From Billing Codes to Specimens: Recovering Structure in Multi-Assignment Pathology Reports

Shaurya Bisht
Mentor: Joshua Levy, PhD
EDIT AI 2026 Mentor Track, Dartmouth Health

WHY THIS MATTERS

One pathology accession can contain multiple tissue specimens, and each label may recur in diagnosis, microscopy, comments, addenda, or special studies. The linked billing file provides only case-level quantities for CPT 88302-88309, so it cannot identify which text span or specimen produced each unit.

Research question
How much specimen structure can conservative, auditable rules recover from the complete linked cohort, and which report sections provide additional evidence for later count prediction?

STUDY DESIGN

Cohort

301,889 linked reports

782 discordant component cases removed

301,107 eligible reports

201,273 target-positive

77,314 multi-assignment

11,751 multiple categories

Prespecified strata

Target-positive: at least one 88302-88309 assignment

Multi-assignment: total target quantity at least 2

Multiple categories: at least two target CPT categories

Billing counts define complexity, not specimen labels.

Validation frame

96 reports prepared for CPT-blinded, candidate-hidden review; 24 prespecified for pathologist double review. Endpoints include specimen count, line partition, overlap-matched boundaries, and reviewer agreement.

Data handling

Row-level text and identifiers remained on Discovery; only aggregate results appear here. Source years were 2018-2020 and 2022, with no 2021 records.

METHOD

The deterministic parser:
recognizes final diagnosis, microscopy, comments, addenda, interpretations, and special-study sections

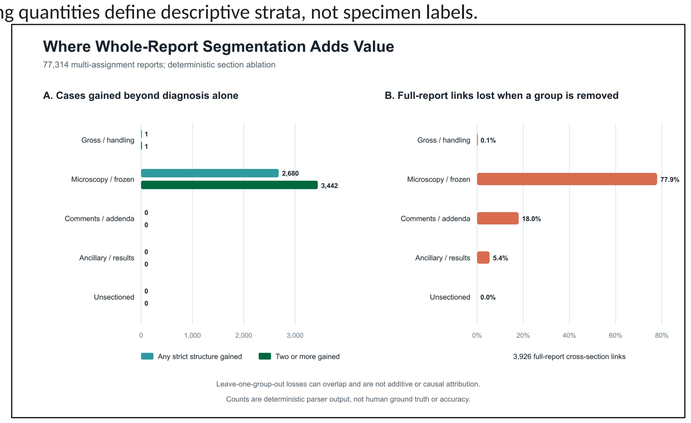
- requires explicit specimen anchors
- normalizes labels while preserving original text
- links repeated labels across report fields

preserves exact character offsets and source sections
flags repeated labels, mixed numbering, and ambiguous links

Fail-closed rule
Supporting sections may extend an anchored specimen, but they cannot create one.

Analyses
1. Whole-cohort structural observability
2. Diagnosis-only versus whole-report ablation
3. Billing-complexity stratification
4. Calendar-year standardization
5. Case-level count correspondence

Billing quantities define descriptive strata, not specimen labels.



Microscopy/frozen content accounted for 2,680 of 2,681 cases gained with any structure and 3,442 of 3,443 cases gained with at least two structures.

Interpretation

- Recoverability rose from 18.3% overall to 57.0% in multiple-category cases.

- Microscopy/frozen content contributed nearly every structure recovered beyond diagnosis alone.

- Most of the 2018-2022 increase remained after standardizing the billing-complexity mix.

- Complex reports are a high-yield validation set, but correspondence cannot replace expert boundary review.

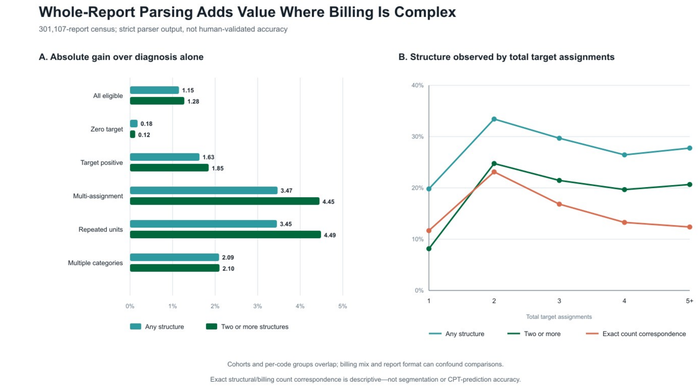
KEY RESULTS

Strict structural coverage
55,235 / 301,107 eligible reports (18.3%)
48,092 / 201,273 target-positive (23.9%)
23,573 / 77,314 multi-assignment (30.5%)
6,695 / 11,751 multiple categories (57.0%)

Whole-report gain
+2,681 cases with any structure
+3,443 with at least two structures
Microscopy/frozen content explained 2,680 and 3,442 of those gains.

Temporal robustness
14.8% in 2018 to 21.8% in 2022
5.3 percentage points within strata
1.7 percentage points from billing-mix change

Count correspondence
13,963 / 77,314 (18.1%)
Descriptive, not accuracy.



TAKE-HOME

Main finding
Whole-report section alignment recovered additional structure that diagnosis-only parsing missed, especially in microscopy and frozen-section content.

Contribution
The project provides a structural atlas, auditable code, exact offsets, section-ablation results, temporal analysis, and a frozen review design.

Remaining gap
Candidate boundaries have not been judged clinically correct, and billing counts do not identify the specimen that produced each unit. No accuracy claim is made before review.

Next step
Finish blinded review, then compare report-level and specimen-aware CPT quantity models on group-safe temporal splits.