AAMD ANNUAL MEETING · 2026

Clinical Alignment for Automated Planning

A structured framework for validating deterministic, blueprint-driven automation — built so dosimetrists adopt through evidence, not assumption.

Peter Mc Loone
Presented for the Clinical Alignment program · Lumonus

Lumonus

1

TODAY'S SESSION

What you'll leave with.

Describe the framework

The components of a Clinical Alignment Framework for benchmarking automated and manual plans — and what each piece protects you against.

Interpret the metrics

The dosimetric and efficiency metrics used to evaluate deterministic automation — what they mean and where they mislead.

Apply the QA scaffolding

Structured QA and documentation methods to validate automation inside your institutional workflow.

Lumonus

2

THE PREMISE

“ The first question a dosimetrist asks when an automated plan lands on their workstation isn't *is the dose calculation correct.* It's **would I have planned it this way?**”

If the answer is no

You can't defend the plan at chart rounds. You can't explain it to the referring physician. You can't take responsibility for it. And you won't trust the next plan it produces.

The bar isn't dosimetric constraint achievement

It's clinical alignment. Alignment is not all constraints passing. It is structured, evidence-based agreement across coverage, sparing, deliverability, and intent.

DEFINITION

Clinical Alignment is the discipline of demonstrating, with **measurable evidence** , that an automated plan **agrees with** the manual plan a competent dosimetrist would produce across coverage, sparing, deliverability, and **clinical intent** .

MEASURABLE

Quantitative, not qualitative.

No "looks good to me." Every comparison is reduced to numbers a third party can re-derive from the source data.

AGREES WITH

Peer, not superior.

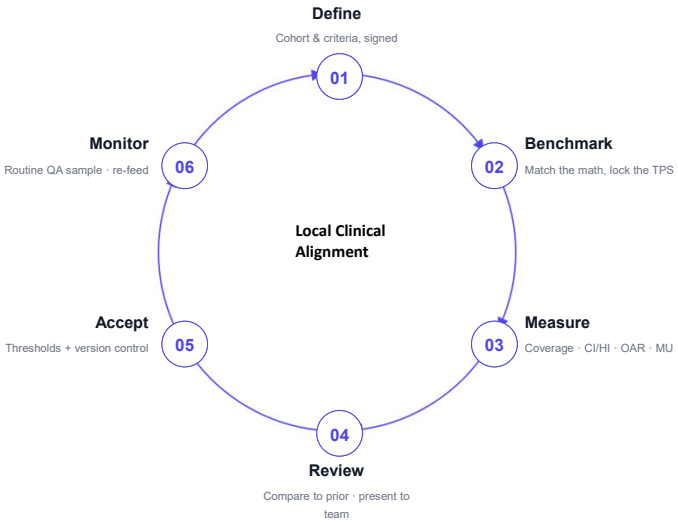
We are not asking whether automation beats a dosimetrist. We are asking whether it produces a plan you'd accept as a colleague's.

CLINICAL INTENT

The part the algorithm can't see.

Patient-specific anatomy, the clinicians' opinions on dose homogeneity. Alignment is what proves the automation respected intent it couldn't read.

THE FRAMEWORK



A loop, not a checklist.

Every stage feeds the next, and stage six (Monitor) feeds back into stage one (Define). The bar shifts as you add cases, new disease sites, and new constraints — the framework is built to shift with it.

Team-anchored. Every loop returns to the local clinical team's judgment — dosimetry, physics, attending.

Transparent. Every claim is reducible to numbers a peer can re-derive.

Living. The QA stream sharpens the framework every month.

Vendor-agnostic. Nothing here requires a particular automation tool.

STAGE 01 · DEFINE

Define success *before* you see any results.

The Cohort Definition Document

Representative and standardised. Same prescription, same anatomy spread, same complexity.

Endpoints up front. Coverage points, OAR ceilings, deliverability thresholds.

Acceptance criteria up front. What percent the local clinical team must accept on per-case review against the prior plans to declare alignment.

STAGE 02 · BENCHMARK

Hold everything constant except the planning workflow your local clinical team would actually use.

Five things that must match exactly

CT & structure set
Dose-calculation algorithm + grid
Machine model
Prescription + fractionation
Constraint set

COMMON FAILURE MODE

Optimization parameters the local team can't trace

Parameters can absolutely differ between manual and automated, that's expected. But they must be structured the way your local clinical team would set them up, so a dosimetrist reading the plan can follow the logic and reproduce it if needed.

STAGE 03 · MEASURE

Every metric, on every plan, every time.

TARGET COVERAGE

PTV D95, D98, D2

PLAN QUALITY

Conformity & Homogeneity indices

SPARING

OAR mean & Dmax (per protocol)

DELIVERABILITY

MU per beam · modulation complexity

STAGE 04 · REVIEW

The previously planned cases are the benchmark. The local clinical team is the reviewer.

How the comparison works

Prior approved plans are the gold standard
Every metric, side by side
Per-case alignment report
Aggregate roll-up across the cohort

COMPARE DISTRIBUTIONS, NOT SINGLE NUMBERS — IQR

For each metric we plot the **interquartile range (Q1–Q3, the middle 50%)** across both the prior approved plans and the automated plans. Overlapping IQRs mean the automation sits inside your clinic's normal spread. Anything past **1.5 × IQR** is flagged as an outlier and reviewed on its own — so a single odd case can't move the verdict.

Results presented back to the local team

Walk-through with dosimetry, physics, and the attending
Acceptability judged against the prior clinical plan
Differences recorded and fed back to Stage 1 (Define)

STAGE 05 · ACCEPT

Set the bar in writing — your clinic's thresholds

Has your previously defined acceptance criteria been met

Coverage threshold. e.g. *PTV D95 ≥ 95% on ≥ 98% of plans.*

OAR ceilings. e.g. *No plan exceeds protocol OAR limits.*

Local team clearance. e.g. *≥ 85% of plans accepted by the local team against the prior clinical plan.*

Variability ceiling. e.g. *Inter-planner OAR spread < 5% Dmean.*

MU policy. e.g. *MU per fraction not greater than manual baseline + 10%.*

These thresholds are your clinic's, not the vendor's. The framework only enforces that you write them down before you measure.

STAGE 06 · MONITOR

Automation isn't a one-time validation — it's an ongoing surveillance problem.

Triggered, not scheduled

Acceptance rate is the signal. When the local team's accept rate on automated plans drops below threshold, surveillance kicks in.

Re-run the pipeline. Same metrics, same comparison against prior clinical plans, local-team review.

Also triggered on change. Any new disease site, new machine, or new constraint set forces a re-benchmark.

Why it's a loop

The data you collect in stage six (Monitor) feeds back into stage one (Define). The cohort sharpens as edge cases surface and are added; cases the framework now handles consistently are retired. Acceptance thresholds tighten as confidence grows. The framework gets stronger the longer you run it.

02

SECTION TWO OF THREE · CASE STUDY

One national protocol, 20 clinics — Lung SABR, 48 Gy / 4 fx.

31 paired cases · 20 clinics across five states · the organisation's own National Lung SABR protocol, then two states walked in full.

CASE STUDY · COHORT

Thirty-one paired cases, 20 clinics across five states, one protocol — benchmarked against each clinic's own approved plans.

PAIRED CASES
31
Clinical plans + automated plans

CLINICS / STATES
20 / 5
States anonymised (State 1–5)

PROTOCOL
Aligned
National Lung SABR · 48 Gy / 4 fr

TARGET COVERAGE
Met
PTV D95 ≥ 95 & ITV D98 ≥ 100 at every clinic

A TEXTBOOK AUTOMATION-READY SITE

- One **national protocol** — Lung SABR 48 Gy / 4 fr
- Standard work instructions for every planning step
- A standard structure-naming convention across all 20 clinics
- Within each clinic: same CT protocols, machine type, and dose engine on both arms

THE QUESTION THIS COHORT ANSWERS

This site did everything you'd ask for before automating. Every plan meets protocol — so does **meeting the constraints** mean the automation shares each clinic's **intent** , especially how hot to run the target? The next four slides separate the two; then we walk two groups in full.

Lumonus

COHORT · TARGET COVERAGE

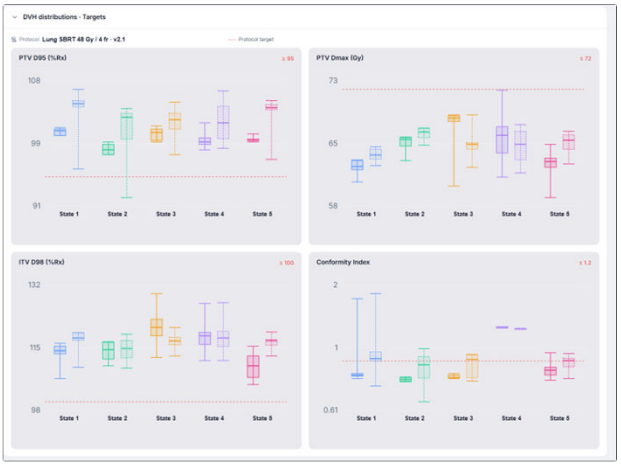
Every plan clears the coverage constraints — but automation runs the target hotter.

PTV D95 (≥ 95%Rx). Both arms clear the floor at every clinic, but the automated median sits **above clinical at every site** . Automation consistently runs coverage hotter than the manual baseline.

ITV D98 (≥ 100%Rx). Comfortably above target everywhere — the internal target is fully covered on both arms.

PTV Dmax (≤ 72 Gy). Hotspot maxima stay under the ceiling on both arms; State 4 shows the widest clinical spread, brushing the limit on its upper whisker.

Conformity index (≤ 1.2). At or under target at four states; State 4 sits on the line on both arms — the one coverage metric to watch.



THE HEADLINE

Every plan passes the protocol, but the automated medians sit **systematically above** the clinical baseline. Constraint compliance is not the same as matching intent. Hold that thought for the next slide.

Lumonus

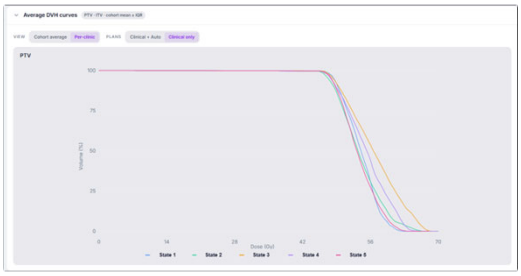
COHORT · TARGET COVERAGE

The medians don't overlap: five states, five intents on target dose falloff

Compare per-clinic mean PTV dose–volume histogram, clinical plans. Every curve holds 100% to the 48 Gy prescription so coverage isn't the issue. It's the **high-dose falloff** that tells the real story.

The falloff fans out. Past the prescription the five state medians separate, they don't overlap. Each state runs a different target hotspot.

State 3 and State 4 run hot; their curves carry higher high-dose tails — a hotter target by intent. **State 1, State 2 and State 5 run cooler,** falling off faster.



WHY IT MATTERS

One national protocol but **not one intent** . The states legitimately disagree on how hot to run the target, and all of them pass. A single blueprint can't encode an intent they don't share — automation has to be calibrated to each clinic, not assumed from the protocol.

COHORT · OAR SPARING

OAR sparing tracks the manual baseline with the expected exceptions.

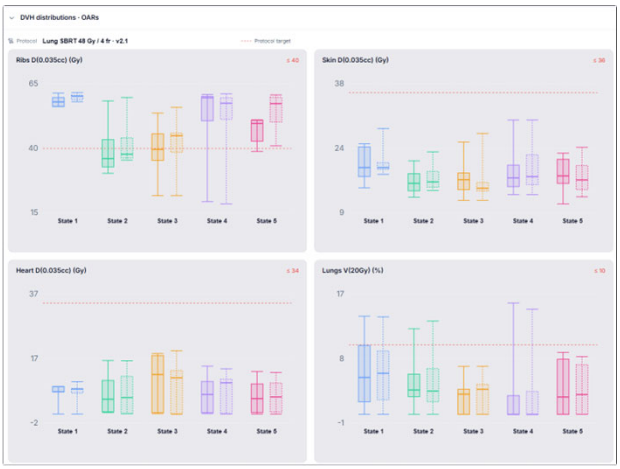
Skin (≤ 36 Gy) & Heart (≤ 34 Gy). Comfortably within tolerance at every clinic, both arms.

Lungs V20Gy (≤ 10%). Medians under target; a few upper whiskers (State 1, State 2, State 4) cross 10% on individual plans, flagged per case, not systematic.

Ribs (≤ 40 Gy). Routinely exceeded on both arms — expected for peripheral targets where ribs overlap the PTV. Recorded at-parity, not a regression; State 1 and State 4 sit highest.

THE PATTERN

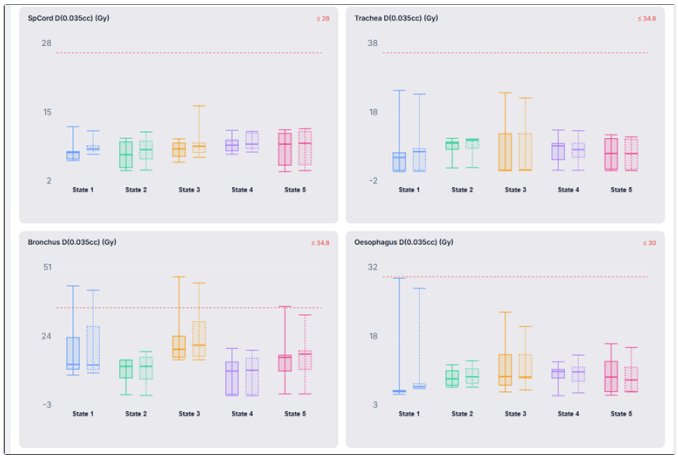
Automated and clinical distributions overlap structure by structure. Where a tolerance is exceeded, it's the anatomy not the automation.

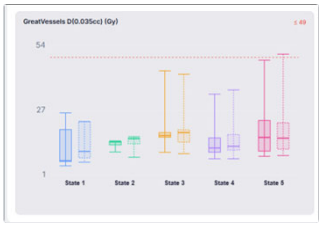


8

COHORT · OAR SPARING

Serial and central structures are within tolerance, clinic to clinic.





Spinal cord (≤ 26 Gy). Large margin everywhere; tight spread.

Trachea (≤ 34.8) & oesophagus (≤ 30). Well under tolerance at all clinics.

Bronchus (≤ 34.8). Medians under target; State 1 and State 3 whiskers cross on individual central cases.

Great vessels (≤ 49). Under tolerance; State 5 the highest whisker.

No systematic separation. Auto tracks clinical on every serial structure; outliers are case-specific.

Lumonus

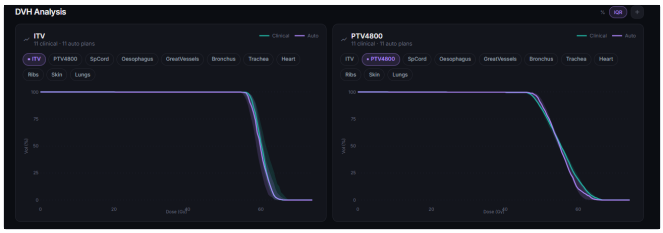
DEEP DIVE · STATE 1

State 1, walked in full — 11 paired plans, Lung SABR 48 Gy / 4 fx

Plan QA verdict · flagged

PRESCRIPTION	48 Gy / 4 fx Lung SABR · peripheral target
PAIRED PLANS	11 clinical · 11 auto State 1 · same TPS export, same constraints
PLAN QA	**91** / 100 Target ≥ 90 · 48 / 52 metrics pass · 4 flags
PTV COVERAGE	**84** / 100 All constraints pass; conformity index degrades
OAR SPARING	**96** / 100 No at-fault failures; informational hotspots
DELIVERABILITY	**95** / 100 2-arc VMAT · MU/Gy within cohort range

DVH ANALYSIS · CLINICAL VS. AUTO, ALL 11 PLANS PER ARM

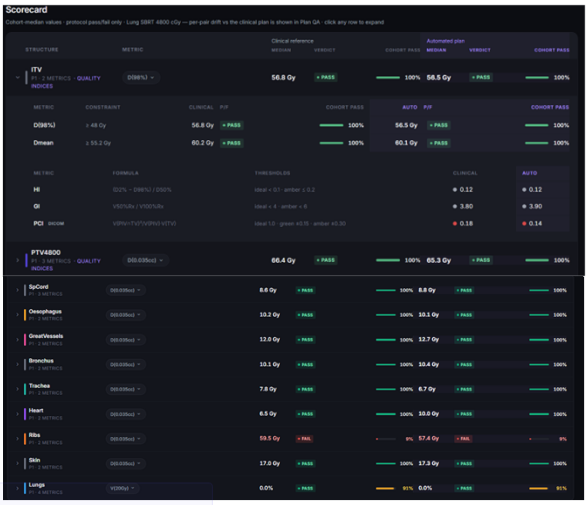


Median lines with shaded IQR bands. On coverage the clinical and auto curves sit on top of each other on both targets — the target is fully covered on both arms. The divergence this plan is flagged for shows up off the coverage curve: in conformity and the OAR hotspots.

Lumonus

DEEP DIVE · STATE 1

Every structure is shown for both arms, nothing is averaged away or hidden



What to read here

Most rows track within ~0.5 Gy between clinical and auto. Equivalence on the structures the protocol cares most about.

Ribs fail on both arms. Expected for peripheral lung SBRT — ~1.6 cc rib overlap with the PTV. Recorded as *at-parity*, not a regression.

Heart D(0.035cc) jumps +3.5 Gy. 6.5 Gy clinical → 10.0 Gy auto. Well within protocol; NTCP for pericarditis still negligible — but a pattern the physicist should *see* , not have buried in an aggregate.

Trachea, bronchus, oesophagus, skin all within tolerance — informational only.

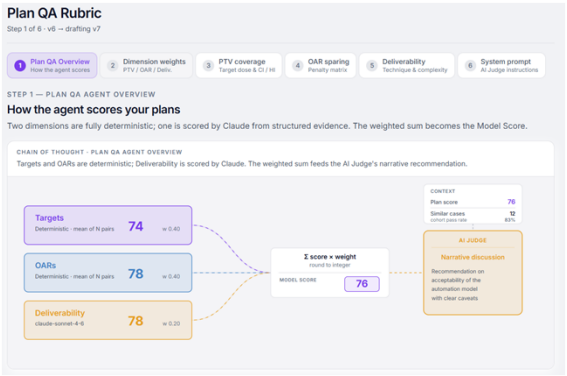
THE DISCIPLINE BEHIND THE LAYOUT

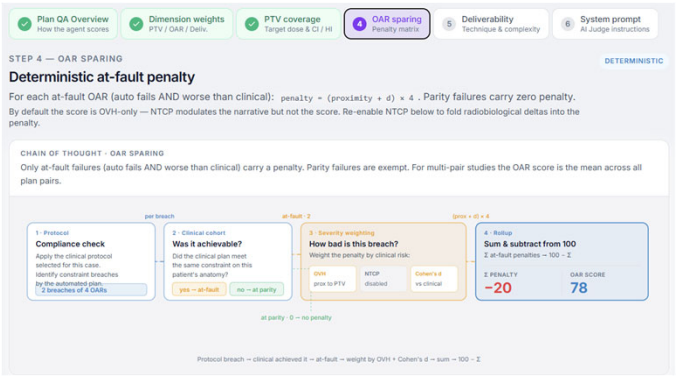
No cherry-picking. The bar is *"no OAR is worse"* not *"the average OAR is better."* When something *is* worse, the framework surfaces it as a flag, not a footnote.

Lumonus

DEEP DIVE · STATE 1

Compare on a case-by-case basis to determine an overall score





Local team defines these metrics and weightings. What is important to your clinic and how harshly should a plan be penalized for deviating from a clinical plan. If both plans fails a dosimetric constraint should the automated plan be penalized.

Lumonus

DEEP DIVE · STATE 4 · THE CONTRAST

The same framework on a different state, and automation that doesn't align.

Plan QA verdict · misaligned

Same protocol, same pipeline — opposite result. In State 1 the automated and clinical DVH bands sat on top of each other. Here they pull apart on both targets.

Automation under-doses the target. The auto curve (purple) falls off well before clinical (teal) on ITV and PTV4800 — the automated plans give up high-dose coverage the manual plans hold.

WHY THIS STATE

State 4 is one of the states whose clinical intent ran **hottest** in the cohort. The generic blueprint can't reach that intent, so the gap between auto and clinical is widest exactly where the clinic's intent is strongest.

DEEP DIVE · STATE 4 · THE CONTRAST

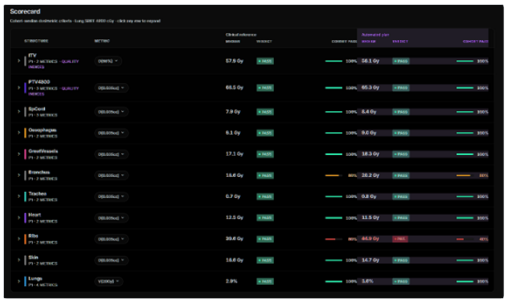
Where it shows up on the scorecard. A structure crosses from pass to fail.

What to read here

Ribs: PASS → FAIL. Clinical median 39.6 Gy clears the 40 Gy ceiling; the automated plan lands at 44.9 Gy and fails. Cohort pass falls 60% → 40%, the automation made a structure *worse*, not at-parity.

The target medians drop. ITV 57.9 → 56.1 Gy and PTV4800 68.5 → 65.3 Gy, the under-dosing the DVH showed, now quantified.

Most rows still pass and that's the trap. A constraint-only check sees mostly green and ships. The framework is what surfaces the rib regression and the coverage drift an aggregate would bury.



THE VERDICT

Do not deploy as-is. Where State 1 earned *accept-with-watch* , State 4's blueprint needs rework with re-weighted optimisation and margin review before this site goes live.

CASE STUDY · LESSONS

What this cohort showed.

The site did everything right and that wasn't the whole answer.

A national protocol, standard work instructions, and a standard naming convention across all 20 clinics — textbook automation-readiness. Every automated plan met protocol. If constraint compliance were the bar, you'd stop here and ship.

Meeting the constraints is not the same as matching intent.

Coverage cleared at every clinic but the per-state PTV DVH medians don't overlap, and automation runs the target systematically hotter than the clinical baseline. The framework surfaced a misalignment a pass/fail constraint check would have waved straight through.

Five states, five intents so one blueprint can't be assumed.

On the ITV hotspot the states split: clinical plans were pushed hotter for State 3 and State 4 while pulled cooler for State 1, State 2 and State 5. There is no single "national intent" to encode so each clinic's blueprint has to be calibrated and accepted locally before automation goes live.

03

SECTION THREE OF THREE

How you take this home — a 90-day playbook for your clinic.

One disease site. One decision. Supported by evidence, then repeat.

PLAYBOOK

A ninety-day Clinical Alignment pass you can run in your own department.

DAYS 1–15

Define

Pick one disease site. Build the cohort definition document. Set endpoints and acceptance thresholds. Get sign-off from physics, dosimetry, attending.

DAYS 16–45

Benchmark + Measure

Run 30–50 paired plans (auto vs. the previously planned clinical case). Collect every metric. Prepare the per-case alignment reports.

DAYS 46–75

Review + Accept

Walk the local team through the per-case alignment reports. Write the acceptance report. Decide go / no-go against the criteria *you wrote on day one*.

DAYS 76–90

Monitor

If go then set up the monthly QA sample, version-control workflow, and drift-detection cadence. If no-go then write what failed, iterate, return to day one.

RESOURCING — BE HONEST ABOUT IT

- **One dosimetrist lead** · ~30% time for the 90 days
- **One medical physicist** · ~10% time for QA scaffolding
- **One attending** · ~3 hours total for the per-case team review
- **Sample size** · 30–50 cases is enough to make a call on one site

AFTER DAY 90

If the answer was go, you now have a versioned blueprint and a monthly QA cadence and you can start the playbook on your next disease site with the scaffolding already in place. Second site is roughly half the effort.

Lumonus

25

CLOSING THOUGHT

The case for automation isn't *"trust the algorithm."* **The case is, adopt through evidence.**

The Clinical Alignment Framework is one way, not the only way, to generate that evidence systematically. It puts the dosimetrist back in the position of judging the work, with quantitative data and per-case review against the prior clinical plans backing the judgment. That's a posture you can defend in chart rounds, at your tumor board, and in front of your patients.

Thanks for listening — I'd love to take questions.

Lumonus

26

Q&A

Lumonus

27