# Computational modeling for circulating cell-free DNA in clinical oncology

Linh Nguyen Phuong, Sébastien Salas, Sébastien Benzekry

# Computational modeling for circulating cell-free DNA in clinical 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 ability to collect circulating cell-free DNA (cfDNA) from blood samples provides real-time insights into tumor biology, enabling its application in clinical practice for cancer screening, diagnosis, minimal residual disease assessment, and prediction and monitoring of treatment response and relapse.

Given the increasing complexity, volume, and longitudinal nature of cfDNA data, there is a growing demand for advanced computational modeling (CM) approaches that can transform these data into clinically actionable insights.

We report on the diverse CM approaches used to analyze cfDNA in oncology. After an overview of the current data derived from cfDNA, the use of CM is detailed for their application in clinical studies, both in processing cfDNA data at a particular time point and in capturing their temporal dynamics. We emphasize on approaches using machine learning and mechanistic modeling embedded within non-linear mixed effects statistical constructs.

This review provides guidance to computational modelers, clinical researchers and healthcare practitioners in effectively utilizing cfDNA data to enhance research and improve patient care.

**Keywords:** cell-free DNA; computational modeling; machine learning biomarker; diagnosis; prognosis; prediction of treatment response

## INTRODUCTION

Cell-free DNA (cfDNA) comprises encapsulated DNA fragments released into body fluids such as blood, urine or cerebrospinal fluid[1] (Figure 1A), allowing easy access to the genetic background of hard-to-reach tissues. Originating from necrosis, apoptosis[2], lysis[3], active secretion of exosomes[4] and the hematopoietic system[5], cfDNA biology in oncology is a current research focus for diagnosis, prognosis, treatment monitoring, and therapy personalization.

CfDNA includes circulating-tumor DNA (ctDNA), i.e. DNA fragments released by tumor cells (primary tumor, circulating tumor cells and metastases) containing specific genetic aberrations[6]. Consequently, cfDNA analyses can offer insights into disease biology, tumor mutations and growth with multiple advantages. First, collection of cfDNA can easily be performed from blood sampling (liquid biopsies, Figure 1B)[7]. These are less invasive than tissue biopsies and provide real-time disease monitoring, reinforced by the short half-life of cfDNA in the blood (15 minutes to 2 hours[8]). CfDNA can also help detect new tumor sites and serves as a surrogate for tissue biopsies when the primary site is unidentified. Furthermore, ctDNA offers insights on both intra- and inter-tumoral genetic heterogeneity, otherwise difficult to access[9].

In the clinic, cfDNA can be applied for screening, diagnosis, detection of minimal residual disease and treatment monitoring. The cfDNA-based colorectal cancer screening test Epi proColon®[10] has been approved by the Food and Drug Administration (FDA) in 2016 and is recommended for adults at risk aged over 50 who are not completing the colorectal cancer screening. In 2022, the European Society for Medical Oncology (ESMO) Precision Medicine Working Group recommended ctDNA as an adjunctive diagnostic tool for several solid cancers (NSCLC, breast, gastric, …)[11]. In recent years, multiple clinical trials have been conducted using ctDNA to guide targeted therapy[12,13]. These hold the premise of transforming precision medicine. However, no cfDNA-based assay has yet reached sufficient level of maturity to be used in routine care.

Sequencing and quantifying cfDNA over time generates extensive data. These include gene aberrations, fragment size profiles and 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 two first are interested in establishing associations and predictive models between cfDNA as input and outcome as output. The latter aims at integrating biological knowledge into mathematical models of the pharmaco-patho-physiological processes at stake, which not only provides predictive tools but also improves our understanding and allows to test mechanistic hypotheses and perform simulations of putative scenarii.

CM-cfDNA emerged around 2015, aligned with the increasing popularity of cfDNA-based studies. It constituted 24% of tumoral cfDNA studies in 2022, rising 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). These methods, including survival analysis for time-to-event data, ML for predictive modeling, and mechanistic modeling for tumor-drug interactions, have proven effective in oncology for early detection, cancer subtype classification, prognosis, treatment monitoring, and prediction of response / relapse in cancers such as melanoma[14], lung[15], breast[16] and colorectal[17] cancers (Figure 1D). Specifically, despite its rare application, collecting cfDNA data over time, such as during treatment, and modeling it with mechanistic approaches enhanced the understanding of cfDNA biology and ensured robust, predictive outcomes.

## CfDNA AND ctDNA DATA

### *CfDNA collection, quantification, sequencing and storage*

CfDNA data collection involves blood collection followed by plasma extraction by centrifugation, storage, DNA extraction, quantification and sequencing (Figure 1B).

Research has explored cfDNA molecular diagnostics for quantifying fragments and detecting tumor aberrations, using targeted and non-targeted approaches[18]. Targeted approaches focus on predefined genes related to the patient's pathology. They comprise digital droplet Polymerase Chain Reaction (PCR), quantitative-PCR, amplification-refractory mutation system, BEAMing-PCR, tagged-amplicon deep sequencing and cancer personalized profiling by deep-sequencing[19], offering good sensitivity and better specificity than non-targeted approaches[18]. Non-targeted approaches assess the entire genome (array-comparative genomic hybridization, whole genome sequencing). They can identify new aberrations through genome-wide screening but require larger cfDNA amounts.

Emerging database platforms (FinaleDB[20], CFEA[21]) provide comprehensive cfDNA datasets from various studies and clinical conditions.

### *CfDNA features*

Concentration is the most easily accessible cfDNA-based feature. It fluctuates with tumor size, number of metastases and presence of circulating tumor cells[22]. In healthy individuals, cfDNA concentrations range from 0 to 100 ng/ml, averaging 30 ng/ml, while in cancer patients, it ranges from 0 to 1,000 ng/ml, averaging 180 ng/ml[23]. Cancer stage also affects ctDNA proportion within cfDNA, being two times smaller in stage I than in stage III patients[24]. However, elevated cfDNA levels can also result from pro-inflammatory or auto-immune diseases, cirrhosis, hepatitis[25], systemic lupus[26], pregnancy[27], or intense physical activity[28], introducing potential biases.

Further, research has investigated fragmentation features such as fragment sizes, end and breakpoint motifs, jagged ends and nucleosome footprints (Figure 1C), grouped together under the term fragmentomics, introduced in 2015[29]. Short fragments (~166 bp) arise from apoptosis[2], whereas longer fragments (~10,000 bp) appear to originate from necrosis[30], much less present in the plasma[2]. Additionally, shorter fragments (90-150 bp) were observed in cancer patients. Subsequently, focusing on these fragments could enhance ctDNA detection[31]. Finally, cancer patients present distinct and more

variable end motifs[32], different proportions of jagged ends (uneven DNA extremities, Figure 1C)[33,34], and various breakpoint motifs[35].

Finally, a large body of literature has examined mutations detectable in cfDNA. When these mutations are tumor-specific, the afferent cfDNA is considered as ctDNA (Figure 1B-C). CtDNA can monitor the mutation level of oncogene-addicted cancers like endothelial growth factor receptor- (EGFR) or Kirsten rat sarcoma viral oncogene homologue- (KRAS) positive lung cancers, allowing treatment adaptation[36–38]. Blood EGFR mutations appear to match tissue EGFR mutations, reinforcing liquid biopsy as a surrogate of tissue biopsy[36].

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).

## CM-CFDNA FOR DIAGNOSIS

A major use of CM-cfDNA is to detect tumors before clinical symptoms by discerning changes in cfDNA characteristics to discriminate pathological to healthy individuals or to distinguish between different tumor subtypes. Moreover, identification of the primary tumor site can be a major challenge. ctDNA specific mutations can help pinpointing the primary tumor location. Consequently, multiple studies conducted pan-cancer analyses.

### *ML for ctDNA-based diagnosis*

Supervised classification methods (Figure 2A, see supplement for details) have been increasingly employed in the last decade for early cancer detection[39]. These statistical analysis techniques enable learning from an initial patient dataset to predict diagnosis for new patients. Table 1 summarizes a list of ML-ctDNA studies for diagnosis.

ML models have enabled the differentiation between lung cancer histologies leveraging copy number profiling of cfDNA[40]. Various classifiers were compared: RF, SVM, LR with ridge, elastic-net (EN) or least absolute shrinkage and selection operator (LASSO) regularizations. The ridge-penalized LR outperformed the other ones with a mean area under the receiver operating characteristic curve (AUC) of 0.936. Another study used SVM to select discriminative differentially methylated blocks for early lung cancer detection[41], achieving a sensitivity (true positive rate) of 52-64-77–81% for stages IA-IB-II-III patients, respectively, for a fixed specificity (true negative rate) of 96% (95% confidence interval (CI) 93–98%). Liu et al.[39] reviewed the most relevant and recent ML studies for early detection and noted four ML-cfDNA studies employing linear models, possibly with EN or LASSO regularization. Three other ML-cfDNA studies employed SVM, one investigating 5-hydroxymethylcytosine density for pancreatic ductal adenocarcinoma early detection[42]. 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[43]. Another identified new biomarkers from the cfDNA methylome in the plasma, able to diagnose and locate gastrointestinal cancers with an AUC 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[44].

Deep learning has been less frequently used due to small sample size of these types of studies. One study employed a convolutional neural network for early stage lung cancer detection[45], using two-dimensional grids representing the sequenced reads. The algorithm then detected base changes, such as deletions, mutations or insertions, focusing on distinguishing artifacts from genuine cancer mutations. After training on 3 patient genomes, they achieved a sensitivity of 0.903 for a specificity of 0.94 for detection and distinguishment of specific lung cancer mutation patterns and systemic sequencing artifacts.

Moser et al.[46] cited over twenty diagnosis studies employing ML, noting that somatic mutations of non-cancerous origin can increase false positives. For instance, age-related clonal hematopoiesis can produce misleading results. To address this, Chabon et al.[47] developed an ensemble classifier composed of 5-nearest neighbor, 3-nearest neighbor, naïve Bayes, LR and decision tree. They linearly combined the scores from these classifiers to distinguish tumor from clonal hematopoiesis mutations, finding the latter in longer cfDNA fragments.

Among supervised learning methods, none seems to significantly outperform the other ones. Simple linear models as well as non-linear methods can achieve high accuracy scores, depending on the dataset, target outcome or cancer. Linear models are the most used methods and have the benefit of simplicity and interpretability. On the other hand, random forests often provide better diagnostic performances.

Eventually, unsupervised learning (Figure 2A) has also been employed for cancer subtypes classification, although less frequently. For example, Luo et al. used hierarchical clustering to distinguish colorectal cancer patients from normal subjects according to methylation markers, and also identified patient subgroups with different overall survival (OS)[17].

***ML for fragmentomics***

Studies began integrating DNA fragmentation patterns, which may also reflect biological characteristics of the tumor. 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), achieving 95.4% sensitivity and 97.8% specificity in the test set[48]. Guo et al. compared LR, deep learning and extreme gradient boosting to detect stage I lung adenocarcinoma, using the 6bp breakpoint

motif (defined as « the 3bp extensions to both directions of the aligned cfDNA 5′ »), achieving 92.5% sensitivity and 90.0% specificity in the external validation cohort[35], outperforming early diagnosis from ctDNA mutations. 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[49]. They reached impressive scores of 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. to classify healthy individuals and cancer patients (from seven different pathologies)[50]. They employed the cfDNA integrity index, defined as the ratio of short (100-150 bp) to long fragments (150-200 bp), across 504 genome bins. They highlighted distinct size variations across different genome regions of cancer patients. They integrated these features into a stochastic gradient tree boosting framework, splitting samples according to a 10-fold cross-validation repeated 10 times, with feature selection at each of the ten steps on the inner-fold training dataset. With a 95% specificity, they detected 80% of the cancer patients. Expanding the framework, they identified tumor tissue origin with 90% specificity and 61% accuracy, reaching 75% accuracy for the top two predictions.

Mathios et. al implemented a comparable ML approach based on similar fragmentation features for lung cancer detection and staging in high-risk symptomatic subjects[51]. They reduced the features' dimensionality by selecting principal components explaining 90% of the fragmentation variance. Subsequently, they used a LASSO-penalized LR to assess the fragmentation features. With 10 replicates of a 5-fold cross-validation, they defined a score able to detect 94% of the cancer patients with an 80% specificity, in a population with 91% early-stage (I-II) cancer patients.

Mouliere et al. developed CM-cfDNA ML (RF and LR) to detect cancer, using proportion of fragments in multiple size ranges, their ratios, and 10 bp periodicity amplitudes occurring before 150 bp[31]. They reached an AUC of 0.891 for cancers with low ctDNA amounts (pancreatic, renal and glioma) with RF, with four fragmentation features, selected as the best feature set in LR and RF among nine.

Renaud et al. used unsupervised learning to detect cancer fragments in cfDNA via fragments lengths measured by shallow whole genome sequencing[52]. They

decomposed the fragment size profiles matrix using a novel non-negative matrix factorization (NMF):

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

with $k$ the number of "sources", to be set (e.g., $k = 2$ for healthy and cancer distributions of the fragment lengths). Integrating the weights into an SVM framework allowed to detect cancer patients with an AUC of 0.95 for $k = 30$.

While ctDNA genomics provide information about tumor gene mutations, fragmentomics offers insight into the architecture of DNA molecules and the non-random patterns of fragmentation. Currently, depending on the cancer type and stage, neither fragmentomics nor ctDNA aberrations have been identified as the best feature for early diagnosis. However, fragmentomics, augmented with ML, is beginning to show better prediction in early stages[35] compared to ctDNA analysis. Ultimately, combination of both approaches may offer the most accurate prediction, providing two complementary sources of information about the tumor[48,50]

***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, in order to determine the optimal screening frequency for early lung cancer detection[53]. 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 an expectation depending on mechanistic parameters (tumo cells growth and death rates, mean shedding rate of a cell death and ctDNA elimination rate). It was found from simulations that ctDNA-based annual screening would yield a median detection size of 2.0 to 2.3 cm of diameter, against 3.5 cm for the standard annual screening based on imaging.

## CM-CFDNA FOR PROGNOSIS AND PREDICTION OF RESPONSE TO TREATMENT

Multiple studies used CM-cfDNA to monitor tumor size and predict therapeutic responses, grouped in Table 2. They aimed to identify signatures enabling early treatment adjustments and prevention of adverse events.

Most of them relied on baseline markers (cfDNA/ctDNA concentration, ctDNA detection, fragmentomics), either post-surgery to correlate with relapse, or before treatment initiation to predict treatment response. Some studies also collected biological markers at multiple time points throughout treatment, providing longitudinal datasets including absolute values and relative changes from baseline data. These datasets enabled analysis of cfDNA dynamics to identify patterns related to time to relapse, progression, treatment response, or mortality.

Initially, classical statistical methods such as survival analysis (Figure 2B, supplement) were used to connect cfDNA measurements to treatment outcomes, before computational methodologies were introduced by the use of ML, non-linear mixed effects models (NLME) and mechanistic modeling.

### *ML-cfDNA from baseline data*

Yang et al. classified breast cancer patients into responders and non-responders to neoadjuvant chemotherapy using hierarchical clustering based on the coverage depth near transcription start sites in cfDNA[54]. On the other side, Panagopoulou et al. compared supervised learning methods to classify breast cancer patients according to fragmentome and methylation patterns[55]. SVM more effectively classified patients into progressive disease, partial response, and stable disease groups (AUC: 0.74, 95% CI: 0.622-0.937), while logistic regression based only on cfDNA concentration was sufficient to distinguish responders from non-responders to chemotherapy (AUC 0.803, 95% CI 0.606, 1.000).

### *Longitudinal ML modeling*

ML is able to leverage longitudinal data by pooling the cfDNA features from multiple time points using, e.g., absolute or relative changes over time. These can serve as inputs to classify patients as responders or non-responders using supervised or semi-supervised methods.

Assaf et al.[56] analyzed 466 NSCLC patients, developing n ML framework to predict immunotherapy response using longitudinal ctDNA data collected at baseline and days 1 of cycle 2 (C2D1) and cycle 3 (C3D1). They integrated 19 ctDNA metrics at each timepoint and 59 relative ctDNA changes of these metrics from baseline into three models: baseline, baseline + C2D1, and baseline + C2D1 + C3D1. The latter exhibited the highest C-index for OS prediction. Then, they combined ctDNA features with clinical features. Using an EN approach with leave-one-out cross-validation (LOOCV) through a 10-folds nested cross-validation process, they retained features selected in over 50% of iterations with a positive gain metric, according to the next-door analysis[57]. This process selected five relevant variables, including global cfDNA concentration at C3D1 (HR: 1.48, 95% CI: 1.06-2.07). The final model categorized patients into high (progressive disease), intermediate (stable disease) and low (responders) risk patients, and demonstrated significant risk stratification in both train and test datasets for OS. In the test set high-risk patients showed shorter OS (median 7.3 months) than low- and intermediate-risk patients (median 25.2 months) (HR: 3.28, 95% CI: 2.2-4.9, $p < 0.001$).

Ding et al.[58] outperformed the previous results by increasing the C-index by 9.8% and 16.2% to predict OS and PFS respectively, using functional principal component analysis[59], an unsupervised learning method, to extract new features from ctDNA trajectories, and selecting high-importance features by random forest. This method allowed to capture and use the entire kinetics rather than only five snapshot time points.

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[60]. They highlighted the C3D1 median number of mutant molecules per mL as the key OS predictor.

***Longitudinal mechanistic modeling***

Few studies have explored the dynamics CM-cfDNA (Figure 2C). Ribba et al. developed a mechanistic modeling of the joint ctDNA–tumor size evolution over time in order to assess atezolizumab response in NSCLC and melanoma patients[61]. They used a bi-exponential system to describe both the log10-transformed number of mutant molecules per mL and the sum of the longest diameters (SLD) of target lesions. A coefficient linking the tumor size decay rate to the ctDNA growth rate was assumed.

This system was able to accurately describe ctDNA and tumor kinetics over time, even when negatively correlated (one increasing while the other decreasing), highlighting the biological link between tumor growth and ctDNA release for immune-checkpoint inhibition (ICI)-treated patients. The parameters were estimated by a population approach, using NLME and Bayesian estimation for individual parameter estimation. The estimated ctDNA growth rate showed a high correlation with the estimated SLD growth rate.

Janssen et al. analyzed ctDNA biomarkers' kinetics to early predict resistance to targeted therapy in NSCLC patients [62]. They developed a NLME model (Figure 2C) to describe the dynamics of EGFR mutations in ctDNA, using a zero-order growth model, i.e.:

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

where $y(t)$ is the change in either L858R or exon19del over time and $R(t) = e^{-\lambda t}$ for driver mutations, and $R(t) = e^{-\lambda.y(t)}$ for the T790M mutation concentration. The model was able to correctly approximate the actual concentrations. Integrating the modeled ctDNA values into survival models for PFS revealed that the relative change in driver mutations concentration was the only predictor statistically significant for stratifying responders and non-responders (p = 0.001, likelihood-ratio test).

Prior to this study, Khan et al. sought to model carcinoembryonic antigen dynamics, which is proportional to the total number of tumor cells[63], and to correlate it with ctDNA dynamics, to assess cetuximab response in colorectal cancer patients. They used a biexponential model for tumor dynamics, while cfDNA mutant frequencies were modeled using a single exponential growth model. The cfDNA model described well the actual dynamics ($R^2 = 0.979$) and demonstrated a correlation between the cfDNA relapse rate and the carcinoembryonic antigen one. It was found that these tow parameters were able to predict the time to relapse. Recently, Li et al. proposed a stochastic birth-death process of tumor cells under targeted therapy, chemotherapy or radiotherapy, shedding ctDNA under varying probabilities, to simulate randomized patient cohorts.[64] From this modeling, they defined new ctDNA biomarkers of response, offering good prediction of response during a 48-hour and two weeks time-period after initiation of treatment for target therapy/chemotherapy and radiotherapy respectively, outperforming the existing ctDNA biomarkers.

Eventually, Esfahani et al. published the only study found to investigate fragmentomics through mechanistic modeling, using cfDNA fragmentation profiles in lung carcinoma and diffuse B cell carcinoma patients[65]. They particularly modeled the nucleosome positions on 2000-bp fragments containing transcription start site and the potential accessible sites for cut during fragmentation. The associated size profiles were then simulated. This modeling allowed to explore the parameters influencing gene expression detection within cfDNA, helping for the finding of specific tumor genes.

## DISCUSSION

In recent years, several studies have explored CM-cfDNA as an innovative biomarker for cancer diagnosis and treatment monitoring. CM-cfDNA has become essential for managing this emerging data. In the past two years, approximately one in four studies has employed CM-cfDNA to support precision medicine in oncology, including ML. Early approaches involved simple linear models to link cfDNA characteristics with cancer presence, type, or stage. More advanced methods, including decision tree-based models and deep learning, were later employed to capture complex relationships. To handle a larger number of cfDNA characteristics than patients, penalized linear models or RF approaches were employed for feature selection. The integration of longitudinal data into ML has gained momentum for cancer screening and treatment monitoring. However, the lack of biological interpretability in complex models brings interest to extend CM-cfDNA to mechanistic modeling.

Research has highlighted the potential of ctDNA tests for screening, reaching a maturity level sufficient to be FDA[66]- or ESMO[11]-approved for several solid cancers (NSCLC, breast, gastric, …). As they can cost a lot and has not been enough validated in large population, they are however not yet widely used in routine. In addition, several were calibrated on symptomatic, diagnosed or high-stage patients. Therefore, further prospective evaluations targeting and validating markers in asymptomatic individuals at risk for cancer[67] or in early-stage patients are necessary for clinical applicability.

Another major limitation of ctDNA-based assays is the need of standardized protocols for data collection across institutions and companies, crucial for ensuring reliable results and reproducibility. Except for quantitative-PCR[10], which is ready for routine care, most of the experimental methods have only been limited to clinical research. To integrate them in clinical practice, these methods require to be standardized and rigorously compared to other molecular techniques[68]. In addition, harmonization issues within and between institutions can affect reproducibility for clinical use[69]. While cfDNA is easy to collect, pre-analytical variables like tube type, agitation, pre-centrifugation delay and temperature, centrifugation steps, sample storage, quantification and sequencing methods, can impact data precision[70]. The US National Cancer Institute recently published guidelines based on literature evidence and validated by a panel of international experts to standardize processing practices[69]. Additionally, high-

throughput next-generation sequencing is costly and time-consuming[71], prompting new methodologies to simplify techniques.

One such interesting approach is the cfDNA fragmentomics field, which also exploits that ctDNA is detectable in less than 80% of cancer patients[72]. Specifically, analyzing the size distribution of cfDNA fragments seems particularly interesting [50]. It can be performed at low cost, by combining hydrodynamic and electrokinetic actuation[73], only needs 10 μL of blood and does not require to extract cfDNA from plasma. Wider, fragmentomics is still relatively new and remains underused, especially for treatment monitoring. Only two studies were found searching for “fragmentomics” AND “chemotherapy” on PubMed. Additionally, only two studies were found searching for “fragmentomics” AND (“immune-checkpoint inhibitors” OR “immunotherapy”).

CM-cfDNA has shown promising results in advancing precision medicine in oncology. Recent studies have compared various ML models, including ensemble methods[50] to enhance predictive performance. Techniques like cross-validation have further demonstrated the robustness of these predictive signatures. However, there is no consensus on a single ML method that is universally well-suited for clinical application, making it necessary to explore different ML methods for each feature-outcome combination. A major challenge remains the lack of interpretability in complex ML frameworks, which can obscure the underlying biological rationale. Two opposite approaches can be undertaken to address this: use simple linear models that offer better interpretability, or turn to mechanistic modeling, parameterized with experimental data and able to leverage biochemical knowledge underlying cfDNA release. For nonlinear ML, the use of SHapley Additive Values is increasingly popular for explainability[74]. In addition, ML studies mainly focused on maximizing specificity first to avoid false positive, leading to high false-negative results[11]. Mutated ctDNA is also less prevalent in early-stage cancers. Thus, integrating ctDNA information with additional biological or multi-omics data may enhance detection sensitivity[75].

Despite these advancements, no CM-cfDNA-based test has yet been approved for clinical practice. For successful integration of these CM-cfDNA methods into routine, several steps are necessary in addition to the previous: rigorously software development and training, certification as medical devices, and ultimately validation in prospective randomized trials [76]. Clinicians also need to be trained to correctly interpret CM tools and predictions.

Currently, most research on cfDNA in clinical oncology has focused on survival and classification learning, with limited application of biologically-based mechanistic models that leverage longitudinal data effectively. Future research should expand cfDNA analyses by incorporating mathematical models that leverage the processes of cfDNA release and interactions across tumors, their microenvironment, and circulating materials (including the immune system). Achieving this will require enhanced collaboration among clinicians, biologists, computational scientists, and mathematicians.

Finally, combining mechanistic modeling with ML and survival analysis (mechanistic learning[77]) holds promise for developing highly informative biologico-computational markers and predictive tools. This approach not only offers a clearer understanding of the biological mechanisms but also complements the predictive power of ML with deeper insights into the biology of cfDNA in oncology.

## FIGURES

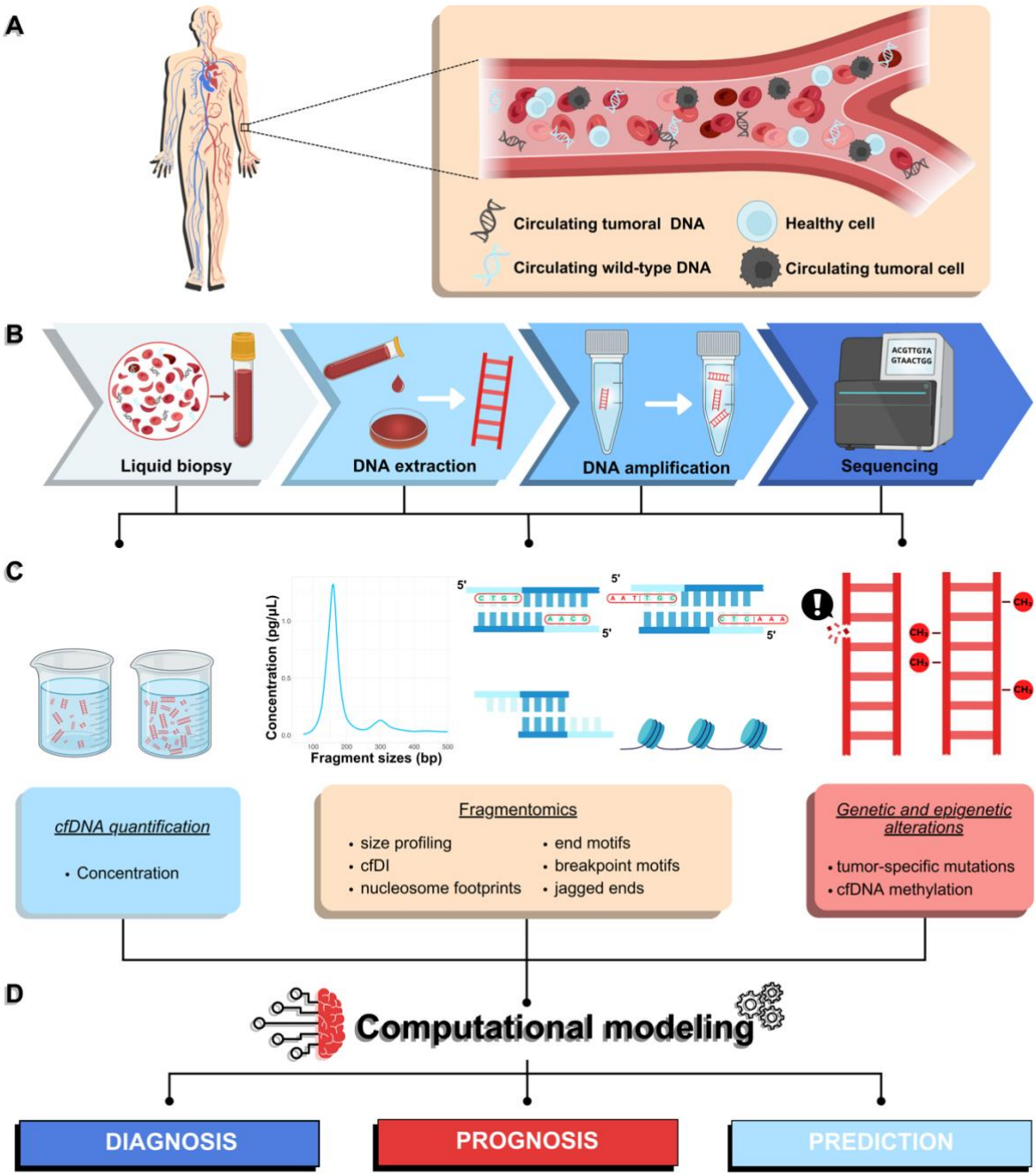


***Figure 1: Cell-free DNA data: a new biological tool for on oncology***

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.

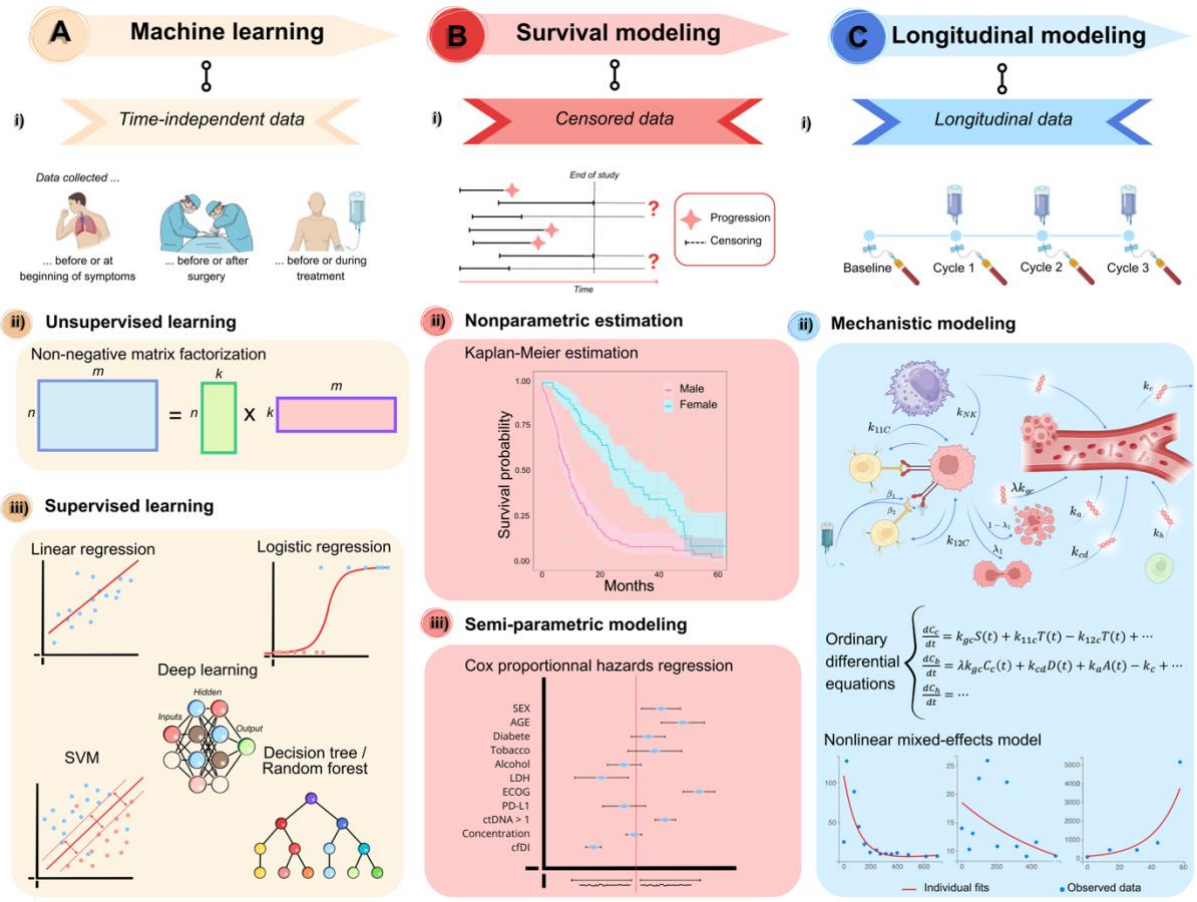


***Figure 2: Cell-free DNA data: a new biological tool for oncology***

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 |
|---|---|---|---|---|
| [43] | Pancancer | RF | ctDNA aberrations | Cancer type classification |
| [78] | Colorectal | LASSO – LR | | Early diagnosis |
| [79] | NSCLC | Linear regression | | |
| [47] | | 5nn – 3nn naïve Bayes – LR – Decision tree | | |
| [45] | Lung | CNN | | |
| [42] | Pancreas | RF – SVM – EN LR | | Stage classification |
| [40] | NSCLC | LASSO – Ridge – EN LR | | Histological classification |
| [80] | Oral | LR | cfDNA quantification | Early diagnosis |
| [53] | Lung | Mechanistic modeling | cfDNA concentration / ctDNA mutations | Early diagnosis |
| [51] | Lung | PCA – LR | Fragmentome | Early diagnosis |
| [31] | Pancancer | RF – LR | | |
| [48] | Hepatocellular carcinoma | LR | | |
| [35] | Lung adenocarcinoma | LR – deep learning, gradient and extreme gradient boosting | | |

| | | | | |
|---|---|---|---|---|
| [49] | Colorectal adenocarcinoma | Generalized linear model, deep learning, RF, extreme gradient boosting | | |
| [50] | Pancancer | Gradient-tree boosting | | Early diagnosis and cancer type classification |
| [81] | Colorectal | LASSO LR | Methylome | Early diagnosis |
| [17] | | Hierarchical clustering | | |
| [41] | Lung | SVM | | |
| [82] | SCLC | PCA | | Histological classification |
| [44] | 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 |
|---|---|---|---|---|---|
| [83] | NSCLC | (Durvalumab ± tremelimumab) + platinum-based chemotherapy | CPH | ctDNA aberrations | Baseline |
| [50] | Pancancer | Anti-EGFR or anti-ERBB2 | KM | | |
| [84] | Hepatocellular carcinoma | Atezolizumab + bevacizumab | KM – CPH | | |
| [15] | NSCLC | Atezolizumab or docetaxel | | | |
| [85] | Pancancer | ICI | | | |
| [86] | Colorectal | Surgery or chemotherapy | | | |

| | | | | | |
|---|---|---|---|---|---|
| [87] | Melanoma | (Pembrolizumab or nivolumab) ± ipilimumab | KM – CPH – LR | | |
| [14] | Melanoma | Ipilimumab | Descriptive statistics | | Longitudinal |
| [88] | Pancancer | ICI | KM | | |
| [89, 90, 91, 92] | NSCLC | / / / | KM – CPH | | |
| [13] | Urothelial carcinoma | Atezolizumab | | | |
| [93] | Melanoma | Nivolumab ± ipilimumab | | | |
| [94] | Pancancer | Durvalumab ± tremelimumab | | | |
| [95] | NSCLC | ICI | KM – CPH – Bayesian probit model | | |
| [96] | Pancancer | Pembrolizumab | KM – CPH – LR | | |
| [61] | NSCLC / Melanoma | ICI ± cobimetinib | NLME | | |
| [60] | NSCLC | Atezolizumab or docetaxel | LASSO linear model | | |
| [62] | NSCLC | Erlotinib / gefitinib | Mechanistic modeling / NLME | | |
| [63] | Colorectal cancer | Cetuximab | Mechanistic modeling | | |
| [56] | NSCLC | (Atezolizumab ± bevacizumab) + carboplatin + paclitaxel | EN linear regression | cfDNA concentration / ctDNA mutations | |
| [37] | Renal | (Ipilimumab + nivolumab) or anti-VEGFR-TKIs | KM – CPH – LR | cfDNA concentration | Baseline |

| | | | | | |
|---|---|---|---|---|---|
| [97] | NSCLC | TKI or pembrolizumab-CT or CT | KM – CPH | | Longitudinal |
| [65] | 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.

## Declaration of Interest statement

None declared.

# SUPPLEMENT

### *SUPERVISED MACHINE LEARNING (ML) ALGORITHMS*

ML encompasses multiple approaches, from linear models to highly nonlinear ones (Figure 2A) [98]. Logistic regression (LR) predicts the probability of a binary outcome, seeking a linear relationship between the log-odds of an event and variables. Support vector machines (SVM) find the hyperplane that maximally separates individuals into classes. Decision trees are nonlinear methods that create a series of interconnected binary choices, enabling regression and classification. Random forests (RF) are an example of ensemble methods that use multiple decision trees trained on different features' and patients' subsets[99]. A second example is XGBoost[100], which constructs a new tree at each iteration to anticipate the prediction errors and integrates regularization to avoid overfitting. The final prediction is the combination of each tree's prediction. Finally, deep learning, a subset of ML involving large neural networks, is suited for complex features patterns and relationships. It consists of interconnected artificial neurons evaluating a weighted sum of inputs and passing the results to the next-layer neurons through nonlinear activation functions.

### *CLASSICAL SURVIVAL ANALYSIS*

Survival analysis (Figure 2B) models time-to-event data, such as progression or death, and specifically accounts for censored data (unreached event). The main methods include univariable/multivariable Cox proportional-hazards regression (CPHR)[101].

In the fragmentome field, Lapin et al. demonstrated, using multivariable CPHR in advanced pancreatic cancer, that higher cfDNA levels were associated with smaller PFS (Hazard Ratio (HR): 3.05, 95% CI: 1.40-6.65) and OS (HR: 2.24, 95% CI: 1.09-4.59), independently from clinical variables, carbohydrate antigen and cfDNA fragment size[102].

Moding et al.[89] monitored ctDNA molecular residual disease in advanced NSCLC patients, collecting ctDNA immediately after chemoradiation therapy (CRT). A second ctDNA sample was collected early during immune-checkpoint inhibition (ICI) treatment for the immunotherapy-treated arm. Undetectable ctDNA before ICI treatment correlated with good prognosis, irrespective of ICI treatment. Detectable ctDNA early during ICI therapy also correlated with shorter PFS. Analyzing ctDNA changes over

time demonstrated that increased ctDNA concentration were associated with worse prognosis compared with decreased levels.

Powles et al.[13] used CPHR to study atezolizumab response in urothelial carcinoma, collecting plasma at baseline and early on treatment (first day of the first and third cycles). They found significant disease-free survival differences according to ctDNA changes over time. Patients with ctDNA clearance had three to four times lower relapse risks, according to univariable, stratified and multivariable CPHRs.

## 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. 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

5. 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

6. 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

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

8. 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

9. 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

10. Shirley M. Epi proColon® for Colorectal Cancer Screening: A Profile of Its Use in the USA. *Mol Diagn Ther*. 2020;24(4):497-503. doi:10.1007/s40291-020-00473-8

11. 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

12. Heinrich MC, Jones RL, George S, et al. Ripretinib versus sunitinib in gastrointestinal stromal tumor: ctDNA biomarker analysis of the phase 3 INTRIGUE trial. *Nat Med*. 2024;30(2):498-506. doi:10.1038/s41591-023-02734-5

13. 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

14. 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

15. 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

16. 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

17. 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

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. Zheng H, Zhu MS, Liu Y. FinaleDB: a browser and database of cell-free DNA fragmentation patterns. *Bioinformatics*. 2021;37(16):2502-2503. doi:10.1093/bioinformatics/btaa999

21. Yu F, Li K, Li S, et al. CFEA: a cell-free epigenome atlas in human diseases. *Nucleic Acids Research*. 2020;48(D1):D40-D44. doi:10.1093/nar/gkz715

22. Gahan PB, Schwarzenbach H, Anker P. The History and Future of Basic and Translational Cell-Free DNA Research at a Glance. *Diagnostics (Basel)*. 2022;12(5):1192. doi:10.3390/diagnostics12051192

23. 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.

24. 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

25. 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

26. 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.

27. 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.

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

29. 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

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. 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

37. 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

38. 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

39. 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

40. 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

41. 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

42. 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

43. 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

44. 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

45. 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.

46. 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

47. 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

48. 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

49. 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

50. 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

51. 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

52. 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

53. 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

54. Yang X, Cai GX, Han BW, et al. Association between the nucleosome footprint of plasma DNA and neoadjuvant chemotherapy response for breast cancer. *npj Breast Cancer*. 2021;7(1):1-12. doi:10.1038/s41523-021-00237-5

55. Panagopoulou M, Karaglani M, Balgkouranidou I, et al. Circulating cell-free DNA in breast cancer: size profiling, levels, and methylation patterns lead to prognostic and predictive classifiers. *Oncogene*. 2019;38(18):3387-3401. doi:10.1038/s41388-018-0660-y

56. 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

57. Guan L, Tibshirani R. Post model-fitting exploration via a “Next-Door” analysis. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*. 2020;48(3):447-470.

58. Ding H, Xu XS, Yang Y, Yuan M. Improving Prediction of Survival and Progression in Metastatic Non–Small Cell Lung Cancer After Immunotherapy Through Machine Learning of Circulating Tumor DNA. *JCO Precis Oncol*. 2024;(8):e2300718. doi:10.1200/PO.23.00718

59. Ramsay JO, Silverman BW, eds. Principal components analysis for functional data. In: *Functional Data Analysis*. Springer; 2005:147-172. doi:10.1007/0-387-22751-2_8

60. 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

61. 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

62. 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

63. 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

64. Li A, Lou E, Leder K, Foo J. Early ctDNA kinetics as a dynamic biomarker of cancer treatment response. Published online July 3, 2024:2024.07.01.601508. doi:10.1101/2024.07.01.601508

65. 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

66. 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

67. 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

68. Pei XM, Yeung MHY, Wong ANN, et al. Targeted Sequencing Approach and Its Clinical Applications for the Molecular Diagnosis of Human Diseases. *Cells*. 2023;12(3):493. doi:10.3390/cells12030493

69. Greytak SR, Engel KB, Parpart-Li S, et al. Harmonizing cell-free DNA Collection and Processing Practices through Evidence-based Guidance. *Clin Cancer Res*. 2020;26(13):3104-3109. doi:10.1158/1078-0432.CCR-19-3015

70. Haselmann V, Ahmad-Nejad P, Geilenkeuser WJ, et al. Results of the first external quality assessment scheme (EQA) for isolation and analysis of circulating

tumour DNA (ctDNA). *Clin Chem Lab Med*. 2018;56(2):220-228. doi:10.1515/cclm-2017-0283

71. Thierry AR. Circulating DNA fragmentomics and cancer screening. *Cell Genomics*. 2023;3(1):100242. doi:10.1016/j.xgen.2022.100242

72. 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

73. Boutonnet A, Pradines A, Mano M, et al. Size and Concentration of Cell-Free DNA Measured Directly from Blood Plasma, without Prior DNA Extraction. *Anal Chem*. 2023;95(24):9263-9270. doi:10.1021/acs.analchem.3c00998

74. Lundberg S, Lee SI. A Unified Approach to Interpreting Model Predictions. *arXiv*. Published online 2017. doi:10.48550/arxiv.1705.07874

75. Campos-Carrillo A, Weitzel JN, Sahoo P, et al. Circulating tumor DNA as an early cancer detection tool. *Pharmacol Ther*. Published online December 18, 2019:107458. doi:10.1016/j.pharmthera.2019.107458

76. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. 2019;25(1):44.

77. 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

78. 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

79. 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

80. 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

81. 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

82. 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

83. 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

84. 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

85. 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

86. 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

87. 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

88. 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

89. 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

90. 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

91. 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

92. 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

93. 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

94. 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

95. 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

96. 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

97. 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

98. Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An introduction to statistical learning. Published online 2013. Accessed August 21, 2024. https://biblio.cerist.dz/index.php/hrbdonf5214/ouvrages/00000000621048000001_2.pdf

99. Breiman L. Random Forests. *Machine Learning*. 2001;45(1):5-32. doi:10.1023/A:1010933404324

100. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. Association for Computing Machinery; 2016:785-794. doi:10.1145/2939672.2939785

101. Frank E. Harrell. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Springer International Publishing; 2015. doi:10.1007/978-3-319-19425-7

102. 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