CASE STUDY

EXPANDING LABORATORY REQUISITION OCR TEST COVERAGE WITH UDA-GENERATE

Client Overview

A U.S. Fortune 500 diagnostic information services provider processes laboratory requisitions received from physician offices, hospitals, patient service centers, and other healthcare organizations.

These requisitions contain information needed to identify the patient, determine the requested tests, associate specimens with the correct order, establish medical necessity, and bill the appropriate party. Information can include patient demographics, provider identifiers, test and diagnosis codes, specimen details, collection dates and times, insurance information, and billing selections.

The organization uses optical character recognition systems to extract this information from scanned laboratory requisitions and convert it into structured data for downstream processing.

Challenge: Testing OCR Across Real-World Document Conditions

Laboratory requisitions vary considerably in layout, content, and completion method. Some forms are system-generated, while others contain handwritten entries or annotations. Different test categories may also use specialized requisitions with unique fields, checkboxes, tables, and clinical-information sections.

The documents can reach the OCR system through multiple channels and in widely varying conditions. Fax transmission, repeated photocopying, blur, stains, low contrast, compression, skew, handwritten content, and other forms of degradation can make critical fields difficult to recognize.

OCR performance must remain reliable across these document templates, field combinations, and image-quality conditions. Even small extraction errors can create downstream exceptions, such as associating a specimen with the wrong patient, selecting an incorrect test, missing a diagnosis code, or capturing incomplete insurance information.

Testing with production requisitions creates privacy and security concerns because the documents contain protected health information (PHI). Production samples also provide limited control over the scenarios available for testing. Rare field combinations and difficult document conditions may not appear often enough to support comprehensive regression testing.

Manually creating and degrading test requisitions was time-consuming, inconsistent, and difficult to reproduce.

The Turning Point: Generating Synthetic Requisitions and Negative Variants

The organization adopted UDA-Generate to create realistic synthetic laboratory requisitions for OCR testing.

UDA-Generate populates production-faithful document templates with synthetic patient demographics, provider information, test codes, diagnosis codes, specimen details, collection dates, insurance identifiers, and billing selections. Because the generated values are synthetic, testing teams can exercise realistic document-processing workflows without exposing actual patient information.

UDA-Generate can also create more than 30 negative document variants representing conditions encountered in real-world OCR processing. These include faxed documents, blurred images, degraded photocopies, stains, handwritten content and annotations, and other image-quality challenges.

Testing teams can generate these variants systematically rather than waiting for suitable examples to appear in production. The same source document can be produced under multiple degradation conditions, allowing teams to isolate how each condition affects OCR performance.

The platform also maintains relationships between fields and documents. For example, patient identifiers appearing on a specimen label can be generated to match the corresponding requisition. Controlled mismatch scenarios can also be produced to verify that validation rules and exception-handling processes work correctly.

Each generated document has a known set of expected values. This provides ground truth that can be compared automatically with the OCR system's extracted output.

Results & Benefits

With UDA-Generate, the organization can:

- Generate large, reusable sets of synthetic laboratory requisitions without using production PHI.
- Test OCR extraction across multiple requisition layouts and laboratory-test categories.
- Evaluate OCR resilience against more than 30 negative document variants.
- Test fax artifacts, blur, photocopy degradation, stains, handwritten content, and other difficult document conditions.
- Validate patient names, dates of birth, addresses, provider identifiers, test codes, diagnosis codes, insurance information, and specimen details.
- Create positive, negative, boundary, and exception scenarios on demand.
- Test matching and mismatching patient identifiers across requisitions and specimen labels.
- Compare OCR output with known expected values to measure accuracy by field, document type, and degradation condition.
- Identify the document conditions and fields that produce the highest recognition-error rates.
- Reproduce failed scenarios consistently for defect investigation and regression testing.
- Reduce the time spent manually preparing, degrading, and reviewing OCR test documents.
- Expand test coverage while reducing dependence on sensitive production documents.

Conclusion and Next Steps

The organization currently uses UDA-Generate to create controlled, realistic laboratory-requisition datasets for OCR testing.

Synthetic documents provide the data variety required to evaluate extraction accuracy, while more than 30 negative variants allow teams to test how OCR behaves under difficult real-world conditions. Known expected values support automated validation, repeatable defect investigation, and consistent regression testing without exposing patient information.

Future priorities include:

- Expanding synthetic generation to additional requisition types and specialized laboratory workflows.
- Increasing coverage of complex combinations of document degradation and incomplete data.
- Testing OCR accuracy thresholds for each document type and negative variant.
- Adding more relationships between requisitions, specimen labels, insurance documents, and supporting clinical forms.
- Integrating synthetic document generation and OCR validation more deeply into automated testing and CI/CD pipelines.