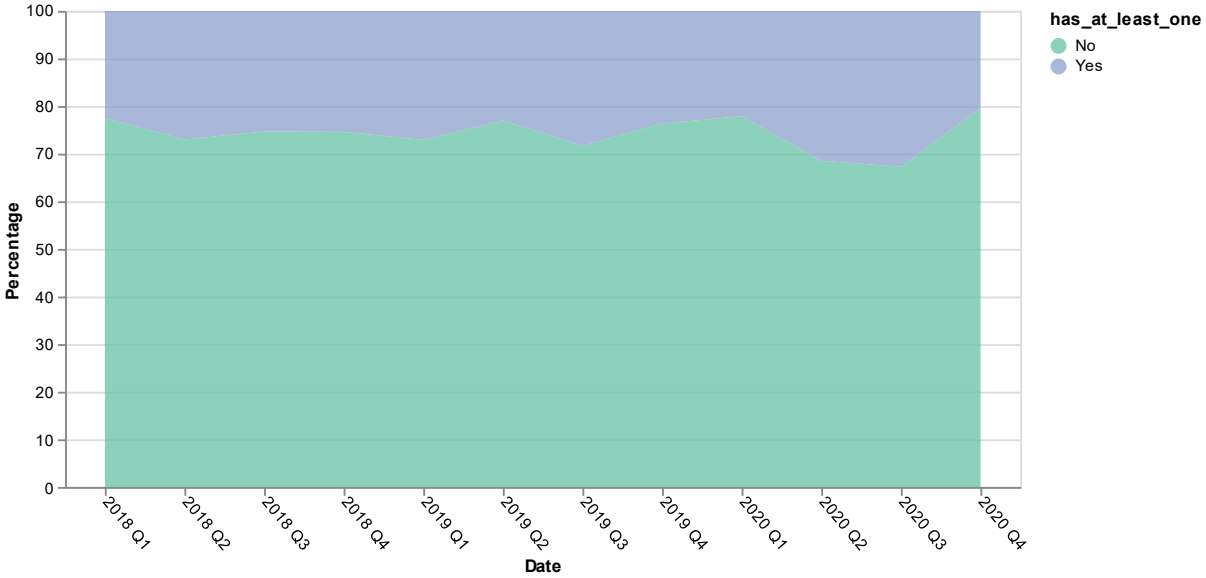
Supplementary Figure 2. Flowchart of identification of patients newly referred between 2018 and 2020 for a colorectal cancer to the Greater Paris University Hospitals, with a medical history available since 2016



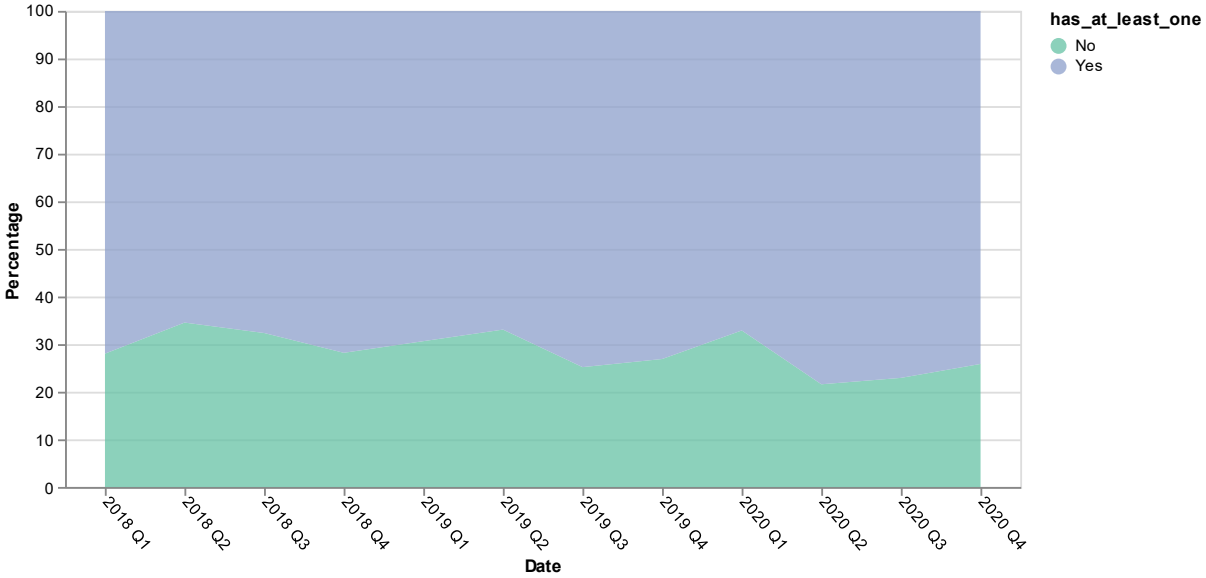
Supplementary Figure 3. Evolution of the cumulative monthly number of new cancer cases of colon (2.A) and rectum (2.B), primary tumor resections for colon (2.C) and rectum (2.D) over time referred to the Greater Paris University Hospitals, in 2018, 2019 and 2020, respectively.



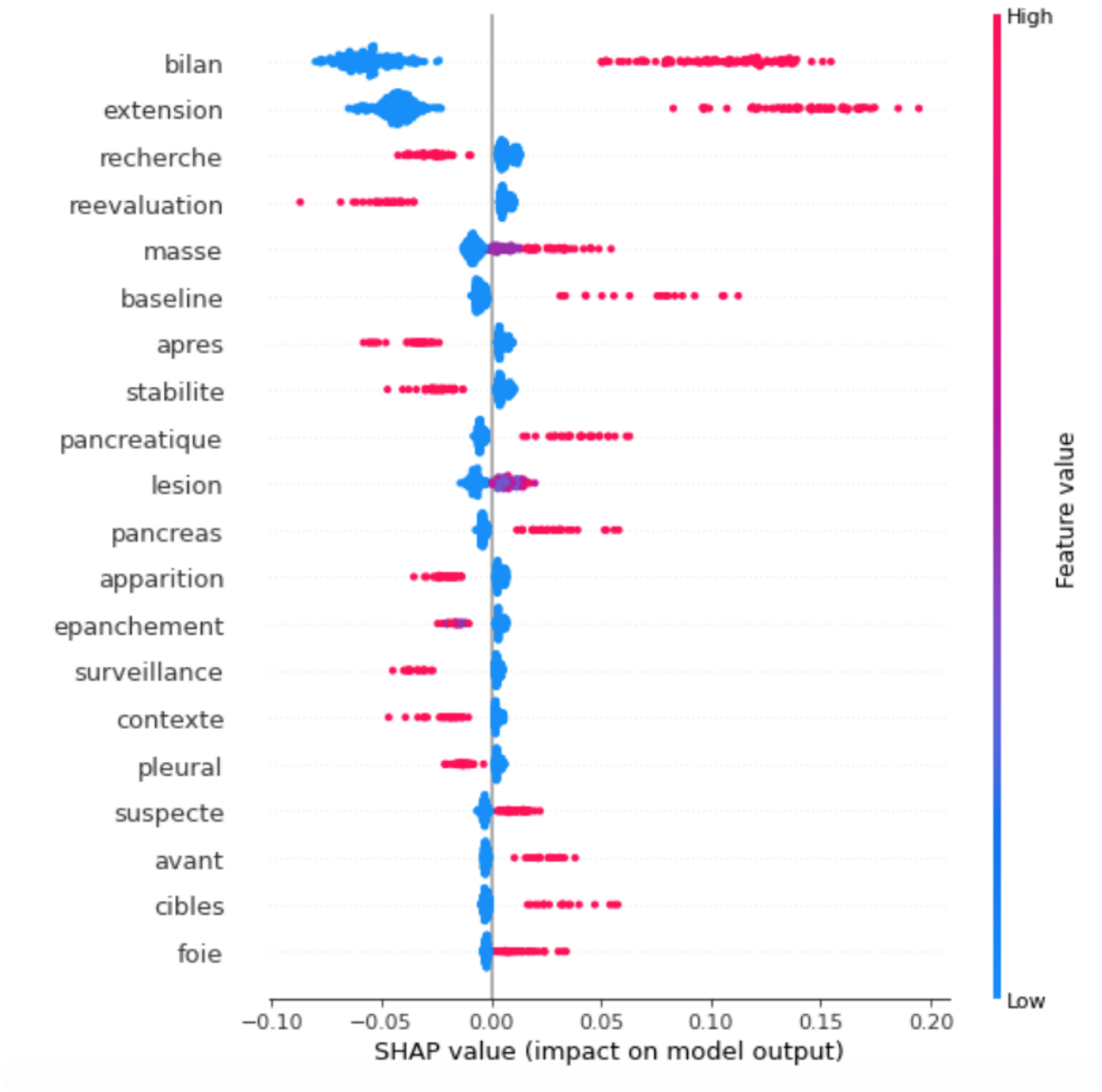
Supplementary Figure 4. Ratio between surgical visits with at least one pathology report available in the CDW and surgical visits without any pathology report available in the CDW for patients with a primary colon resection.



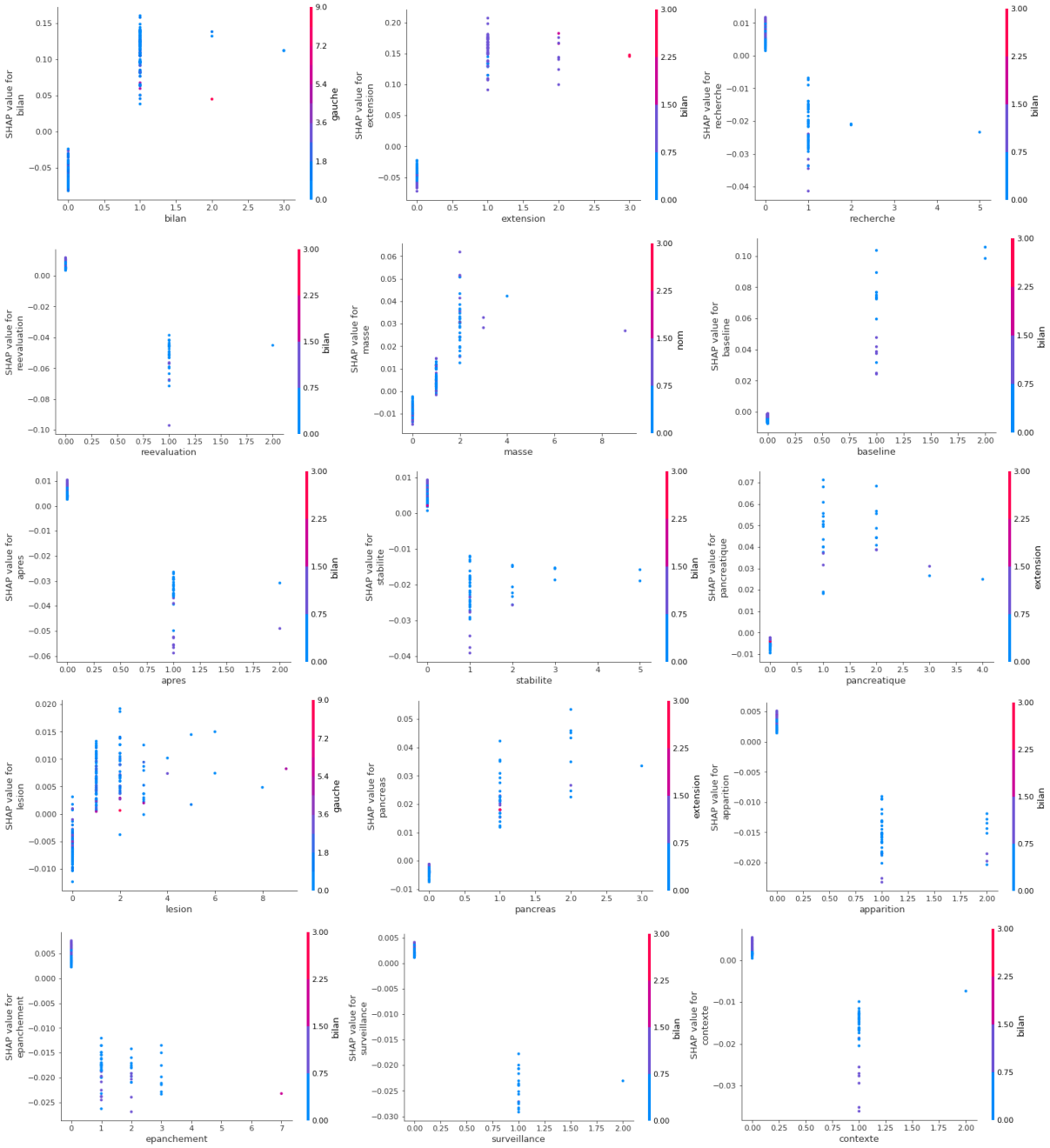
Supplementary Figure 5. Ratio of patients with at least one CT scan report available in the CDW



Supplementary Figure 6. Ratio of patients with at least one MDM report available in the CDW



Supplementary Figure 7. SHAP summary plot to rank the words used in the CT-scan classification algorithm, according to their respective weights



Supplementary Figure 8. SHAP dependence plots of the 15 words with the highest weights in the classification of CT-scans

Supplement to Kempf 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. Accepted Author Manuscript. https://doi.org/10.1002/ijc.33928

Appendix

Appendix 1. Classification Commune des Actes médicaux (CCAM) list codes related to the resection of colorectal primary tumors

HJFA004	HJFA001	HJFA002	HJFA006	HJFA007	HJFA011
HJFA012	HJFA014	HJFA017	HJFA019	HJFC023	HJFC031
HJFA008	HHFA018	HHFA009	HHFA026	HHFA023	HHFA006
HHFA022	HHFA008	HHFA021	HHFA017	HHFA010	HHFA014
HHFA005	HHFA024	HHFA004	HHFA002	HJFC023	HJFA012
HHFC296	HHFC040	HHFA030	HHFA031	HJFA006	HJFA007
HJFA019	HJFA005	HJFA003	HJFA018	HJFD002	HJFA004
HJFA002	HJFA001	HJFA015	HJFA016	HHFA029	HJFA017
HHFA028					

Classification Commune des Actes médicaux (CCAM) list codes related to the resections of colorectal secondary tumors

HLFC003	HLFA019	HLFA010	HLFA017	HLFA018	HLFC037
HLFA004	HLFA007	HLFC801	HLFA005	HLFA009	HLFA011
HLFA003	HLFA006	HLFC004	HLFC027	HLFC032	HLFC002
HLFA020	HLFA019	HLFA014	GFFA019	GFFA026	GFFA015
GFFA034	GFFA033	GFFA029	GFFA013	GFFA030	GFFA031
GFFA010	GFFA009	GFFA016	GFFA018	GFFA022	GFFA004
GFFA008	GFFA023	GFFA006	GFFA027	GFFC002	GFFA001
GFFA011	GFFA007	GFFA002	GFFA012	GFFA025	GFFA024
GFFA028	GFFA021	HPFC001	HPFA003		

Classification Commune des Actes médicaux (CCAM) list codes related to the destruction of colorectal secondary tumors by radiofrequence

HLNM001	HLNN900	HLNK001	HLNA007	HLNC003

Classification Commune des Actes médicaux (CCAM) list codes related to the destruction of colorectal secondary tumors by intraarterial antitumor agents

EDLL001

EDLF014

EDLF015

EDLF016

EDLF017

EDLL002