Quality Held. Then It Improved.

What Eleven Hospitals Learned From Two Years of AI-Assisted NSQIP Abstraction Audits

The Baseline

Eleven hospitals in a single health system submit surgical cases to the American College of Surgeons NSQIP registry. Every submission cycle is audited for inter-rater reliability against a 98 percent benchmark, and in 2025 all eleven were clearing it.

Inter-rater reliability answers one question. When a second qualified abstractor works the same chart against the same registry definitions, does the value come out the same? It measures reproducibility rather than correctness, and reproducibility is what an accreditation reviewer tests when they pull a chart and ask the hospital to defend what it submitted.

Nine of those hospitals abstracted manually in 2025. Two used AI-assisted abstraction through Carta Healthcare's Lighthouse platform. The manual sites averaged 99.42 percent agreement and the AI-assisted sites averaged 99.24 percent, a gap of less than two tenths of a percentage point.

The Risk

When reliability already sits above 99 percent, every change to how the work gets done carries obvious risk and invisible reward. Expanding AI-assisted abstraction across more hospitals meant putting a number in play that surgeons, quality leaders, and accreditation reviewers all depend on.

The audit itself offered thin protection against that possibility. Across most hospitals and most abstraction vendors, inter-rater reliability review means re-abstracting three percent of cases a month, because checking more than that by hand costs more than the budget allows. Carta Healthcare's own standard was five percent, which is better and still a spot check.

A sample that small finds the disagreements it happens to land on and says nothing about the ninety-five percent it never opened. It also yields an estimate rather than a measurement, and it surfaces isolated mistakes rather than the patterns underneath them, which leaves education to guesswork.

The Shift

Carta Healthcare's abstraction team changed two things at once. The first was coverage. AI-assisted review moves fast enough that complete coverage costs less than a hand-checked sample used to, so every case in the IQVIA workstation went under review every cycle. The second was posture. The audit became education rather than enforcement, and abstractors reviewed and corrected their own flagged cases.

The Proof

AI-assisted abstraction expanded from two hospitals to six, and reliability held. The 2026 audit results across all eleven hospitals came in as follows.

- 2026 AI-assisted IRR average: **99.36%**
- 2026 manual IRR average: **99.32%**
- Year over year change in overall averages: **less than 0.2%**
- Hospitals at or above the 98% IRR benchmark: **all eleven, both years**

What the audits establish is equivalence rather than superiority. Inter-rater reliability measures agreement, so these figures say that AI-assisted abstraction reaches the same value a second qualified abstractor would, as dependably as manual abstraction does. They do not claim the AI is more correct than a person. That is a different study with a different design, and this was not it.

Equivalence is the permission slip, though. It is not the reason a health system does this. The reason is what speed makes possible once reliability is no longer in question. That combination, machine speed paired with abstractor judgment, is Hybrid Intelligence at work.

The Payoff

Sampling was never a methodology. It was a budget. Hand review is slow, so hospitals and vendors check what they can afford to check and infer the rest. AI-assisted review is not slow, which changes what an audit is allowed to be.

- Industry standard IRR sample: **three percent of cases, monthly**
- Carta Healthcare manual standard: **five percent**
- Carta Healthcare with Lighthouse: **100% of cases, every cycle**
- Reliability cost of that expansion: **none**

Complete review is not an incremental improvement on a five percent sample. It is a different category of assurance, and no sampling-based process reaches it at any staffing level. Consistency is the other half. The reviewer on Tuesday is not the reviewer on Friday, while AI applies one standard on every pass, which is what makes findings comparable across eleven hospitals and across cycles.

*Three percent gives you an estimate. Every case gives you a **measurement**.*