Autonomous Coding You Can Verify

An auditable accuracy benchmark for autonomous E/M coding.

Medmio Research · July 2026 · medmio.com

98.1%	92.9%	10.9 s	100%
billing accuracy on real, paid charts	of codes **finalized autonomously**	median **processing time** per chart	of figures **traceable** to a certification artifact

Executive summary

- **Two numbers, not one:** how accurate the coding is — **98.1%** — and how much of it the software finalizes on its own — **92.9%**, with accuracy on the automated portion held at or above 95% (definitions in §3).
- **Measured against the codes a real practice billed and payers paid** — 945 charts, every headline figure certified on a 376-chart **held-out test set** kept out of all development work. Where the billed record itself is imperfect, the disagreement still counts as our error, pending the pre-registered independent audit.

1 Every figure in this report is verifiable

Every autonomous-coding vendor advertises a big number. In a July 2026 review of publicly available vendor materials, we found none that publishes the evidence behind it: a held-out test set, confidence intervals, calibration, the accuracy-coverage curve. This report publishes all four, for every figure.

How the field reports accuracy	Headline number	What is actually disclosed
Advertised autonomous-coding claims (2024–26)†	"95–99%+ accurate · 85–96% automated"	Self-reported · no held-out test set · no intervals · no calibration · no error direction
General-purpose LLMs, peer-reviewed (Soroush et al., *NEJM AI* 2024[10])	33.9% exact ICD-10-CM · 49.8% CPT	Full method
Domain-fine-tuned LLM[13]	69.2% exact-match	Peer-reviewed, full method
RAG-augmented LLMs vs ED coders[14]	17.6–26.4% exact-match	Preprint, full method
Purpose-built neural coders[7,8,9]	macro-F1 0.09–0.10 on the full code set (micro-F1 ≈ 0.5–0.6); zero recall on the rare half of ICD-10 codes[9]	Public benchmark, full method
Human coders, re-abstraction studies[11,12]	82–87% inter-coder agreement	Same charts, two certified coders (ICD-10 variants)
CodeSight™ (this report)	**The full accuracy–coverage curve**	**Held-out test set · CIs · calibration · billed ground truth · error direction**

The number that matters is the one shown **with its coverage, on a held-out test set, against real billed charts.**

† Drawn from publicly available vendor marketing materials reviewed July 2026: self-reported figures, against vendor-defined denominators. No vendor is identified and no head-to-head comparison is made.

2 The system: a calibrated ensemble

CodeSight™ is an **ensemble of specialized models** — each owning one decision — behind a calibrated confidence gate:

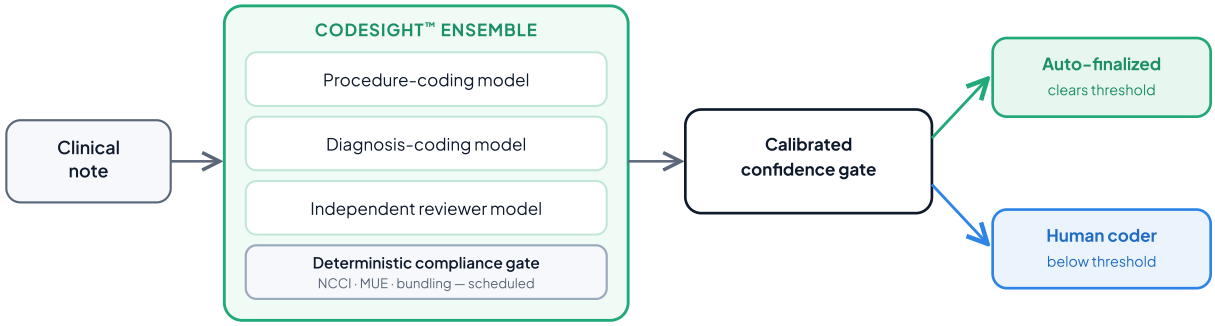

Fig. 1 — The CodeSight™ ensemble. Specialized models author the codes; an independent reviewer flags suspected errors without authority to edit; deterministic checks apply written rules, not model judgment (NCCI/MUE/bundling — scheduled; the practice-policy E/M check in Methods — certified).

The benchmark (Medmio EM-945). 945 billed office-visit (E/M) encounters from a de-identified outpatient specialty practice — every runnable E/M chart the corpus contains, not a curated subset. Ground truth: **the codes the practice billed and payers paid.**

Methods at a glance	
Corpus	945 complete SOAP-structured office notes (one empty note excluded); born-digital text in — no OCR confound
Ground truth	The codes billed and paid, scored per encounter — a record outside our control
Split	Random split locked before certification: development n = 569 · **held-out test set n = 376**, kept out of all development work
Calibration	Isotonic map fit on the development split only, applied unchanged to the test set
Deterministic E/M check	Where a note documents total visit time, the practice's own written coding policy overrides the model's proposed level — a written rule, not a model guess. Certified in two passes: 93.4% before the check, 98.1% with it; the models and prompts were identical in both passes
Case mix	Corpus over-weights low acuity vs the practice's 28,112-encounter billing mix — **case-mix-reweighted** figures reported alongside raw; the highest-acuity bands are thin, so high acuity is not certified (§5)
Privacy	Operated under HIPAA safeguards and Business Associate Agreements; every figure in this report is aggregate and de-identified
Scheduled	Redaction-controlled arm masking codes that appear verbatim in a note's Assessment/Plan — separates reasoning from transcription

The gate. Every code carries a confidence score. Diagnosis confidence is recalibrated by isotonic regression[5] — a standard technique that rescales stated confidence so a stated "90%" means about 90% in reality, because modern models are systematically overconfident.[4] E/M confidence is used as-is: at 98% accuracy there is too little error left for recalibration to act on. Above the confidence bar, a code is finalized; below it, a certified human coder decides — shown the draft code and the reason it was flagged.[6] **The models were not fine-tuned** — no custom training on this practice's data; these are baseline numbers.

Safety by design. Autonomy is gated by guardrails: an **independent reviewer model** that can flag a code but never change it · a **calibrated confidence gate** that routes anything uncertain to a certified human coder · **zero wrong-family codes** across every chart tested · and **published error direction** — 4 over / 3 under on the held-out test set. Built for healthcare: operated under HIPAA safeguards — accountable, auditable, human-in-the-loop.

3 Results: the curve, not a number

n = 945 charts · development 569 / held-out test 376 · every figure traces to a versioned certification artifact.

Certified results — held-out test set (n = 376)	Value	95% CI
E/M level — exact match with the billed code	**98.1%**	96.2–99.1
E/M level — reweighted to the practice's full billing mix	93.8%	—
Diagnosis codes — F1	0.885	0.870–0.899
VAR@95 — condition & symptom codes	**92.9%** (F1 0.906)	—
VAR@95 — all codes	87.9%	—
VAR@97 — all codes	73.9%	—
VAR@99 — all codes	26.9%	—

E/M is one code per encounter, scored as exact match; Wilson intervals. Diagnosis is multi-label micro-F1 (precision 0.867 / recall 0.904); bootstrap intervals. The 1,872 test-set codes = 376 E/M levels + 1,496 diagnosis codes. Across all 945 charts, development split included: E/M 98.7%, diagnosis F1 0.878, VAR@95 82.3% all-codes / 88.3% condition-&-symptom. Ease the accuracy bar to ≥94% and 90% of codes clear the gate, at 94.2% measured accuracy on that slice. Coverage falls steeply past the 97% bar; the pre-registered model upgrade targets that region.

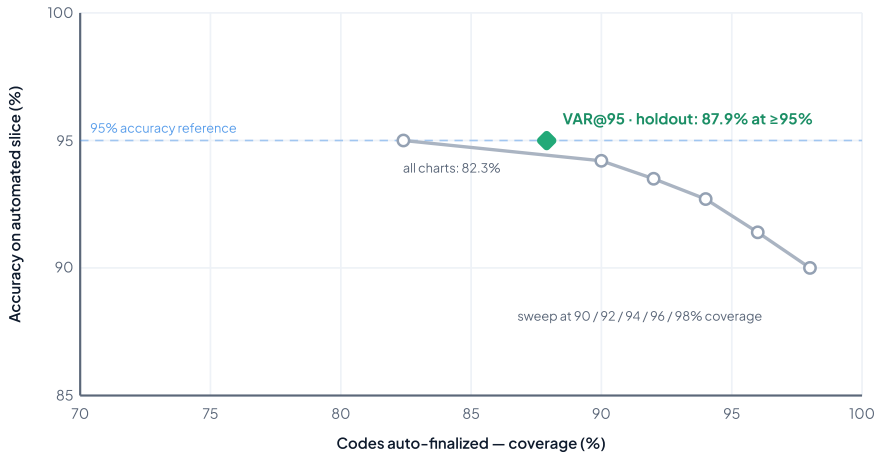

Fig. 2 — How much we automate, and how accurate it stays. Gray: the measured single-gate sweep across all 945 charts (4,703 codes) — coverage opened from the VAR@95 point (82.3% at ≥95%) to 98% coverage (90.0% slice accuracy). Green diamond: the certified held-out VAR@95 operating point — **87.9% of codes auto-finalized at ≥95% accuracy on the automated slice**. All points are measured values from the certification artifacts; nothing on this chart is extrapolated. The condition-&-symptom operating point (92.9% at ≥95%) is tabulated above; this chart plots the all-codes gate.

Verified Automation Rate (VAR@95). VAR@95 answers one question: how many codes can be finalized with no human touch while the finalized group stays at least 95% accurate? Both halves are measured on the held-out test set — the coverage is counted, the accuracy is scored.[1,2,3] The same trade-off is reported at stricter bars as VAR@97 and VAR@99.

Two ways to count. The headline 92.9% scores **condition & symptom codes** — the codes that describe what is clinically wrong with the patient. Administrative status descriptors (ICD-10 "Z" codes: "history of," "long-term medication use," and similar bookkeeping), which practices bill inconsistently as a matter of policy rather than clinical substance, are left out of **both sides** of the comparison — prediction and billed truth alike (1,748 of 1,872 test-set codes remain). Counting every code including status descriptors: 87.9%. Neither number replaces the other; both trace to certification artifacts.

3a · Where the errors live

Across all 945 charts, CodeSight's E/M level disagreed with the billed code on just 12 — and in 6 of those, **the documentation supports CodeSight's level over the billed one**: under-billed revenue the practice left behind. On the held-out test set: 7 disagreements — **4 one level above the billed code, 3 below, 0 outside the code family** (all 945: 7 over / 5 under / 0 wrong-family). Ground truth is still what was billed: every disagreement is scored as our error until the pre-registered independent CPC audit reports.

3b · Why the gate can be trusted

Fig. 3 — Reliability diagram, diagnosis confidence (held-out test set: 376 charts, 1,496 scored diagnosis codes in bins). Each dot is a confidence bin (area ∝ bin size); the diagonal is perfect calibration. After isotonic recalibration (fit on the development split, applied unchanged), expected calibration error is **0.037** on the test set (0.026 across all charts) — and the top bin, where the auto-finalize gate operates, runs slightly *conservative*: codes stated at 0.945 confidence were right 98.1% of the time.

These results predate model fine-tuning and the pre-registered frontier-model upgrade; the next certification run will report whether the curve climbs.

3c · Accuracy by E/M level — including the thin ones

E/M level	Test-set n	Exact accuracy	95% CI
99202	109	100.0%	96.6–100.0
99203	33	100.0%	89.6–100.0
99204	4	100.0%	51.0–100.0
99211	30	96.7%	83.3–99.4
99212	136	99.3%	96.0–99.9
99213	47	95.7%	85.8–98.8
99214	11	90.9%	62.3–98.4
99215*	6	66.7%	30.0–90.3

We publish the thin rows too. 99204 (n = 4), 99214 (n = 11) and 99215 (n = 6) carry intervals too wide to certify, and 99205 does not occur in this corpus — **high-acuity performance is explicitly not certified by this benchmark** (§5). This is why the headline is also reported case-mix-reweighted (93.8%) against the practice's 28,112-encounter billing distribution.

* Both 99215 misses are **conservative under-calls** — the model proposed 99214 rather than award the top level without support. At the highest-value level the system errs downward, never upward; at n = 6, two charts decide the row.

3d · Speed

10.9s	**13.4**s	**~300**/hr
median end-to-end per chart, full 945-chart run	95th percentile	charts **per lane** — add lanes to scale

10.9 seconds is one chart on **one** processing lane — capacity is simply lanes × ~300 charts/hour, and lanes scale horizontally, so enterprise volume is an infrastructure setting, not a staffing plan. (Manual production coding typically runs 20–30 charts per coder-hour.) These are retrospective batch timings, not a production SLA; one chart in 945 hit a cloud-infrastructure stall, disclosed in the run log and since hardened. Marginal compute cost per chart is a small fraction of manual coding.

4 How to read these numbers

The human ceiling. Certified coders disagree with each other. In re-abstraction studies of ICD-10 coding, two trained coders agree ~82–87% of the time at billing-level specificity — and ~47% at full terminal-code specificity (κ ≈ 0.42).[11,12] That bounds every accuracy metric scored against a single biller's labels, including ours: a system matching one biller's choices at the rates reported here is already inside the band where two trained humans disagree — and some scored "misses" are defensible

judgments, or documentation the biller under-credited. Settling those requires expert review: hence the pre-registered independent CPC audit.

Our setting is deliberately different from the benchmarks in §1 — **focused, specialty-scoped, calibrated, human-in-the-loop**. A raw frontier model, however capable, is not a coding system: published exact-match rates for general-purpose LLMs sit near 34%.[10] The distance from there to the numbers in this report is the system — specialization, domain calibration, and a **calibrated reject option**[1,2,6] that automates only what it has earned, at a precision you can verify.

The disclosure checklist

What a verifiable accuracy claim discloses	This report	Typical advertised claim
Sample size (charts and codes)	**945 charts · 4,703 codes**	rarely stated
Held-out test set	**n = 376, locked before certification**	not stated
Ground-truth definition	**codes billed and paid**	vendor-defined or unstated
Confidence intervals on headline figures	**Wilson + bootstrap, shown inline**	absent
Calibration of the gating confidence	**reliability diagram + ECE**	absent
Error direction (over- vs under-coding)	**published: 4 over / 3 under**	absent
Case-mix adjustment	**reweighted to real billing mix**	absent
The full accuracy–coverage curve	**Fig. 2, measured points only**	absent

5 Limitations

- **One practice, one specialty.** A single outpatient vascular practice, one EMR, one billing team's coding culture as ground truth. Generalization to other specialties is untested here (a procedure/ultrasound benchmark on ~12,000 notes from the same corpus is in progress as a separate track).
- **High acuity is not certified.** 99205 is absent from the corpus; 99204/99214/99215 are too thin for tight intervals. The per-level table shows the thin rows.
- **Scope is E/M office visits.** No procedure codes, modifiers, laterality, or units in this benchmark; the deterministic NCCI/MUE compliance gate is scheduled and was not part of this certification.

Bottom line

Do not buy an accuracy number. Buy the curve — and the evidence under it.

The certified answer: **92.9% of condition & symptom codes finalized autonomously at ≥95% accuracy on the automated slice**, measured on a held-out test set of real billed-and-paid charts, with confidence intervals, calibration, and error direction published.

Ask any vendor for the disclosure table above.

How to cite this report. Medmio Research. *Autonomous Coding You Can Verify: An Auditable Accuracy Benchmark for E/M Coding (Medmio EM-945).* Technical Report, July 2026 edition. `https://www.medmio.com/case-studies/codesight-accuracy-benchmark/`
All figures derive from versioned certification artifacts (n = 945, held-out test set n = 376), available for inspection under NDA. This is a living benchmark: each scheduled certification run is published as a new dated edition; earlier editions remain available at their original URLs.
Contact: press@medmio.com · medmio.com · linkedin.com/company/medmio

References

1. C. K. Chow. "On Optimum Recognition Error and Reject Tradeoff." *IEEE Trans. Information Theory*, 1970. doi:10.1109/TIT.1970.1054406
2. Y. Geifman, R. El-Yaniv. "Selective Classification for Deep Neural Networks." *NeurIPS* 2017. arXiv:1705.08500
3. R. El-Yaniv, Y. Wiener. "On the Foundations of Noise-free Selective Classification." *JMLR* 11, 2010. jmlr.org/papers/v11/el-yaniv10a
4. C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger. "On Calibration of Modern Neural Networks." *ICML* 2017. arXiv:1706.04599
5. B. Zadrozny, C. Elkan. "Transforming Classifier Scores into Accurate Multiclass Probability Estimates" (isotonic calibration). *KDD* 2002. doi:10.1145/775047.775151
6. H. Mozannar, D. Sontag. "Consistent Estimators for Learning to Defer to an Expert." *ICML* 2020. arXiv:2006.01862
7. J. Mullenbach, S. Wiegreffe, J. Duke, J. Sun, J. Eisenstein. "Explainable Prediction of Medical Codes from Clinical Text" (CAML). *NAACL* 2018. arXiv:1802.05695
8. C.-W. Huang, S.-C. Tsai, Y.-N. Chen. "PLM-ICD: Automatic ICD Coding with Pretrained Language Models." *ClinicalNLP* 2022. arXiv:2207.05289
9. J. Edin et al. "Automated Medical Coding on MIMIC-III and MIMIC-IV: A Critical Review and Replicability Study." *SIGIR* 2023. doi:10.1145/3539618.3591918
10. A. Soroush, B. S. Glicksberg, E. Zimlichman, et al. "Large Language Models Are Poor Medical Coders — Benchmarking of Medical Code Querying." *NEJM AI* 1(5), 2024. doi:10.1056/AIdbp2300040
11. J. Stausberg, N. Lehmann, D. Kaczmarek, M. Stein. "Reliability of diagnoses coding with ICD-10." *Int. J. Medical Informatics* 77(1), 2008. PMID 17185030
12. M. Peng et al. "Coding reliability and agreement of ICD-10 codes in emergency department data." *Int. J. Population Data Science*, 2018. PMC7299481
13. Hou et al. "Enhancing medical coding efficiency through domain-specific fine-tuned large language models." *npj Health Systems*, 2025. doi:10.1038/s44401-025-00018-3
14. E. Klang, I. Tessler, D. U. Apakama, et al. "Assessing Retrieval-Augmented Large Language Model Performance in Emergency Department ICD-10-CM Coding Compared to Human Coders." *medRxiv*, 2024. doi:10.1101/2024.10.15.24315526