Computational modeling approaches for circulating cell-free DNA in oncology

Linh Nguyen Phuong, Sébastien Salas, Sébastien Benzekry

Computational modeling approaches for circulating cell-free DNA in oncology

Linh Nguyen Phuong[1], Sébastien Salas[1,2] and Sébastien Benzekry[1]

[1] COMPutational pharmacology and clinical Oncology Department, Inria Sophia Antipolis-Méditerranée, Cancer Research Centre of Marseille, Inserm UMR1068, CNRS UMR7258, Aix Marseille University UM105, Marseille, France

[2] Assistance Publique-Hôpitaux de Marseille, Timone Hospital, Aix Marseille University, Marseille, France

Social media handles:

Authors: @SBenzekry

Institutions: @aphm_actu, @inria_sophia, @crcm_marseille

ABSTRACT

Liquid biopsy has emerged as a powerful tool for cancer early diagnosis, prognosis, and treatment monitoring across a wide range of cancer types. The non-invasive collection of blood markers enables real-time insights into the disease biology. Cell-free circulating DNA (cfDNA) offers a potential window into various biological and genetic processes, especially circulating tumor DNA directly originated from tumor cells.

Considering the attributes of cfDNA data, their inherent complexity, and the ease of collecting them over time, employing statistical modeling analyses appears necessary to extract relevant information. This review explores the diverse modeling approaches used to analyze cfDNA in oncology, emphasizing its role in oncology. After an overview of the current knowledge of cfDNA biology, the use of statistical analysis, machine learning, and non-linear mixed effects models is detailed for their application in clinical studies, both in processing cfDNA data at a particular time point and in capturing their temporal dynamics.

Overall, this review provides a comprehensive overview of the diverse modeling approaches applied to cfDNA in oncology, with a focus on dynamic approaches.

Keywords: cell-free DNA, circulating tumor DNA, fragmentomics, cancer, computational modeling, biomarker, early detection, prognosis, treatment monitoring

INTRODUCTION

Cell-free DNA (cfDNA) consists of fragments of encapsulated DNA released in body fluids such as blood, urine or cerebrospinal fluid[1] (Figure 1A), allowing easy access to the genetic background of hardly-reachable tissues. Originating from events like necrosis, apoptosis[2] or lysis[3], the understanding of cfDNA biology, particularly in oncology, is a current research focus. At bedside, applications are relevant for diagnosis, prognosis and treatment monitoring. They also allow for adapting and personalizing the therapeutic strategy.

The interest in studying cfDNA lies is its ability to encapsulate and share biological and genetic information. Notably, circulating-tumor DNA (ctDNA), which is a part of cfDNA, comprises fragments originated from tumor cells (primary tumor, circulating tumor cells and micro- or macro-metastases) and may contain specific tumor aberrations[4]. Consequently, cfDNA analyses can offer insights into the disease biology, tumor genetic mutations and tumor growth.

Furthermore, cfDNA offers the benefit of being collected through liquid biopsies[5] (Figure 1B), less invasive than tissue biopsies. This provides frequent biological markers enabling real-time monitoring of the disease. Additionally, cfDNA is representative of genetic information coming from all parts of the body, including primary or metastatic tumor sites. This feature can help for the detection of new sites and serve as a surrogate for tissue biopsies in case the primary site has not been identified. Furthermore, liquid biopsies provide the unique ability to capture intra-tumoral genetic heterogeneity[6]. This intricate genetic landscape is often missed by tissue biopsies, which only sample a small portion of the tumor. Eventually, liquid biopsies offer the valuable advantage of detecting genetic and genomic adaptations during the treatment course.

The sequencing and quantification of cfDNA over time generated extensive datasets, offering valuable insights into the biological dynamics between tumors and their microenvironment. These data include the detectability or level of aberrations in specific genes, fragment size profiles, ratio between short and long fragments, as well as the description of base pair patterns (Figure 1C). To manage these data, studies have turned to mathematical and computational modeling of cfDNA (CM-cfDNA), comprising statistical, machine learning (ML) and mechanistic modeling. The latter

refers to simplifying real-world physical and biological concepts throughout mathematical systems and computer programs. These approaches allow data mining, predictive modeling and *in silico* simulations that have become necessary as the data complexity continues to grow. Since the early 21st century, CM has played a role in understanding individual variability in drug responses, paving the way to personalized medicine. The emergence of CM-cfDNA around 2015, aligning with the increasing popularity of cfDNA, constituted 24% of tumoral cfDNA studies in 2022 and raised to 28% in 2023 ((“modeling” OR “computational modeling” OR “machine learning” OR “survival analysis”) AND “tumoral cfDNA” / “tumoral cfDNA” PubMed search, 249/883 entries in 2023). CM-cfDNA embraces three primary modeling approaches: survival analysis, investigating time-to-event data (such as progression); ML, which exhibits a substantial surge over the past two decades and contributed significantly to medical decision-making (e.g., diagnosis and prediction of treatment response); and mechanistic modeling, employing mathematical systems to describe interactions between tumors, drugs, and biological systems.

Within the last eight years, CM-cfDNA has demonstrated efficacy in oncology for personalized early detection, cancer subtype classification, prognosis, treatment monitoring and prediction of response in multiple cancer types, including melanoma[7], lung[8], breast[9] and colorectal[10] cancers (Figure 1D).

BIOLOGY OF cfDNA AND ctDNA

The history of cfDNA began with its discovery in 1948 by Mendel and Métais[11], who first described the presence of nucleic acids in the blood. Later in 1977, it was observed that patients with various types of cancer exhibited elevated levels of cfDNA[12]. A dozen years later, Stroun et al. demonstrated the presence of cancer-derived DNA fragments in the blood of cancer patients, encouraging the exploration of new blood-borne markers but also in other body fluids[13,14].

As previously mentioned, the biology of cfDNA is not fully understood, although it appears that cfDNA may originate from apoptosis, necrosis, active secretion of exosomes[15] or the hematopoietic system[16]. The half-life of cfDNA ranges from 15 minutes to 2 hours[17], allowing for a representative monitoring of the actual released amounts over time. Nevertheless, cfDNA quantities and genetic aberrations provide access to biological characteristics.

CfDNA quantification and sequencing assays

Initially, numerous studies have explored cfDNA genotyping to quantify fragments, detect specific tumor aberrations or discover new ones. These methods fall into two categories; targeted and non-targeted approaches[18]. Targeted approaches rely on a set of predefined genes, mainly associated to the patient's pathology. They comprise digital droplet Polymerase Chain Reaction (PCR), Q-PCR, amplification-refractory mutation system, BEAMing-PCR, tagged-amplicon deep sequencing and cancer personalized profiling by deep-sequencing[19]. They present good sensitivity and better specificity than non-targeted approaches[18]. Non-target approaches assess the entire genome (array-comparative genomic hybridization, whole exome sequencing, whole genome sequencing). While these methods have the potential to identify new genetic aberrations through genome-wide screening, they require larger cfDNA amounts.

These quantification and sequencing methodologies have enabled the extraction of relevant cfDNA characteristics under specific medical conditions.

CfDNA concentration

For instance, the proportion of cfDNA coming from a single cancer depends on the tumor's size. For a 100 grammes tumor (approximatively 3 x 10^{10} tumor cells), 3.3% of its DNA is released daily into the bloodstream[20]. The presence of metastases and circulating tumor cells also impacts on these quantities. In healthy individuals, cfDNA concentrations typically range from 0 to 100 ng/ml of blood, with an average of 30 ng/ml, whereas the concentration of cfDNA in the blood of cancer patients varies from 0 to 1,000 ng/ml, with an average of 180 ng/ml[12]. The cancer stage also impacts on the proportion of detected ctDNA within cfDNA, being approximately two times smaller in stage I patients compared to stage III patients[21].

However, elevated cfDNA levels, while not exclusive to cancer, can also result from conditions like pro-inflammatory or auto-immune diseases, cirrhosis, hepatitis[22], or systemic lupus[23]. Furthermore, quantities of cfDNA may also be influenced by pregnancy. Studies have detected maternal and fetal cfDNA in the mother's blood, with a rapid decrease of cfDNA levels after birth (mean ~ 16.3 min)[24]. Lastly, intense physical activity may increase cfDNA amounts due to inflammation[25]. Consequently, the origin of cfDNA could introduce a potential bias in the results.

CfDNA fragmentomics

Another category of research, called fragmentomics, focuses on the study of the fragments of cfDNA. This notion has been introduced in 2015[26]. The authors showed that fragment patterns differed across cancer patients and tissue of origin. Several studies have since investigated various fragmentation features, such as fragment sizes, end motifs, breakpoint motifs, jagged ends or nucleosome footprints (Figure 1C).

First, it was revealed that fragment sizes are relative to the clinical condition of a patient. Healthy individuals typically exhibit cfDNA size distributions ranging from 130 to 200 base pairs (bp), with a peak at 166 bp[27]. This phenomenon seems to be linked to the nucleosome footprint (nucleosome positions)[28]. This specific length corresponds to the DNA fragments wrapped around a nucleosome core (~147 bp) added to a linker fragment of ~20 bp[29]. These fragments appear to originate from a double process of apoptosis-induced proliferation and proliferation-induced apoptosis[2]. In contrast, longer fragments such as 10,000 bp fragments appear to originate from necrosis or phagocytosis of necrotic tumor cells by macrophages[30], and are much less present in the plasma[2]. Additionally, cancer patients appear to have an enrichment in shorter

fragments (90-150 bp). Subsequently, focusing on these fragments could enhance the detection of ctDNA[31].

Second, some end motifs appear more frequently in healthy individuals, whereas cancer patients present distinct and more variable end motifs[32], which can be relevant to find the tissue of origin. The proportion of jagged ends (uneven DNA extremities, Figure 1C) was found to vary between tumor and wild-type fragments. It appeared to increase in hepatocellular patients[33] and decrease in bladder cancer patients[34], compared with healthy individuals.

Last, breakpoint motifs also vary between healthy individuals to cancer patients. The proportion of AATTGC motifs is larger in cancer patients, whereas the GCAGTA, GCACTT and CTCAAA motifs proportions are smaller[35].

CfDNA mutations and ctDNA

A large body of literature examined mutations that can be detected on cfDNA fragments using sequencing techniques. When these mutations are tumor specific, the afferent cfDNA is then considered as part of ctDNA. These studies historically preceded fragmentomics but examine data at a smaller scale, i.e. molecular alterations detected by sequencing techniques (Figure 1B-C).

The detection of ctDNA quantities can be influenced by the primary tumor type. For instance, ctDNA is detected in most patients with metastatic bladder, colorectal, gastroesophageal or ovarian cancer[21]. Conversely, cancers such as prostate or thyroid cancers appear to have few or undetectable ctDNA[21]. Detectability can also vary depending on the investigated body fluid. In brain cancer, plasma ctDNA typically accounts for less than 1% of cfDNA due to the blood brain barrier. Therefore, cerebrospinal fluid or urine collection has proven effectiveness for identifying ctDNA[36].

The analyze of cfDNA has shown power to monitor, e.g., the mutation level of oncogene addicted cancers such as endothelial growth factor receptor (EGFR) or Kirsten rat sarcoma viral oncogene homologue (KRAS) positive lung cancers, in the adaptation of treatment lines.

Blood EGFR mutations seem to match with tissue EGFR mutations, suggesting a transition from tissue sampling to liquid biopsy[37]. For example, anti-EGFR gefitinib has shown efficacy in blood EGFR-positive NSCLC patients in first line [37]. Similarly, renal

cancer patients carrying at least one cfDNA TP53 mutation had shorter progression-free survival[38]. In colorectal cancer, patients initially treated with an anti-EGFR antibody, the on-treatment monitoring of genetic alterations in, e.g., the KRAS or EGFR genes could reveal the emergence of resistance, enabling the early adaptation of treatment[39]. Finally, the detection of more than four KRAS mutant copies per milliliter in peripheral blood of pancreatic ductal adenocarcinoma has been linked with shorter PFS (HR=3.4 (1.2-09.7)) at pre- and post- resection, as well as in portal venous blood (HR=4.6 (1.6-13.3))[40].

All this genetic and biological variability, together with the data type diversity calls for the use of CM-cfDNA to develop new tools for personalized diagnosis, treatment setup and monitoring (Figure 1D).

CFDNA COMPUTATIONAL MODELING FOR DIAGNOSIS

One of the primary goals of CM-cfDNA is to detect tumors before clinical symptoms. Some studies aimed to discern changes in cfDNA characteristics to discriminate pathological to healthy individuals (Table 1) or to distinguish between different tumor subtypes. Moreover, identification of the primary tumor site is a major challenge. In the plasma, ctDNA can originate from any body part, making specific ctDNA mutations essential for providing clues about the primary tumor location. Consequently, multiple studies conducted pan-cancer analyses.

CtDNA-based machine learning

Classical cfDNA studies primarily employ conventional statistical hypothesis testing to identify dependencies between cfDNA markers and categorical outcomes, such as diagnosis, or cancer type/subtype. Additionally, ML models (Figure 2A), especially supervised classification methods, have been increasingly employed in the last decade for early cancer detection[41]. These statistical analysis techniques enable learning from an initial patient dataset to predict diagnosis for new patients.

Logistic regression (LR) models and predicts the probability of a binary data, seeking a linear relationship between the log-odds of the event occurring and variables. Support vector machines (SVM) (Figure 2A) aim to find the hyperplane that maximally separates individuals into two or more classes. Decision trees are a series of interconnected binary choices, enabling regression and classification. Random forests (RF) learn from multiple decision trees trained on different features subset. Finally, deep learning, a subset of machine learning based on neural networks is mostly used for complex features patterns and relationships and consists of interconnected artificial neurons that evaluate a weighted sum of inputs and pass the results to the next-layer neurons through a nonlinear activation function.

ML models have enabled the differentiation between lung adenocarcinoma, lung squamous carcinoma and squamous cell lung cancer, leveraging copy number profiling of cfDNA[42]. This distinction was achieved comparing five classifiers: RF, SVM, LR with ridge, elastic-net (EN) or least absolute shrinkage and selection operator (LASSO) regularizations. Another study used SVM to select discriminative differentially methylated blocks for early lung cancer detection[43], detecting 52–81% of the stages

IA to III patients, with a specificity (true negative rate) of 96% (95% confidence interval (CI) 93–98%).Liu et al.[41] reviewed the most relevant and recent early detection ML-based studies, noting four CM-cfDNA studies employing linear models, possibly with elastic-net (EN) or least absolute shrinkage and selection operator (LASSO) regularization. One identified elevated plasmatic cfDNA levels in oral cancer patients, compared to control subjects[44]. Three CM-cfDNA studies employed SVM, one investigating a significant mammalian DNA epigenetic modification (5-hydroxymethylcytosine) for pancreatic ductal adenocarcinoma early detection[45]. Five studies used RF, including one exploring 5-hydroxymethylcytosine patterns to discriminate among seven cancer types, achieving 87.5% and 92% accuracy for two datasets[46]. Another one identified new biomarkers in the plasma cfDNA methylome profiling to diagnose and locate gastrointestinal cancers, resulting in an area under the curve of 0.96 ± 0.04 (mean ± standard deviation), 0.89 ± 0.06, 0.91 ±0.07 for hepatocellular carcinoma, colorectal cancer and pancreatic cancer respectively[47].

Deep learning has been less frequently used due to small sample size of these types of studies. Nevertheless, Liu et al.[41] cited one study using it for early stage lung cancer detection[48]. They employed a convolutional neural network with two-dimensional grids representing the sequenced reads. Each column was representing a read and exclusively colored according to the row corresponding to its respective base (A, C, G, T, or N if unrecognized). The algorithm then detected base changes, such as deletions, mutations or insertions, focusing on distinguishing artifacts from genuine cancer mutations.

Moser et al.[49] cited over twenty diagnosis studies employing ML models or combinations of them. A common issue involves somatic mutations of non-cancerous origin, increasing the number of false positives. For instance, age can induce the development of somatic mutations during clonal hematopoiesis, resulting in misleading results. Chabon et al.[50] developed a classification framework (5-nearest neighbor, 3-nearest neighbor, naïve Bayes, LR and decision tree) to distinguish tumor from clonal hematopoiesis mutations and matched them with risk-matched controls. They demonstrated that clonal hematopoiesis mutations tended to occur in longer cfDNA fragments.

Eventually, unsupervised learning (Figure 2A) has also been employed to classify individuals into cancer subtypes. Luo et al. used hierarchical clustering to distinguish

colorectal cancer patients from normal subjects according to methylation markers, and finally classify patients into two subgroups with different overall survival (OS)[10].

Fragmentomics-based machine learning

Chen et al. used LR to differentiate hepatocellular carcinoma from liver cirrhosis and healthy controls using four cfDNA fragmentome features: genome-wide 5-hydroxymethylcytosine, nucleosome footprint, 5′ end motif and fragmentation profiles. This led to a 95.4% sensitivity and a 97.8% specificity in the test set[51]. Similarly, Duo et al. compared LR, deep learning and extreme gradient boosting to early detect lung adenocarcinoma, using the 6bp breakpoint motif, defined as « the 3bp extensions to both directions of the aligned cfDNA 5′ », achieving a 92.5% sensitivity and a 90.0% specificity in the external validation cohort[35]. Ma et al. compared five ML algorithms (generalized linear model, deep learning, RF, gradient and extreme gradient boosting), integrating fragment size ratio and distribution, end and breakpoint motif, and copy number variation[52]. They reached a 94.8% specificity and 98% sensitivity to distinguish healthy individuals from early-stage colorectal adenocarcinoma.

An important and influential work has been performed by Cristiano et al.[53]. They employed the cfDNA integrity index, defined as the ratio of short fragments (100-150 bp) to long fragments (150-200 bp), across 504 genome bins, to classify healthy individuals and cancer patients (with seven different pathologies). They highlighted distinct variations in fragment size profiles across different genome regions of cancer patients. The features were then integrated into a stochastic gradient tree boosting framework. Samples were split according to a 10-fold cross-validation repeated 10 times, with features selected at each of the ten steps on the inner-fold training dataset. At each iteration, the model was estimated on the training dataset and evaluated on the test dataset. Predictions were made based on the average of predictions over the hundred steps. With a 95% specificity, they detected 80% of the cancer patients. Expanding the framework, they identified tumor tissue origin with 90% specificity, 61% accuracy and reaching 75% accuracy when looking at the top two predictions.

Mathios et. al implemented a comparable ML approach based on similar fragmentation features for lung cancer detection and stage identification in high-risk symptomatic patients[54]. Initially, they reduced dimensionality of the fragmentation features by selecting principal components explaining 90% of the fragmentation variance.

Subsequently, they used a LASSO-penalized LR to assess the fragmentation components along with 39 chromosomal arms Z-scores (number of standard deviations from the mean of the mapped read fraction). With 10 replicates of a 5-fold cross-validation, they defined a score able to detect 94% of the cancer patients with 80% specificity, in a population with 91% early-stage (I-II) cancer patients.

Mouliere et al. also developed CM-cfDNA ML (RF and LR)[31]. They used the proportion of fragments in multiple size ranges, ratios of some of these proportions, and 10 bp periodicity oscillation amplitudes occurring before 150 bp to detect cancer patients, even for pathologies presenting weaker levels of ctDNA. They reached an area under the receiver operating characteristic curve (AUC) of 0.891 for cancers with low amounts of ctDNA (pancreatic, renal and glioma) with RF, having selected 4 features among 9.

Unsupervised learning is less commonly encountered in the literature. Renaud et al. performed such analysis to detect the presence of cancer fragments in the overall cfDNA thanks to fragments lengths, measured by shallow whole genome sequencing[55]. The aim was to decompose the matrix of the fragment size profiles through the non-negative matrix factorization (NMF) method:

$$sample_{n\times m} = weights_{n\times k} \times signatures_{k\times m},$$

where:

- each row i of $sample_{n\times m}$ represents a sample; each column j represents a fragment size. One matrix cell indicates the relative fragments frequency with length represented in the column j, for the sample i.

- $signatures_{k\times m}$ represents the signature matrix. Each row corresponds to a typical profile of fragment lengths, according to the source of the fragments. The hyperparameter k is the number of sources used in the factorization and must be set. As an example, $k = 2$ sources could represent typical healthy and cancer distributions of the fragment lengths.

- $weights_{n\times k}$ are the weights associated to each signature profile for each sample.

They initially calibrated the number of sources to $k = 2$ and inferred signatures by the NMF method. They found correlations between the weights and ctDNA ratios (variant allele fractions) (r = 0 .75). Eventually, NMF was applied to a cohort of healthy and pathological individuals with diverse cancer types[53]. The weights were integrated into

an SVM framework to detect cancer patients. By increasing the number of sources to 30, they reached an AUC of 0.95.

Mechanistic modeling for annual screening

Using longitudinal data, Avanzini et al. developed a mechanistic model of ctDNA shedding during apoptosis linked to the tumor size evolution over time to determine the optimal screening time for early lung cancer detection[56]. They modeled the expected number of ctDNA haploid genomic equivalent (hGE) circulating in the bloodstream for a tumor with size M as a Poisson-distributed random variable, with mean:

$$C = M \times \frac{d \times q_d}{\epsilon + r}$$

where:

- d is the tumor death rate per day. $d = b - r$ where b is the cell division rate per day and r is the net tumor growth rate per day.

- q_d is the mean shedding rate of a cell death. On average, $q_d \approx 0.1$ hGE per cell death.

- ϵ is the ctDNA elimination rate per day.

Using mechanistic modeling and considering various sources of biological and technical errors, they could predict the expected tumor detection size. They demonstrated that ctDNA-based annual screening had a median detection size of 2.0 to 2.3 cm of diameter, against 3.5 cm for usual annual screening, highlighting the optimal marker-frequency combination for cancer screening.

Taken together, there has been in recent years a dwealth of studies and results related to CM-cfDNA-based early cancer detection, mainly focusing on cfDNA aberrations. Drug and medical regulatory organizations have begun to recommend the use of cfDNA. In 2020, the FoundationOne Liquid CDx test[57], analyzing cfDNA-based genes, has been approved by the Food and Drug Administration. In 2022, the European Society for Medical Oncology Precision Medicine Working Group published an article warranting the use of ctDNA as an adjunctive diagnostic tool[58].

In summary, ML is increasingly integrated to find associations between fragments features and diagnosis outcomes such as cancer detection, stage or histological type. Fewer researchers modeled cfDNA kinetics by integrating biological mechanisms to

prove the effectiveness of these fragments as novel biological markers. Finally, limitations include the necessity to evaluate prospectively these methods on asymptomatic individuals at risk for cancer[59], in order to generalize their applicability in real-world conditions and detect cancer before symptoms arise.

CFDNA COMPUTATIONAL MODELING FOR PROGNOSIS AND TREATMENT PREDICTION

Many studies used CM-cfDNA to monitor tumor size and predict therapeutic responses (Table 2). Their aim was to identify new signatures enabling early treatment adjustments and preventing adverse events.

Most of them relied on baseline markers (cfDNA/ctDNA concentration, ctDNA positivity, fragmentomics). These markers are typically assessed at one specific time, either following surgical resection to establish associations with time to relapse, or just before treatment initiation to determine correlations with imaging-evaluated treatment response. Sometimes, these studies also rely on biological markers collected at various time points, including at some or all treatment cycles. They provide longitudinal datasets including absolute values and/or relative changes from baseline data. These datasets enable comprehensive analysis of cfDNA dynamics, aiming to discover new patterns associated with time to relapse, disease progression, treatment response, or mortality.

These studies first employed classical statistical analysis (e.g., survival analysis) to establish connections between cfDNA measurements and relapse or response to treatment. However, computational methodologies have evolved to ML, non-linear mixed effects models (NLME) and mechanistic modeling.

Classical survival analysis

Survival analysis (Figure 2B) analyzes the duration before a specific event, such as progression or death, and specifically accounts for censored data (unreached event). The main methods include Kaplan-Meier (KM) estimation and univariable /multivariable Cox proportional-hazards regression (CPHR).

In the fragmentome field, Lapin et al. demonstrated an association of fragment sizes smaller than 147 bp and high cfDNA levels pre-treatment with shorter PFS and OS, using Kaplan-Meier estimation in advanced pancreatic cancer patients[60]. Multivariable CPHR demonstrated that cfDNA levels could predict PFS (Hazard Ratio (HR): 3.05, 95% CI: 1.40-6.65) and OS (HR: 2.24, 95% CI: 1.09-4.59).

Moding et al.[61] monitored ctDNA molecular residual disease in advanced non-small cell lung cancer (NSCLC) patients, collecting ctDNA immediately after chemoradiation therapy (CRT), later followed or not by consolidation with immune-checkpoint inhibition (ICI). A second ctDNA sample was collected early during ICI treatment for the immunotherapy-treated arm. Undetectable ctDNA before ICI treatment correlated with good prognosis, irrespective of ICI treatment. Detectable ctDNA early on ICI also correlated with a shorter progression-free survival (PFS). Analyzing ctDNA changes over time demonstrated that increased ctDNA levels were associated with worse prognosis compared to a decrease.

Powles et al.[62] performed CPHR to predict atezolizumab response in urothelial carcinoma. They collected plasma onset and early on treatment (first day of the first and third cycles). They revealed significant disease-free survival differences according to ctDNA changes over time. Patients with ctDNA clearance appeared to have three to four times lower relapse risks, according to univariable, stratified and multivariable CPHRs.

Longitudinal dynamics modeling

Few studies have explored dynamical CM-cfDNA (Figure 2C). One study developed of a mechanistic modeling of the joint ctDNA–tumor size evolution over time, to assess atezolizumab response in NSCLC and melanoma patients[63]. The authors used a bi-exponential system to independently describe both the log10-transformed number of mutant molecules per mL and the sum of the longest diameters (SLD) of lesions, assessed by the RECIST 1.1 criteria. Parameters included the model-estimated value at the first time point, the growth rate and the decay rate. The estimated ctDNA growth rate showed high correlation with the estimated SLD growth rate. The final joint system is:

$$SLD(t) = SLD_0 \left(e^{-k_{sT}.t} + e^{k_{gT}.t} - 1\right)$$

$$ctDNA(t) = ctDNA_0 \left(e^{-\zeta.k_{sT}.t} + e^{k_g.t} - 1\right),$$

where:

- SLD_0 and $ctDNA_0$ are the baseline values of SLD and $ctDNA$ respectively.

- k_{gT} is the growth rate of the tumor size.

- k_{gT} is the decay rate of the tumor size.

- k_g is the decay rate of the ctDNA level.

ζ is the coefficient linking the tumor size growth rate to the ctDNA growth rate.

This system fitted well the ctDNA and tumor kinetics over time, even when negatively correlated (one increasing while the other decreasing). Thus, they highlighted the mechanistic link between tumor growth and ctDNA release for patients under ICI.

Janssen et al. employed NLME (Figure 2C) for ctDNA biomarkers analysis to predict early treatment responses[64]. They used a NLME model to describe the dynamics of EGFR mutations in ctDNA from NSCLC patients treated with erlotinib or gefitinib. In this model, the concentration of three mutations was modeled by a zero-order growth model, chosen between baseline, turnover and first-order growth models. The first model described both L858R or exon19del (driver) mutations concentrations, while the second one described T790M mutation concentrations. The equation writes:

$$\frac{dy}{dt} = k_{in} - k_{out}.y(t).R(t) \qquad (1)$$

where $R(t) = e^{-\lambda t}$ for driver mutations, and $R(t) = e^{-\lambda.y(t)}$ for the T790M mutation concentration. Here:

- $y(t)$ is the change in either L858R or exon19del over time.

- k_{in} represents the zero-order increase in mutations concentrations.

- k_{out} represents the drug-driven decrease in mutations concentrations.

- $R(t)$ accounts for the time-dependent resistance development where λ is the progression rate.

Another parameter was estimated to consider the baseline mutations concentration, unavailable in this study. The growth model was fitted to observed concentrations to identify each parameter. Subsequently, the predicted time-course of mutations concentrations were compared to the observed ones, revealing that a zero-order increase and a first-order elimination model (equation 1) best approximated the actual concentrations.

These predicted ctDNA values were integrated into parametric survival models to predict PFS. Among exponential, Weibull and Gompertz hazard models, Weibull was found to be the more efficient. Considering all the predictors of disease progression at

random timepoints post-treatment initiation, including relative changes from baseline and absolute values of the three mutations concentrations, only the relative change in driver mutations was statistically significant for stratifying responders and non-responders (p = 0.001, likelihood-ratio test). This significance was validated using stratified KM curves. Consequently, patients with a predicted relative change from baseline greater than zero (median value) experienced a shorter disease progression.

Prior to this study, Khan et al. sought to model the carcinoembryonic antigen dynamics, which is proportional to the total number of tumor cells[65], to assess cetuximab response in colorectal cancer patients. They used the following equation for the tumor burden:

$$N(t) = n_s \mathrm{e}^{-\lambda_s t} + n_r e^{\lambda_r t},$$

where the cells number $N(t)$ is divided into a population of treatment sensitive cells n_s, dying under treatment at rate λ_s, and a population of treatment resistant cells n_r, growing at rate λ_r.

CfDNA mutant frequencies was modeled using a single exponential growth model. Rates were estimated for each patient with at least three time points. The cfDNA model well described the real dynamics ($R^2 = 0.979$). Particularly, the cfDNA relapse rate was shown to be correlated with the tumor one. By comparing relapse rates with RECIST v1.1 criteria, the model could precisely predict the time to relapse.

Fragmentomics data are emerging as prognosis markers but remain underused for treatment monitoring. Some studies have investigated the cfDNA integrity index variations in breast cancer during neoadjuvant or adjuvant chemotherapy. Only two studies were found searching for “fragmentomics” AND “chemotherapy” on PubMed, and only one searching for “fragmentomics” AND “immune-checkpoint inhibitors” (no results for “fragmentomics” AND “immunotherapy”). The latter investigated cfDNA fragmentation profiles in lung carcinoma and diffuse B cell carcinoma patients[66]. They calculated an « expression inference from cfDNA-sequencing » score, based on a statistic of expression levels changes of genes before ICI-treatment and after ~ 4 weeks. This score allowed to identify both patients with durable clinical benefit of ICI and shorter PFS (HR: 11.38, Wald test: p = 0.006). They also developed a mechanistic model of the nucleosome accessibility at transcription start sites regions. The model established a connection between this accessibility and fragmentation profiles and

expression levels. The model was used to perform simulations and explore the parameters influencing the detection of a specific gene expression within cfDNA.

Longitudinal ML modeling

ML leverages longitudinal data by merging the features from each time point, comprising absolute and relative changes over time and integrating them into ML models to classify patients as either responders or non-responders, employing supervised or semi-supervised classification methods.

Assaf et al.[67] conducted such a study with 466 NSCLC patients. They developed a ML framework to predict immunotherapy response using longitudinal ctDNA data. CtDNA was collected at baseline (before treatment) and at day 1 of cycle 2 (C2D1), and cycle 3 (C3D1). The models integrated 19 ctDNA levels metrics and 59 relative ctDNA changes from baseline. Three models were compared: only baseline features, baseline + C2D1 and baseline + C2D1 + C3D1. The latter was selected for OS prediction, as it exhibited the highest C-index. Then, the ctDNA features were combined with baseline clinical features, including ECOG status, metastases count, age group, sex, smoking history, PD-L1 status, and the sum of lesion diameters. Baseline and C2D1 tumor size was also considered. They employed an Elastic Net (EN) approach with leave-one-out cross-validation (LOOCV) conducted through a 10-nested cross-validation process. Features were retained if they were selected in over 50% of cross-validation iterations and if the gain metric was positive according to the next-door analysis. This analysis involves fitting the same model after removing one predictor and comparing the error rate to the one of the full model. Consequently, the feature set was ultimately reduced to five features, including the global cfDNA concentration at C3D1. The final model categorized patients into three groups: high risk (progressive disease), intermediate risk (stable disease) and low risk (responders), and KM curves demonstrated significant risk stratification in both training and testing datasets.

Similar to the previous study but in a smaller cohort of 94 NSCLC patients treated with atezolizumab or docetaxel, Zou et al. applied LOOCV LASSO-penalized regression, linking ctDNA metrics (collected at baseline, C2D1 and C3D1) to OS[68]. The model highlighted the C3D1 median number of mutant molecules per mL as the most important predictor for OS.

CONCLUSION

Recent studies have delved into the use of cfDNA as an innovative biological marker for cancer detection and treatment monitoring in the last years. Most of these studies focused on early detection and demonstrated sufficient level of evidence to prompt regulators to recommend cfDNA as a complementary diagnostic tool[57,58]. There is, however, a pressing need for standardizing ctDNA detection and cfDNA quantification methods. Furthermore, over the past two years, approximately one in four studies employed CM-cfDNA as a support for precision medicine. While the majority concentrated on classical survival and classification analyses, there is a growing need to model biological mechanisms over time to improve precision, which has been addressed by only few studies.

To enhance the reliability of findings and assess the methods' reproducibility, it is also crucial to conduct studies on larger datasets, with validation on external cohorts. Currently, most assays train and test their models on data from symptomatic or diagnosed patients, or even high stage patients. Additional studies that specifically target and validate markers on data from asymptomatic individuals are warranted.

In contrast to its use for diagnosis, CM-cfDNA for patient prognosis and treatment monitoring is less common. Nevertheless, it is particularly well suited for such purposes. Specifically, mechanistic modeling parameterized on experimental data facilitates the better understanding of biological mechanisms behind cfDNA release in body fluids. Combining mechanistic modeling with ML methods and survival analysis (mechanistic learning[69]) holds the premise to develop genuinely informative biologico-computational markers and associated predictive tools.

Most research efforts have focused on ctDNA analysis, targeting mutations of a specific cancer and requiring a proper understanding of pathology-associated genetic aberrations. But ctDNA is only detectable in less than 80%[70] of the cancer patients, making it worthwhile to further explore the use of alternative cfDNA characteristics, such as global concentration, methylation, fragment size profiles and end motifs, in the quest for pan-cancer biomarkers.

To this regard, optimizing data acquisition, improving data relevance, and diversifying data types will be the first keys to finally better adapt and improve CM-cfDNA.

FIGURE LEGENDS

Figure 1: Cell-free DNA data: a new biological tool for on oncology

A. Cell-free DNA (cfDNA) are fragments of encapsulated data released in the human body fluids, such as blood, urine, cerebrospinal liquid. Some of these fragments are originated from tumoral cells, which can be primary, metastatic or circulating tumor cells. Plasmatic cfDNA is the most analyzed because of its ease of collection.

B. After the blood collection, cfDNA is extracted and amplified, usually by Polymerase Chain Reaction (PCR) methods. Fragments are then sequenced by various methods, targeted ones (which target specific genes) and non-targeted approaches (sequencing the whole genome) to provide information at smaller case of the molecular alterations.

C. PCR and sequencing processes yield diverse cfDNA data. a) Global concentration is the first main quantitative feature describing cfDNA. b) Fragmentomics study a wide range of data, focusing on the fragment sizes and patterns. It provides a profile of fragment size distribution, enabling the extraction of quantities of fragments of various sizes. The sequencing of the fragments also provides nucleotide base patterns from end motifs, jagged ends, and breakpoint motifs. Additionally, fragmentomics work on nucleosome footprints and cfDNA integrity index (cfDI), calculated as the ratio between short and long fragments at a same locus. c) At a smaller case, ctDNA mutations are predominantly analyzed, using features such as ctDNA positivity (number of mutations detected greater than x mutations), ctDNA concentration (copies number per milliliter), or variant allele frequency.

D) These cfDNA data are incorporated into computational modeling frameworks to identify associations with clinical outcomes. This enables the establishment of cfDNA as new biological marker in cancer research, serving for diagnosis, prognosis, and prediction.

Figure 2: Cell-free DNA data: a new biological tool for oncology

A) i) To early diagnose, evaluate the cancer type or stage, make prognosis or predict response to treatment, studies compute machine learning methods. CfDNA data are collected at different moments of the cancer progression according to the outcome to predict. Data can be collected over time, but they are not considered as time-dependent during modeling. Machine learning methods can be divided into two major

groups: ii) the unsupervised learning and iii) the supervised learning models. ii) First ones are built to discriminate k groups within the complete set of individuals or reduce dimensionality of the features space. The idea is to find individuals that are closed into the space of features. A typical unsupervised learning method is the hierarchical clustering, which build a hierarchy of individuals groups. Another one is the non-negative matrix factorization, which decomposed a matrix of non-negative elements into two matrices, for example by factorizing a matrix of cfDNA size profiles into a coefficient's matrix and a matrix of size profile signatures. iii) Supervised learning methods learn outcome's individuals on a train set to then predict outcomes of a new cohort of patients. Most common supervised algorithms in cfDNA modeling are the logistic regression, support vector machines (SVM), decision trees and random forest. Neural networks are used mostly in the case of complex patterns and relationships between features.

B) i) Classical survival modeling gathers technical tools that enable the modeling of a duration until the occurrence of an event. Progression and death are the main events modeled in the medical domain. Thus, individuals may be censored as the event never occur during the study's time, due to the track loss of the patient, or the end of follow-up by the study. In those cases, the event of progression or death is not observed: patients are referred to as censored.

ii) A usual nonparametric estimation is the Kaplan-Meier one, which allows to visualize and check hypothesis about the ability of a variable to discriminate long to short survival.

iii) The Cox proportional-hazard regression is a widely used method for the analyze of censored time data in survival modeling. This method assumes that the effect of predictor variables on the hazard rate remains constant over time. Cox regression helps to identify significant features as machine learning regressions do, estimate the hazard ratios, which indicate the proportional changes in the hazard for one unit change in a predictor variable. Additionally, it may generate survival curves (survival probability over time for different levels of the feature).

C) i) Longitudinal data may be modelled as time-dependent data, to follow the evolution of cfDNA kinetics during treatment. ii) Mechanistic modeling integrates biological hypothesis and fundamental principles, known to induce the observed kinetics, into a dynamic system. The models are then parameterized on experimental data thanks to

non-linear mixed effect models, which allows a better understanding of the biological mechanisms and the validation of hypothesis.

TABLES

Table 1: Summary of cfDNA computational modeling studies for early diagnosis, organ, stage and histological classification

Source	Cancer	Modeling	Marker	Purpose
[46]	Pancancer	RF	ctDNA aberrations	Cancer type classification
[71]	Colorectal	LASSO – LR		Early diagnosis
[72]	NSCLC	Linear regression		
[50]		5nn – 3nn naïve Bayes – LR – Decision tree		
[48]	Lung	CNN		
[45]	Pancreas	RF – SVM – EN LR		Stage classification
[42]	NSCLC	LASSO – Ridge – EN LR		Histological classification
[44]	Oral	LR	cfDNA quantification	Early diagnosis
[56]	Lung	Mechanistic modeling	cfDNA concentration / ctDNA mutations	Early diagnosis
[54]	Lung	PCA – LR	Fragmentome	Early diagnosis
[31]	Pancancer	RF – LR		
[51]	Hepatocellular carcinoma	LR		
[35]	Lung adenocarcinoma	LR – deep learning, gradient and extreme gradient boosting		

[52]	Colorectal adenocarcinoma	Generalized linear model, deep learning, RF, extreme gradient boosting		
[53]	Pancancer	Gradient-tree boosting		Early diagnosis and cancer type classification
[73]	Colorectal	LASSO LR	Methylome	Early diagnosis
[10]		Hierarchical clustering		
[43]	Lung	SVM		
[74]	SCLC	PCA		Histological classification
[47]	Gastrointestinal	RF		

NSCLC: non-small cell lung cancer; SCLC: small cell lung cancer; RF: random forest; LASSO: least absolute shrinkage and selection operator; LR: logistic regression; X-nn: X nearest neighbors; CNN: convolutional neural network; SVM: support vector machine; EN: elastic-net; PCA: principal component analysis; ctDNA: circulating tumoral DNA; cfDNA: cell-free DNA.

Table 2: Summary of cfDNA modeling assays for treatment monitoring

Source	Cancer	Treatment	Modeling	Marker	Baseline / Longitudinal
[76]	NSCLC	(Durvalumab ± tremelimumab) + platinum-based chemotherapy	CPH	ctDNA aberrations	Baseline
[53]	Pancancer	Anti-EGFR or anti-ERBB2	KM		
[77]	Hepatocellular carcinoma	Atezolizumab + bevacizumab	KM – CPH		
[8]	NSCLC	Atezolizumab or docetaxel			
[78]	Pancancer	ICI			
[79]	Colorectal	Surgery or chemotherapy			

[80]	Melanoma	(Pembrolizumab or nivolumab) ± ipilimumab	KM – CPH – LR		
[7]	Melanoma	Ipilimumab	Descriptive statistics		Longitudinal
[81]	Pancancer	ICI	KM		
[61, 82, 83, 84]	NSCLC	/ / /	KM – CPH		
[62]	Urothelial carcinoma	Atezolizumab			
[85]	Melanoma	Nivolumab ± ipilimumab			
[86]	Pancancer	Durvalumab ± tremelimumab			
[87]	NSCLC	ICI	KM – CPH – Bayesian probit model		
[88]	Pancancer	Pembrolizumab	KM – CPH – LR		
[63]	NSCLC / Melanoma	ICI ± cobimetinib	NLME		
[68]	NSCLC	Atezolizumab or docetaxel	LASSO linear model		
[64]	NSCLC	Erlotinib / gefitinib	Mechanistic modeling / NLME		
[65]	Colorectal cancer	Cetuximab	Mechanistic modeling		
[67]	NSCLC	(Atezolizumab ± bevacizumab) + carboplatin + paclitaxel	EN linear regression	cfDNA concentration / ctDNA mutations	
[38]	Renal	(Ipilimumab + nivolumab) or anti-VEGFR-TKIs	KM – CPH – LR	cfDNA concentration	Baseline

[89]	NSCLC	TKI or pembrolizumab-CT or CT	KM – CPH		Longitudinal
[66]	Lung adenocarcinoma & B cell carcinoma	PD-(L)1 ICI	KM – CPH – LR Mechanistic modeling	Fragmentomics	Longitudinal

NSCLC: non-small cell lung cancer: CPH: Cox proportional hazards (model); KM: Kaplan-Meier (estimation); LR: logistic regression; NLME: nonlinear mixed effects (model): ctDNA: circulating tumoral DNA; cfDNA: cell-free DNA; ICI: immune checkpoint inhibitors; EGFR: epidermal growth factor receptor; ERBB2: erythroblastic oncogene B 2; VEGFR: vascular EGFR; TKI: tyrosine kinase inhibitors.

ACKNOWLEDGEMENTS

This work received support from the French government under the France 2030 investment plan, as part of the Initiative d'Excellence d'Aix-Marseille Université - A*MIDEX (AMX-19-IET-001 & AMX-21-IET-017).

Figures were partially created with BioRender.com.

The authors declare no conflict of interest.

REFERENCES

1. Dasari A, Morris VK, Allegra CJ, et al. ctDNA applications and integration in colorectal cancer: an NCI Colon and Rectal–Anal Task Forces whitepaper. *Nat Rev Clin Oncol*. 2020;17(12):757-770. doi:10.1038/s41571-020-0392-0

2. Heitzer E, Auinger L, Speicher MR. Cell-Free DNA and Apoptosis: How Dead Cells Inform About the Living. *Trends in Molecular Medicine*. 2020;26(5):519-528. doi:10.1016/j.molmed.2020.01.012

3. Hu Z, Chen H, Long Y, Li P, Gu Y. The main sources of circulating cell-free DNA: Apoptosis, necrosis and active secretion. *Crit Rev Oncol Hematol*. 2021;157:103166. doi:10.1016/j.critrevonc.2020.103166

4. Goebel G, Zitt M, Zitt M, Müller HM. Circulating Nucleic Acids in Plasma or Serum (CNAPS) as Prognostic and Predictive Markers in Patients with Solid Neoplasias. *Dis Markers*. 2005;21(3):105-120. doi:10.1155/2005/218759

5. Thompson JR, Menon SP. Liquid Biopsies and Cancer Immunotherapy. *Cancer J*. 2018;24(2):78-83. doi:10.1097/PPO.0000000000000307

6. Coto-Llerena M, Benjak A, Gallon J, et al. Circulating Cell-Free DNA Captures the Intratumor Heterogeneity in Multinodular Hepatocellular Carcinoma. *JCO Precis Oncol*. 2022;6:e2100335. doi:10.1200/PO.21.00335

7. Lipson EJ, Velculescu VE, Pritchard TS, et al. Circulating tumor DNA analysis as a real-time method for monitoring tumor burden in melanoma patients undergoing treatment with immune checkpoint blockade. *J Immunother Cancer*. 2014;2(1):42. doi:10.1186/s40425-014-0042-0

8. Gandara DR, Paul SM, Kowanetz M, et al. Blood-based tumor mutational burden as a predictor of clinical benefit in non-small-cell lung cancer patients treated with atezolizumab. *Nat Med*. 2018;24(9):1441-1448. doi:10.1038/s41591-018-0134-3

9. Magbanua MJM, Swigart LB, Ahmed Z, et al. Clinical significance and biology of circulating tumor DNA in high-risk early-stage HER2-negative breast cancer receiving neoadjuvant chemotherapy. *Cancer Cell*. 2023;41(6):1091-1102.e4. doi:10.1016/j.ccell.2023.04.008

10. Luo H, Zhao Q, Wei W, et al. Circulating tumor DNA methylation profiles enable early diagnosis, prognosis prediction, and screening for colorectal cancer. *Sci Transl Med*. 2020;12(524):eaax7533. doi:10.1126/scitranslmed.aax7533

11. Mandel P, Metais P. [Nuclear Acids In Human Blood Plasma]. *C R Seances Soc Biol Fil*. 1948;142(3-4):241-243.

12. Leon SA, Shapiro B, Sklaroff DM, Yaros MJ. Free DNA in the serum of cancer patients and the effect of therapy. *Cancer Res*. 1977;37(3):646-650.

13. Stroun M, Anker P, Lyautey J, Lederrey C, Maurice PA. Isolation and characterization of DNA from the plasma of cancer patients. *Eur J Cancer Clin Oncol*. 1987;23(6):707-712. doi:10.1016/0277-5379(87)90266-5

14. Stroun M, Anker P, Maurice P, Lyautey J, Lederrey C, Beljanski M. Neoplastic characteristics of the DNA found in the plasma of cancer patients. *Oncology*. 1989;46(5):318-322. doi:10.1159/000226740

15. Thakur BK, Zhang H, Becker A, et al. Double-stranded DNA in exosomes: a novel biomarker in cancer detection. *Cell Res*. 2014;24(6):766-769. doi:10.1038/cr.2014.44

16. Lui YY, Chik KW, Chiu RW, Ho CY, Lam CW, Lo YD. Predominant Hematopoietic Origin of Cell-free DNA in Plasma and Serum after Sex-mismatched Bone Marrow Transplantation. *Clin Chem*. 2002;48(3):421-427. doi:10.1093/clinchem/48.3.421

17. Khier S, Lohan L. Kinetics of circulating cell-free DNA for biomedical applications: critical appraisal of the literature. *Future Science OA*. 2018;4(4):FSO295. doi:10.4155/fsoa-2017-0140

18. Alix-Panabières C, Pantel K. Clinical Applications of Circulating Tumor Cells and Circulating Tumor DNA as Liquid Biopsy. *Cancer Discov*. 2016;6(5):479-491. doi:10.1158/2159-8290.CD-15-1483

19. Nikanjam M, Kato S, Kurzrock R. Liquid biopsy: current technology and clinical applications. *J Hematol Oncol*. 2022;15(1):131. doi:10.1186/s13045-022-01351-y

20. Diehl F, Li M, Dressman D, et al. Detection and quantification of mutations in the plasma of patients with colorectal tumors. *Proc Natl Acad Sci U S A*. 2005;102(45):16368-16373. doi:10.1073/pnas.0507904102

21. Bettegowda C, Sausen M, Leary RJ, et al. Detection of Circulating Tumor DNA in Early- and Late-Stage Human Malignancies. *Sci Transl Med*. 2014;6(224):224ra24. doi:10.1126/scitranslmed.3007094

22. Shapiro B, Chakrabarty M, Cohn EM, Leon SA. Determination of circulating DNA levels in patients with benign or malignant gastrointestinal disease. *Cancer*. 1983;51(11):2116-2120. doi:10.1002/1097-0142(19830601)51:11<2116::aid-cncr2820511127>3.0.co;2-s

23. Raptis L, Menard HA. Quantitation and characterization of plasma DNA in normals and patients with systemic lupus erythematosus. *J Clin Invest*. 1980;66(6):1391-1399.

24. Lo YM, Zhang J, Leung TN, Lau TK, Chang AM, Hjelm NM. Rapid clearance of fetal DNA from maternal plasma. *Am J Hum Genet*. 1999;64(1):218-224.

25. Breitbach S, Tug S, Simon P. Circulating Cell-Free DNA. *Sports Med*. 2012;42(7):565-586. doi:10.2165/11631380-000000000-00000

26. Ivanov M, Baranova A, Butler T, Spellman P, Mileyko V. Non-random fragmentation patterns in circulating cell-free DNA reflect epigenetic regulation. *BMC Genom*. 2015;16(Suppl 13):S1. doi:10.1186/1471-2164-16-S13-S1

27. Jiang P, Chan CWM, Chan KCA, et al. Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients. *Proc Natl Acad Sci U S A*. 2015;112(11):E1317-E1325. doi:10.1073/pnas.1500076112

28. Snyder MW, Kircher M, Hill AJ, Daza RM, Shendure J. Cell-free DNA Comprises an In Vivo Nucleosome Footprint that Informs Its Tissues-Of-Origin. *Cell*. 2016;164(1):57-68. doi:10.1016/j.cell.2015.11.050

29. Lo YMD, Chan KCA, Sun H, et al. Maternal Plasma DNA Sequencing Reveals the Genome-Wide Genetic and Mutational Profile of the Fetus. *Sci Transl Med*. 2010;2(61). doi:10.1126/scitranslmed.3001720

30. Jahr S, Hentze H, Englisch S, et al. DNA Fragments in the Blood Plasma of Cancer Patients: Quantitations and Evidence for Their Origin from Apoptotic and Necrotic Cells1. *Cancer Res*. 2001;61(4):1659-1665.

31. Mouliere F, Chandrananda D, Piskorz AM, et al. Enhanced detection of circulating tumor DNA by fragment size analysis. *Sci Transl Med*. 2018;10(466):eaat4921. doi:10.1126/scitranslmed.aat4921

32. Jiang P, Sun K, Peng W, et al. Plasma DNA End-Motif Profiling as a Fragmentomic Marker in Cancer, Pregnancy, and Transplantation. *Cancer Discov*. 2020;10(5):664-673. doi:10.1158/2159-8290.CD-19-0622

33. Jiang P, Xie T, Ding SC, et al. Detection and characterization of jagged ends of double-stranded DNA in plasma. *Genome Res*. 2020;30(8):1144-1153. doi:10.1101/gr.261396.120

34. Avgeris M, Marmarinos A, Gourgiotis D, Scorilas A. Jagged Ends of Cell-Free DNA: Rebranding Fragmentomics in Modern Liquid Biopsy Diagnostics. *Clinical Chem*. 2021;67(4):576-578. doi:10.1093/clinchem/hvab036

35. Guo W, Chen X, Liu R, et al. Sensitive detection of stage I lung adenocarcinoma using plasma cell-free DNA breakpoint motif profiling. *eBioMedicine*. 2022;81:104131. doi:10.1016/j.ebiom.2022.104131

36. Wadden J, Ravi K, John V, Babila CM, Koschmann C. Cell-Free Tumor DNA (cf-tDNA) Liquid Biopsy: Current Methods and Use in Brain Tumor Immunotherapy. *Front Immunol*. 2022;13:882452. doi:10.3389/fimmu.2022.882452

37. Douillard JY, Ostoros G, Cobo M, et al. First-line gefitinib in Caucasian EGFR mutation-positive NSCLC patients: a phase-IV, open-label, single-arm study. *Br J Cancer*. 2014;110(1):55-62. doi:10.1038/bjc.2013.721

38. Del Re M, Crucitta S, Paolieri F, et al. The amount of DNA combined with TP53 mutations in liquid biopsy is associated with clinical outcome of renal cancer patients treated with immunotherapy and VEGFR-TKIs. *J Transl Med*. 2022;20(1):371. doi:10.1186/s12967-022-03557-7

39. Siravegna G, Mussolin B, Buscarino M, et al. Monitoring clonal evolution and resistance to EGFR blockade in the blood of metastatic colorectal cancer patients. *Nat Med*. 2015;21(7):795-801. doi:10.1038/nm.3870

40. Nitschke C, Markmann B, Walter P, et al. Peripheral and Portal Venous KRAS ctDNA Detection as Independent Prognostic Markers of Early Tumor Recurrence in Pancreatic Ductal Adenocarcinoma. *Clinical Chemistry*. 2023;69(3):295-307. doi:10.1093/clinchem/hvac214

41. Liu L, Chen X, Petinrin OO, et al. Machine Learning Protocols in Early Cancer Detection Based on Liquid Biopsy: A Survey. *Life (Basel)*. 2021;11(7):638. doi:10.3390/life11070638

42. Raman L, Van der Linden M, Van der Eecken K, et al. Shallow whole-genome sequencing of plasma cell-free DNA accurately differentiates small from non-small cell lung carcinoma. *Genome Med*. 2020;12:35. doi:10.1186/s13073-020-00735-4

43. Liang N, Li B, Jia Z, et al. Ultrasensitive detection of circulating tumour DNA via deep methylation sequencing aided by machine learning. *Nat Biomed Eng*. 2021;5(6):586-599. doi:10.1038/s41551-021-00746-5

44. Lin LH, Chang KW, Kao SY, Cheng HW, Liu CJ. Increased Plasma Circulating Cell-Free DNA Could Be a Potential Marker for Oral Cancer. *Int J Mol Sci*. 2018;19(11):3303. doi:10.3390/ijms19113303

45. Guler GD, Ning Y, Ku CJ, et al. Detection of early stage pancreatic cancer using 5-hydroxymethylcytosine signatures in circulating cell free DNA. *Nat Commun*. 2020;11(1):5270. doi:10.1038/s41467-020-18965-w

46. Song CX, Yin S, Ma L, et al. 5-Hydroxymethylcytosine signatures in cell-free DNA provide information about tumor types and stages. *Cell Res*. 2017;27(10):1231-1242. doi:10.1038/cr.2017.106

47. Wang Y, Zheng J, Li Z, et al. Development of a novel liquid biopsy test to diagnose and locate gastrointestinal cancers. *J Clin Oncol*. 2020;38(15_suppl):1557-1557. doi:10.1200/JCO.2020.38.15_suppl.1557

48. Kothen-Hill ST, Zviran A, Schulman R, et al. Deep learning mutation prediction enables early stage lung cancer detection in liquid biopsy. *International Conference on Learning Representations*. February 15, 2018.

49. Moser T, Kühberger S, Lazzeri I, Vlachos G, Heitzer E. Bridging biological cfDNA features and machine learning approaches. *Trends Genet*. 2023;39(4):285-307. doi:10.1016/j.tig.2023.01.004

50. Chabon JJ, Hamilton EG, Kurtz DM, et al. Integrating genomic features for noninvasive early lung cancer detection. *Nature*. 2020;580(7802):245-251. doi:10.1038/s41586-020-2140-0

51. Chen L, Abou-Alfa GK, Zheng B, et al. Genome-scale profiling of circulating cell-free DNA signatures for early detection of hepatocellular carcinoma in cirrhotic patients. *Cell Res*. 2021;31(5):589-592. doi:10.1038/s41422-020-00457-7

52. Ma X, Chen Y, Tang W, et al. Multi-dimensional fragmentomic assay for ultrasensitive early detection of colorectal advanced adenoma and adenocarcinoma. *J Hematol Oncol*. 2021;14:175. doi:10.1186/s13045-021-01189-w

53. Cristiano S, Leal A, Phallen J, et al. Genome-wide cell-free DNA fragmentation in patients with cancer. *Nature*. 2019;570(7761):385-389. doi:10.1038/s41586-019-1272-6

54. Mathios D, Johansen JS, Cristiano S, et al. Detection and characterization of lung cancer using cell-free DNA fragmentomes. *Nat Commun*. 2021;12(1):5060. doi:10.1038/s41467-021-24994-w

55. Renaud G, Nørgaard M, Lindberg J, et al. Unsupervised detection of fragment length signatures of circulating tumor DNA using non-negative matrix factorization. *eLife*. 11:e71569. doi:10.7554/eLife.71569

56. Avanzini S, Kurtz DM, Chabon JJ, et al. A mathematical model of ctDNA shedding predicts tumor detection size. *Sci Adv*. 2020;6(50):eabc4308. doi:10.1126/sciadv.abc4308

57. Woodhouse R, Li M, Hughes J, et al. Clinical and analytical validation of FoundationOne Liquid CDx, a novel 324-Gene cfDNA-based comprehensive genomic profiling assay for cancers of solid tumor origin. *PLoS One*. 2020;15(9):e0237802. doi:10.1371/journal.pone.0237802

58. Pascual J, Attard G, Bidard FC, et al. ESMO recommendations on the use of circulating tumour DNA assays for patients with cancer: a report from the ESMO Precision Medicine Working Group. *Ann Oncol*. 2022;33(8):750-768. doi:10.1016/j.annonc.2022.05.520

59. LeeVan E, Pinsky P. Predictive Performance of Cell-Free Nucleic Acid-Based Multi-Cancer Early Detection Tests: A Systematic Review. *Clinical Chem*. Published online October 4, 2023:hvad134. doi:10.1093/clinchem/hvad134

60. Lapin M, Oltedal S, Tjensvoll K, et al. Fragment size and level of cell-free DNA provide prognostic information in patients with advanced pancreatic cancer. *J Transl Med*. 2018;16:300. doi:10.1186/s12967-018-1677-2

61. Moding EJ, Liu Y, Nabet BY, et al. Circulating Tumor DNA Dynamics Predict Benefit from Consolidation Immunotherapy in Locally Advanced Non-Small Cell Lung Cancer. *Nat Cancer*. 2020;1(2):176-183. doi:10.1038/s43018-019-0011-0

62. Powles T, Assaf ZJ, Davarpanah N, et al. ctDNA guiding adjuvant immunotherapy in urothelial carcinoma. *Nature*. 2021;595(7867):432-437. doi:10.1038/s41586-021-03642-9

63. Ribba B, Roller A, Helms HJ, Stern M, Bleul C. Circulating tumor DNA: Opportunities and challenges for pharmacometric approaches. *Front Pharmacol*. 2023;13. Accessed October 26, 2023. https://www.frontiersin.org/articles/10.3389/fphar.2022.1058220

64. Janssen JM, Verheijen RB, van Duijl TT, et al. Longitudinal nonlinear mixed effects modeling of EGFR mutations in ctDNA as predictor of disease progression in treatment of EGFR-mutant non-small cell lung cancer. *Clin Transl Sci*. 2022;15(8):1916-1925. doi:10.1111/cts.13300

65. Khan KH, Cunningham D, Werner B, et al. Longitudinal liquid biopsy and mathematical modelling of clonal evolution forecast waiting time to treatment failure in the PROSPECT-C phase II colorectal cancer clinical trial. *Cancer Discov*. 2018;8(10):1270-1285. doi:10.1158/2159-8290.CD-17-0891

66. Esfahani MS, Hamilton EG, Mehrmohamadi M, et al. Inferring gene expression from cell-free DNA fragmentation profiles. *Nat Biotechnol*. 2022;40(4):585-597. doi:10.1038/s41587-022-01222-4

67. Assaf ZJF, Zou W, Fine AD, et al. A longitudinal circulating tumor DNA-based model associated with survival in metastatic non-small-cell lung cancer. *Nat Med*. Published online March 16, 2023:1-10. doi:10.1038/s41591-023-02226-6

68. Zou W, Yaung SJ, Fuhlbrück F, et al. ctDNA Predicts Overall Survival in Patients With NSCLC Treated With PD-L1 Blockade or With Chemotherapy. *JCO Precis Oncol*. 2021;(5):827-838. doi:10.1200/PO.21.00057

69. Ciccolini J, Barbolosi D, André N, Barlesi F, Benzekry S. Mechanistic Learning for Combinatorial Strategies With Immuno-oncology Drugs: Can Model-Informed Designs Help Investigators? *JCO Precis Oncol*. 2020;(4):486-491. doi:10.1200/PO.19.00381

70. Nagasaka M, Uddin MH, Al-Hallak MN, et al. Liquid biopsy for therapy monitoring in early-stage non-small cell lung cancer. *Mol Cancer*. 2021;20:82. doi:10.1186/s12943-021-01371-1

71. Cohen JD, Li L, Wang Y, et al. Detection and localization of surgically resectable cancers with a multi-analyte blood test. *Science*. 2018;359(6378):926-930. doi:10.1126/science.aar3247

72. Jamal-Hanjani M, Wilson GA, Horswell S, et al. Detection of ubiquitous and heterogeneous mutations in cell-free DNA from patients with early-stage non-small-cell lung cancer. *Annals of Oncology*. 2016;27(5):862-867. doi:10.1093/annonc/mdw037

73. Wu X, Zhang Y, Hu T, et al. A novel cell-free DNA methylation-based model improves the early detection of colorectal cancer. *Mol Oncol*. 2021;15(10):2702. doi:10.1002/1878-0261.12942

74. Chemi F, Pearce SP, Clipson A, et al. cfDNA methylome profiling for detection and subtyping of small cell lung cancers. *Nat Cancer*. 2022;3(10):1260-1270. doi:10.1038/s43018-022-00415-9

75. Weiss GJ, Beck J, Braun DP, et al. Tumor Cell–Free DNA Copy Number Instability Predicts Therapeutic Response to Immunotherapy. *Clin Cancer Res*. 2017;23(17):5074-5081. doi:10.1158/1078-0432.CCR-17-0231

76. Si H, Kuziora M, Quinn KJ, et al. A Blood-based Assay for Assessment of Tumor Mutational Burden in First-line Metastatic NSCLC Treatment: Results from the MYSTIC Study. *Clin Cancer Res*. 2021;27(6):1631-1640. doi:10.1158/1078-0432.CCR-20-3771

77. Matsumae T, Kodama T, Myojin Y, et al. Circulating Cell-Free DNA Profiling Predicts the Therapeutic Outcome in Advanced Hepatocellular Carcinoma Patients Treated with Combination Immunotherapy. *Cancers*. 2022;14(14):3367. doi:10.3390/cancers14143367

78. Khagi Y, Goodman AM, Daniels GA, et al. Hypermutated Circulating Tumor DNA: Correlation with Response to Checkpoint Inhibitor–Based Immunotherapy. *Clinical Cancer Research*. 2017;23(19):5729-5736. doi:10.1158/1078-0432.CCR-17-1439

79. Luo H, Zhao Q, Wei W, et al. Circulating tumor DNA methylation profiles enable early diagnosis, prognosis prediction, and screening for colorectal cancer. *Sci Transl Med*. 2020;12(524):eaax7533. doi:10.1126/scitranslmed.aax7533

80. Lee JH, Long GV, Boyd S, et al. Circulating tumour DNA predicts response to anti-PD1 antibodies in metastatic melanoma. *Ann Oncol*. 2017;28(5):1130-1136. doi:10.1093/annonc/mdx026

81. Cabel L, Riva F, Servois V, et al. Circulating tumor DNA changes for early monitoring of anti-PD1 immunotherapy: a proof-of-concept study. *Ann Oncol*. 2017;28(8):1996-2001. doi:10.1093/annonc/mdx212

82. Ricciuti B, Jones G, Severgnini M, et al. Early plasma circulating tumor DNA (ctDNA) changes predict response to first-line pembrolizumab-based therapy in non-small cell lung cancer (NSCLC). *J Immunother Cancer*. 2021;9(3):e001504. doi:10.1136/jitc-2020-001504

83. Goldberg SB, Narayan A, Kole AJ, et al. Early Assessment of Lung Cancer Immunotherapy Response via Circulating Tumor DNA. *Clin Cancer Res*. 2018;24(8):1872-1880. doi:10.1158/1078-0432.CCR-17-1341

84. Anagnostou V, Ho C, Nicholas G, et al. ctDNA response after pembrolizumab in non-small cell lung cancer: phase 2 adaptive trial results. *Nat Med*. 2023;29(10):2559-2569. doi:10.1038/s41591-023-02598-9

85. Herbreteau G, Vallée A, Knol AC, et al. Circulating Tumor DNA Early Kinetics Predict Response of Metastatic Melanoma to Anti-PD1 Immunotherapy: Validation Study. *Cancers (Basel)*. 2021;13(8):1826. doi:10.3390/cancers13081826

86. Zhang Q, Luo J, Wu S, et al. Prognostic and Predictive Impact of Circulating Tumor DNA in Patients with Advanced Cancers Treated with Immune Checkpoint Blockade. *Cancer Discov*. 2020;10(12):1842-1853. doi:10.1158/2159-8290.CD-20-0047

87. Nabet BY, Esfahani MS, Moding EJ, et al. Noninvasive Early Identification of Therapeutic Benefit from Immune Checkpoint Inhibition. *Cell*. 2020;183(2):363-376.e13. doi:10.1016/j.cell.2020.09.001

88. Bratman SV, Yang SYC, Iafolla MAJ, et al. Personalized circulating tumor DNA analysis as a predictive biomarker in solid tumor patients treated with pembrolizumab. *Nat Cancer*. 2020;1(9):873-881. doi:10.1038/s43018-020-0096-5

89. Gristina V, Barraco N, La Mantia M, et al. Clinical Potential of Circulating Cell-Free DNA (cfDNA) for Longitudinally Monitoring Clinical Outcomes in the First-Line Setting of Non-Small-Cell Lung Cancer (NSCLC): A Real-World Prospective Study. *Cancers (Basel)*. 2022;14(23):6013. doi:10.3390/cancers14236013

Figure 1

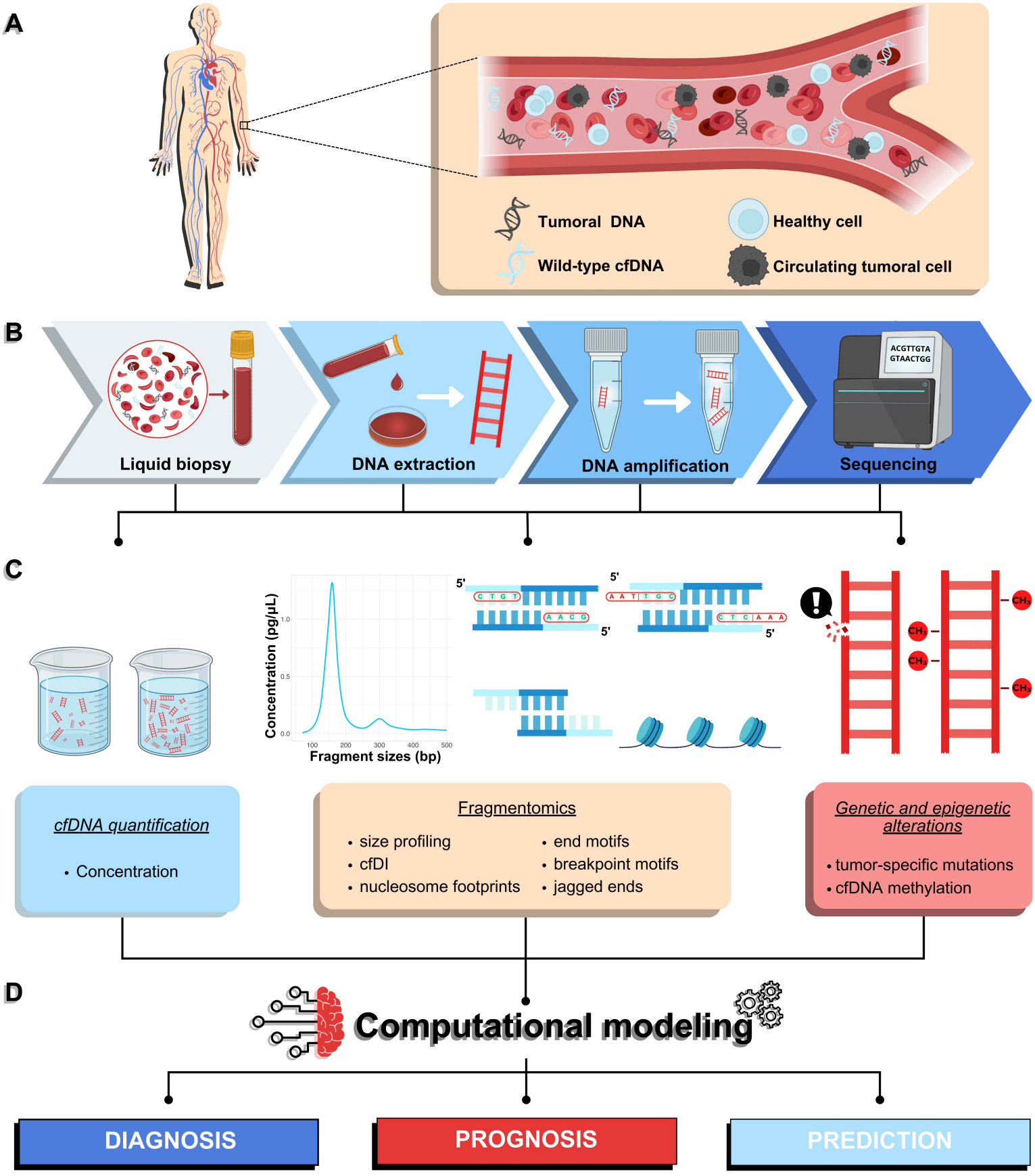
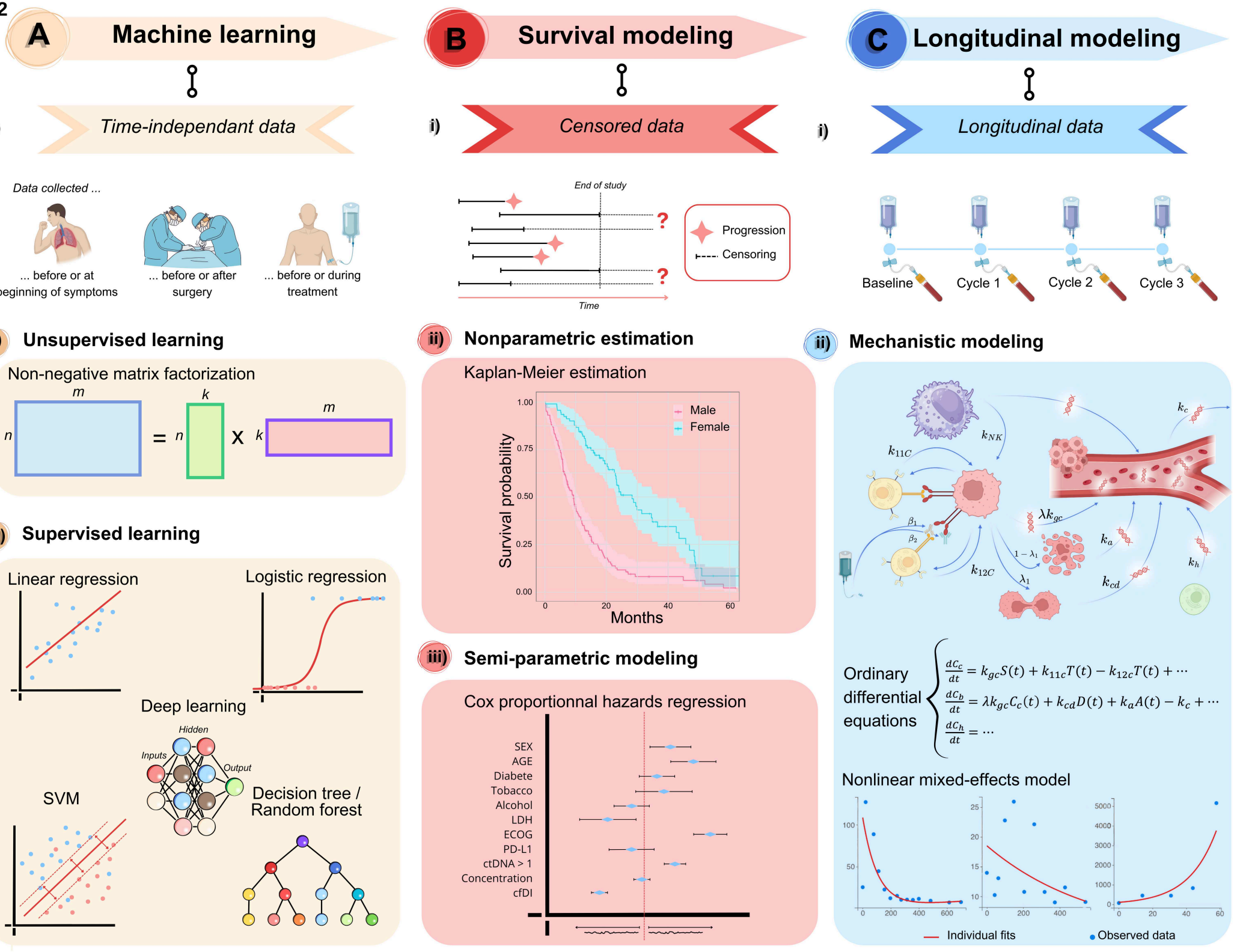

A
Tumoral DNA
Healthy cell
Wild-type cfDNA
Circulating tumoral cell
B
Liquid biopsy
DNA extraction
DNA amplification
ACGTTGTA
GTAACTGG
Sequencing
C
Concentration (pg/µL)
Fragment sizes (bp)
5'
CTGT
AACG
AATTGC
CTCAAA
CH3
cfDNA quantification
• Concentration
Fragmentomics
• size profiling
• cfDI
• nucleosome footprints
• end motifs
• breakpoint motifs
• jagged ends
Genetic and epigenetic alterations
• tumor-specific mutations
• cfDNA methylation
D
Computational modeling
DIAGNOSIS
PROGNOSIS
PREDICTION

Figure 2


A
Machine learning
i)
Time-independant data
Data collected ...
... before or at beginning of symptoms
... before or after surgery
... before or during treatment
ii)
Unsupervised learning
Non-negative matrix factorization
m
n
k
n
k
m
iii)
Supervised learning
Linear regression
Logistic regression
Deep learning
Hidden
Inputs
Output
SVM
Decision tree / Random forest
B
Survival modeling
i)
Censored data
End of study
Time
Progression
Censoring
ii)
Nonparametric estimation
Kaplan-Meier estimation
Survival probability
1.00
0.75
0.50
0.25
0.00
0
20
40
60
Months
Male
Female
iii)
Semi-parametric modeling
Cox proportionnal hazards regression
SEX
AGE
Diabete
Tobacco
Alcohol
LDH
ECOG
PD-L1
ctDNA > 1
Concentration
cfDI
C
Longitudinal modeling
i)
Longitudinal data
Baseline
Cycle 1
Cycle 2
Cycle 3
ii)
Mechanistic modeling
k_{NK}
k_{11C}
k_c
λk_{gc}
β_1
β_2
k_a
$1-\lambda_1$
k_{12C}
λ_1
k_{cd}
k_h
Ordinary differential equations
$\frac{dC_c}{dt} = k_{gc}S(t) + k_{11c}T(t) - k_{12c}T(t) + \cdots$
$\frac{dC_b}{dt} = \lambda k_{gc}C_c(t) + k_{cd}D(t) + k_aA(t) - k_c + \cdots$
$\frac{dC_h}{dt} = \cdots$
Nonlinear mixed-effects model
Individual fits
Observed data