Tumörinfiltrerande lymfocyter (TILs) har visat sig vara en viktig biomarkör för både prognos och behandlingseffekt vid trippelnegativ bröstcancer (TNBC). Högre nivåer av TILs korrelerar med förbättrad överlevnad, både med och utan adjuvant kemoterapi. Traditionellt har bedömningen av TILs utförts manuellt av patologer, men denna metod uppvisar stor variabilitet mellan olika bedömare. Artificiell intelligens (AI) har potential att radikalt förbättra bildanalys inom patologin genom att standardisera och automatisera bedömningen av TILs.

AI förbättrar bedömning av tumörinfiltrerande lymfocyter i bröstcancer

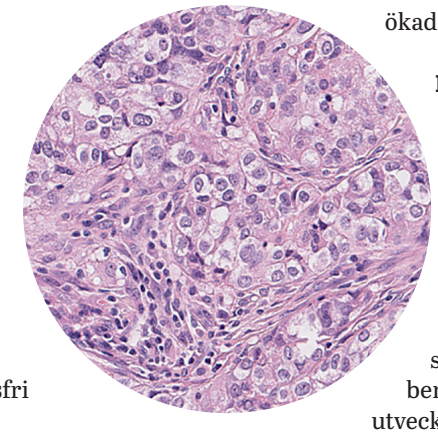
Under de senaste åren har olika maskininlärningsmetoder utvecklats för att bedöma anti-tumörimmunitet, vilket har resulterat i ett flertal TIL-biomarkörer med potentiell klinisk tillämpning.

Utmaningar vid implementering av AI inom patologi

Trots förekomsten av flera kommersiella AI-system som fokuserar på histopatologi, har implementeringen av digital patologi inom vården varit långsam. Högkvalitativa AI-system för exempelvis cancerdetektion har funnits i flera år och har erhållit både CE- och FDA-godkännande. Ändå har deras införande i sjukvården på global nivå varit begränsad. Under de senaste åren har olika maskininlärningsmetoder utvecklats för att bedöma anti-tumörimmunitet, vilket har resulterat i ett flertal TIL-biomarkörer med potentiell klinisk tillämpning.

Studiedesign och metodik

I denna studie utvärderades tio icke-kommersiella AI-modeller för bedömning av TILs, med fokus på deras analytiska och prognostiska validitet. Modellerna testades på två kohorter: en retrospektiv analytisk kohort från USA (Yale School of Medicine) och en prospektiv kohort från Sverige (SCAN-B), bestående av patienter med TNBC. Den analytiska validiteten utvärderades genom korrelation med patologers manuella bedömningar av TILs, medan den prognostiska validiteten bedömdes genom modellernas förmåga att prediktera invasiv sjukdomsfri överlevnad (IDFS).



Resultat: Variation i analytisk och prognostisk validitet

Resultaten visade på betydande variationer i den analytiska validiteten mellan de olika AI-modellerna. Korrelationskoefficienterna varierade från 0,63 till 0,73 i den externa valideringskohorten, vilket indikerar en måttlig till god korrelation med manuella bedömningar. Gällande prognostisk validitet visade åtta av de tio modellerna statistiskt signifikanta resultat i den externa kohorten, med liknande och överlappande hazardkvoter (HR) för IDFS.

TILs som robust biomarkör

Resultaten tyder på att AI-modeller har robust prognostisk styrka och kan bedöma TILs i TNBC, även om de utvecklats med begränsade patientkohorter. TILs är en robust biomarkör, vilket sannolikt förklarar varför även modeller med mindre omfattande träning presterade väl. Dock uppstod skillnader i den analytiska prestandan vid jämförelse av korrelationen mellan AI-genererade TILs-poäng och patologers manuella bedömningar i interna och externa valideringskohorter. Alla AI-modeller uppvisade god korrelation i den interna valideringskohorten, men resultaten försämrades i den externa kohorten, vilket resulterade i endast måttlig korrelation. Dessutom ledde en ökning av träningskohortens storlek inte till förbättrade korrelationsresultat, vilket antyder att andra faktorer påverkar modellernas generaliserbarhet.

Utmaningar med datakvalitet och teknisk transparens

En ytterligare utmaning är skillnaderna mellan olika bildskanningsplattformar, som kan påverka modellernas prestanda. Transparens i AI-modellernas funktionalitet är avgörande, särskilt i fall av falska resultat. Många av de undersökta AI-modellerna använder en cellvisualiseringsmetod för att möjliggöra tolkning och verifiering av upptäckta celler, vilket bidrar till ökad tillit bland patologer och onkologer.

Behovet av bred validering och benchmarking

För att möjliggöra klinisk implementering av AI-modeller krävs validering i breda, realistiska kliniska sammanhang. Många studier begränsar sina valideringar till smala kontexter, och upphovsrättsskyddade modeller offentliggörs sällan, vilket försvårar jämförelse och benchmarking. En möjlig lösning är att utveckla ett stort, inkluderande och multicentrisk benchmark-dataset, där modeller kan valideras på ett enhetligt sätt utan att kompromettera immateriella rättigheter. Detta skulle stärka trovärdigheten och identifiera respektive modells styrkor och svagheter.

Sammanfallningsvis visar denna studie att det finns variationer i både analytisk och prognostisk validitet mellan olika AI-baserade TIL-bedömningsmodeller. Det finns ett akut behov av ett tillgängligt, multicentrisk benchmark-dataset som omfattar flera kliniska prövningskohorter. Detta skulle säkerställa jämförbarhet och klinisk användbarhet hos olika AI-metodologier innan de implementeras. Den multiinstitutionella CATALINA-utmaningsstudien kan utgöra ett viktigt steg i denna riktning.

Text BALAZS ACS
Associate Professor; Professor Johan Hartman's group, Department of Oncology-Pathology, Karolinska Institutet,
balazs.acs@ki.se