How We Know Our Training Labels Are Trustworthy

The Challenge

We're building a model to predict which patients will be diagnosed with colorectal cancer within the next 6 months. Identifying positive cases is straightforward—if a patient receives a CRC diagnosis (C18 colon, C19 rectosigmoid, C20 rectal), we see it in the record.

The harder question: *How do we confirm a patient stayed cancer-free?* Patients leave health systems, get diagnosed elsewhere, or simply don't return for follow-up. If we only used patients with perfect follow-up, we'd throw out most of our data.

Our Solution: Tiered Label Confidence

Rather than treating all "no diagnosis" patients the same, we categorize negative labels by how confident we are that they're truly negative:

Tier 1 – High Confidence (47% of negatives)

Patient returned 7-12 months after the observation date with no CRC diagnosis. This fully covers our 6-month prediction window—we *know* they were cancer-free.

Tier 2 – Medium Confidence (23% of negatives)

Patient returned 4-6 months after observation *and* has an established PCP relationship. The return visit covers most of the window, and the ongoing care relationship means their PCP would likely document a cancer diagnosis made elsewhere.

Tier 3 – Assumed Negative (30% of negatives)

Patient didn't return during the observation period, but has an established PCP. We assume no news is good news—their PCP relationship provides a documentation pathway for outside diagnoses.

Why Include Tier 3?

A reasonable question: *"Isn't it risky to assume patients without follow-up are cancer-free?"*

Three reasons we include them:

1. **Statistically:** Even if some Tier 3 labels are incorrect, the error rate is low enough that these patients add predictive signal rather than noise. A small amount of label noise in a large dataset is less harmful than a large reduction in training data.
2. **Pragmatically:** Excluding Tier 3 would discard 30% of our negative cases. That loss of training data hurts model performance more than the occasional mislabeled case.
3. **Clinically:** Patients with active PCP relationships who get diagnosed at outside facilities typically have that diagnosis communicated back to their primary care provider. The PCP relationship isn't perfect confirmation, but it's a reasonable proxy for ongoing health documentation.

The Bottom Line

Our training data includes approximately 2 years of patient observations with a 0.41% positive rate (about 1 CRC diagnosis per 250 patient observations). The three-tier system lets us maximize usable training data while being explicit about our confidence levels—rather than pretending all negatives are equally certain or discarding valuable data that doesn't meet an artificially strict threshold.