OXFORD

Briefings in Bioinformatics, 2024, **25(2)**, 1–16

https://doi.org/10.1093/bib/bbae028

Review

New techniques to identify the tissue of origin for cancer of unknown primary in the era of precision medicine: progress and challenges

Wenyuan Ma, Hui Wu, Yiran Chen, Hongxia Xu, Junjie Jiang, Bang Du, Mingyu Wan, Xiaolu Ma, Xiaoyu Chen, Lili Lin, Xinhui Su, Xuanwen Bao, Yifei Shen, Nong Xu, Jian Ruan, Haiping Jiang and Yongfeng Ding

Corresponding authors. Yongfeng Ding, Department of Medical Oncology, the First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou 310006, China. Tel.: +86-0571-87235841. E-mail: dingyongfeng@zju.edu.cn; Haiping Jiang, Department of Medical Oncology, the First Affiliated Hospital, Zhejiang University School of Medicine, Hangzhou 310006, China. Tel.: +86-0571-87235841. E-mail: jianghaiping@zju.edu.cn

Abstract

Despite a standardized diagnostic examination, cancer of unknown primary (CUP) is a rare metastatic malignancy with an unidentified tissue of origin (TOO). Patients diagnosed with CUP are typically treated with empiric chemotherapy, although their prognosis is worse than those with metastatic cancer of a known origin. TOO identification of CUP has been employed in precision medicine, and subsequent site-specific therapy is clinically helpful. For example, molecular profiling, including genomic profiling, gene expression profiling, epigenetics and proteins, has facilitated TOO identification. Moreover, machine learning has improved identification accuracy, and non-invasive methods, such as liquid biopsy and image omics, are gaining momentum. However, the heterogeneity in prediction accuracy, sample requirements and technical fundamentals among the various techniques is noteworthy. Accordingly, we systematically reviewed the development and limitations of novel TOO identification methods, compared their pros and cons and assessed their potential clinical usefulness. Our study may help patients shift from empirical to customized care and improve their prognoses.

Keywords: cancer of unknown primary (CUP); tissue of origin (TOO); gene expression profiling; precision medicine; liquid biopsy; image omics

Wenyuan Ma is a student in the Department of Medical Oncology, The First Affiliated Hospital, Zhejiang University School of Medicine. His research interests include data integration analysis and cancer of unknown primary.
Hui Wu is a student of the Department of Medical Oncology, The First Affiliated Hospital, Zhejiang University School of Medicine. Her research interests include gastric cancer and immunotherapy.
Yiran Chen is a research fellow of the Department of Surgical Oncology, The First Affiliated Hospital, Zhejiang University School of Medicine. Her research interests include bioinformatics and phenotyping and treatment of gastric cancer.
Hongxia Xu is a research fellow of the Zhejiang University-University of Edinburgh Institute (ZJU-UoE Institute), Zhejiang University School of Medicine. Her research interests are in the field of bioinformatics and cancer-related research.
Junjie Jiang is a physician in the Department of Gastroenterology, Affiliated Hangzhou First People's Hospital, Zhejiang University School of Medicine. His research interests include bioinformatics and gastrointestinal neoplasms.
Bang Du is a researcher at the Real Doctor AI Research Center, School of Medicine, Zhejiang University. Her research interests include medical and artificial intelligence applications.
Mingyu Wan is a student of the Department of Medical Oncology, The First Affiliated Hospital, School of Medicine, Zhejiang University. Her research interests include bioinformatics and cancer-related research.
Xiaolu Ma is a student in the Department of Medical Oncology, The First Affiliated Hospital, Zhejiang University School of Medicine. Her research interests include bioinformatics and cancer-related research.
Xiaoyu Chen is a student of the Department of Medical Oncology, The First Affiliated Hospital, Zhejiang University School of Medicine. Her research interests include machine learning with applications in bioinformatics.
Lili Lin is a doctor of the Department of Nuclear Medicine, The First Affiliated Hospital, Zhejiang University School of Medicine. Her research interests include the development and biological application of molecular imaging probes.
Xinhui Su is a professor in the Department of Nuclear Medicine, The First Affiliated Hospital, Zhejiang University School of Medicine. His laboratory focuses on the basic and clinical study of PET/CT imaging and the synthesis.
Xuanwen Bao is a researcher in the Department of Medical Oncology, The First Affiliated Hospital, Zhejiang University School of Medicine. His laboratory focuses on exploring tumor microenvironment and tumor immune heterogeneity through computational biology and bioinformatics.
Yifei Shen is a researcher in the Department of Laboratory Medicine, The First Affiliated Hospital, Zhejiang University School of Medicine. His laboratory focuses on the use of machine learning and data mining techniques for early screening and prognostic assessment of diseases using multidimensional omics data.
Nong Xu is a professor in the Department of Medical Oncology, The First Affiliated Hospital, Zhejiang University School of Medicine. His laboratory focuses on the mechanism and medical treatment of gastrointestinal tumor and lung cancer.
Jian Ruan is a professor in the Department of Medical Oncology, The First Affiliated Hospital, Zhejiang University School of Medicine. His laboratory focuses on the mechanism and treatment of lung cancer.
Haiping Jiang is a professor in the Department of Medical Oncology, The First Affiliated Hospital, Zhejiang University School of Medicine. Her laboratory focuses on bioinformatics, therapeutic and molecular mechanism research of gastric cancer.
Yongfeng Ding is a professor in the Department of Medical Oncology, The First Affiliated Hospital, Zhejiang University School of Medicine. His laboratory focuses on image omics, therapeutic and molecular mechanism exploration of gastric cancer and CUP.
Received: July 16, 2023. **Revised:** December 10, 2023. **Accepted:** January 11, 2024

INTRODUCTION

Despite standardized diagnostic workups, cancers of unknown primary (CUPs) comprise a heterogeneous collection of metastatic tumors with unknown primary tumors [1]. Epidemiologically, CUPs are estimated to account for 2–5% of all diagnosed tumors worldwide [2–4]. Indeed, most patients (80–90%) with CUPs fall into an unfavorable subset, with median overall survival (OS) durations ranging from 3 to 11 months and a 1-year OS of 25–40% [5–7]. Clinically aggressive, early-spreading and unpredictable metastases define these tumors [2, 4, 8]. Due to the lack of standard treatment, most CUP patients receive empirical chemotherapy, including platinum–taxane regimens [1, 9].

Patients diagnosed with CUP have a worse prognosis than those with metastatic cancer of known origin [10, 11]. This suggests that tissue of origin (TOO) identification and subsequent site-specific therapy for CUP may enhance survival and prognosis. Data from various clinical trials suggest the possibility of this hypothesis [12–15]. Varadhachary *et al.* [14] were the pioneers in demonstrating, using a small sample of patients with CUP associated with a colon-cancer profile (CCP-CUP), and found that patients with CCP-CUP derive substantial benefits from using specific treatments developed for colon cancer. Similarly, Hainsworth *et al.*'s [15] assay-directed site-specific therapy yielded a median survival time of 12.5 months, surpassing outcomes associated with empiric CUP regimens. Furthermore, regarding whether patients with CUP benefit from site-specific therapy, Ding *et al.* [16] conducted a meta-analysis and concluded that identifying the TOO and administration of site-directed therapy is effective, specifically for CUP patients with responsive tumor types. Accordingly, identifying the TOO of CUP is critical for optimizing and pinpointing treatment. The meta-analysis by Ding *et al.* [16] also showed that improving the accuracy of TOO identification could significantly improve patient prognosis. From a psychological point of view, CUP patients have more psychological perplexity and stress than those with recognized TOO [17–21], which concerns both clinicians and patients. Therefore, it is of great clinical significance to develop TOO identification techniques for patients diagnosed with CUP.

A comprehensive diagnostic approach for CUP typically includes physical examination, medical history review, hematology assessment, endoscopy, imaging studies and pathological analysis, as depicted in Figure 1 [1, 2, 7]. However, these methods may not consistently identify the TOO in all suspected CUP patients. For example, conventional imaging techniques exhibit a TOO detection rate of only 20–27%, while positron emission tomography (PET) improves this rate slightly to 37% [22, 23]. Immunohistochemical analysis, although necessary, is a labor-intensive and often inefficient method for TOO identification in malignant tumors, with a notable 27% of cases remaining undetermined. The concordance rate between pathological and clinical diagnoses was only 59% [24, 25]. Notably, multiple immunostainings consume a limited amount of the tumor tissue. The limitations of traditional clinical methods underscore the need for innovative and effective TOO identification techniques.

Identifying TOO in metastatic tumors is the cornerstone of clinical work in oncology. However, a notable challenge emerges in a small subset of cases, particularly within poorly differentiated carcinomas and squamous cell carcinomas, where the diagnostic process is significantly hampered by the absence of specific site-specific immunohistochemical markers [24]. New TOO detection techniques for CUP are being proposed and validated as technology develops (Figure 2). On the one hand, the prediction accuracy, sample requirements and technical principles of the different identification techniques show significant diversity. Contrarily, a comprehensive evaluation comparing the merits, demerits and future prospects of these methods is yet to be undertaken. Thus, we systematically assessed emerging TOO detection methods for CUP to determine their clinical utility.

MAIN TEXT

Methods of literature search and criteria for article selection

A systematic literature search was performed using PubMed, Web of Science, Embase, Cochrane Library and ClinicalTrials.gov from 1 January 2000 to 1 May 2023, with English language restrictions. Conference abstracts from the American Society of Clinical Oncology (ASCO) and European Society of Medical Oncology (ESMO) meetings were also included. Search terms were as follows: [(cancer* OR carcinom* OR neoplas* OR malignan*) AND ('unknown primary' OR 'occult primary' OR 'primary metastatic') AND (origin* OR type) AND (trace* OR infer* OR classif* OR identif* OR predict*) AND (accuracy OR sensitivity OR specificity)]. Only studies conducted on CUP patients were included. The primary lesion was identified using non-routine clinical diagnostic methods, and the research was limited only to human model. Case reports, editorials and commentaries were excluded (Supplementary Figure 1).

A total of 14 369 potentially eligible studies were initially identified from the systematic literature search, as shown in Supplementary Figure 1. After removing the duplicates from the different databases ($n = 7249$), irrelevant studies ($n = 7018$) were excluded by title and abstract screening. A total of 102 studies were assessed for eligibility. Eight articles were excluded because they did not introduce the technique to identify TOO, 16 articles were excluded because they did not focus on patients with CUP, 20 articles were excluded because accuracy data were not available and 15 articles were excluded because their sample size was less than 30 cases. Due to the limited number of comparative studies available, one comparative study conducted by Chen *et al.* was still included in our research, despite its small sample size (Supplementary Figure 1).

New techniques of identifying TOO of CUP

Based on genomic profiling

The TOO of CUP can be identified at the deoxyribonucleic acid (DNA) level. DNA copy number variations (CNVs) [26, 27], somatic and germline mutation [28–30], expression quantitative trait loci (eQTL) [31] and single-nucleotide polymorphisms (SNPs) [32, 33] have been used to identify TOO in tumor tissues (Table 1).

CNV, a genetic marker of the genome, is a variation of DNA fragments ranging in size from 1 kb to 3 Mb. CNVs are critical in affecting gene function through gene dosage, breakage, fusion and position effects and have a strong association with tumors [34, 35]. The machine learning (ML) model utilized genomic data to identify TOO of CUP. One noteworthy tool in this context is CNAOrigin, developed by Liang *et al.*, which harnesses a convolutional neural network (CNN) model. After subjecting the model to rigorous 10-fold cross-validations, CNAOrigin demonstrated an impressive predictive accuracy of 83.81% on internal datasets and 79% on independent datasets [26]. Another useful DNA-level method for TOO identification is eQTL, which explains the association between SNPs and gene expression levels. A recent study by Miao *et al.* explicitly integrated eQTL into the eXtreme Gradient Boosting (XGBoost) classification model. This integration yielded a remarkable prediction accuracy of over 96% in 10-fold

Table 1: Comparison of approaches used in the identification of the site of origin in CUP

	Year published	Authors	Sample size	Source	Type	Materials	NT	Accuracy	External	NE	CUP	AE	Prospective	Therapy	Time	Details of the method
Genomic profiling																
1	2013	Felix Dietlein *et al.* [32]	905	TCGA, CCLE	FF, FFPE	DNA	23	79.00%	Yes	431	No	71.00%	No	Yes	NA	Single-nucleotide polymorphism (SNP) + second generation sequencing
2	2015	Andrea Marion Marquard *et al.* [30]	2820	COSMIC	NA	DNA	28	87.60%	Yes	993	No	85.00%	No	No	NA	Somatic mutation, copy number profiles + random forest (RF)
3	2020	Bingsheng He *et al.* [33]	4909	ICGC	FFPE	DNA	13	88.22%	No				No	No	NA	SNP + RF
4	2020	Ying Liang *et al.* [26]	3480	TCGA	FFPE	DNA	6	83.31%	Yes	NA	No	79.00%	No	No	NA	Copy number alteration (CNA) + RF
5	2020	Xiaojun Liu *et al.* [29]	3374	ICGC	FFPE	DNA	13	81.00%	No				No	No	NA	Somatic mutation + RF
6	2020	Yulin Zhang *et al.* [27]	4566	TCGA	FFPE	DNA	10	89.13%	Yes	1262	No	74.21%	No	No	NA	Copy number variations (CNVs) + eXtreme Gradient Boosting (XGBoost)
7	2022	Yongchang Miao1 *et al.* [31]	Over 7000	TCGA	NA	DNA	20	96.00%	Yes	87	No	90.00%	No	No	NA	Expression quantitative trait loci (eQTL) + XGBoost
8	2022	Luan Nguyen *et al.* [28]	6756	PCAWG	FFPE	DNA	35	90.00%	Yes	141	No	58.00%	No	No	NA	Genome-wide mutation features + RF
Gene expression profiling (GEP)																
9	2005	Richard W. Tothill *et al.* [42]	229	TCGA	FFPE	RNA	14	89.00%	Yes	13	Yes	84.60%	No	No	NA	Quantitative PCR, microarray + support vector machine (SVM)
10	2007	Gauri R. Varadhachary *et al.* [37]	104	Self-collected cases	FFPE	RNA	13	61.00%	No				Yes	Yes	NA	10 genes Quantitative reverse-transcription PCR (qRT-PCR)
11	2008	Nitzan Rosenfeld *et al.* [68]	253	Self-collected cases	FF, FFPE	RNA	22	>90.00%	Yes	83	No	89.00%	No	No	NA	MicroRNA array
12	2009	Agendia BV *et al.* [43]	633	Self-collected cases	FF, FFPE	RNA	30	81.00%	Yes	229	No	82.00%	Yes	No	NA	Microarray
13	2010	Rosetta Genomics Ltd *et al.* [63]	356	Self-collected cases	FFPE	RNA	25	90.00%	Yes	204	No	85.00%	Yes	No	NA	MicroRNA array + qRT-PCR
14	2013	Rolf Søkilde *et al.* [65]	208	Self-collected cases	FFPE	RNA	15	85.00%	Yes	48	No	88.00%	Yes	No	NA	MicroRNA array + microarray profiling
15	2013	John D. Hainsworth *et al.* [15]	252	Self-collected cases	FFPE	RNA	NA	98.00%	No				Yes	Yes	2 to 3 weeks	CancerTYPE ID (92-gene qRT-PCR)
16	2015	Richard W. Tothill *et al.* [48]	450	Self-collected cases	FFPE	RNA	18	82.00%	Yes	58	No	78.00%	Yes	No	NA	Microarray
17	2017	Yuanyuan Li *et al.* [59]	9096	TCGA	FFPE	RNA	31	95.60%	No				No	No	NA	RNA sequencing
18	2017	Marcos Tadeu dos Santos *et al.* [41]	4429	Self-collected cases	FFPE	RNA	25	86.60%	Yes	102	Yes	83.80%	No	No	2 weeks	95-gene qRT-PCR
19	2019	Mengyao Li *et al.* [60]	1007	GEO	FFPE	RNA	1	99.36%	Yes	24	Yes	100.00%	Yes	Yes	NA	Relative gene expression orderings (REOs)
20	2020	Qing Ye *et al.* [39]	609	Self-collected cases	FFPE	RNA	21	89.80%	Yes	141	No	71.60%	Yes	No	NA	90-gene qRT-PCR
21	2020	Yue Zhao *et al.* [56]	18,217	TCGA, ICGC	FFPE	RNA	32	98.54%	Yes	486	Yes	96.70%	No	No	NA	RNA sequencing +1D-Inception
22	2021	Sijie Chen *et al.* [45]	5759	TCGA	NA	RNA	21	96.38%	Yes	42	No	83.30%	No	No	NA	Microarray + XGBoost
23	2021	Noemi Laprovitera *et al.* [64]	159	Self-collected cases	FFPE	RNA	17	95.00%	Yes	53	No	90.00%	Yes	No	2 days	MicroRNA array + prediction analysis of microarrays, the least absolute shrinkage and selection operator
24	2021	Julien Vibert *et al.* [54]	20,918	TCGA, GTEX, HPA	FF	RNA	39	96.26%	Yes	48	Yes	79.00%	Yes	Yes	NA	RNA sequencing + RF, K-nearest neighbors (KNN)
25	2021	Ruixi Li *et al.* [52]	7713	TCGA	FFPE	RNA	20	96.10%	Yes	79	No	83.50%	No	No	NA	RNA sequencing + gradient boosting (GBDT)
26	2021	Yifei Shen *et al.* [53]	10,553	TCGA	FFPE	RNA	24	93.00%	No				Yes	No	NA	RNA sequencing + rank-based majority vote algorithm

(Continued)

Table 1: Continued

	Year published	Authors	Sample size	Source	Type	Materials	NT	Accuracy	External	NE	CUP	AE	Prospective	Therapy	Time	Details of the method
27	2022	James Hong *et al.* [55]	1528	TCGA	FFPE	RNA	14	97.00%	No				No	No	NA	RNA sequencing + deep learning
28	2022	Qingfeng Lu *et al.* [44]	5708	TCGA	FFPE	RNA	15	96.90%	No				No	No	NA	Microarray + XGBoost
29	2022	Wei Sun *et al.* [40]	1417	Self-collected cases	FFPE	RNA	21	94.40%	No				Yes	No	NA	90-gene qRT-PCR
30	2022	Weiqin Jiang *et al.* [58]	17	TCGA, self-collected cases	FFPE	RNA	4	94.40%	Yes	7	Yes	100.00%	No	No	NA	RNA sequencing + naive Bayes algorithm
31	2023	Jackson Michuda *et al.* [57]	52,936	TCGA	FFPE	RNA	68	91.10%	Yes	1708	No	80.00%	No	No	NA	RNA sequencing + multinomial logistic regression classifier with L2 regularization
Epigenetics																
32	2016	Sebastian Moran *et al.* [12]	7691	TCGA	FFPE	DNA	38	99.60%	Yes	216	No	87.00%	Yes	Yes	NA	Microarray DNA methylation signatures
33	2022	Agustin F. Fernandez *et al.* [72]	42	Self-collected cases	NA	DNA	19	69.00%	No				No	No	NA	DNA methylation fingerprint—1505 CpG sites
34	2023	Ze Zhang *et al.* [73]	7735	TCGA, GEO	FFPE	DNA	30	99.00%	Yes	1775	Yes	93.00%	No	No	NA	DNA methylation + multilayer perceptron
Multi-omics																
35	2016	Wei Tang *et al.* [96]	Over 5000	TCGA	FFPE	RNA	14	87.78%	No				No	No	NA	miRNA array expression
						DNA		96.43%					No	No	NA	DNA methylation profiles + maximum relevance maximum distance
						DNA		97.06%					No	No	NA	DNA methylation profiles + principal component analysis
36	2020	Binsheng He *et al.* [101]	7008	TCGA, GEO	FFPE	RNA	20	89.28%	Yes	19	No	NA	No	No	NA	Gene expression + multi-classifier RF
						DNA		53.51%				NA	No	No	NA	Gene mutation + multi-classifier RF
						RNA + DNA		95.77%				89.47%	No	No	NA	Gene expression + gene mutation + multi-classifier RF
37	2021	Haiyan Liu *et al.* [102]	7244	TCGA	FFPE	RNA	21	94.63%	No				No	No	NA	Gene expression
						DNA		43.33%					No	No	NA	Somatic mutation
						DNA		86.13%					No	No	NA	DNA methylation
						RNA + DNA		94.02%					No	No	NA	DNA methylation + gene expression + somatic mutation
38	2022	Kaiyan Chen *et al.* [103]	16	Self-collected cases	FFPE	DNA	1	43.30%	No				No	No	NA	Genomic profiling
						DNA		100.00%					No	No	NA	DNA methylation
						DNA		100.00%					No	No	NA	Genomic profiling + DNA methylation
Liquid biopsy																
39	2015	Ronald Lebofsky *et al.* [76]	34	Self-collected cases	Blood samples	DNA	18	97.00%	No				Yes	No	NA	Cell-free tumor DNA
40	2015	Myron G. Best *et al.* [80]	283	Self-collected cases	Blood samples	RNA	6	71.00%	No				Yes	No	NA	RNA sequencing of tumor-educated platelets
41	2020	Ayuko Hoshino *et al.* [79]	34	Self-collected cases	FFPE	Protein	4	100.00%	Yes	9	No	100.00%	No	No	NA	Extracellular vesicle (EV) and particle Biomarkers
			37	Self-collected cases	Blood samples	Protein	5	100.00%	Yes	10	No	100.00%	No	No	NA	EV and particle biomarkers
Others																
42	2022	Enrico Moiso *et al.* [94]	11,744	TCGA, MOCA	FFPE	RNA	33	97.40%	Yes	52	Yes	95.00%	Yes	Yes	NA	Developmental deconvolution
43	2021	Ming Y. Lu *et al.* [95]	32,537	Self-collected cases	Whole-slide images	Images	18	83.00%	Yes	317	Yes	61.00%	No	No	NA	Pathomics + CNN encoder, attention -based multiple-instance learning algorithm

Predictive accuracy was generally determined by comparison with designations made using gold-standard clinicopathological criteria. Accuracy, accuracy of internal data; AE, accuracy of external data; CUP, presence or absence of CUP in the external data; External, presence or absence of validation of external data; Materials, materials used in the study; Source, source of selected samples; Type, type of tissue; NT, number of tumor types; NE, presence or absence of number of external data; Prospective, whether prospective or not; Therapy, whether site-specific therapy received or not; Time, time required to identify primary; Year, year of publication. CCLE, Cancer Cell Line Encyclopedia; CNN, convolutional neural network; COSMIC, the Catalogue of Somatic Mutations in Cancer; CUP, cancer of unknown primary site; FF, fresh frozen; FFPE, formalin-fixed paraffin-embedded; GEO, Gene Expression Omnibus; GTEx, Genotype-Tissue Expression; HPA, Human Protein Atlas; ICGC, International Cancer Genome Consortium; miRNA, microRNA; MOCA, Mouse Organogenesis Cell Atlas; NA, not available; PCAWG, Pan-Cancer Analysis of Whole Genomes; PCR, polymerase chain reaction; TCGA, The Cancer Genome Atlas; TOO, tissue of origin.

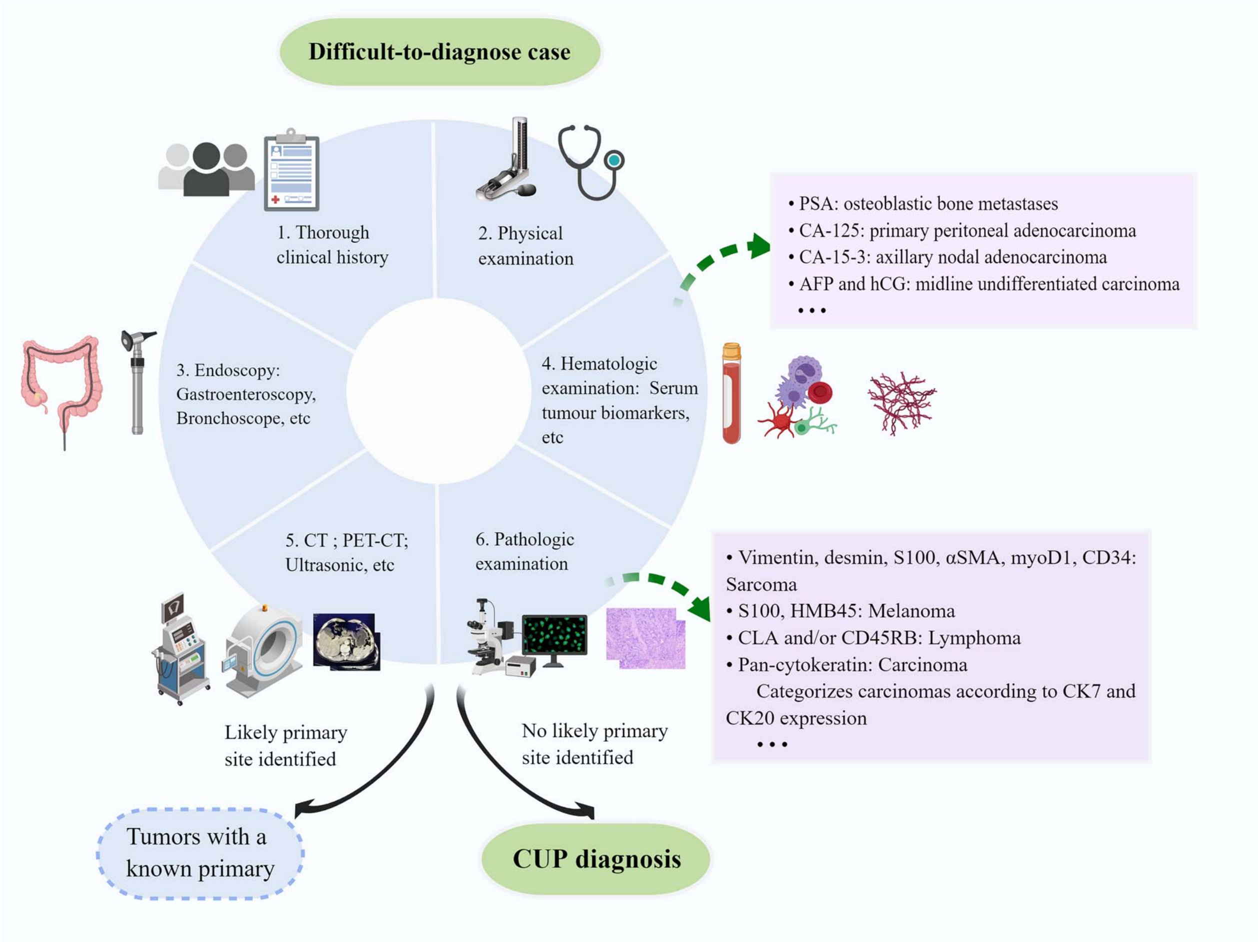

Figure 1. Diagnostic methods recommended for the anticipatory diagnosis of CUP patients. The clinical evaluation of CUP begins with a thorough tumor history, family history and physical examination. This is followed by analysis, including a basic hematologic examination; CT and PET scans of the chest, abdomen and pelvis; and determination of tumor biomarkers. Endoscopy, like gastrointestinal endoscopy and laryngoscopy, can not only visualize the location of the tumor but also provide the tissue samples needed for pathological examination, and various immunohistochemical combinations can play a role in identifying the tumor category. If the location of the primary tumor cannot be determined, the diagnosis of CUP remains. αSMA, α-smooth muscle actin; AFP, α-fetoprotein; CA, cancer antigen; CD, cluster of differentiation; CK, cytokeratin; CLA, cutaneous lymphocyte-associated antigen; CT, computed tomography; CUP, cancer of unknown primary; hCG, human chorionic gonadotropin; HMB45, human melanoma black 45; myoD1, myoblast determination protein 1; PET, positron emission tomography; PSA, prostate-specific antigen; S100, calcium-binding protein G. Potential cancer type designation is determined by marker positivity unless otherwise noted (Figure 1 belongs to the Introduction section).

cross-validation using The Cancer Genome Atlas (TCGA) data [31]. Combining multiple DNA-level methods may help detect the TOO of CUP more accurately. For instance, Marquard *et al.* [30] performed a comprehensive study in which an approach using only point mutation had an accuracy of only 69% and an approach that integrated point mutation and CNV significantly improved accuracy to 85%.

Based on gene expression profiling

Metastatic tumors may retain gene expression patterns from cell-type-specific tumors [36]. Therefore, gene expression profiling (GEP) is vital for TOO detection. Currently, several GEP methods (Table 1), such as reverse transcription-polymerase chain reaction (RT-PCR) [15, 37–41], microarrays [42–51], second-generation sequencing of ribonucleic acid (RNA) [38, 52–59] and relative gene expression orderings (REOs) [60], are available to aid in the search for the TOO of CUP.

RT-PCR was one of the pioneering methods used for discrimination and is still utilized today. Several commercial platforms, such as CancerType ID (a 92-gene RT-PCR-based cancer classifier), have been developed. However, RT-PCR is limited compared to microarrays and second-generation sequencing. Microarrays and second-generation sequencing offer the advantage of identifying a broader spectrum of tumor types, assessing older samples with preservation periods extending up to a decade and facilitating targeted therapies [15, 39]. Our research group extracted approximately 1000 signature molecules from the TCGA and RNA sequencing of clinical samples from our institution to create the Bayes algorithm for tissue origin diagnosis (TOD-Bayes algorithm) to diagnose the TOO of hepatobiliary pancreatic malignancies [58]. The accuracy rate of our internal data exceeded 95%, and the external validation corroborated an accuracy rate of 94.4% [58]. Sample REO is stable, which minimizes the impact of experimental batch, data conversion, RNA degradation and tumor tissue sampling site randomization [60, 61]. For example, Li *et al.* used five gene pairs as markers to predict the TOO of metastatic colorectal cancer (CRC), achieving accuracy rates of 99.36% and 100% for internal and external data, respectively [60].

MicroRNA (miRNA) is a non-coding family of 22-nucleotide single-stranded RNA molecules encoded by endogenous genes [62]. MiRNAs are persistent and resistant to ribonuclease (RNase) degradation in compromised clinical samples, making miRNA

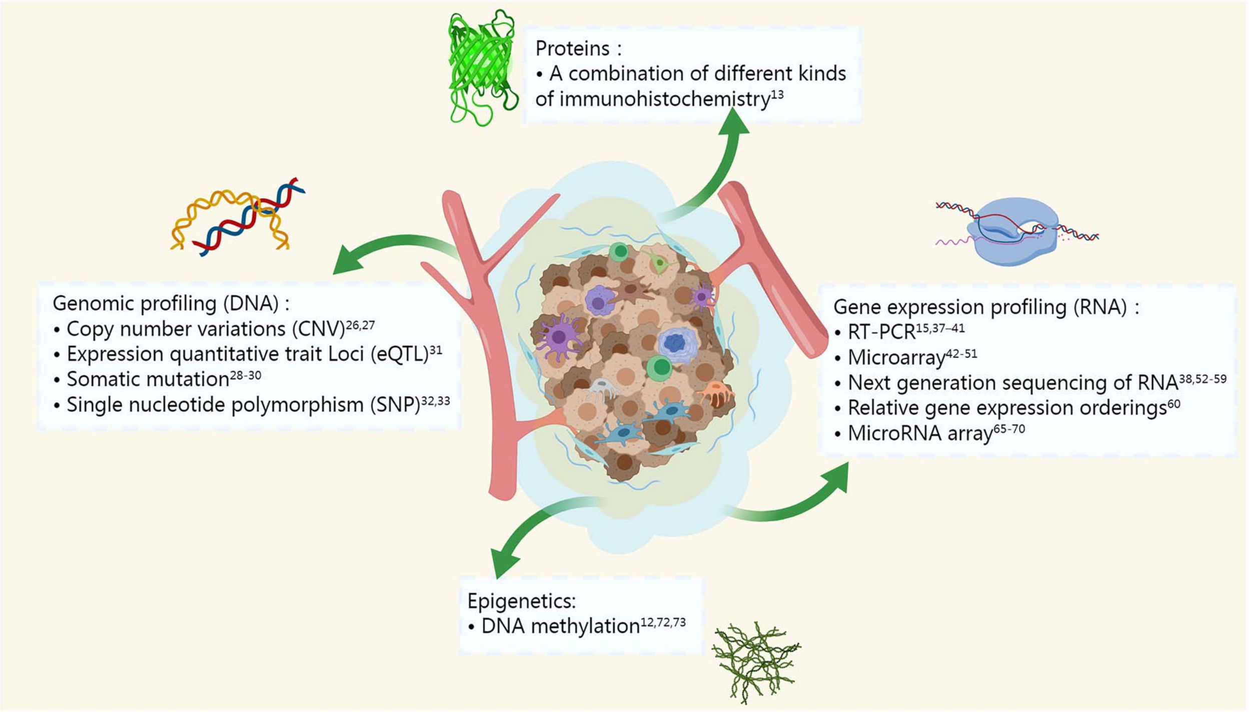

Figure 2. New techniques developed at different molecular levels for the detection of TOO in CUP are emerging. CUP, cancer of unknown primary; RT-PCR, reverse transcription-polymerase chain reaction; TOO, tissue of origin (Figure 2 belongs to the New techniques of identifying TOO of CUP section).

array a reliable TOO detection technique [63–68]. Laprovitera and colleagues used 89 miRNAs to deduce the TOO of CUPs. The miRNA expression was evaluated in 159 samples using digital droplet PCR and the least absolute shrinkage and selection operator (LASSO) model combined with the predictive analysis of microarrays (PAMR) nearest shrinkage center of mass method. This integrated approach yielded an internal data accuracy of 95% and increased OS in CUP patients [64]. This study highlights the potential utility of miRNA array in identifying the TOO of CUP.

Based on epigenetics

Epigenetics mechanisms, including DNA methylation, histone modification and chromosomal remodeling, regulate gene expression independently of changes in the DNA sequence [69, 70]. Studies have shown that CUP is characterized by a substantial overall loss of DNA methylation, resulting in a decrease in 5-methylcytosine levels ranging from 20% to 60%, making DNA methylation an ideal biomarker for identifying the TOO of CUP [71–73]. Recent research has used improved DNA methylation platforms to detect 10 481 tumor samples with 99.6% specificity and 97.7% sensitivity using approximately 450 000 CpG sites in the human genome [12].

Based on proteins

Several proteomic methods are available, including tandem mass tagging/isobaric tags for relative and absolute quantification (TMT/iTRAQ) and data-independent acquisition/sequential window acquisition of all theoretical fragment ions (DIA/SWATH) [24]. Nonetheless, no research currently employs rigorous proteomic techniques to identify the TOO of CUP. Hasegawa *et al.* conducted a retrospective analysis of 90 patients with an unfavorable subset of CUP using a combination of immunohistochemical markers to identify TOO. Fifty-six patients (62.2%) with predicted TOO using this technique received site-specific therapy and had a median OS of 20.3 months, a significant improvement in survival compared to 10.7 months for patients receiving empiric chemotherapy [13]. Although this may not meet the criteria of a 'stringent' proteomic technique, it nonetheless underscores the considerable prospective utility of proteomics.

Based on liquid biopsy

Liquid biopsy, a non-invasive methodology, has the potential to revolutionize the diagnosis, treatment and prognosis of CUP [74, 75]. Key biomarkers employed in this approach encompass circulating tumor cells (CTCs), circulating tumor DNA (ctDNA) [76–78], extracellular vesicles (EVs) [79], peripheral blood circulating RNA and tumor-educated platelets (TEPs) [80] (Table 1, Figure 3).

CTCs, originating from the primary tumor and circulating within the bloodstream [81], contrast with ctDNA, which comprises DNA fragments shed by cancer cells through apoptosis or necrosis [82]. CTCs and ctDNA can reveal cancer genetic and phenotypic traits and predict TOO [77, 82–85]. Lebofsky and colleagues reported a remarkable 97% concordance between ctDNA analysis and the accurate identification of TOO across 34 patients encompassing 18 distinct tumor types [76]. Nonetheless, it is important to note that the effectiveness may be somewhat limited in detecting early-stage tumors or in older patients [86] because of the diverse metastatic nature of the tumor [74, 75, 87].

EVs refer to a heterogeneous population of small, membrane-bound vesicles found in various body fluids, which contain diverse biomolecules [88]. These EVs can be categorized into subgroups such as exosomes, endosomes, microbubbles and apoptotic bodies are EV subgroups, distinguished by their size and morphological characteristics [88–91]. Hoshino *et al.* employed EV protein patterns derived from tumor tissue and plasma to differentiate melanoma, colorectal, pancreatic and lung cancers, with a 100% accuracy rate. Moreover, the study showed that the specificity of the EV-based diagnostic method remained consistent across

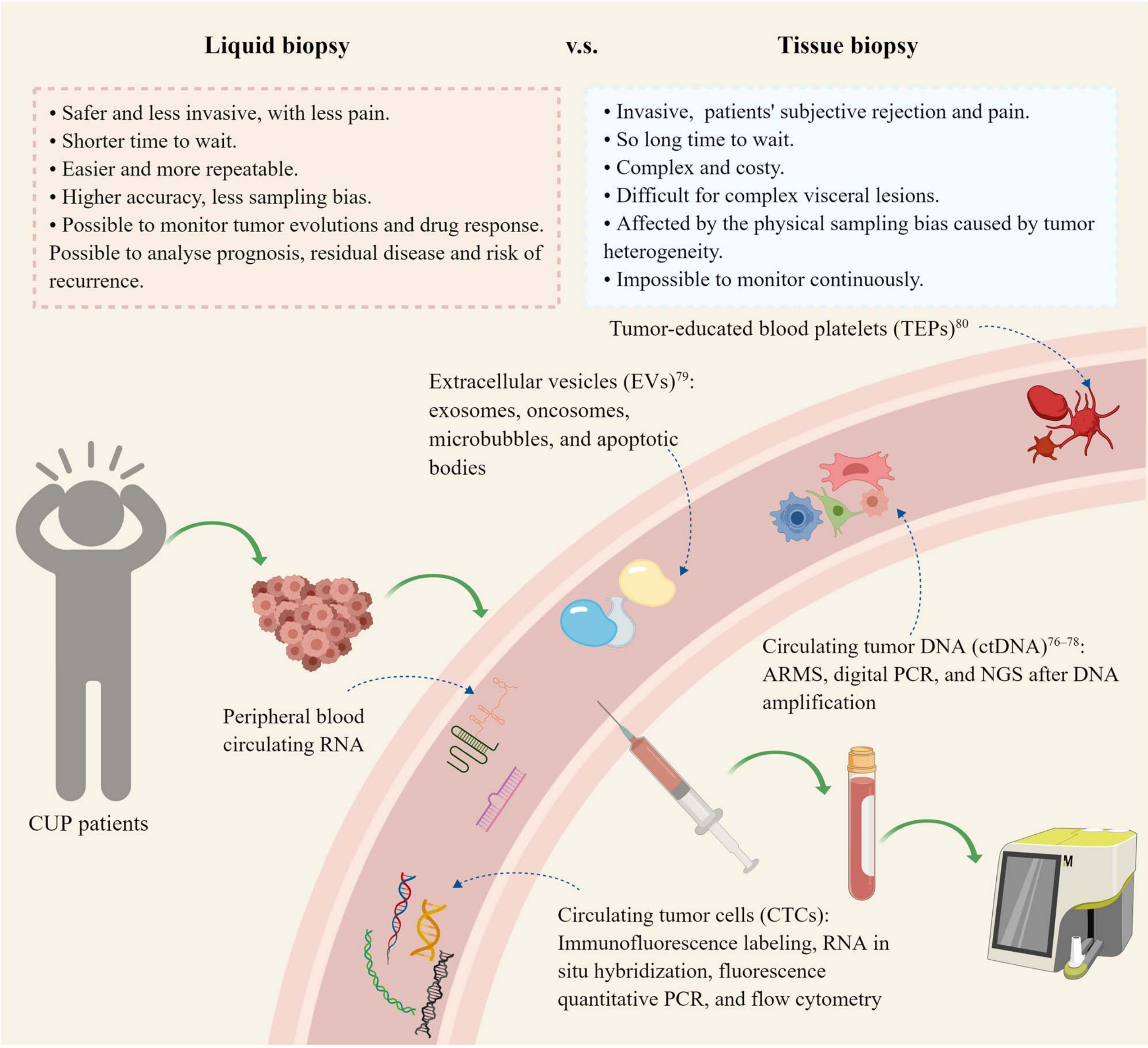

Figure 3. Comparison between liquid biopsy and tissue biopsy for the detection of TOO in CUP. ARMS, amplification refractory mutation system; CUP, cancer of unknown primary; NGS, next-generation sequencing; PCR, polymerase chain reaction; TOO, tissue of origin (Figure 3 belongs to the Based on liquid biopsy section).

different stages of cancer and could even detect cancers in their early stages [79].

TEPs are important in the systemic and local responses to tumor growth, thereby altering their RNA profile. Best *et al.* determined the diagnostic potential of TEPs by mRNA sequencing of 283 platelet samples. The TOO was accurately identified in 71% of cases across six different tumor types [80]. These findings suggest that blood platelets are a valuable platform for detecting TOO of CUP.

Circulating RNA in the peripheral blood has the potential to aid in the diagnosis and treatment of CUP [92], although its utilization in research remains limited. In a recent study by Yao *et al.*, the authors demonstrated the potential of this approach by analyzing miRNA profiles from plasma samples obtained from individuals with gastric cancer and non-cancer patients using two independent gene expression synthesis datasets. Three miRNAs-hsa-miR-320a, 1260b and 6515-5p have demonstrated exceptional specificity in distinguishing primary gastric tumors [93]. However, further research is needed to determine the efficacy of this method for CUP patients.

Based on other techniques

In addition to the above techniques for TOO detection, tumor developmental atlases and image omics show considerable potential. Moiso's team has constructed a comprehensive human tumor development atlas by analyzing and comparing single-cell data from TCGA tumor samples with the Mouse Organogenesis Cell Atlas (MOCA). The atlas aims to establish correlations between cancer biology and development. The team used a developmental multilayer perceptron (D-MLP) classifier constructed from this atlas, which showed remarkable accuracy of 0.974 in identifying TOO [94]. Image omics could also determine the TOO of CUP. Lu *et al.* developed an artificial intelligence–based (AI-based) pathology training model capable of simultaneously predicting the metastatic status and identifying the origin of 18 different tumor types. On the known primary tumor test set, the model demonstrated outstanding performance, achieving a maximum level 1 accuracy of 0.83 and a level 3 accuracy of 0.96. On the external test set, it also achieved the highest levels 1 and 3 accuracy of 0.80 and 0.93, respectively [95].

ML for identifying TOO of CUP

The basic flowchart of building a TOO classifier with an ML algorithm

The process of developing a classifier for TOO utilizing an ML algorithm entails the following steps (Figure 4): initially, the training set must be created by obtaining multimodal data either from public databases like TCGA and International Cancer Genome Consortium (ICGC) or through self-collection of the data. The data collected may include genomic profiling, gene expression profiling and proteins data from tumor tissue, CTCs and EVs data from plasma and CT images or pathological images. Using genomic profiling data as an example, bioinformatics and ML algorithms are applied to score and rank the most relevant genes for creating tumor–gene associations and constructing TOO classifiers. Several ML algorithms to identify the TOO of CUP have been applied in this context [28, 29, 33, 42, 44, 45, 52–54, 57, 58, 64, 73, 96] (Supplementary Figure 3 and Table 2). These associations are subsequently assessed through independent validation sets, and the classifier's efficacy is further verified with challenging clinical cases. Finally, the classifier can calculate the 'tissue origin score' when applied to CUP patients and then choose the tissue source with the highest score for site-specific therapy [95, 97, 98].

The algorithm underlying these experimental techniques for identifying TOO of CUP

Our literature review indicates a current concentration on applying supervised learning algorithms [27, 29, 30, 42, 54], with limited exploration of unsupervised learning methods [96]. As representatives of supervised learning algorithms, the **r**andom **f**orest (RF) model and the XGBoost model are frequently applied algorithms for identifying TOO of CUP [29–31]. The algorithms possess high accuracy, incorporating strategies for handling missing feature data and thus provide an advantage in processing DNA**-** and RNA-related information (Supplementary Figure 3). Among the reviewed studies, only one article utilized the **p**rincipal **c**omponent **a**nalysis (PCA) algorithm within the unsupervised learning realm [96]. Notably, there is a conspicuous absence of discussions regarding the application of reinforcement learning algorithms, highlighting a research gap at the algorithmic level.

The feature extraction process is a crucial preliminary step in model construction, involving selecting a subset of the most relevant features from the original set. Different types of data can adopt different types of feature selection strategies. For text data, simple statistical methods like the Pearson correlation algorithm employed by Zhang *et al.* [27] and Hoshino *et al.* [79] can filter features. Yet, complex gene interactions challenge traditional methods assuming feature independence. Many studies have employed decision tree models (such as RF) to address feature selection [28, 30, 33]. In addition, Laprovitera *et al.* [64] used the LASSO algorithm, and Jiang *et al.* [58] used correlation-based feature selection (CFS), considering the correlation between the target variable and features. Tang *et al.* [96] proposed a two-tier feature selection strategy, with the first tier based on miRNA differential expression and DNA differential methylation analysis and the second tier mainly employing mathematical algorithms like the PCA algorithm. Traditional ML algorithms, which often rely on a limited set of genes or characteristics, may be constrained in their capacity to discern numerous cancer subtypes. To overcome this limitation, deep learning (DL) algorithms that use various image features to achieve higher accuracy rates have been introduced [99]. For image data, the CNN is a practical feature extraction method, with Lu *et al.* [95] segmenting images on this basis and extracting local feature descriptors to learn essential features in the images.

Selecting the appropriate algorithm poses a challenge due to significant variations in their advantages, limitations and application areas (Supplementary Figure 3 and Table 2). Using DL algorithmic models is a necessary approach when working with image data. The most elementary of these models is the CNN model. DL models, including Transformer and ResNet, can be employed depending on the objective, such as image detection, classification or segmentation [97]. For textual or sequential data, such as DNA, RNA and proteins, employing XGBoost and lightGBM classification models can produce the desired outcomes.

DISCUSSION

Progress over the past two decades

Over the past two decades, methods for TOO identification have changed drastically (Figure 5). Firstly, the broadening of research perspectives: whereas in earlier years, researchers focused on traditional DNA and mRNA levels, the focus has shifted to the novel, multifunctional analytes such as non-coding RNAs [63–68], proteins [13], epigenetic markers [12, 71–73], ctDNA [76–78] and EVs [79]. Another hallmark is the merging of multiple perspectives and unique insights [96, 99–101]. Some studies including eQTL [31] and REO [60] now analyze multiple genes simultaneously, increasing dimensionality of research. Secondly, the expansion of research tools: with the rapid changes in relevant technologies, the tools used by researchers have changed from PCR to second-generation sequencing [32, 52–56, 59] and tumor developmental atlas [94], thus achieving greater efficiency and accuracy. Thirdly, the expansion of materials used: research has expanded beyond traditional tumor tissue. Liquid biopsy techniques have enabled the shift toward plasma samples [76, 79, 80], whereas image omics have also empowered pathomics to discover the TOO of CUP [95]. Fourthly, expanding the scope of research: anticipated tumor diversity is expanding, and researchers are entering previously unreachable areas (Supplementary Figure 2A). Fifthly, advances in accuracy: accuracy rates have increased from an average of about 80% to nearly 100% in 20 years (Supplementary Figure 2B and C). This progress is, in part, attributed to the proliferation of ML, fostering the growth of bioinformatics (Supplementary Figure 2D) and enabling the analysis of extensive biological datasets, holding significant promise [102–104].

Challenges of the current studies on experiential methods

We identified some challenges in the current study by summarizing all the studies. The following issues may need to be addressed to improve TOO detection in CUP prognosis. Many techniques used to identify TOO are highly accurate, but whether this 'digital' accuracy can be translated into clinical benefits remains to be discussed. Despite a predictive rate of 78.6% for TOO, site-specific therapy based on microarray analysis did not significantly enhance 1-year survival, according to a study by Hidetoshi Hayashi and colleagues [51].

Compared to empirical chemotherapy, identifying TOO and pursuing organ-specific treatment will inevitably result in some

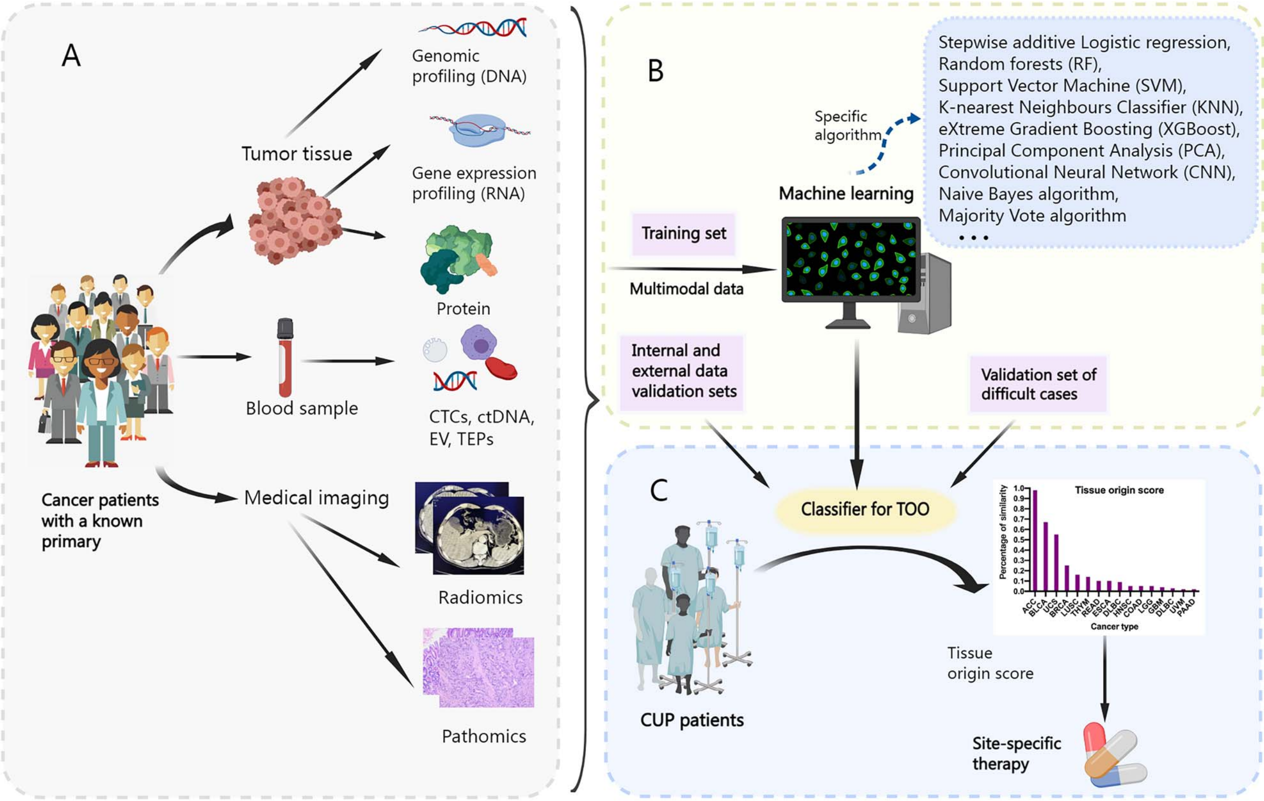

Figure 4. The basic flowchart of building a TOO classifier with an ML algorithm. (**A**) Multimodal data used to identify the TOO of CUP. (**B**) The ML algorithm for identifying TOO of CUP. (**C**) The Construction and Application of TOO Classifier. CTCs, circulating tumor cells; ctDNA, circulating tumor DNA; CUP, cancer of unknown primary; EVs, extracellular vesicles; ICGC, International Cancer Genome Consortium; TCGA, The Cancer Genome Atlas; TEPs, tumor-educated platelets; TOO, tissue of origin (Figure 4 belongs to the The basic flowchart of building a TOO classifier with an ML algorithm section).

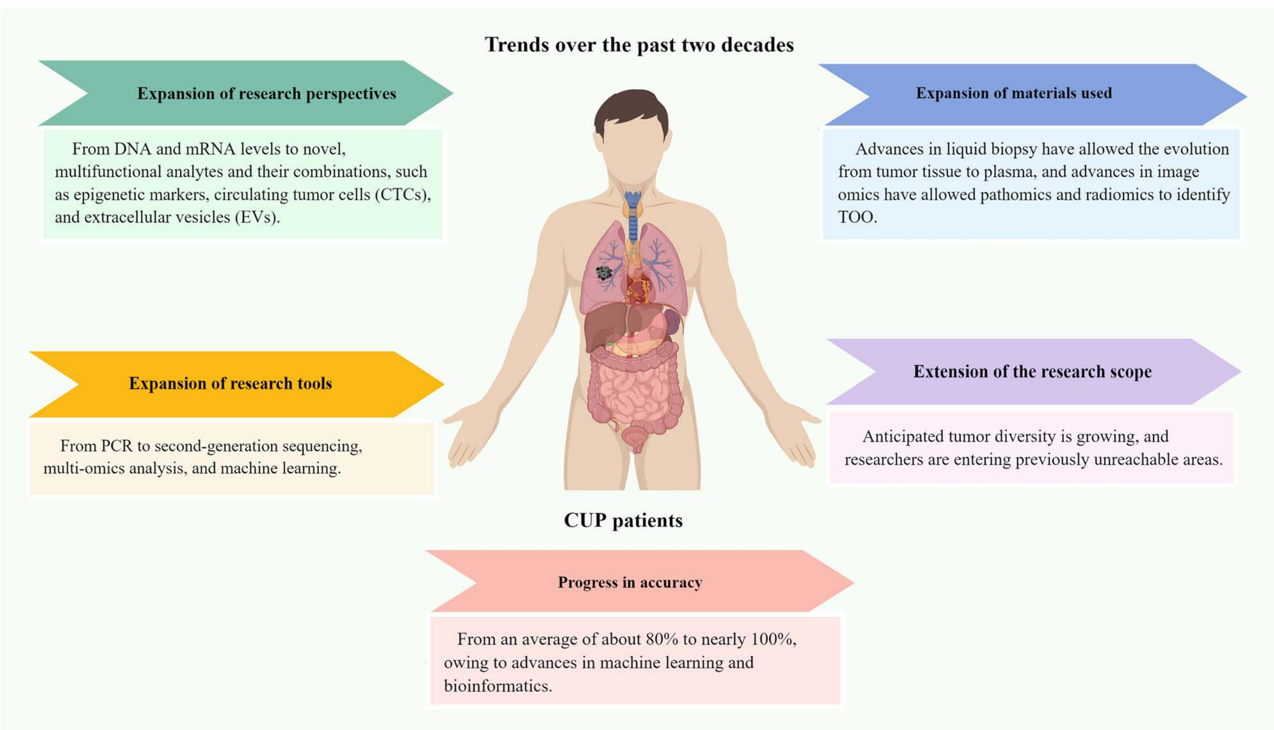

Figure 5. Progress for the detection of TOO in CUP over the past two decades. CUP, cancer of unknown primary; PCR, polymerase chain reaction; TOO, tissue of origin (Figure 5 belongs to the Discussion section).

Table 2: Comparison of ML algorithms used in the identification of the TOO in CUP

Machine learning algorithms [111]	Strength	Weaknesses	Range of application
Stepwise additive logistic regression [130]	1. Automatic feature selection: It can automatically select features that have significant predictive power for the response variable, simplifying the model.	1. Possibility for suboptimal models: Stepwise selection or elimination of features may not always find the best model.	1. Suitable for classification problems, especially when interpretability is crucial.
	2. Model interpretability: By reducing unnecessary features, a more interpretable model can be obtained.	2. Instability in selection: Small changes in the data can result in significant variations in the selected feature set.	2. Applicable when the dataset has numerous features.
	3. Control overfitting.	3. High computational cost of iterative calculations.	3. Can serve as a preliminary feature selection stage in a multifaceted modeling procedure.
Random forest (RF) [131, 132]	1. High predictive accuracy.	1. Lack of interpretability.	1. Suitable for both classification and regression problems.
	2. Robust to overfitting: RF is less prone to overfitting due to the ensemble averaging effect.	2. Computational complexity: It can be computationally expensive.	2. Applicable to datasets with a mixture of numerical and categorical features.
	3. Feature importance estimation: RF can provide information about the relative importance of different features.	3. Bias toward features with more categories: Potentially leading to biased feature importance rankings.	3. Not suitable when interpretability is a primary concern or when computational resources are limited.
	4. Robust to outliers and missing data.		
Support vector machine (SVM) [133, 134]	1. Effective in high-dimensional spaces.	1. Computationally intensive and time-consuming.	1. Effective in complex classification problems with datasets exhibiting complex distributions or clear boundaries.
	2. Robust to overfitting: SVM uses a regularization parameter to control overfitting.	2. Requires feature scaling: SVM is sensitive to the scale of input features and often requires feature standardization.	2. Effective when dealing with various features, especially when the number of features exceeds samples.
	3. Versatility in kernel selection: Different kernel functions can handle non-linear relationships between features.	3. Lack of interpretability.	
	4. Effective in small sample sizes.	4. No direct probability estimation: Computing the probability of instances belonging to a class requires additional steps.	
K-nearest neighbors classifier (KNN) [135]	1. Simplicity: Easy to understand and implement.	1. Computational complexity.	1. Small dataset scenarios.
	2. Non-parametric: KNN makes no assumptions about the underlying data distribution.	2. Sensitivity to feature scaling.	2. Non-linear dataset scenarios.
	3. No training phase: New data points can be classified immediately.	3. Lack of robustness to noisy data.	3. Choosing an appropriate distance metric is crucial for accurate classification.
	4. Interpretable results: KNN provides a transparent decision-making process.	4. Boundedness of dimensionality: KNN performance deteriorates as the number of dimensions increases, due to the sparsity of data in high-dimensional spaces.	
eXtreme Gradient Boosting (XGBoost) [136]	1. High performance.	1. Complexity: It requires careful tuning of hyperparameters to achieve optimal performance.	1. Large dataset scenarios, various classification and regression problems.
	2. Flexibility: It can handle various types of data, including numerical and categorical features.	2. Computationally expensive.	2. More suitable for structured data tasks.
	3. Handling missing values: It can reduce the need for extensive preprocessing.	3. Lack of interpretability.	
	4. Cross-validation: Allow cross-validation to easily obtain the optimal number of boosting iterations.		

(Continued)

Table 2: Continued

Machine learning algorithms [111]	Strength	Weaknesses	Range of application
	1. Dimensionality reduction.	1. Loss of interpretability.	1. It can be used for data dimensionality reduction, visualization and preprocessing.
Principal component analysis (PCA) [137]	2. Feature extraction: PCA can extract a smaller set of features (principal components) that capture the maximum variance in the data.	2. Assumption of linearity: If the underlying data have complex non-linear relationships, PCA may not capture the most important features accurately.	2. It captures the maximum variance by searching for the principal components of the data, thereby simplifying the data structure.
	3. Noise reduction: PCA can help remove noise or irrelevant features from the dataset by focusing on the components with the highest variance.	3. Sensitive to outliers.	
	4. Data visualization: PCA can be used to visualize high-dimensional data in lower-dimensional spaces.	4. Boundedness of dimensionality: Its performance deteriorates as the number of dimensions increases.	
Naive Bayes algorithm [138]	1. Simplicity and efficiency.	1. Strong independence assumption: In real-world scenarios, features may have dependencies, leading to suboptimal performance.	Naive Bayes is a simple probability-based classifier that is particularly suitable for high-dimensional data and text classification tasks.
	2. Scalability: It performs well with a small amount of training data, real-time or streaming data.	2. Limited expressiveness: Due to its simplicity, it may struggle with capturing complex relationships.	
	3. Robust to irrelevant features: It assumes that features are conditionally independent given the class label, making it robust to irrelevant features and helping avoid overfitting.	3. Data scarcity issue: When data are scarce, it may result in poor performance and unreliable predictions.	
	4. Interpretability.	4. Sensitive to feature distributions.	
Majority vote algorithm [53]	1. Simplicity.	1. Increase computational burden.	The algorithm is an integrated technology that can combine the prediction results of multiple models to achieve better performance and stability.
	2. Reduction of bias: By combining the predictions of multiple classifiers, it can improve the overall accuracy and robustness of the ensemble.	2. Not always providing improvements: If some of the models have poor performance, integration may not provide any benefits.	
	3. Stability: It is less sensitive to small changes in the training data.	3. Not applicable to all problems.	
	4. Interpretability.	4. Limited decision boundaries: It may struggle to capture complex or non-linear relationships.	
Convolutional neural network (CNN) [139, 140]	1. Effective feature extraction.	1. Computationally expensive.	1. Suitable for image-related image classification, object detection and semantic segmentation tasks.
	2. Spatial invariance: CNNs are able to capture spatial relationships in data, making them robust to translations, rotations and scale variations.	2. Large memory footprint.	2. Suitable for scenarios with other spatially structured data and a large number of training samples.
	3. Parameter sharing: CNNs utilize weight sharing across different spatial locations, reducing the number of parameters.	3. Limited interpretability.	
		4. Data requirements: CNNs typically require a large amount of labeled training data to generalize well.	

time delay. John D. Hainsworth *et al.* found that TOO takes 2–3 weeks to identify [15], which may not be feasible for CUP patients with short OS. In this regard, reducing the time delay is critical, and the time spent searching for TOO should be a vital criterion for evaluating the technology's efficacy. However, despite its clinical importance [7, 105], few studies have reported the TOO identification time [15, 64]. This requires researchers' attention and effort in the future.

Moreover, in clinical practice, tumor tissue from patients with CUP is limited (coarse needle aspiration/biopsy specimens) and can only meet the needs of routine immunohistochemistry in the clinic. Conducting TOO-related tests may require a second tissue biopsy due to insufficient samples, a procedure patients often avoid due to its inherent risks [74, 75, 87, 106, 107]. This emphasizes the need for non-intrusive methods.

Notably, there are few comparative studies on different techniques [96, 99–101, 108]. Atara Posner *et al.* used DNA features to identify TOO in 51 out of 61 CUP cases, with GEP proving useful in only 21 cases. Their study concluded that DNA mutation analysis outperformed GEP in TOO identification [108]. The authors also found that GEP had lower classification accuracy for cholangiocarcinoma because its transcriptional profile resembles that of pancreatic or upper gastrointestinal tract tumors [56]. On the other hand, DNA mutation profiling is particularly useful because some gene mutations (alterations in *IDH1*, *FGFR2* and *BAP1*) are highly enriched and have diagnostic significance [108, 109]. While Wei Tang *et al.* applied miRNA expression and DNA methylation profiles to identify the TOO, the overall accuracy was 87.78% based on the miRNA dataset and 97.06% based on DNA methylation [96]. Haiyan Liu *et al.* discovered that DNA methylation, GEP and somatic mutation data were best classified by GEP (mean accuracy 94.63%) and worst classified by somatic mutation (mean accuracy 43.33%) [100]. However, no consensus has been regarding the superiority of the different techniques.

Challenges of the current studies on computational methods

In addition to experimental technical obstacles, computational methods have significant pitfalls.

The first statistical challenge, known as the 'curse of dimensionality', is a common concern among bioinformatics experts [110]. This term refers to the overfitting problem caused by the excessive number of features, resulting in poor predictive performance on validation sets [111]. Due to the high dimensionality of omics data, the problem of dimensional curse is particularly prominent. Moreover, Chen *et al.* [110] pointed out the presence of feature redundancy or lack of correlation, introducing noise in high-dimensional space and making it more challenging for models to extract genuine signals. One solution discussed in the paper by Michuda *et al.* [57] is using regularization indicators to penalize prediction models with too many features, thus supporting simpler models with a relatively minor set of features. Simultaneously, it is necessary to divide the dataset into three subsets: the training, model selection and test set. The model selection set is to identify models with optimal generalization performance. However, new challenges arise, such as the current high cost of next-generation sequencing for liquid biopsy samples, leading to datasets often insufficient for three subsets [76, 79, 80, 90].

Though promising, developing models that integrate prior biological knowledge (e.g. known gene regulatory pathways for specific types of tumors) has not been extensively explored due to the limited availability of such prior knowledge [110]. Selecting the most appropriate features from a multitude of features may also alleviate the issue of the curse of dimensionality. Despite these efforts, organically selecting features from multimodal data and enhancing the interpretability of selected algorithms remain significant challenges in the future.

Future perspective

Different research methods are complex, and each has its own advantages and disadvantages. However, with improvement of medical quality, simple, applicable and accurate research may be the future direction [112, 113]. Liquid biopsy, a safer, cost-effective and less invasive alternative, has emerged as a novel diagnostic, predictive and prognostic window for CUP. Blood is widely believed to be a reservoir for tumor cells *in vivo*. Thus, liquid biopsy can potentially reduce the sampling bias of tissue biopsy and ultimately provide greater predictive accuracy. Liquid biopsy evaluates prognosis, disease load, risk of recurrence, therapeutic alternatives and dynamic mutational processes [74, 87, 114, 115]. Although the concordance between tissue and liquid biopsy in CUP patients has not been fully evaluated [116–119], the performance of liquid biopsy is a promising direction for predicting TOO in CUP (Figure 3). Besides the above-mentioned materials for CUP, future studies can be conducted on peripheral blood circulating RNA [93, 120] and circulating tumor vascular endothelial cells (cTECs) [121].

Lu *et al.* demonstrated the enormous benefit of pathomics in identifying TOO of CUP [95], and another potential area is radiomics [122–125]. Conventional tumor evaluation through radiography relies primarily on qualitative features, also known as 'semantic' features, like tumor density, enhancement pattern, intratumoral cellular and acellular composition, regularity of tumor margins and anatomical relationships with surrounding tissues [102, 126]. Radiomics allows radiographic images to be quantified according to their shape, size and texture patterns [103, 127, 128]. In cases of extremely high accuracy and integration of multiple data sources, CT- or PET-based imaging may be desirable.

As shown above, comparing various strategies is difficult owing to the significant variability of the tumor types selected in each study, the diverse model development methods and the limited data of the selected samples. Perhaps studies are also needed to compare the ability of different biomarkers under the same conditions, including the same dataset, preprocessing scheme and classification algorithm.

Multi-omics is still intriguing as sequencing costs decrease and technology advances, but its potential to enhance prediction accuracy requires further investigation. Haiyan Liu *et al.* downloaded GEP, somatic mutation and DNA methylation data of 7224 samples from TCGA and generated seven different feature matrices through various combinations. They found that the best accuracy was 94.63% for the single method and 94.02% after combination, revealing that simply combining multiple biomarkers did not do much to improve prediction accuracy [100]. In contrast, He *et al.* employed the RF model and integrated gene mutations and expression (TOOme) to infer tumor TOO, which differs from Liu. Their approach yielded higher accuracy (95.77%) compared to using somatic mutations (53.51%) or GEP data alone (89.28%) [99]. While these findings indicate potential, it's clear that simply stacking multi-omics data is insufficient [129]. A more integrated approach using ML models is likely necessary. Accordingly, further studies are required to determine which omics approaches work best and how to combine them to predict TOO.

CONCLUSIONS

In the era of precision medicine, the endless stream of new technologies has led to rapid advances in TOO identification of CUP: accuracy rates are increasing by leaps and bounds; molecular profiling, including techniques based on genomic profiling, gene expression profiling and epigenetics, is flourishing; and ML is rising. Liquid biopsy and image omics enable non-invasive methods for TOO detection. However, it remains to be confirmed whether the current technological advances have improved patient prognosis. Large-scale clinical studies, multi-institutional collaborations and a unified standard database may need more work.

Key Points

- Tissue of origin (TOO) identification for cancer of unknown primary and subsequent site-specific therapy can improve prognosis of CUP.
- Of these, techniques to identify the TOO are the critical part.
- We systematically review the development and limitations of novel TOO identification methods, compare their pros and cons and assess their potential clinical usefulness in the future.

SUPPLEMENTARY DATA

Supplementary data are available online at https://academic.oup.com/bib.

FUNDING

This study is kindly supported by the Natural Science Foundation of Zhejiang (LQ20H160043); Clinical Research Fund Project of Zhejiang Medical Association (2019ZYC-A83); and National Natural Science Foundation of China (82303950).

DATA AVAILABILITY

The data underlying this article are available in the article and in its online supplementary material.

REFERENCES

1. Fizazi K, Greco FA, Pavlidis N, *et al.* Cancers of unknown primary site: ESMO clinical practice guidelines for diagnosis, treatment and follow-up. *Ann Oncol* 2015;**26**(Suppl 5):v133–8.
2. Lee MS, Sanoff HK. Cancer of unknown primary. *BMJ* 2020;**371**:m4050.
3. Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. *CA Cancer J Clin* 2023;**73**:17–48.
4. Pavlidis N, Pentheroudakis G. Cancer of unknown primary site. *Lancet Lond Engl* 2012;**379**:1428–35.
5. Huebner G, Link H, Kohne CH, *et al.* Paclitaxel and carboplatin vs gemcitabine and vinorelbine in patients with adeno- or undifferentiated carcinoma of unknown primary: a randomised prospective phase II trial. *Br J Cancer* 2009;**100**:44–9.
6. Hess KR, Abbruzzese MC, Lenzi R, *et al.* Classification and regression tree analysis of 1000 consecutive patients with unknown primary carcinoma. *Clin Cancer Res* 1999;**5**:3403–10.
7. Rassy E, Pavlidis N. Progress in refining the clinical management of cancer of unknown primary in the molecular era. *Nat Rev Clin Oncol* 2020;**17**:541–54.
8. Varghese AM, Arora A, Capanu M, *et al.* Clinical and molecular characterization of patients with cancer of unknown primary in the modern era. *Ann Oncol* 2017;**28**:3015–21.
9. Kato S, Alsafar A, Walavalkar V, *et al.* Cancer of unknown primary in the molecular era. *Trends Cancer* 2021;**7**:465–77.
10. Kim CS, Hannouf MB, Sarma S, *et al.* Survival outcome differences based on treatments used and knowledge of the primary tumour site for patients with cancer of unknown and known primary in Ontario. *Curr Oncol Tor Ont* 2018;**25**:307–16.
11. Rassy E, Parent P, Lefort F, *et al.* New rising entities in cancer of unknown primary: is there a real therapeutic benefit? *Crit Rev Oncol Hematol* 2020;**147**:102882.
12. Moran S, Martínez-Cardús A, Sayols S, *et al.* Epigenetic profiling to classify cancer of unknown primary: a multicentre, retrospective analysis. *Lancet Oncol* 2016;**17**:1386–95.
13. Hasegawa H, Ando M, Yatabe Y, *et al.* Site-specific chemotherapy based on predicted primary site by pathological profile for carcinoma of unknown primary site. *Clin Oncol R Coll Radiol G B* 2018;**30**:667–73.
14. Varadhachary GR, Raber MN, Matamoros A, Abbruzzese JL. Carcinoma of unknown primary with a colon-cancer profile-changing paradigm and emerging definitions. *Lancet Oncol* 2008;**9**:596–9.
15. Hainsworth JD, Rubin MS, Spigel DR, *et al.* Molecular gene expression profiling to predict the tissue of origin and direct site-specific therapy in patients with carcinoma of unknown primary site: a prospective trial of the Sarah Cannon research institute. *J Clin Oncol* 2013;**31**:217–23.
16. Ding Y, Jiang J, Xu J, *et al.* Site-specific therapy in cancers of unknown primary site: a systematic review and meta-analysis. *ESMO Open* 2022;**7**:100407.
17. Wolyniec K, O'Callaghan C, Fisher K, *et al.* A qualitative study of patients with cancer of unknown primary: perceptions of communication, understanding of diagnosis and genomic testing, and information needs. *Psychooncology* 2023;**32**:589–96.
18. Hyphantis T, Papadimitriou I, Petrakis D, *et al.* Psychiatric manifestations, personality traits and health-related quality of life in cancer of unknown primary site. *Psychooncology* 2013;**22**:2009–15.
19. Wagland R, Bracher M, Drosdowsky A, *et al.* Differences in experiences of care between patients diagnosed with metastatic cancer of known and unknown primaries: mixed-method findings from the 2013 cancer patient experience survey in England. *BMJ Open* 2017;**7**:e017881.
20. Wolyniec K, Sharp J, Fisher K, *et al.* Psychological distress, understanding of cancer and illness uncertainty in patients with cancer of unknown primary. *Psychooncology* 2022;**31**:1869–76.
21. Richardson A, Wagland R, Foster R, *et al.* Uncertainty and anxiety in the cancer of unknown primary patient journey: a multiperspective qualitative study. *BMJ Support Palliat Care* 2015;**5**:366–72.
22. Kwee TC, Kwee RM. Combined FDG-PET/CT for the detection of unknown primary tumors: systematic review and meta-analysis. *Eur Radiol* 2009;**19**:731–44.
23. Ambrosini V, Nanni C, Rubello D, *et al.* 18F-FDG PET/CT in the assessment of carcinoma of unknown primary origin. *Radiol Med (Torino)* 2006;**111**:1146–55.
24. Selves J, Long-Mira E, Mathieu M-C, *et al.* Immunohistochemistry for diagnosis of metastatic carcinomas of unknown primary site. *Cancer* 2018;**10**:108.
25. Mokhtari M, Safavi D, Soleimani N, *et al.* Carcinoma of unknown primary origin: application of immunohistochemistry with emphasis to different cytokeratin 7 and 20 staining patterns. *Appl Immunohistochem Mol Morphol* 2022;**30**:623–34.

26. Liang Y, Wang H, Yang J, *et al.* A deep learning framework to predict tumor tissue-of-origin based on copy number alteration. *Front Bioeng Biotechnol* 2020;**8**:701.
27. Zhang Y, Feng T, Wang S, *et al.* A novel XGBoost method to identify cancer tissue-of-origin based on copy number variations. *Front Genet* 2020;**11**:585029.
28. Nguyen L, Van Hoeck A, Cuppen E. Machine learning-based tissue of origin classification for cancer of unknown primary diagnostics using genome-wide mutation features. *Nat Commun* 2022;**13**:4013.
29. Liu X, Li L, Peng L, *et al.* Predicting cancer tissue-of-origin by a machine learning method using DNA somatic mutation data. *Front Genet* 2020;**11**:674.
30. Marquard AM, Birkbak NJ, Thomas CE, *et al.* TumorTracer: a method to identify the tissue of origin from the somatic mutations of a tumor specimen. *BMC Med Genomics* 2015;**8**:58.
31. Miao Y, Zhang X, Chen S, *et al.* Identifying cancer tissue-of-origin by a novel machine learning method based on expression quantitative trait loci. *Front Oncol* 2022;**12**:946552.
32. Dietlein F, Eschner W. Inferring primary tumor sites from mutation spectra: a meta-analysis of histology-specific aberrations in cancer-derived cell lines. *Hum Mol Genet* 2014;**23**:1527–37.
33. He B, Dai C, Lang J, *et al.* A machine learning framework to trace tumor tissue-of-origin of 13 types of cancer based on DNA somatic mutation. *Biochim Biophys Acta Mol Basis Dis* 2020;**1866**:165916.
34. Poduri A, Evrony GD, Cai X, Walsh CA. Somatic mutation, genomic variation, and neurological disease. *Science* 2013;**341**:1237758.
35. Redon R, Ishikawa S, Fitch KR, *et al.* Global variation in copy number in the human genome. *Nature* 2006;**444**:444–54.
36. Dermawan JK, Rubin BP. The role of molecular profiling in the diagnosis and management of metastatic undifferentiated cancer of unknown primary: molecular profiling of metastatic cancer of unknown primary. *Semin Diagn Pathol* 2021;**38**:193–8.
37. Varadhachary GR, Talantov D, Raber MN, *et al.* Molecular profiling of carcinoma of unknown primary and correlation with clinical evaluation. *J Clin Oncol* 2008;**26**:4442–8.
38. Greco FA, Lennington WJ, Spigel DR, Hainsworth JD. Molecular profiling diagnosis in unknown primary cancer: accuracy and ability to complement standard pathology. *J Natl Cancer Inst* 2013;**105**:782–90.
39. Ye Q, Wang Q, Qi P, *et al.* Development and clinical validation of a 90-gene expression assay for identifying tumor tissue origin. *J Mol Diagn* 2020;**22**:1139–50.
40. Sun W, Wu W, Wang Q, *et al.* Clinical validation of a 90-gene expression test for tumor tissue of origin diagnosis: a large-scale multicenter study of 1417 patients. *J Transl Med* 2022;**20**:114.
41. Santos MTD, de Souza BF, Cárcano FM, *et al.* An integrated tool for determining the primary origin site of metastatic tumours. *J Clin Pathol* 2018;**71**:584–93.
42. Tothill RW, Kowalczyk A, Rischin D, *et al.* An expression-based site of origin diagnostic method designed for clinical application to cancer of unknown origin. *Cancer Res* 2005;**65**:4031–40.
43. van Laar RK, Ma X-J, de Jong D, *et al.* Implementation of a novel microarray-based diagnostic test for cancer of unknown primary. *Int J Cancer* 2009;**125**:1390–7.
44. Lu Q, Chen F, Li Q, *et al.* A machine learning method to trace cancer primary lesion using microarray-based gene expression data. *Front Oncol* 2022;**12**:832567.
45. Chen S, Zhou W, Tu J, *et al.* A novel XGBoost method to infer the primary lesion of 20 solid tumor types from gene expression data. *Front Genet* 2021;**12**:632761.
46. Monzon FA, Medeiros F, Lyons-Weiler M, Henner WD. Identification of tissue of origin in carcinoma of unknown primary with a microarray-based gene expression test. *Diagn Pathol* 2010;**5**:3.
47. Ades F, de Azambuja E, Daugaard G, *et al.* Comparison of a gene expression profiling strategy to standard clinical work-up for determination of tumour origin in cancer of unknown primary (CUP). *J Chemother* 2013;**25**:239–46.
48. Tothill RW, Shi F, Paiman L, *et al.* Development and validation of a gene expression tumour classifier for cancer of unknown primary. *Pathology (Phila)* 2015;**47**:7–12.
49. Staub E, Buhr H-J, Gröne J. Predicting the site of origin of tumors by a gene expression signature derived from normal tissues. *Oncogene* 2010;**29**:4485–92.
50. Ojala KA, Kilpinen SK, Kallioniemi OP. Classification of unknown primary tumors with a data-driven method based on a large microarray reference database. *Genome Med* 2011;**3**:63.
51. Hayashi H, Kurata T, Takiguchi Y, *et al.* Randomized phase II trial comparing site-specific treatment based on gene expression profiling with carboplatin and paclitaxel for patients with cancer of unknown primary site. *J Clin Oncol* 2019;**37**:570–9.
52. Li R, Liao B, Wang B, *et al.* Identification of tumor tissue of origin with RNA-Seq data and using gradient boosting strategy. *Biomed Res Int* 2021;**2021**:6653793.
53. Shen Y, Chu Q, Yin X, *et al.* TOD-CUP: a gene expression rank-based majority vote algorithm for tissue origin diagnosis of cancers of unknown primary. *Brief Bioinform* 2021;**22**:2106–18.
54. Vibert J, Pierron G, Benoist C, *et al.* Identification of tissue of origin and guided therapeutic applications in cancers of unknown primary using deep learning and RNA sequencing (TransCUPtomics). *J Mol Diagn* 2021;**23**:1380–92.
55. Hong J, Hachem LD, Fehlings MG. A deep learning model to classify neoplastic state and tissue origin from transcriptomic data. *Sci Rep* 2022;**12**:9669.
56. Zhao Y, Pan Z, Namburi S, *et al.* CUP-AI-dx: a tool for inferring cancer tissue of origin and molecular subtype using RNA gene-expression data and artificial intelligence. *EBioMedicine* 2020;**61**:103030.
57. Michuda J, Breschi A, Kapilivsky J, *et al.* Validation of a transcriptome-based assay for classifying cancers of unknown primary origin. *Mol Diagn Ther* 2023;**27**:499–511.
58. Jiang W, Shen Y, Ding Y, *et al.* A naive Bayes algorithm for tissue origin diagnosis (TOD-Bayes) of synchronous multifocal tumors in the hepatobiliary and pancreatic system. *Int J Cancer* 2018;**142**:357–68.
59. Li Y, Kang K, Krahn JM, *et al.* A comprehensive genomic pan-cancer classification using the cancer genome atlas gene expression data. *BMC Genomics* 2017;**18**:508.
60. Li M, Li H, Hong G, *et al.* Identifying primary site of lung-limited cancer of unknown primary based on relative gene expression orderings. *BMC Cancer* 2019;**19**:67.
61. Berry S, Pelkmans L. Mechanisms of cellular mRNA transcript homeostasis. *Trends Cell Biol* 2022;**32**:655–68.
62. He B, Zhao Z, Cai Q, *et al.* miRNA-based biomarkers, therapies, and resistance in cancer. *Int J Biol Sci* 2020;**16**:2628–47.
63. Rosenwald S, Gilad S, Benjamin S, *et al.* Validation of a microRNA-based qRT-PCR test for accurate identification of tumor tissue origin. *Mod Pathol* 2010;**23**:814–23.
64. Laprovitera N, Riefolo M, Porcellini E, *et al.* MicroRNA expression profiling with a droplet digital PCR assay enables molecular

diagnosis and prognosis of cancers of unknown primary. *Mol Oncol* 2021;**15**:2732–51.

65. Søkilde R, Vincent M, Møller AK, *et al.* Efficient identification of miRNAs for classification of tumor origin. *J Mol Diagn* 2014;**16**:106–15.
66. Ferracin M, Pedriali M, Veronese A, *et al.* MicroRNA profiling for the identification of cancers with unknown primary tissue-of-origin. *J Pathol* 2011;**225**:43–53.
67. Varadhachary GR, Spector Y, Abbruzzese JL, *et al.* Prospective gene signature study using microRNA to identify the tissue of origin in patients with carcinoma of unknown primary. *Clin Cancer Res* 2011;**17**:4063–70.
68. Rosenfeld N, Aharonov R, Meiri E, *et al.* MicroRNAs accurately identify cancer tissue origin. *Nat Biotechnol* 2008;**26**:462–9.
69. Casado-Pelaez M, Bueno-Costa A, Esteller M. Single cell cancer epigenetics. *Trends Cancer* 2022;**8**:820–38.
70. Dawson MA, Kouzarides T. Cancer epigenetics: from mechanism to therapy. *Cell* 2012;**150**:12–27.
71. Moran S, Martinez-Cardús A, Boussios S, Esteller M. Precision medicine based on epigenomics: the paradigm of carcinoma of unknown primary. *Nat Rev Clin Oncol* 2017;**14**:682–94.
72. Fernandez AF, Assenov Y, Martin-Subero JI, *et al.* A DNA methylation fingerprint of 1628 human samples. *Genome Res* 2012;**22**:407–19.
73. Zhang Z, Lu Y, Vosoughi S, *et al.* HiTAIC: hierarchical tumor artificial intelligence classifier traces tissue of origin and tumor type in primary and metastasized tumors using DNA methylation. *NAR Cancer* 2023;**5**:zcad017.
74. Ignatiadis M, Sledge GW, Jeffrey SS. Liquid biopsy enters the clinic—implementation issues and future challenges. *Nat Rev Clin Oncol* 2021;**18**:297–312.
75. Li W, Liu J-B, Hou L-K, *et al.* Liquid biopsy in lung cancer: significance in diagnostics, prediction, and treatment monitoring. *Mol Cancer* 2022;**21**:25.
76. Lebofsky R, Decraene C, Bernard V, *et al.* Circulating tumor DNA as a non-invasive substitute to metastasis biopsy for tumor genotyping and personalized medicine in a prospective trial across all tumor types. *Mol Oncol* 2015;**9**:783–90.
77. Lu S-H, Tsai W-S, Chang Y-H, *et al.* Identifying cancer origin using circulating tumor cells. *Cancer Biol Ther* 2016;**17**:430–8.
78. Laprovitera N, Salamon I, Gelsomino F, *et al.* Genetic characterization of cancer of unknown primary using liquid biopsy approaches. *Front Cell Dev Biol* 2021;**9**:666156.
79. Hoshino A, Kim HS, Bojmar L, *et al.* Extracellular vesicle and particle biomarkers define multiple human cancers. *Cell* 2020;**182**:1044–1061.e18.
80. Best MG, Sol N, Kooi I, *et al.* RNA-Seq of tumor-educated platelets enables blood-based Pan-cancer, multiclass, and molecular pathway cancer diagnostics. *Cancer Cell* 2015;**28**:666–76.
81. Batth IS, Mitra A, Rood S, *et al.* CTC analysis: an update on technological progress. *Transl Res J Lab Clin Med* 2019;**212**:14–25.
82. Cheng ML, Pectasides E, Hanna GJ, *et al.* Circulating tumor DNA in advanced solid tumors: clinical relevance and future directions. *CA Cancer J Clin* 2021;**71**:176–90.
83. Dang DK, Park BH. Circulating tumor DNA: current challenges for clinical utility. *J Clin Invest* 2022;**132**:e154941.
84. Pellini B, Chaudhuri AA. Circulating tumor DNA minimal residual disease detection of non-small-cell lung cancer treated with curative intent. *J Clin Oncol* 2022;**40**:567–75.
85. Schuster E, Taftaf R, Reduzzi C, *et al.* Better together: circulating tumor cell clustering in metastatic cancer. *Trends Cancer* 2021;**7**:1020–32.
86. Hemminki K, Pavlidis N, Tsilidis KK, *et al.* Age-dependent metastatic spread and survival: cancer of unknown primary as a model. *Sci Rep* 2016;**6**:23725.
87. von Felden J, Garcia-Lezana T, Schulze K, *et al.* Liquid biopsy in the clinical management of hepatocellular carcinoma. *Gut* 2020;**69**:2025–34.
88. EL Andaloussi S, Mäger I, Breakefield XO, *et al.* Extracellular vesicles: biology and emerging therapeutic opportunities. *Nat Rev Drug Discov* 2013;**12**:347–57.
89. Cheng L, Hill AF. Therapeutically harnessing extracellular vesicles. *Nat Rev Drug Discov* 2022;**21**:379–99.
90. Nikanjam M, Kato S, Kurzrock R. Liquid biopsy: current technology and clinical applications. J. Hematol. Oncol. *J Hematol Oncol* 2022;**15**:131.
91. Yu W, Hurley J, Roberts D, *et al.* Exosome-based liquid biopsies in cancer: opportunities and challenges. *Ann Oncol* 2021;**32**:466–77.
92. Mugoni V, Ciani Y, Nardella C, Demichelis F. Circulating RNAs in prostate cancer patients. *Cancer Lett* 2022;**524**:57–69.
93. Yao Y, Ding Y, Bai Y, *et al.* Identification of serum circulating MicroRNAs as novel diagnostic biomarkers of gastric cancer. *Front Genet* 2020;**11**:591515.
94. Moiso E, Farahani A, Marble HD, *et al.* Developmental deconvolution for classification of cancer origin. *Cancer Discov* 2022;**12**:2566–85.
95. Lu MY, Chen TY, Williamson DFK, *et al.* AI-based pathology predicts origins for cancers of unknown primary. *Nature* 2021;**594**:106–10.
96. Tang W, Wan S, Yang Z, *et al.* Tumor origin detection with tissue-specific miRNA and DNA methylation markers. *Bioinforma Oxf Engl* 2018;**34**:398–406.
97. Kleppe A, Skrede O-J, De Raedt S, *et al.* Designing deep learning studies in cancer diagnostics. *Nat Rev Cancer* 2021;**21**:199–211.
98. Howard J. Algorithms and the future of work. *Am J Ind Med* 2022;**65**:943–52.
99. He B, Lang J, Wang B, *et al.* TOOme: a novel computational framework to infer cancer tissue-of-origin by integrating both gene mutation and expression. *Front Bioeng Biotechnol* 2020;**8**:394.
100. Liu H, Qiu C, Wang B, *et al.* Evaluating DNA methylation, gene expression, somatic mutation, and their combinations in inferring tumor tissue-of-origin. *Front Cell Dev Biol* 2021;**9**:619330.
101. Chen K, Zhang F, Yu X, *et al.* A molecular approach integrating genomic and DNA methylation profiling for tissue of origin identification in lung-specific cancer of unknown primary. *J Transl Med* 2022;**20**:158.
102. Huang S, Yang J, Fong S, Zhao Q. Artificial intelligence in cancer diagnosis and prognosis: opportunities and challenges. *Cancer Lett* 2020;**471**:61–71.
103. Bi WL, Hosny A, Schabath MB, *et al.* Artificial intelligence in cancer imaging: clinical challenges and applications. *CA Cancer J Clin* 2019;**69**:127–57.
104. Jiang Y, Yang M, Wang S, *et al.* Emerging role of deep learning-based artificial intelligence in tumor pathology. *Cancer Commun Lond Engl* 2020;**40**:154–66.
105. Pentheroudakis G, Briasoulis E, Pavlidis N. Cancer of unknown primary site: missing primary or missing biology? *Oncologist* 2007;**12**:418–25.
106. Brezgyte G, Shah V, Jach D, Crnogorac-Jurcevic T. Non-invasive biomarkers for earlier detection of pancreatic cancer-a comprehensive review. *Cancer* 2021;**13**:2722.

107. Fitzgerald RC, Antoniou AC, Fruk L, Rosenfeld N. The future of early cancer detection. *Nat Med* 2022;**28**:666–77.
108. Posner A, Prall OW, Sivakumaran T, *et al.* A comparison of DNA sequencing and gene expression profiling to assist tissue of origin diagnosis in cancer of unknown primary. *J Pathol* 2023;**259**:81–92.
109. Arai Y, Totoki Y, Hosoda F, *et al.* Fibroblast growth factor receptor 2 tyrosine kinase fusions define a unique molecular subtype of cholangiocarcinoma. *Hepatol Baltim Md* 2014;**59**: 1427–34.
110. Chen P-HC, Liu Y, Peng L. How to develop machine learning models for healthcare. *Nat Mater* 2019;**18**:410–4.
111. Greener JG, Kandathil SM, Moffat L, Jones DT. A guide to machine learning for biologists. *Nat Rev Mol Cell Biol* 2022;**23**: 40–55.
112. Olivier T, Fernandez E, Labidi-Galy I, *et al.* Redefining cancer of unknown primary: is precision medicine really shifting the paradigm? *Cancer Treat Rev* 2021;**97**:102204.
113. Rassy E, Labaki C, Chebel R, *et al.* Systematic review of the CUP trials characteristics and perspectives for next-generation studies. *Cancer Treat Rev* 2022;**107**:102407.
114. Ye Q, Ling S, Zheng S, Xu X. Liquid biopsy in hepatocellular carcinoma: circulating tumor cells and circulating tumor DNA. *Mol Cancer* 2019;**18**:114.
115. Kato S, Krishnamurthy N, Banks KC, *et al.* Utility of genomic analysis in circulating tumor DNA from patients with carcinoma of unknown primary. *Cancer Res* 2017;**77**: 4238–46.
116. Chae YK, Davis AA, Carneiro BA, *et al.* Concordance between genomic alterations assessed by next-generation sequencing in tumor tissue or circulating cell-free DNA. *Oncotarget* 2016;**7**: 65364–73.
117. Chae YK, Davis AA, Jain S, *et al.* Concordance of genomic alterations by next-generation sequencing in tumor tissue versus circulating tumor DNA in breast cancer. *Mol Cancer Ther* 2017;**16**:1412–20.
118. Zugazagoitia J, Ramos I, Trigo JM, *et al.* Clinical utility of plasma-based digital next-generation sequencing in patients with advance-stage lung adenocarcinomas with insufficient tumor samples for tissue genotyping. *Ann Oncol* 2019;**30**: 290–6.
119. Schwaederlé MC, Patel SP, Husain H, *et al.* Utility of genomic assessment of blood-derived circulating tumor DNA (ctDNA) in patients with advanced lung adenocarcinoma. *Clin Cancer Res* 2017;**23**:5101–11.
120. Hamam R, Hamam D, Alsaleh KA, *et al.* Circulating microRNAs in breast cancer: novel diagnostic and prognostic biomarkers. *Cell Death Dis* 2017;**8**:e3045.
121. Lin PP. Aneuploid circulating tumor-derived endothelial cell (CTEC): a novel versatile player in tumor neovascularization and cancer metastasis. *Cell* 2020;**9**:1539.
122. Lambin P, Leijenaar RTH, Deist TM, *et al.* Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol* 2017;**14**:749–62.
123. Mayerhoefer ME, Materka A, Langs G, *et al.* Introduction to Radiomics. *J Nucl Med* 2020;**61**:488–95.
124. Bera K, Braman N, Gupta A, *et al.* Predicting cancer outcomes with radiomics and artificial intelligence in radiology. *Nat Rev Clin Oncol* 2022;**19**:132–46.
125. Guiot J, Vaidyanathan A, Deprez L, *et al.* A review in radiomics: making personalized medicine a reality via routine imaging. *Med Res Rev* 2022;**42**:426–40.
126. Aerts HJWL, Velazquez ER, Leijenaar RTH, *et al.* Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun* 2014;**5**:4006.
127. Hosny A, Parmar C, Quackenbush J, *et al.* Artificial intelligence in radiology. *Nat Rev Cancer* 2018;**18**:500–10.
128. Huynh E, Hosny A, Guthier C, *et al.* Artificial intelligence in radiation oncology. *Nat Rev Clin Oncol* 2020;**17**:771–81.
129. Zheng B, Fang L. Spatially resolved transcriptomics provide a new method for cancer research. *J Exp Clin Cancer Res* 2022;**41**:179.
130. Padoa-Schioppa C. Logistic analysis of choice data: a primer. *Neuron* 2022;**110**:1615–30.
131. Amaratunga D, Cabrera J, Lee Y-S. Enriched random forests. *Bioinforma Oxf Engl* 2008;**24**:2010–4.
132. Chai Z, Zhao C. Multiclass oblique random forests with dual-incremental learning capacity. *IEEE Trans Neural Netw Learn Syst* 2020;**31**:5192–203.
133. Nedaie A, Najafi AA. Support vector machine with Dirichlet feature mapping. *Neural Netw* 2018;**98**:87–101.
134. Wang H, Shao Y, Zhou S, *et al.* Support vector machine classifier via L0/1 soft-margin loss. *IEEE Trans Pattern Anal Mach Intell* 2022;**44**:7253–65.
135. Shi J, Chen X, Xie Y, *et al.* Delicately reinforced k -nearest neighbor classifier combined with expert knowledge applied to abnormity forecast in electrolytic cell. *IEEE Trans Neural Netw Learn Syst* 2023;1–11.
136. Li F, Zuo Y, Lin H, Wu J. BoostXML: gradient boosting for extreme multilabel text classification with tail labels. *IEEE Trans Neural Netw Learn Syst* 2023;1–14.
137. Mi J-X, Zhang Y-N, Lai Z, *et al.* Principal component analysis based on nuclear norm minimization. *Neural Netw* 2019;**118**: 1–16.
138. Ding X, Zhang H, Ma C, *et al.* User identification across multiple social networks based on naive Bayes model. *IEEE Trans Neural Netw Learn Syst* 2022;1–12.
139. Sarıgül M, Ozyildirim BM, Avci M. Differential convolutional neural network. *Neural Netw* 2019;**116**:279–87.
140. Anwar SM, Majid M, Qayyum A, *et al.* Medical image analysis using convolutional neural networks: a review. *J Med Syst* 2018;**42**:226.