Supplementary Table 1 ROBINS-I assessment of included non-randomized studies

Domain	Judgment	Rationale
Bias due to confounding	Serious	Most included studies were observational, predominantly retrospective, and compared patients undergoing liver resection with patients managed non-surgically. Surgical candidacy is inherently selective and strongly influenced by resectability, metastatic burden, performance status, comorbidities, response to systemic therapy, tumor biology, and referral to experienced centers. Although most studies reported multivariable-adjusted estimates, residual confounding by indication is highly likely
Bias in selection of participants into the study	Serious	Inclusion into surgical versus non-surgical cohorts was not randomized and was frequently influenced by baseline prognosis and treatment feasibility. This creates an important risk of selection bias, including immortal-time and time-dependent biases in some retrospective comparisons
Bias in classification of interventions	Moderate	Classification of surgery versus no surgery was generally straightforward; however, definitions of surgical strategies, timing of chemotherapy, curative versus palliative intent, and inclusion of conversion therapy varied across studies. This may have introduced some misclassification at the intervention level
Bias due to deviations from intended interventions	Moderate	Co-interventions, especially systemic therapy regimens and sequencing, were heterogeneous and incompletely reported in a proportion of studies. Because treatment pathways were not protocolized, deviations from intended intervention strategies could have influenced outcomes
Bias due to missing data	Moderate	Several studies lacked complete reporting of important variables, including follow-up duration, systemic therapy exposure, R0 resection rates, and biological characteristics.

Domain	Judgment	Rationale
		Missing follow-up and incompletely reported prognostic factors may have affected the precision and comparability of estimates
Bias in measurement of outcomes	Low to moderate	Overall survival is a relatively objective endpoint and is less prone to measurement bias. However, PFS is more vulnerable to bias because imaging intervals, surveillance schedules, and progression definitions were inconsistently reported across studies
Bias in selection of the reported result	Moderate	Selective reporting could not be excluded, especially in retrospective studies with heterogeneous reporting of adjusted and unadjusted estimates, subgroup analyses, and incomplete publication of non-significant findings
Overall risk of bias (OS)	Serious	The body of evidence for OS is affected mainly by serious confounding and participant selection bias, despite the use of adjusted HRs in most studies
Overall risk of bias (PFS)	Serious to critical	In addition to the limitations affecting OS, PFS was available in only a limited subset of studies and was more susceptible to heterogeneous assessment schedules and measurement bias

Supplementary Table 2 GRADE certainty of evidence for the main outcomes

Outcome	Studies (n)	Participants (n)	Effect estimate	Certainty assessment	Overall certainty
Overall survival	67	368380	HR 0.38 (95%CI: 0.34-0.43)	Downgraded for serious risk of bias, serious inconsistency, and serious publication bias	LOW

Outcome	Studies (n)	Participants (n)	Effect estimate	Certainty assessment	Overall certainty
Progression-free survival	9	Not consistently reported across all studies	HR 0.46 (95% CI: 0.31-0.66)	Downgraded for serious risk of bias, serious inconsistency, serious imprecision, and suspected publication bias	VERY LOW