Interactive exploration of a global clinical network from a large breast cancer cohort

Nadir Sella[a,b,c,1], Anne-Sophie Hamy[b,d,e,2], Vincent Cabeli[c,3], Lauren Darrigues[e], Marick Laé[f,g], Fabien Reyal[b,e], Hervé Isambert[c*]

a. Institut Roche, Boulogne-Billancourt, France

b. Residual Tumor & Response to Treatment Laboratory, RT2Lab, INSERM, U932 Immunity and Cancer, Institut Curie, Paris, F-75248, France.

c. Laboratoire Physico Chimie Curie, Institut Curie, PSL Research University, CNRS UMR168, 75005, Paris, France

d. Department of Medical Oncology, Institut Curie, Saint-Cloud, F-92230, France.

e. Department of Surgery, Institut Curie, Université Paris, Paris, F-75248, France.

f. Department of Tumor Biology, Institut Curie, Paris, F-75248, France.

g. Department of Pathology, Henri Becquerel Cancer Center, INSERM U1245, UniRouen Normandy University

Corresponding Author information: Hervé Isambert, herve.isambert@curie.fr, +33 1 56 24 64 74

Co-corresponding Author information : Pr Fabien REYAL, 0033144324087, 0033615271980, fabien.reyal@curie.fr

Abstract

Despite unprecedented amount of information now available in medical records, health data remain underexploited due to their heterogeneity and complexity. Simple charts and hypothesis-driven statistics can no longer apprehend the content of information-rich clinical data. There is, therefore, a clear need for powerful interactive visualization tools enabling medical practitioners to perceive the patterns and insights gained by state-of-the-art machine learning algorithms. Here, we report an interactive graphical interface for use as the front end of a machine learning causal inference server (MIIC), to facilitate the visualization and comprehension by clinicians of relationships between clinically relevant variables. The widespread use of such tools, facilitating the interactive exploration of datasets, is crucial both for data visualization and for the generation of research hypotheses. We demonstrate the utility of the MIIC

interactive interface, by exploring the clinical network of a large cohort of breast cancer patients treated with neoadjuvant chemotherapy (NAC). This example highlights, in particular, the direct and indirect links between post-NAC clinical responses and patient survival. The MIIC interactive graphical interface has the potential to help clinicians identify actionable nodes and edges in clinical networks, thereby ultimately improving the patient care pathway.

Keywords: Machine learning - data visualization- residual cancer burden - neoadjuvant chemotherapy – breast cancer.

INTRODUCTION

The availability of health data from patient medical records is increasing, and these data constitute, in theory, a rich resource for research purposes. However, despite the unprecedented amount of information now available, health data remain underexploited due to their heterogeneity and complexity. There is, therefore, an urgent need for innovative tools, based on intuitive and interactive graphical interfaces, specifically designed for the exploration of health data by medical practitioners. Data visualization is gradually emerging as a new field of research, and graphical representations are used for two main purposes: (i) explanatory illustration, to highlight novel scientific insights graphically and to ensure efficient communication between scientists [1–4]; and (ii) exploratory analysis, searching for relationships previously overlooked and leading to new discoveries, thereby maximizing the potential of information-rich databases. We present here an *exploratory analysis* of a global clinical network from a large breast cancer cohort, with a novel interactive graphical interface for the exploration of health data.

We previously developed an advanced computational method for graphical analyses, including causal relationships, from multivariate data[5]. The underlying MIIC (Multivariate Information-based Inductive Causation) algorithm, which was released as an online server[6], uses a machine learning method combining constraint-based and information theory approaches to reconstruct causal, non-causal or mixed networks from large datasets. The MIIC algorithm was first developed to analyze categorical

genomic data[5, 6] and has recently been extended to the analysis of more challenging heterogeneous datasets, such as medical records, combining both categorical and continuous variables, in which interdependence is notoriously difficult to assess[7].

Breast cancer (BC) clinical datasets are particularly suitable for the type of exploratory analysis presented here, as BC is a complex heterogeneous disease highly variable in its aggressiveness and prognosis. BC remains one of the leading causes of cancer-related death among women. The BC patients included in the cohort analyzed here were treated with neoadjuvant (or preoperative) chemotherapy (NAC). NAC was originally restricted to patients with inflammatory or locally advanced BC, but is now the standard care for aggressive early-stage breast cancers, *i.e.* triple-negative (TNBC) and *HER2*-positive BCs[8, 9]. From the patient's viewpoint, the benefits of the neoadjuvant strategy include a greater feasibility of breast-conserving surgery and the prognostic stratification of risk obtained after analyses of the residual tumor burden at surgery. From the research and development standpoint, the neoadjuvant setting makes it possible to monitor the chemosensitivity of the tumor *in vivo*, and provides an opportunity for the rapid validation of research hypotheses and the acceleration of drug approval.

RESULTS

The global network displayed in Fig. 1 is accessible at https://miic.curie.fr/job_results_showcase.php?id=NEOREP. We discuss below some of the links inferred in the NEOREP network after grouping according to several clinically relevant concepts identified from published studies on BC.

MIIC performs quality control

MIIC first identifies relationships between a disease and the corresponding treatment. ER positivity — which is predictive of efficacy for anti-hormonal treatment[10]— is associated with the use of endocrine therapy (Fig. S3A), and a similar association is observed for *HER2*-positivity and trastuzumab use (Fig. S3B)[11]. Beyond cancer, significant associations are also found between depression and the use of psycholeptics (Fig. S3C), between thyroid disorders and thyroid hormone use (Fig. S3D),

and between hypertension and drugs for the treatment of cardiovascular diseases (Fig. S3E). More generally, comedication use is associated with the type of NAC (Fig. S3F), reflecting the greater likelihood of less toxic regimens being prescribed to fragile patients (patients on other types of medication) than to patients without comedication[12–14].

MIIC then identifies clinical factors known to be epidemiologically related (Fig. S4A). Menopause, a process occurring in older women, is directly linked to age (Fig. S4B) (median age: 43 years for premenopausal, versus 58 years for postmenopausal women). Postmenopausal status is associated with dyslipidemia (Fig. S4C)[15]. Consistent with these associations, body mass index (BMI) increases with age (Fig. S4A, S4D) and both factors, which have been reported to increase cardiovascular risks, are linked to hypertension (Fig. S4A, S4E). The number of drugs taken by a patient (comedication) increases with the number of comorbidities (Fig. S4A, S4F).

MIIC identifies inherent associations between variables

The duration of neoadjuvant treatment is directly linked to the type of NAC regimen delivered (Fig. 2A) reflecting the fact that anthracycline-based (AC) regimens usually include four cycles (median of 106 days, Fig. 2B), whereas sequential regimens in which anthracyclines are followed by taxanes are generally administered over six or eight cycles (median of 147 days, Fig. 2B). The number of nodes retrieved is associated with the type of axillary surgery (Fig. 2C), consistent with the fact that sentinel node (SLN) biopsy procedures were developed to reduce the number of lymph nodes removed during dissection (LND) (Fig. 2D)[16]. MIIC correctly represents the direct links between residual cancer burden (RCB) (Fig.2E) and the patterns making up this score, derived from measurements on the primary tumor bed (size, fraction of invasive cancer, cellularity) and the regional lymph nodes (number of positive lymph nodes).

MIIC identifies intra- and inter-modality associations

For the variables derived from pathology records, MIIC found associations between tumor grade, Ki67, and mitotic index (Fig. S5A-B-C), all of which are markers of tumor proliferation[17]. MIIC can also visualize links between patterns assessed in different ways. Measurements of pre-NAC tumor size evaluated clinically, by mammography and by MRI, were found to be closely related (Fig. S5C-E) as previously reported[18, 19]. Similarly, the response to treatment assessed clinically at NAC completion was found to be associated with histological size based on the surgical specimen (Fig. S5F).

MIIC provides insight into tumor biology and response to treatment

The presence of lymphovascular invasion (LVI) in the post-NAC specimen is associated with a higher RCB index, consistent with the strong resistance to chemotherapy of these tumors[20] (Fig. S6A). TNBCs and *HER2*-positive tumors have a higher pre-NAC mitotic index and more stromal TIL infiltration (Fig. S6B-C) than luminal BCs[21, 22]. Consistently, high TIL levels are significantly associated with histological grade 3 tumors (Fig. S6D).

MIIC reflects clinical practice

Several associations highlighted in the network reflect clinical practice decisions applied throughout BC centers. For example, the likelihood of performing conservative breast surgery depends on tumor histology (higher rates of mastectomy have been reported for patients with lobular or other histological types of tumor less likely to respond to NAC) [23, 24] (Fig. S7A) and is positively associated with the practice of oncoplastic surgery[25] (Fig. S7B). Similarly, lumpectomy is more frequently associated with radiation therapy than with mastectomy (Fig. S7C)[26–29]. After surgery, the addition of a second line of treatment by adjuvant chemotherapy, to decrease the risk of relapse, is driven by the identification of factors associated with a poor prognosis[30], such as high levels of lymph node involvement (Fig. S7D).

Beyond these well-established practices, MIIC also identified differences in clinical practices between the two centers of the cohort (Fig. 3A). For example, oncoplastic surgery and adjuvant chemotherapy were performed at only one of the two centers (Fig.3B-C); the NAC regimen also differed between

centers, with the Curie St Cloud center using more AC regimens than AC-taxane combinations, resulting in a shorter duration of NAC treatment (Fig. 3D-E).

MIIC traces the natural course of the disease

The natural course of BC may include local relapse, possibly followed by distant metastases, the trigger events leading to death[31–35] (Fig. 4A-C). Contralateral BC is often used in composite survival endpoints, such as distant relapse-free survival[36], but MIIC clearly identifies contralateral BC as an event being independent of other oncologic events and almost totally isolated from the rest of the network (Fig.1). Luminal BC is known to recur and develop metastases later than *HER2*-positive BC and TNBC (Fig. 4D)[21, 22, 37, 38]. The link between has also been found between PR negativity and a higher risk of brain metastasis[39–43] (Fig. 4E).

MIIC identifies unexpected associations, leading to new discoveries

With more than 15 associations involving treatment center (Fig. 3A), MIIC unmasked an unexpected "batch" effect relating to the site of BC treatment in this cohort. The observed differences reflect not only differences in therapeutic practice, but also in the characteristics of the population (differences in the proportion of women with psychological disorders, difference in incomes), in tumor presentation (tumor size), in pathological variable scoring (grade, presence of pre-NAC LVI, tumor cellularity, TILs), and in time to treatment within the care pathway.

MIIC identifies factors likely to improve prediction or prognosis

MIIC also favors new insights, e.g. comedication appears to protect against local relapse (Fig. 5A). Several retrospective studies have reported this association, with the use of statins[44], NSAIDs[45], or beta-blockers[46] found to have indirect anticarcinogenic effects. It has recently been suggested that these non-oncological treatments may have immunomodulatory and chemosensitizing effects[47].

MIIC suggests relevant combinations of predictive of prognostic biomarkers

MIIC may provide clues to combinations of new prognostic biomarkers likely to improve the prediction of response to chemotherapy, or post-NAC prognosis. Pre-NAC lymphovascular invasion (LVI) was found to be associated with both lower rates of clinical response (Fig. 5B) and shorter relapse-free survival (Fig. 5C). Both RCB (Fig. 5D-E) and post-NAC mitotic index (Fig.5D-F), a parameter rarely used in practice but nevertheless reported to be a predictor of BC recurrence[48, 49], appear to be strongly associated with the risk of death. MIIC may, therefore, be an efficient tool for identifying features likely to improve prognosis, by combining gold standard indicators with other parameters, such as post-NAC mitotic index, and post-NAC LVI, for example. Finally, MIIC also makes it possible to optimize the binning of residual cancer burden (RCB). RCB is a post-NAC histological score calculated as an increasing continuous index, and then subdivided into four classes (0, I, II, and III)[50]. Our analysis based on information maximization principles suggested a new unsupervised classification of RCB scores into three categories (Fig. 5E), with RCB=0 with low RCB values merged, in particular, into a single class associated with a good prognosis.

DISCUSSION

When applied to a large cohort of BC patients, the MIIC algorithm successfully (i) performed quality controls; (ii) identified intra- and inter-modality correlations; (iii) highlighted differences in clinical practice, including center specificities; (iv) traced the natural course of the disease; (v) highlighted unsuspected and hidden associations, leading to new discoveries. The interactive visualization and causal analyses provided by this algorithm make it a promising tool for fast and effective explorations of the increasing amount of available health data.

The amount of exploitable health data is increasing exponentially. The best known health data resource for cancer studies remains the SEER (Surveillance, Epidemiology, and End Results) database, which collects data from population-based cancer registries covering approximately 34.6% of the US population[51, 52]. By 2016, the National Cancer Database (NCDB) had amassed more than 34 million hospital records from cancer patients (almost four times the size of the SEER database), to become the

largest clinical cancer registry in the world[53]. In France, the French administrative health care database, the SNDS (*Système National des Données de Santé*), is one of the largest administrative databases in the domain of medicine, providing many opportunities for medical research[54, 55], as it covers 99% of the French population (about 66 million people). The French government is planning to ease access to this almost exhaustive population research resource, through release as part of the "Health data hub" project. Finally, beyond these structured databases, the largest mine of untapped data worldwide remains the content of electronic health records (EHRs), encompassing a full range of data (clinical notes, laboratory results, imaging, genetic data, etc.) relating to patient care. Recent advances in information technology have made it easier for both hospitals and healthcare institutions to collect large amounts of healthcare data.

Biomedical scientists are now facing new challenges in the management and analysis of massive, heterogeneous datasets[56]. These challenges include the development of tools for exploration and visualization, analytical methods, integration into a comprehensive overview, and translation of the findings into public health impact. The visualization of information makes it possible for users to find profound patterns in clinical data, through visual recognition. Simple charts cannot represent the complexity of big data analyses and fail to support multifaceted tasks effectively[3, 4]. There is, therefore, a need for sophisticated visualization tools dealing with many elements simultaneously and enabling users to perceive the patterns and insight generated by the algorithm[57]. Supplementary Table 1 shows the main data visualization tools used to present medical data. Many of the visual methods have been adopted directly from the field of data mining, but others, specific to the healthcare domain, have also been designed (Supplementary Table 2). For example, Happe and Drezen built the ePEPs toolbox, which displays relevant patterns extracted by eye from patient reimbursement data in the SNDS database, and supporting interactive exploration by researchers[58]. CARRE provides web-based components for interactive health data (fitness and biomarkers) visualization and risk analysis for the management of cardiorenal diseases[59]. The MITRE Corporation has also developed a web-based solution that provides an overview of an individual's health through graphical representations of EHR data, highlighting abnormal values[60]. None of these visualization programs has yet managed to bridge the gap between of

the large amounts of clinical data available and the discovery of clinical knowledge or paths for scientific research. By processing large heterogeneous sets of variables inherent to clinical records, MIIC provides physicians with a full picture of BC disease.

In addition to this use for visualization, the MIIC algorithm presents several other advantages for analyses, including its unsupervised nature, overcoming the need for training or human involvement. This feature makes it possible to obtain new knowledge through the automatic identification of patterns and dependences in the data, highlighting new interactions, and it may be of use for feature selection in machine learning models.

In conclusion, MIIC, an open-access, interactive, multitask tool, is designed to visualize datasets to help clinicians and researchers to understand the relationships between the variables within them. It opens up promising perspectives for guiding the generation of new hypotheses, helping clinicians identify actionable nodes and edges in clinical networks, and revealing new clues to relationships of interest for research purposes. Its widespread use in the field of health data could increase the accuracy of prediction for treatment responses and prognosis. This tool has the potential to improve the care pathway and, ultimately, the survival of patients.

METHODS

Patients and treatment

We analyzed a cohort of 1197 patients with non-metastatic BC treated by NAC, with or without trastuzumab, followed by surgery, at either of the two Institut Curie sites (Paris and Saint Cloud) between 2002 and 2012 (NEOREP Cohort, CNIL declaration number 1547270). We included unilateral, non-recurrent, non-inflammatory, non-metastatic tumors, and excluded T4 tumors. This study was conducted in accordance with institutional and ethical rules regarding research on tissue specimens and patients. Information on family history, clinical characteristics (age; menopausal status; body mass index) and tumor characteristics (clinical tumor stage and grade; histology; clinical nodal status; ER, PR and *HER2* status; BC subtype; mitotic index; Ki67; lymphovascular invasion) was retrieved from

electronic medical records. All the patients of the cohort received NAC, and additional treatments were decided in accordance with national guidelines.

Tumor samples and pathological review

In accordance with the guidelines used in France (Group for Evaluation of Prognostic Factors using Immunohistochemistry in Breast Cancer[61]), cases were considered estrogen receptor (ER)-positive or progesterone receptor (PR)-positive if at least 10% of the tumor cells expressed estrogen and/or progesterone receptors (ER/PR). Endocrine therapy was prescribed when this threshold was exceeded. *HER2*-negative status was defined as a score of 0 or 1+ for the tissue section stained by immunohistochemistry (IHC). Tissue sections with scores of IHC 2+ or IHC 3+ were then analyzed by fluorescence *in situ* hybridization (FISH) to confirm *HER2* positivity. BC tumors were classified into subtypes (TNBC, *HER2*-positive, and luminal *HER2*-negative [referred to hereafter as "luminal"]). BC subtypes were defined as follows: luminal, ER^+ or PR^+/ $HER2^-$; TNBC, $ER^-/PR^-/HER2^-$; *HER2*-positive BC, $HER2^+$. Pretreatment core needle biopsy specimens and/or the corresponding post-NAC surgical specimens were reviewed independently by breast disease experts for research purposes, to assess residual cancer burden index, and the levels of tumor-infiltrating lymphocytes. The pathological reviews of these specimens are described in detail elsewhere[62, 20, 63]. Pathological complete response (pCR) was defined as the absence of residual invasive cancer cells in the breast and axillary lymph nodes (ypT0/is þ/ypN0).

Survival endpoints

Relapse-free survival (RFS) was defined as the time from surgery to death, loco-regional recurrence or distant recurrence, whichever occurred first. Overall survival (OS) was defined as the time from surgery to death. The date of last known contact was retained for patients for whom none of these events were recorded. The cutoff date for survival analysis was March, 13th, 2019.

Variables of interest

The care pathway of BC patients eligible for neoadjuvant chemotherapy can be summarized as follows: *i)* pretreatment biopsy for BC diagnosis; *ii)* administration of chemotherapy as the first-line treatment; *iii)* removal of the tumor by surgery; *iv)* histological analysis of the specimens obtained; *v)* prescription of adjuvant treatments, if indicated (radiotherapy, hormonotherapy, chemotherapy); (*vi*) patients follow-up to monitor for relapse or death. We identified 94 clinically relevant variables from clinical, radiological, pathological and outcome data, which we grouped into 14 categories (hospital, history, co-medication, comorbidities, clinical baseline, baseline histology, pre-NAC pathology, treatment response, surgery, treatment, changes during NAC, post-NAC pathology, delayed relapse/survival, metastasis). For composite variables derived from raw variables (e.g. BC subtype, constructed from a combination of ER status, PR status, *HER2* status), both derived and raw variables were represented on the network.

MIIC algorithm

The functioning of the algorithm has been described in detail elsewhere[5, 7]. Briefly, starting from a fully connected network, the MIIC algorithm first removes dispensable edges by iteratively subtracting the most significant information contributions from indirect paths between each pair of variables. The remaining edges, the underlying effect of which cannot be explained by indirect paths, are then oriented based on the causality signature in the data, corresponding to the simultaneous head-to-head orientations of so-called "v-structures", X→Z←Y. In principle, propagation of v-structure orientations to downstream edges can also be implemented to fulfill underlying model class assumptions[64, 65] but are not applied on the NEOREP clinical network to ensure that MIIC algorithmic decisions are only based on information actually contained in the data.

Each edge corresponds to a "direct" association between two variables, that is, a statistical association that cannot be entirely explained by indirect effects involving other variables. Red and blue edges correspond to positive and negative (*i.e.* anti-correlated) associations, respectively. Four types of edge orientations are distinguished by the MIIC online server: *i)* directed edges with a gray arrowhead

represent inferred causal relationships; *ii)* bidirected edges (drawn with dashed lines) reflect the presence of a latent common cause (L) unobserved in the available dataset, *i.e.* $X \leftarrow (L) \rightarrow Y$; *iii)* directed edges with a colored (red or blue) arrowhead are consistentent with either a causal or a latent common cause relationship; and *iv)* undirected edges, whose orientation if it exists, cannot be inferred from non-perturbative data. The original algorithm was restricted to categorical variables[5], but MIIC has recently been extended to include continuous variables, the values of which are partitioned into optimal bins, maximizing mutual information with another (continuous or categorical) variable of interest, while preventing the overfitting of datasets of finite size due to the use of too many bins[7]. In particular, each continuous variable may have different information-maximizing partitions depending on the associated variable of interest. For instance, MIIC finds three maximally informative bins for the residual cancer burden (RCB) score in association with patient survival status (**Fig. S1A**), whereas eight RCB bins are required to estimate its mutual information with post-NAC cellularity correctly (**Fig. S1B).**

MIIC online server

The MIIC online server is freely accessible at https://miic.curie.fr and can be used with the Google Chrome, Mozilla Firefox, Edge, and Safari browsers. The user guide summarizing the main steps for running the MIIC algorithm is accessible at https://miic.curie.fr/user_guide.php, and an online video tutorial is available at: https://miic.curie.fr/tutorial.php. The workbench is available from https://miic.curie.fr/workbench.php. As input data, the user can upload a dataset formatted as a table with commas, semicolons, tabs, pipes or colons, as separators, without row names. Each variable can be either categorical or quantitative (discrete or continuous). Variables can be grouped into families, identified with different colors on the network. Missing values are allowed in the dataset and their possible statistical biases are taken into account by MIIC[7]. They should be indicated as "NA" in the dataset table. Once the dataset has been prepared, the user runs the algorithm, and an e-mail is sent when the job is completed.

MIIC output

The MIIC online server generates a visualization of the global network of the dataset. An example based on the NEOREP dataset is displayed in Fig. 1, and is accessible as an interactive network at https://miic.curie.fr/job_results_showcase.php?id=NEOREP.

Interactive exploration of the network

The distributions and neighborhoods of each node and edge of the inferred network can be explored through an interactive interface, through the mouse-over right- or left-click buttons on the browser page, as detailed in the online tutorials. Briefly, any variable can be highlighted by clicking on the network or through the "Search" toolbox (Fig. S2A). The corresponding plots can be downloaded as .png or .svg images. Each node can be explored individually in terms of counts (categorical variables, Fig. S2B-C) or distribution (continuous variables Fig. S2D-E). Each edge can be explored by a right click and the choice of "plot join distribution" or "plot discretization". The resulting plots are (i) proportion plots, with the edge representing the total association between two categorical variables (Fig. S2F); (ii) distribution histograms (Fig. S2G) or boxplots (Fig. S2H), in which the edge represents the total association between a categorical and a continuous variable or (iii) scatter plots (Fig. S2I), in which the edge represents the total association between two continuous variables. Additional options include inverting the *x* and *y* axes, the choice of frequency or absolute counts, or NA removal (proportion plots), and faceting or superimposing the variables (distribution histograms). All the figures presented here were generated with the MIIC online interactive visualization tool.

Data availability statement: All images and the associated network are publicly available at: https://miic.curie.fr/job_results_showcase.php?id=NEOREP. Data corresponding to the NEOREP cohort study will be available upon reasonable request.

Acknowledgements: NS acknowledges support from Sorbonne University (ATER), VC from ARC foundation and HI from ITMO Cancer, Institut Curie and CNRS

Author contributions: NS, VC and HI designed and implemented the machine learning and interactive exploration tools; ASH, LD, ML performed research; ASH and FR verified the data; NS, VC, HI, LD, FR and ASH contributed to data analysis; ASH and FR contributed to expert review. All authors contributed to data interpretation. LD, ASH, NS and HI wrote the paper. NS worked on the paper while being affiliated at b. and c. All authors had full access to all the data in the study and had final responsibility for the decision to submit for publication. [1]N.S., [2]A-S.H. and [3]V.C. **contributed equally** to this work.

Competing interests: The authors declare no competing interests.

References

1. Bärtschi M: Health Data Visualization-A review ∗ Seminar Collaborative Data Visualization, in 2015
2. Luo J, Wu M, Gopukumar D, et al: Big Data Application in Biomedical Research and Health Care: A Literature Review. Biomed Inform Insights 8:1–10, 2016
3. Ola O, Sedig K: Beyond simple charts: Design of visualizations for big health data [Internet]. Online J Public Health Inform 8, 2016[cited 2019 Aug 14] Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5302463/
4. Shneiderman B, Plaisant C, Hesse BW: Improving Healthcare with Interactive Visualization. Computer 46:58–66, 2013
5. Verny L, Sella N, Affeldt S, et al: Learning causal networks with latent variables from multivariate information in genomic data. PLoS Comput Biol 13:e1005662, 2017
6. Sella N, Verny L, Uguzzoni G, et al: MIIC online: a web server to reconstruct causal or non-causal networks from non-perturbative data. Bioinformatics 34:2311–2313, 2018
7. Cabeli V, Verny L, Sella N, et al: Learning clinical networks from medical records based on information estimates in mixed-type data [Internet]. PLoS Comput Biol 16, 2020[cited 2021 Feb 4] Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7259796/
8. Brandão M, Reyal F, Hamy A-S, et al: Neoadjuvant treatment for intermediate/high-risk HER2-positive and triple-negative breast cancers: no longer an “option” but an ethical obligation. ESMO Open 4:e000515, 2019
9. Reyal F, Hamy AS, Piccart MJ: Neoadjuvant treatment: the future of patients with breast cancer. ESMO Open 3:e000371, 2018
10. Burstein HJ, Temin S, Anderson H, et al: Adjuvant Endocrine Therapy for Women With Hormone Receptor–Positive Breast Cancer: American Society of Clinical Oncology Clinical Practice Guideline Focused Update. J Clin Oncol 32:2255–2269, 2014
11. Wilson FR, Coombes ME, Wylie Q, et al: Herceptin® (trastuzumab) in HER2-positive early breast

cancer: protocol for a systematic review and cumulative network meta-analysis. Syst Rev 6:196, 2017
12. Aaldriks AA, Maartense E, Nortier HJWR, et al: Prognostic factors for the feasibility of chemotherapy and the Geriatric Prognostic Index (GPI) as risk profile for mortality before chemotherapy in the elderly. Acta Oncol 55:15–23, 2016
13. van Leeuwen RWF, Swart EL, Boven E, et al: Potential drug interactions in cancer therapy: a prevalence study using an advanced screening method. Ann Oncol 22:2334–2341, 2011
14. Popa MA, Wallace KJ, Brunello A, et al: POTENTIAL DRUG INTERACTIONS AND CHEMOTOXICITY IN OLDER PATIENTS WITH CANCER RECEIVING CHEMOTHERAPY. J Geriatr Oncol 5:307–314, 2014
15. Wang N, Qin MZ, Cui J: Lipid profile comparison between pre- and post-menopausal women. Zhonghua Xin Xue Guan Bing Za Zhi 44:799–804, 2016
16. Veronesi U, Paganelli G, Viale G, et al: Sentinel-lymph-node biopsy as a staging procedure in breast cancer: update of a randomised controlled study. Lancet Oncol 7:983–990, 2006
17. Weidner N, Moore DH, Vartanian R: Correlation of Ki-67 antigen expression with mitotic figure index and tumor grade in breast carcinomas using the novel "paraffin"-reactive MIB1 antibody. Hum Pathol 25:337–342, 1994
18. Cortadellas T, Argacha P, Acosta J, et al: Estimation of tumor size in breast cancer comparing clinical examination, mammography, ultrasound and MRI—correlation with the pathological analysis of the surgical specimen. Gland Surg 6:330–335, 2017
19. Berg WA, Gutierrez L, NessAiver MS, et al: Diagnostic accuracy of mammography, clinical examination, US, and MR imaging in preoperative assessment of breast cancer. Radiology 233:830–849, 2004
20. Hamy A-S, Lam G-T, Laas E, et al: Lymphovascular invasion after neoadjuvant chemotherapy is strongly associated with poor prognosis in breast carcinoma. Breast Cancer Res Treat 169:295–304, 2018
21. Meyers MO, Klauber-Demore N, Ollila DW, et al: Impact of breast cancer molecular subtypes on locoregional recurrence in patients treated with neoadjuvant chemotherapy for locally advanced breast cancer. Ann Surg Oncol 18:2851–2857, 2011
22. Lowery AJ, Kell MR, Glynn RW, et al: Locoregional recurrence after breast cancer surgery: a systematic review by receptor phenotype. Breast Cancer Res Treat 133:831–841, 2012
23. Waljee JF, Hu ES, Newman LA, et al: Predictors of re-excision among women undergoing breast-conserving surgery for cancer. Ann Surg Oncol 15:1297–1303, 2008
24. Truin W, Vugts G, Roumen RMH, et al: Differences in Response and Surgical Management with Neoadjuvant Chemotherapy in Invasive Lobular Versus Ductal Breast Cancer. Ann Surg Oncol 23:51–57, 2016
25. Munhoz AM, Montag E, Gemperli R: Oncoplastic breast surgery: indications, techniques and perspectives. Gland Surg 2:143–157, 2013
26. Buchholz TA: Radiation Therapy for Early-Stage Breast Cancer after Breast-Conserving Surgery. New England Journal of Medicine 360:63–70, 2009
27. Carlson RW, Allred DC, Anderson BO, et al: Invasive breast cancer. J Natl Compr Canc Netw 9:136–222, 2011
28. Eifel P, Axelson JA, Costa J, et al: National Institutes of Health Consensus Development Conference Statement: adjuvant therapy for breast cancer, November 1-3, 2000. J Natl Cancer Inst 93:979–989, 2001
29. Halberg FE, Shank BM, Haffty BG, et al: Conservative surgery and radiation in the treatment of stage I and II carcinoma of the breast. American College of Radiology. ACR Appropriateness Criteria. Radiology 215 Suppl:1193–1205, 2000
30. Masuda N, Lee S-J, Ohtani S, et al: Adjuvant Capecitabine for Breast Cancer after Preoperative Chemotherapy. N Engl J Med 376:2147–2159, 2017
31. Dent R, Valentini A, Hanna W, et al: Factors associated with breast cancer mortality after local recurrence. Curr Oncol 21:e418–e425, 2014
32. Whelan T, Clark R, Roberts R, et al: Ipsilateral breast tumor recurrence postlumpectomy is predictive of subsequent mortality: results from a randomized trial. Investigators of the Ontario

Clinical Oncology Group. Int J Radiat Oncol Biol Phys 30:11–16, 1994
33. Kurtz JM, Spitalier JM, Amalric R, et al: The prognostic significance of late local recurrence after breast-conserving therapy. Int J Radiat Oncol Biol Phys 18:87–93, 1990
34. Sopik V, Nofech-Mozes S, Sun P, et al: The relationship between local recurrence and death in early-stage breast cancer. Breast Cancer Res Treat 155:175–185, 2016
35. Witteveen A, Kwast ABG, Sonke GS, et al: Survival after Locoregional Recurrence or Second Primary Breast Cancer: Impact of the Disease-Free Interval. PLOS ONE 10:e0120832, 2015
36. Hudis CA, Barlow WE, Costantino JP, et al: Proposal for standardized definitions for efficacy end points in adjuvant breast cancer trials: the STEEP system. J Clin Oncol 25:2127–2132, 2007
37. Voduc KD, Cheang MCU, Tyldesley S, et al: Breast Cancer Subtypes and the Risk of Local and Regional Relapse. JCO 28:1684–1691, 2010
38. Wu X, Baig A, Kasymjanova G, et al: Pattern of Local Recurrence and Distant Metastasis in Breast Cancer By Molecular Subtype [Internet]. Cureus 8, 2016[cited 2021 Feb 4] Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5222631/
39. Snell CE, Gough M, Middleton K, et al: Absent progesterone receptor expression in the lymph node metastases of ER-positive, HER2-negative breast cancer is associated with relapse on tamoxifen. J Clin Pathol 70:954–960, 2017
40. Nishimura R, Osako T, Okumura Y, et al: Changes in the ER, PgR, HER2, p53 and Ki-67 biological markers between primary and recurrent breast cancer: discordance rates and prognosis. World J Surg Oncol 9:131, 2011
41. Nishimura R, Osako T, Nishiyama Y, et al: Evaluation of factors related to late recurrence--later than 10 years after the initial treatment--in primary breast cancer. Oncology 85:100–110, 2013
42. Darlix A, Griguolo G, Thezenas S, et al: Hormone receptors status: a strong determinant of the kinetics of brain metastases occurrence compared with HER2 status in breast cancer. J Neurooncol 138:369–382, 2018
43. Zhou L, Zhou W, Zhang H, et al: Progesterone suppresses triple-negative breast cancer growth and metastasis to the brain via membrane progesterone receptor α. Int J Mol Med 40:755–761, 2017
44. Ahern TP, Pedersen L, Tarp M, et al: Statin prescriptions and breast cancer recurrence risk: a Danish nationwide prospective cohort study. J Natl Cancer Inst 103:1461–1468, 2011
45. Kwan ML, Habel LA, Slattery ML, et al: NSAIDs and Breast Cancer Recurrence in a Prospective Cohort Study. Cancer Causes Control 18:613–620, 2007
46. Powe DG, Voss MJ, Zänker KS, et al: Beta-Blocker Drug Therapy Reduces Secondary Cancer Formation in Breast Cancer and Improves Cancer Specific Survival. Oncotarget 1:628–638, 2010
47. Hamy A-S, Derosa L, Valdelièvre C, et al: Comedications influence immune infiltration and pathological response to neoadjuvant chemotherapy in breast cancer. OncoImmunology 9:1677427, 2020
48. Farrugia DJ, Landmann A, Diego E, et al: Mitotic index to predict breast cancer recurrence after neoadjuvant systemic therapy. JCO 34:e23265–e23265, 2016
49. Pattali S, Harding N, Visotcky A, et al: Value of mitotic index in residual tumors following neoadjuvant therapy for breast cancer: Single institution experience. JCO 34:548–548, 2016
50. Symmans WF, Peintinger F, Hatzis C, et al: Measurement of Residual Breast Cancer Burden to Predict Survival After Neoadjuvant Chemotherapy. Journal of Clinical Oncology 25:4414–4422, 2007
51. Duggan MA, Anderson WF, Altekruse S, et al: The Surveillance, Epidemiology and End Results (SEER) Program and Pathology: Towards Strengthening the Critical Relationship. Am J Surg Pathol 40:e94–e102, 2016
52. James B. Yu MD: NCI SEER Public-Use Data: Applications and Limitations in Oncology Research [Internet]. Cancer Network , 2009[cited 2019 Aug 27] Available from: https://www.cancernetwork.com/oncology-journal/nci-seer-public-use-data-applications-and-limitations-oncology-research
53. Boffa DJ, Rosen JE, Mallin K, et al: Using the National Cancer Database for Outcomes Research: A Review. JAMA Oncol 3:1722–1728, 2017
54. Bezin J, Duong M, Lassalle R, et al: The national healthcare system claims databases in France, SNIIRAM and EGB: Powerful tools for pharmacoepidemiology. Pharmacoepidemiol Drug Saf

26:954–962, 2017
55. Tuppin P, Rudant J, Constantinou P, et al: Value of a national administrative database to guide public decisions: From the système national d'information interrégimes de l'Assurance Maladie (SNIIRAM) to the système national des données de santé (SNDS) in France [Internet]. /data/revues/03987620/v65sS4/S0398762017304315/ , 2017[cited 2019 Aug 13] Available from: https://www.em-consulte.com/en/article/1140905
56. Margolis R, Derr L, Dunn M, et al: The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. J Am Med Inform Assoc 21:957–958, 2014
57. Keim D, Andrienko G, Fekete J-D, et al: Visual Analytics: Definition, Process, and Challenges [Internet], in Kerren A, Stasko JT, Fekete J-D, et al (eds): Information Visualization. Berlin, Heidelberg, Springer Berlin Heidelberg, 2008, pp 154–175[cited 2019 Aug 14] Available from: http://link.springer.com/10.1007/978-3-540-70956-5_7
58. Happe A, Drezen E: A visual approach of care pathways from the French nationwide SNDS database - from population to individual records: the ePEPS toolbox [Internet], 2018[cited 2019 Aug 18] Available from: https://hal-univ-rennes1.archives-ouvertes.fr/hal-01697626
59. Zhao Y, Parvinzamir F, Wei H, et al: Visual Analytics for Health Monitoring and Risk Management in CARRE. E-Learning and Games; 10th International Conference, Edutainment 2016, Hangzhou, China, April 14-16, 2016, Revised Selected Papers 9654:380–391, 2016
60. Ledesma A, Al-Musawi M, Nieminen H: Health figures: an open source JavaScript library for health data visualization [Internet]. BMC Med Inform Decis Mak 16, 2016[cited 2019 Aug 14] Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4802654/
61. [Recommendations for the immunohistochemistry of the hormonal receptors on paraffin sections in breast cancer. Update 1999. Group for Evaluation of Prognostic Factors using Immunohistochemistry in Breast Cancer (GEFPICS-FNCLCC)]. Ann Pathol 19:336–343, 1999
62. Hamy A-S, Pierga J-Y, Sabaila A, et al: Stromal lymphocyte infiltration after neoadjuvant chemotherapy is associated with aggressive residual disease and lower disease-free survival in HER2-positive breast cancer. Ann Oncol 28:2233–2240, 2017
63. Hamy-Petit A-S, Belin L, Bonsang-Kitzis H, et al: Pathological complete response and prognosis after neoadjuvant chemotherapy for HER2-positive breast cancers before and after trastuzumab era: results from a real-life cohort. Br J Cancer 114:44–52, 2016
64. Affeldt S, Isambert H: Robust reconstruction of causal graphical models based on conditional 2-point and 3-point information, in Proceedings of the 31th conference on Uncertainty in Artificial Intelligence (UAI), 2015
65. Affeldt S, Verny L, Isambert H: 3off2: A network reconstruction algorithm based on 2-point and 3-point information statistics. BMC Bioinformatics 17 Suppl 2:12, 2016

Figures Legends

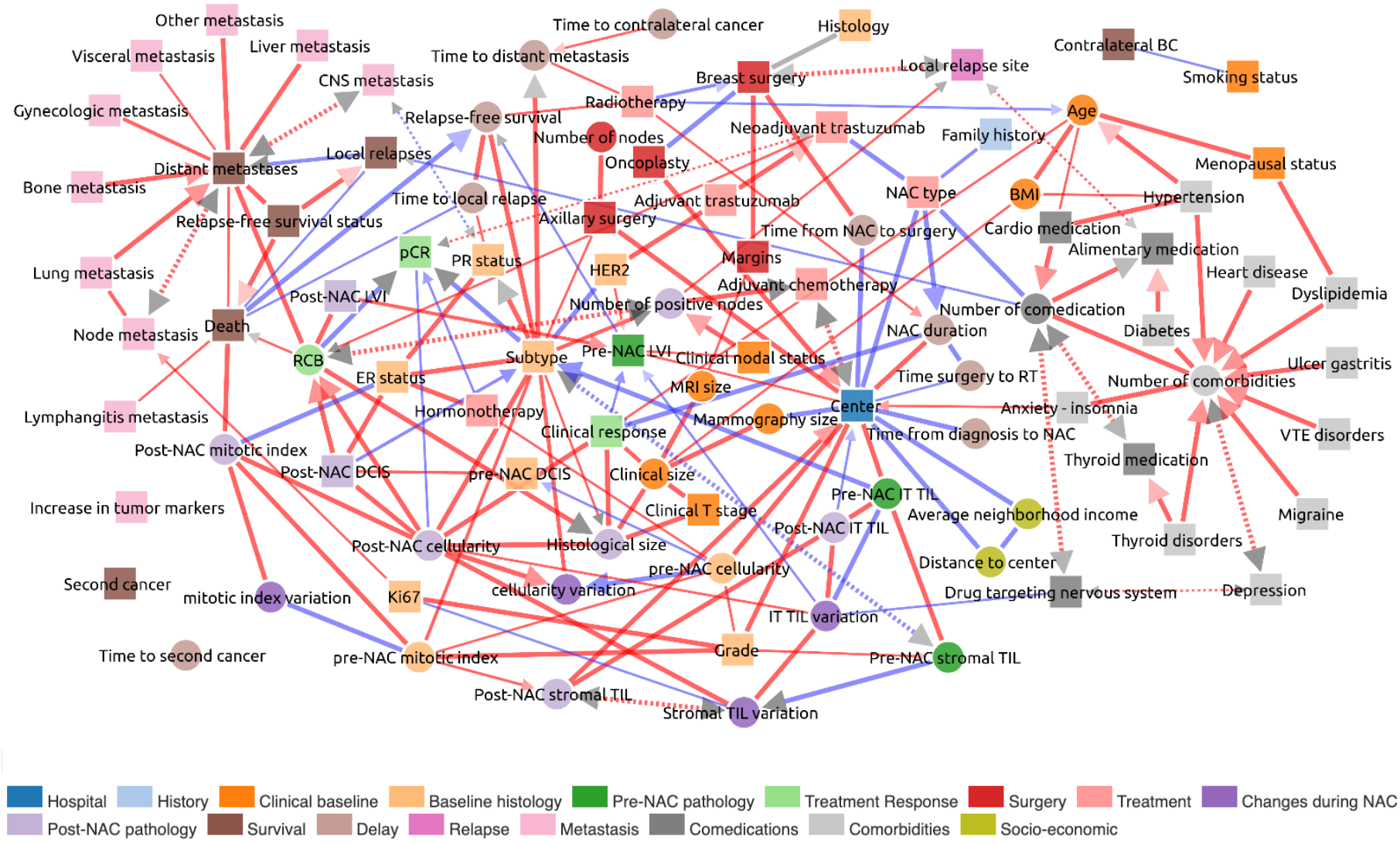
Figure 1: MIIC global network for the NEOREP breast cancer cohort. Each node corresponds to a variable of the dataset, with circles indicating continuous variables and squares indicating categorical variables. The colors define a category of variables, as detailed under the figure. Each edge corresponds to a "direct" association between two variables with different types of orientation described in Methods. *BC= breast cancer, BMI= body mass index, DCIS=ductal carcinoma* in situ*, ER= estrogen receptor status, LVI= lymphovascular invasion, NAC= neoadjuvant chemotherapy, CNS= central nervous system, pCR= pathological complete response, PR=progesterone receptor status, RCB= residual cancer burden, TILs= tumor-infiltrating lymphocytes*. Blue edges indicate negative partial correlations, red edges indicate positive partial correlations.

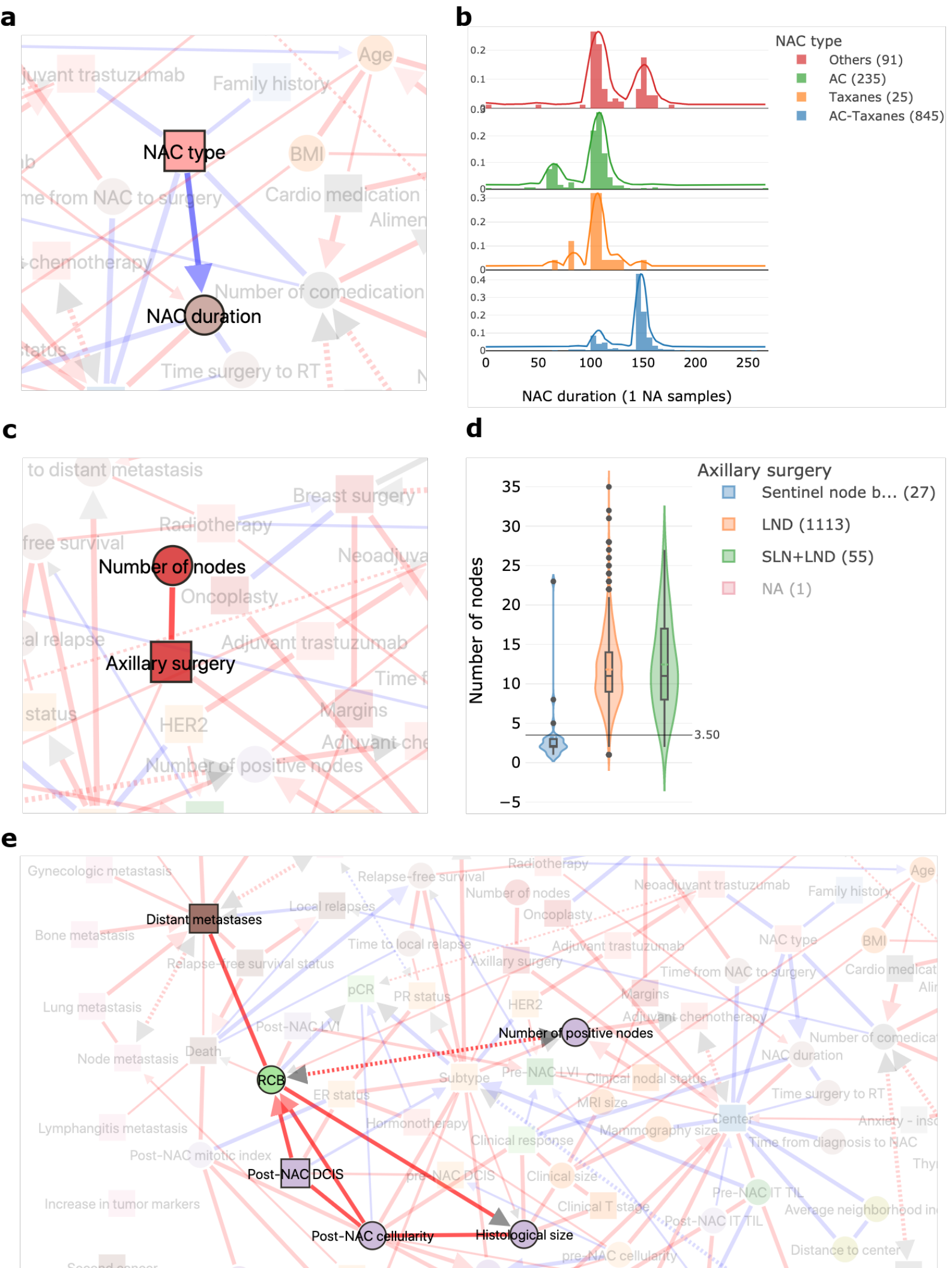
Figure 2. The MIIC interactive online interface identifies inherent associations between variables.
a) NAC type is directly correlated with NAC duration. NAC=neoadjuvant chemotherapy b) Distribution of neoadjuvant chemotherapy (NAC) duration (in days) according to the NAC regimen administered: anthracyclines (AC), taxanes or sequential AC-taxanes c) The number of axillary nodes in the histological specimen depends on the type of axillary surgery performed d) Boxplot showing the number of axillary nodes removed according to the type of surgery performed: lymph node dissection (LND), sentinel lymph node biopsy (SLN) or both e) Network interactions of the RCB node with the five patterns making up the RCB score.

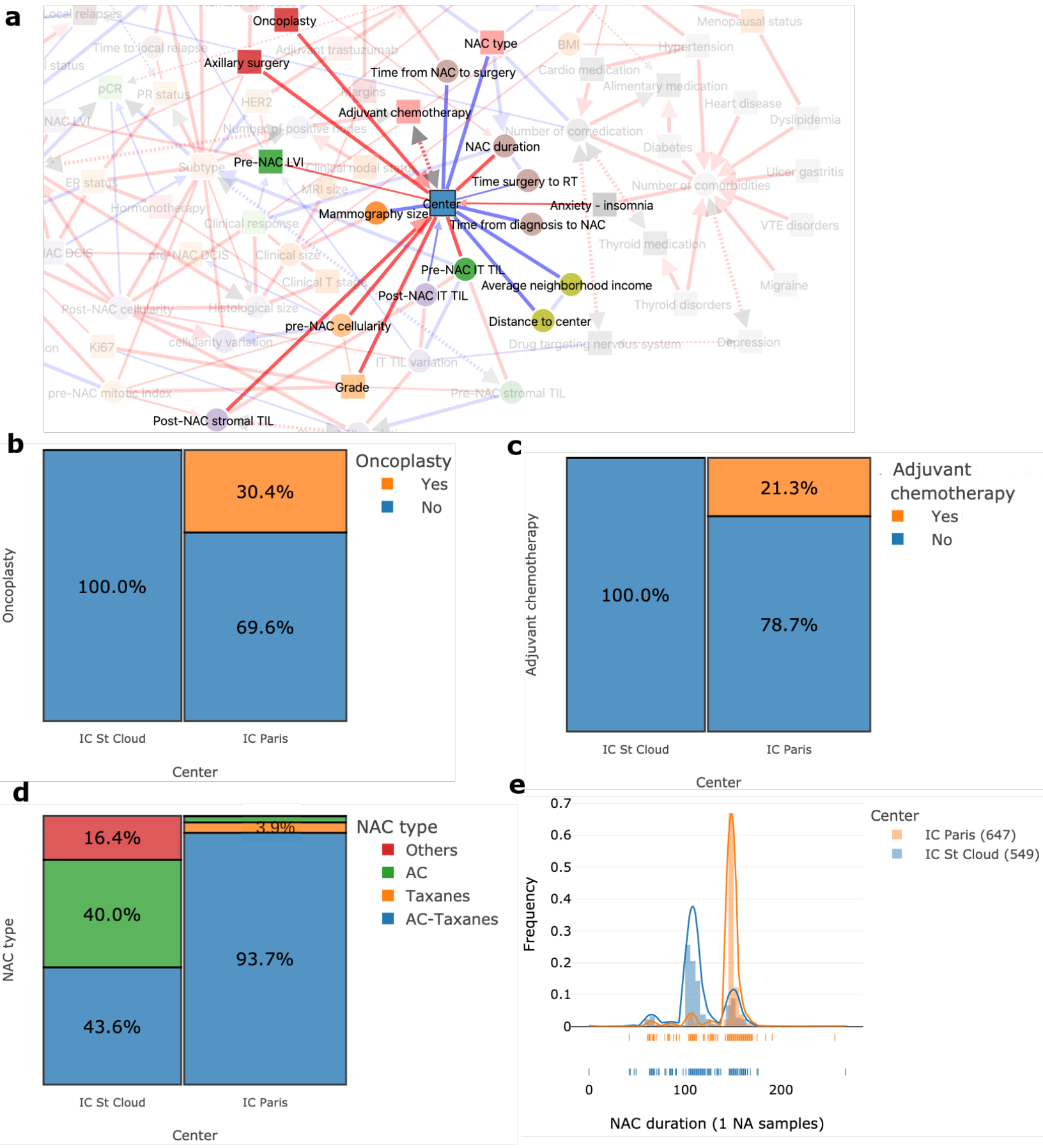
Figure 3. MIIC identifies differences in clinical practices between the two centers of the cohort
a) Network interactions around the node "center" of treatment. b) Proportion of patients undergoing oncoplastic surgery, according to treatment center: Paris or St Cloud c) Proportion of patients receiving adjuvant chemotherapy according to treatment center: Paris or St Cloud. d) Proportion of the various NAC regimens according to treatment center. e) Distribution plot for NAC duration in days, according to treatment center.

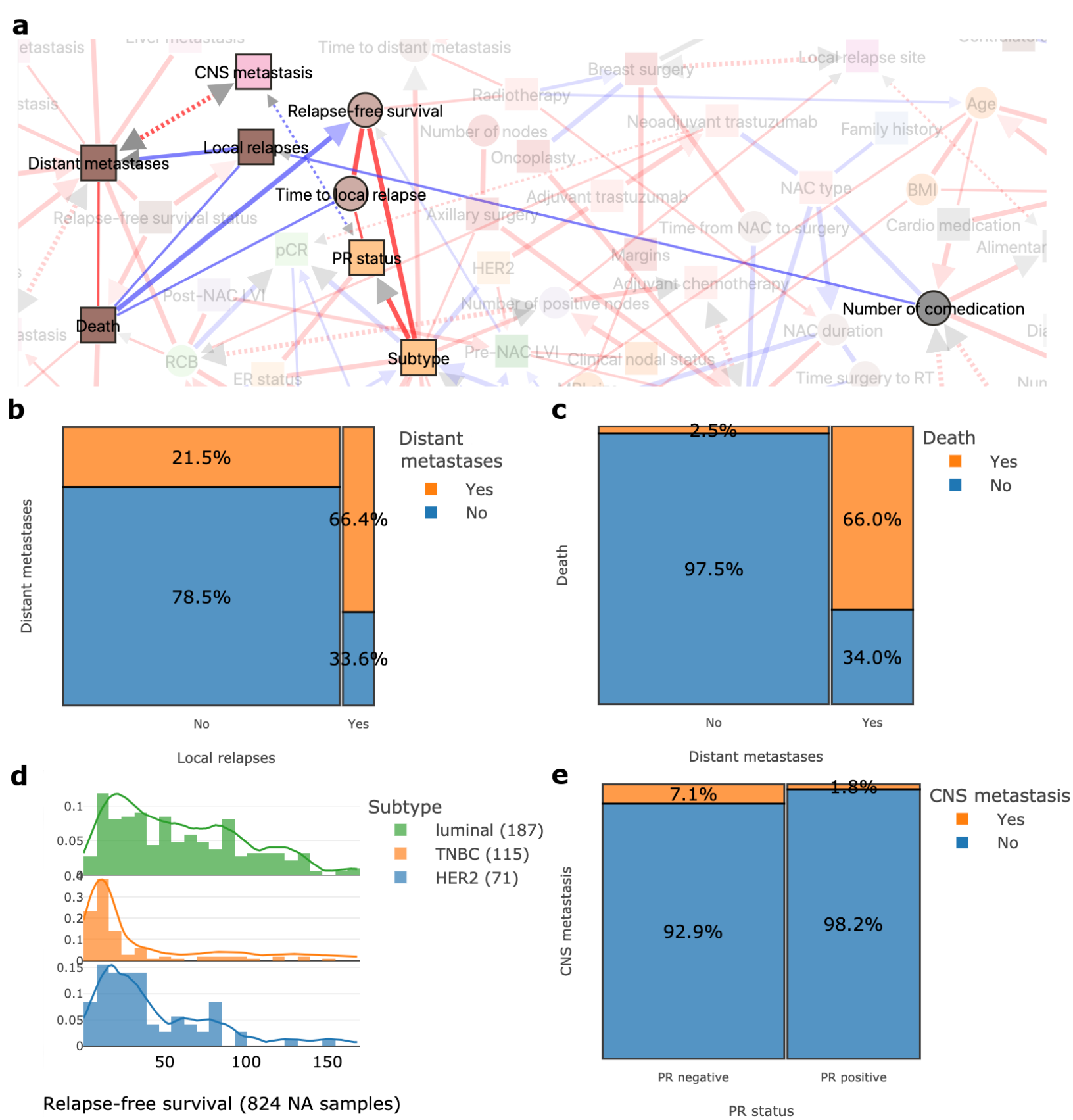
Figure 4. MIIC traces the natural course of the disease
a) Network interactions showing links between relapses, metastases and death in breast cancer. b) Proportion of distant metastases according to the occurrence or absence of local relapses. c) Proportion of deaths according to distant metastasis status. d) Distribution plot for relapse-free survival (in months) according to breast cancer subtype. e) Proportion plot displaying the relationship between central nervous system (CNS) metastasis and progesterone receptor (PR) status.

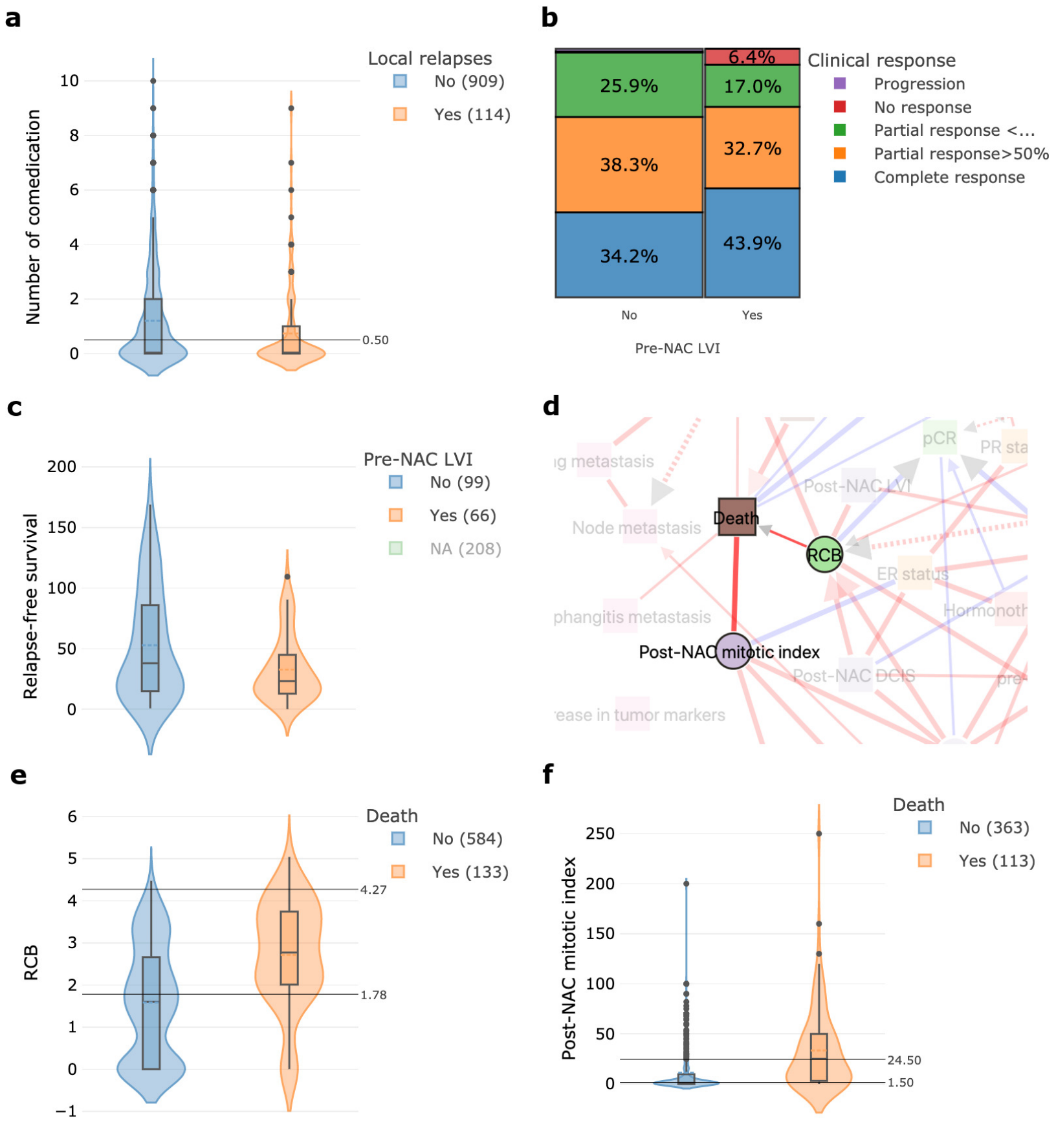
Figure 5 MIIC identifies factors likely to improve prediction or prognosis.
a) Network interaction displaying the link between local relapse occurrence and the number of drugs taken (comedication). b) Proportion plot showing the percentage of different clinical responses according to the presence or absence of pre-NAC lymphovascular invasion. c) Boxplot of relapse-free survival according to the presence or absence of pre-NAC lymphovascular invasion. d) Network interaction displaying the link between death, RCB and post-NAC mitotic index. e) Boxplot of RCB values according to vital status. f) Boxplot of post-NAC mitotic index according to vital status.


Other metastasis
Visceral metastasis
Liver metastasis
Time to contralateral cancer
Histology
Time to distant metastasis
Contralateral BC
Smoking status
Local relapse site
Breast surgery
CNS metastasis
Gynecologic metastasis
Radiotherapy
Age
Relapse-free survival
Number of nodes
Neoadjuvant trastuzumab
Family history
Distant metastases
Local relapses
Oncoplasty
Menopausal status
Bone metastasis
Time to local relapse
Adjuvant trastuzumab
NAC type
BMI
Hypertension
Relapse-free survival status
Axillary surgery
Time from NAC to surgery
Cardio medication
Alimentary medication
pCR
PR status
Margins
Lung metastasis
HER2
Heart disease
Adjuvant chemotherapy
Dyslipidemia
Post-NAC LVI
Number of positive nodes
Number of comedication
Node metastasis
Death
NAC duration
Diabetes
Pre-NAC LVI
Clinical nodal status
Subtype
RCB
Ulcer gastritis
ER status
Time surgery to RT
MRI size
Number of comorbidities
Center
Lymphangitis metastasis
Hormonotherapy
Anxiety - insomnia
Mammography size
Clinical response
Time from diagnosis to NAC
VTE disorders
Post-NAC mitotic index
Thyroid medication
Post-NAC DCIS
pre-NAC DCIS
Clinical size
Pre-NAC IT TIL
Increase in tumor markers
Clinical T stage
Average neighborhood income
Migraine
Post-NAC IT TIL
Thyroid disorders
Post-NAC cellularity
Histological size
Distance to center
pre-NAC cellularity
Second cancer
mitotic index variation
Ki67
cellularity variation
Drug targeting nervous system
Depression
IT TIL variation
Time to second cancer
Grade
pre-NAC mitotic index
Pre-NAC stromal TIL
Post-NAC stromal TIL
Stromal TIL variation
Hospital
History
Clinical baseline
Baseline histology
Pre-NAC pathology
Treatment Response
Surgery
Treatment
Changes during NAC
Post-NAC pathology
Survival
Delay
Relapse
Metastasis
Comedications
Comorbidities
Socio-economic


a
NAC type
NAC duration
Age
Family history
BMI
Cardio medication
Number of comedication
Time surgery to RT
b
NAC type
Others (91)
AC (235)
Taxanes (25)
AC-Taxanes (845)
NAC duration (1 NA samples)
c
Number of nodes
Axillary surgery
Breast surgery
Radiotherapy
Oncoplasty
Adjuvant trastuzumab
HER2
Margins
Number of positive nodes
d
Axillary surgery
Sentinel node b... (27)
LND (1113)
SLN+LND (55)
NA (1)
Number of nodes
3.50
e
Distant metastases
RCB
Post-NAC DCIS
Post-NAC cellularity
Histological size
Number of positive nodes


a
Oncoplasty
Axillary surgery
NAC type
Time from NAC to surgery
Adjuvant chemotherapy
NAC duration
Pre-NAC LVI
Time surgery to RT
Center
Mammography size
Anxiety - insomnia
Time from diagnosis to NAC
Pre-NAC IT TIL
Average neighborhood income
Post-NAC IT TIL
Distance to center
pre-NAC cellularity
Grade
Post-NAC stromal TIL
b
Oncoplasty
Yes
No
30.4%
100.0%
69.6%
IC St Cloud
IC Paris
Center
c
Adjuvant chemotherapy
Yes
No
21.3%
100.0%
78.7%
IC St Cloud
IC Paris
Center
d
NAC type
Others
AC
Taxanes
AC-Taxanes
16.4%
3.9%
40.0%
93.7%
43.6%
IC St Cloud
IC Paris
Center
e
Center
IC Paris (647)
IC St Cloud (549)
Frequency
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
100
200
NAC duration (1 NA samples)


a
CNS metastasis
Relapse-free survival
Distant metastases
Local relapses
Time to local relapse
PR status
Death
Subtype
Number of comedication
Time to distant metastasis
Breast surgery
Local relapse site
Radiotherapy
Age
Number of nodes
Neoadjuvant trastuzumab
Family history
Oncoplasty
NAC type
BMI
Adjuvant trastuzumab
Axillary surgery
Relapse-free survival status
Time from NAC to surgery
Cardio medication
pCR
Margins
HER2
Adjuvant chemotherapy
Post-NAC LVI
Number of positive nodes
NAC duration
RCB
Pre-NAC LVI
Clinical nodal status
ER status
Time surgery to RT
b
Distant metastases
21.5%
78.5%
66.4%
33.6%
No
Yes
Local relapses
Distant metastases
Yes
No
c
Death
2.5%
97.5%
66.0%
34.0%
No
Yes
Distant metastases
Death
Yes
No
d
Subtype
luminal (187)
TNBC (115)
HER2 (71)
0.1
0.05
0.4
0.3
0.2
0.1
0
0.15
0.1
0.05
0
50
100
150
Relapse-free survival (824 NA samples)
e
CNS metastasis
7.1%
92.9%
1.8%
98.2%
PR negative
PR positive
PR status
CNS metastasis
Yes
No


a
Local relapses
No (909)
Yes (114)
Number of comedication
10
8
6
4
2
0
0.50
b
Clinical response
Progression
No response
Partial response <...
Partial response>50%
Complete response
25.9%
38.3%
34.2%
6.4%
17.0%
32.7%
43.9%
No
Yes
Pre-NAC LVI
c
Pre-NAC LVI
No (99)
Yes (66)
NA (208)
Relapse-free survival
200
150
100
50
0
d
Death
RCB
Post-NAC mitotic index
Node metastasis
Post-NAC LVI
pCR
ER status
Post-NAC DCIS
e
Death
No (584)
Yes (133)
RCB
6
5
4
3
2
1
0
−1
4.27
1.78
f
Death
No (363)
Yes (113)
Post-NAC mitotic index
250
200
150
100
50
0
24.50
1.50