MAIA

WHITEPAPER · MAIA'S INDEPENDENT AUDIT RESULTS W KZA

Getting Good at Orthopaedic Coding

What an independent 41-year coding authority found when it audited our AI, and what the published literature says about the baseline

Independently audited by Karen Zupko & Associates, Inc.

KZA

Maia · usemaia.com · November 2026

The Number Nobody Wants to Talk About

In 2014, the HHS Office of Inspector General published the results of a medical record review of a random sample of Medicare Part B evaluation and management claims from 2010. Certified professional coders adjudicated each claim against the underlying documentation.

42% were incorrectly coded. 19% lacked documentation. Together, 55% of E/M claims were miscoded, undocumented, or both. The errors ran both directions: 26% upcoded, 15% downcoded. Medicare paid $6.7 billion improperly, roughly a fifth of everything it spent on E/M in 2010.

That is not a story about bad actors. It shows how hard E/M coding was for trained humans working in real practices under real time pressure. It is also a snapshot: Medicare Part B claims only, from more than fifteen years ago, under E/M guidelines that have since been rewritten. It does not measure today's professional coders.

The specialty-specific data is no more reassuring.

In 2020, Greenky and colleagues at Rothman Orthopaedic Institute published a mock coding survey in *Orthopedics* using commonly encountered orthopaedic clinical scenarios, scored across visit type, modifiers, E/M, and CPT identification. **Practicing attending surgeons averaged 72.8% coding accuracy. Senior residents managed 51.0% accuracy. Junior residents, just 47.4% accuracy.**

These are the surgeons doing the operations. When it came to coding questions about their own specialty, more than one in four answers was wrong. Note the limits: the survey used hypothetical vignettes rather than real charts, and it tested surgeons and residents, not professional coders.

And the money leaks in both directions. When AAPC Audit Services began auditing office and outpatient visits under the 2021 E/M guidelines, they reported that nationwide **up to 19% of visit levels were being undercoded**. Practices are doing the work, in some cases even documenting the work, and then not submitting the correct codes for it. [Source: AAPC Knowledge Center, 2023.]

None of these studies measures today's professional coders directly, and their methods are too different to compare with one another. Together they show that orthopaedic coding is hard for trained people working under real conditions. That is the context in which any AI coding vendor, including us, should be judged.

So we hired the toughest grader we could find.

In June 2026, Maia engaged Karen Zupko & Associates, Inc. (KZA) to audit our AI's output. KZA has advised orthopaedic and surgical specialists since 1985. They are who practices call when a claim is complicated. They wrote the education that a lot of the coders reading this paper trained on.

KZA's certified consultants adjudicate Maia's coding against source documentation, not against what the practice billed. We do not get to have a human coder check the cases in advance. We do not get to argue with the findings. KZA reaches its findings using its own professional judgment, and every audit figure in this paper is reproduced from KZA's reports unchanged.

We did this because an accuracy claim is only as good as the test behind it, and we wanted ours checked by someone outside Maia.

The most useful question to ask about any accuracy number is what it was measured against. Agreement with previously billed encounters measures how closely a system matches past billing, errors included (see the studies above). Adjudication against source documentation measures whether the codes are actually supported.

29%	94%	5
Round 1 surgical overall accuracy	Round 3 surgical overall accuracy	KZA audits to date, including two weaker rounds summarized here

Source: KZA Orthopaedic Surgical Reviews, July to October 2026.

Round 1: We Failed

The first audit came back July 16, 2026.

E/M: 65% overall accuracy. Level assignment correct in 16 of 25 cases. Five overcoded by a level. Four carried an E/M that KZA found missing, unsupported, or in the wrong category. Audited RVU came in 42% below what we produced.

Surgical: 29% overall accuracy. Twenty-nine percent. CPT/HCPCS correct in 25 of 67 codes. ICD-10 correct in 34 of 78. Modifiers correct in 23 of 136.

The modifier findings were specific and they were ugly: assistant-at-surgery modifiers 80 and AS appended to the primary surgeon's codes, modifiers 51 and 59 stacked where one or neither belonged, modifiers hung on add-on code 29826, payment modifiers dropped into the second position instead of the first, missing FA-F9 finger modifiers on phalangeal work.

One detail matters more than the headline number. In Round 1 surgical, **aggregate RVU variance was near zero despite substantial coding errors.** (KZA calculates RVU variance as audited total RVUs minus our reported total RVUs, divided by what we reported. Overstated and understated codes can offset, which is how a near-zero variance can sit on top of real coding errors.) On a claim, that is the difference between getting paid and getting a post-payment audit. A model tuned to agreement with billed charges would have scored well here. KZA scored what the documentation supported, and we got a 29%.

We are publishing that number.

What Changed Between Rounds

After KZA's initial audit, Maia traced each finding back to its root cause and strengthened both halves of Maia's neuro-symbolic architecture. *"Medical coding looks like one AI problem, but it's really two: reading a clinical note is probabilistic inference, and applying CMS payment policy is a constraint problem," says Anson Antony, VP of AI at Maia. Large language models and fine-tuned orthopaedic models still read each note and*

interpret the narrative. Maia moved the mechanical coding decisions out of model judgment and into a deterministic policy layer built on CMS data. Machine learning engineers build and evaluate the models, coding experts rule on the hard calls, and KZA continues to independently audit the output every round, making for a feedback loop that continuously improves accuracy.

For surgical claims, Maia's rules now set modifiers on every line. Payment modifiers go first, and -51 and -59 can no longer attach to add-on codes or stack on each other. Digit and laterality modifiers are filled in from the documented procedure. Primary surgeon and assistant claims are now separated, so assistant modifiers appear only on the assistant's claim. Each assistant line must also pass the CMS payment indicator for its code, which blocks codes CMS never pays for an assistant. NCCI validation now checks every -59 against a real bundling partner on the claim, and the CMS Medically Unlikely Edits table caps units. Diagnosis laterality now comes from the official ICD-10 code book, and supported diagnoses are reported at the claim level. On the E/M side, imaging that is billed separately no longer also counts toward the visit's Data element. Modifier -25 no longer attaches to an E/M reported only with imaging, and modifiers follow one consistent sequence.

Round 2: The Retest

Same auditors, same standards, new charts.

E/M: 82% overall accuracy. Level assignment correct in 21 of 24. The other three were not level errors: in two, KZA found the visit inherent to a same-day office procedure and not separately reportable, and in one, a new-patient visit should have been coded as an established-patient visit. **Zero overcoded. Zero undercoded.** ICD-10 accuracy 42 of 47. RVU variance narrowed from -42% to -15%.

The zeros are the result we care about. An AI that overcodes is not a productivity tool, it is a compliance exposure with a monthly subscription. Across 24 E/M codes adjudicated by a CPMA against source documentation, KZA found no level assigned too high and none too low.

Surgical: 74% overall accuracy, up from 29%. CPT/HCPCS 34 of 44. Modifiers 34 of 37. ICD-10 38 of 62. Part of that gain reflects a different case mix (Round 2 included arthroplasty and spine, which need far fewer modifiers than Round 1's hand and upper-extremity work), so the improvement cannot be attributed entirely to remediation.

Remaining findings concentrated in diagnosis coding: Z96.6xx status codes reported on index arthroplasty rather than subsequent procedures, and incorrect 7th-character selection on injury codes, including one case reporting active treatment, subsequent encounter, and sequela characters simultaneously on a single encounter, which is self-contradictory on its face.

That surfaced a real defect. We implemented a fix after Round 2, but implementing a fix is not the same as proving it works. Round 3 partly confirms it. KZA found no Z96.6xx errors in the October sample, which included hip and knee arthroplasty, but one fracture case was again given a subsequent-encounter 7th character (D) where active treatment (A) applied, and KZA again recommended laterality and 7th-character edits. That part of the defect is not closed.

Round 3: Success

Round 3 is KZA's October 2026 audit, reported October 29, 2026. The surgical cases span hip and knee arthroplasty, knee arthroscopy, shoulder, cervical spine, peripheral nerve and distal radius procedures. E/M: 90% overall accuracy, with all 13 E/M levels correct, CPT/HCPCS 10 of 10, modifiers 9 of 9 and zero RVU variance. Surgical: 94% overall accuracy, with CPT/HCPCS 27 of 27, modifiers 31 of 31 and zero RVU variance. As with Round 2, the result reflects the cases in the sample as well as our changes, so the improvement cannot be attributed entirely to remediation, and samples this small can move a lot from one month to the next.

Diagnosis coding is still our weakest area. E/M ICD-10 accuracy was 75% in Round 3, down from 89% in Round 2 (it was 82% in August and 84% in September). None of the 20 diagnoses we reported was wrong: KZA found five documented pain diagnoses that we did not report, and a missed diagnosis counts against accuracy the same way a wrong one does. Surgical ICD-10 was 71%, 10 of 14 codes, with two laterality or 7th-character errors, one unsupported diagnosis and one missed diagnosis.

Full Results: Three Rounds

Metric	Round 1	Round 2	Round 3 (Oct.)
Evaluation & Management			
Overall accuracy	65%	82%	**90%**
E/M level correct	16/25 (64%)	21/24 (88%)	**13/13 (100%)**
Levels overcoded	5 (20%)	0	**0**
Levels undercoded	0	0	**0**
CPT/HCPCS correct	34/53 (64%)	29/39 (74%)	**10/10 (100%)**
Modifiers correct	38/58 (66%)	37/48 (77%)	**9/9 (100%)**
ICD-10-CM correct	46/72 (64%)	42/47 (89%)	**15/20 (75%)**
Total RVU variance	-42%	-15%	**0%**
Surgical			
Overall accuracy	29%	74%	**94%**
CPT/HCPCS correct	25/67 (37%)	34/44 (77%)	**27/27 (100%)**
Modifiers correct	23/136 (17%)	34/37 (92%)	**31/31 (100%)**
ICD-10-CM correct	34/78 (44%)	38/62 (61%)	**10/14 (71%)**
Total RVU variance	-0%	-6%	**0%**

What We Are and Are Not Claiming

We are going to be precise here, because the whole argument of this paper is that precision is what the category lacks.

We are not running a scoreboard. The OIG reviewed a national random sample of Medicare claims. Greenky surveyed 67 clinicians on hypothetical vignettes. KZA adjudicated 65 E/M dates of service and 60 operative reports of ours across the three rounds in the table. Three different instruments, three different populations, three different questions. Lining those percentages up side by side to declare a winner would be misleading, and we are not going to do it.

What the literature establishes is the difficulty of the task, not a bar we have cleared. Orthopaedic coding defeats attending surgeons, defeats residents, and produced a 42% error rate in a national sample of 2010 Medicare claims. That is the context in which any accuracy number, ours or anyone's, should be read.

Our sample is small. Sample sizes vary by round and by discipline. Rounds 1 and 2 each covered 25 E/M dates of service and 25 operative reports; Round 3 covered 15 and 10. Within those, the number of scored elements depends on the cases: 208, 158 and 52 E/M elements, and 281, 143 and 72 surgical elements, across Rounds 1, 2 and 3. KZA's reports state that findings are for probe and education purposes and that extrapolation beyond the examples is not supported. These results describe the charts reviewed. They are not a population estimate, and we will not present them as one.

Our surgical rounds are not directly comparable, and we are telling you rather than letting you find out. Round 1 was upper extremity: trigger finger, carpal tunnel, shoulder arthroscopy, hand and wrist. Round 2 included total knee and total hip arthroplasty, knee arthroscopy, and spine. Multi-digit hand surgery is modifier-dense in a way arthroplasty simply is not, and the denominators show it: 136 modifiers across 25 Round 1 cases, 37 across 25 Round 2 cases. Part of the move from 17% to 92% is better modifier logic. Part of it is a case mix requiring fewer modifiers. We cannot cleanly separate the two from these rounds. A like-for-like upper extremity re-audit is how we settle it, and we have asked for one. Round 3 mixed both, with hip and knee arthroplasty, knee arthroscopy, shoulder, cervical spine, peripheral nerve and distal radius procedures across only 10 cases, so the same caveat applies.

Five audits, three of them in the table, is still a small sample, not a settled trend.

Surgical ICD-10 at 61% is not good enough. We are not going to bury that under the 74%. Diagnosis coding is also where Round 3 slipped (see Round 3 above).

The auditor is a paid commercial partner. KZA is engaged under a paid monthly retainer to perform these audits and to consult on remediation, and KZA holds approval rights over this document. Separately, Maia and KZA have a commercial partnership that includes joint go-to-market activity, education and podcast collaboration. KZA determines its audit findings without our input, and every audit figure in this paper, including the 29%, is KZA's and is reproduced unchanged. Weigh the relationship for yourself.

Five Questions for Any Vendor, Including Us

1	**Who computed your accuracy number?** If the answer is "we did," ask who checked it, and how.
2	**Against what?** Billed codes, or source documentation? These are different tests and only one of them is hard.
3	**What was the sample, and how was it constructed?** Stratified? Cherry-picked? How many charts?
4	**Show me the bad round.** Every model has one. A vendor with only good numbers is a vendor showing you selected numbers.
5	**Will you publish audits before you've seen it?** We have committed to it at Maia.

What Happens Next

Audits continue monthly. Round 3 is complete, and it is reported above alongside a summary of the August and September audits. We intend to publish future rounds on the same terms as this one, including if they go badly. We think that commitment is worth more than any single percentage in this document.

About Maia

Maia is an AI-powered autonomous medical coding and revenue cycle management platform built exclusively for orthopaedic practices in the United States. Maia reads clinical and operative documentation to recommend CPT, ICD-10, and HCPCS codes with modifiers and clinical justification, and supports prior authorization, denial appeals, and collections. The platform is self-hosted on AWS and integrates with athenahealth and eClinicalWorks. usemaia.com

About KZA – Maia's Audit Partner

Founded in 1985, Karen Zupko & Associates, Inc. (KZA) is a nationally recognized leader in healthcare consulting and education, with more than 40 years of experience supporting physicians and healthcare organizations. KZA specializes in medical coding, documentation, reimbursement, revenue cycle management, and physician practice operations, with extensive expertise across surgical specialties, including orthopaedics. Through independent audits, expert consulting, and industry-leading education, KZA helps healthcare organizations strengthen compliance, improve financial performance, and navigate the complexities of today's healthcare environment.

References

1. Office of Inspector General, U.S. Department of Health and Human Services. *Improper Payments for Evaluation and Management Services Cost Medicare Billions in 2010.* OEI-04-10-00181, May 2014. https://oig.hhs.gov/reports/all/2014/improper-payments-for-evaluation-and-management-services-cost-medicare-billions-in-2010/

2. Greenky MR, Winters BS, Bishop ME, McDonald EL, Rogero RG, Shakked RJ, Raikin SM, Daniel JN, Pedowitz DI. Coding Education in Residency and in Practice Improves Accuracy of Coding in Orthopedic Surgery. *Orthopedics.* 2020;43(6):380-383. doi:10.3928/01477447-20200827-10. https://pubmed.ncbi.nlm.nih.gov/32882048/

3. AAPC. *Finding the Lost Dollar.* AAPC Knowledge Center, 2023. https://www.aapc.com/blog/87649-finding-the-lost-dollar/

4. Karen Zupko & Associates, Inc. *Orthopedic Evaluation and Management Coding Assessment, Maia Medical Billing Corp.* July 16, 2026; August 3, 2026; September 8, 2026; September 29, 2026 (final revised); and October 29, 2026.

5. Karen Zupko & Associates, Inc. *Orthopaedic Surgical Review, Maia Medical Billing Corp.* July 16, 2026; August 17, 2026; September 14, 2026; October 19, 2026; and October 29, 2026.