Prescription database construction by data extraction from scanned files

[1]Noah Hallberg, [1]Ryan Kwong, [2]Ankur Malik, [1]Saul Means, [1]Udayan Pandey, [1]Bharath Sadagopan, [2]Kabir Snell, [1]Varon Srinivasan, [1]Margaret Wang

[1]Purdue University, [2]University of California Santa Barbara

PURDUE UNIVERSITY The Data Mine

inogen

Introduction

- Our corporate mentors tasked the joint Purdue and UCSB research group this semester to help them evaluate different methods in analyzing bulk data
- The goal was to help them chose the most optimal OCR solution for their needs
- Optical Character Recognition (OCR): using computer models to analyze data that tends to be handwritten

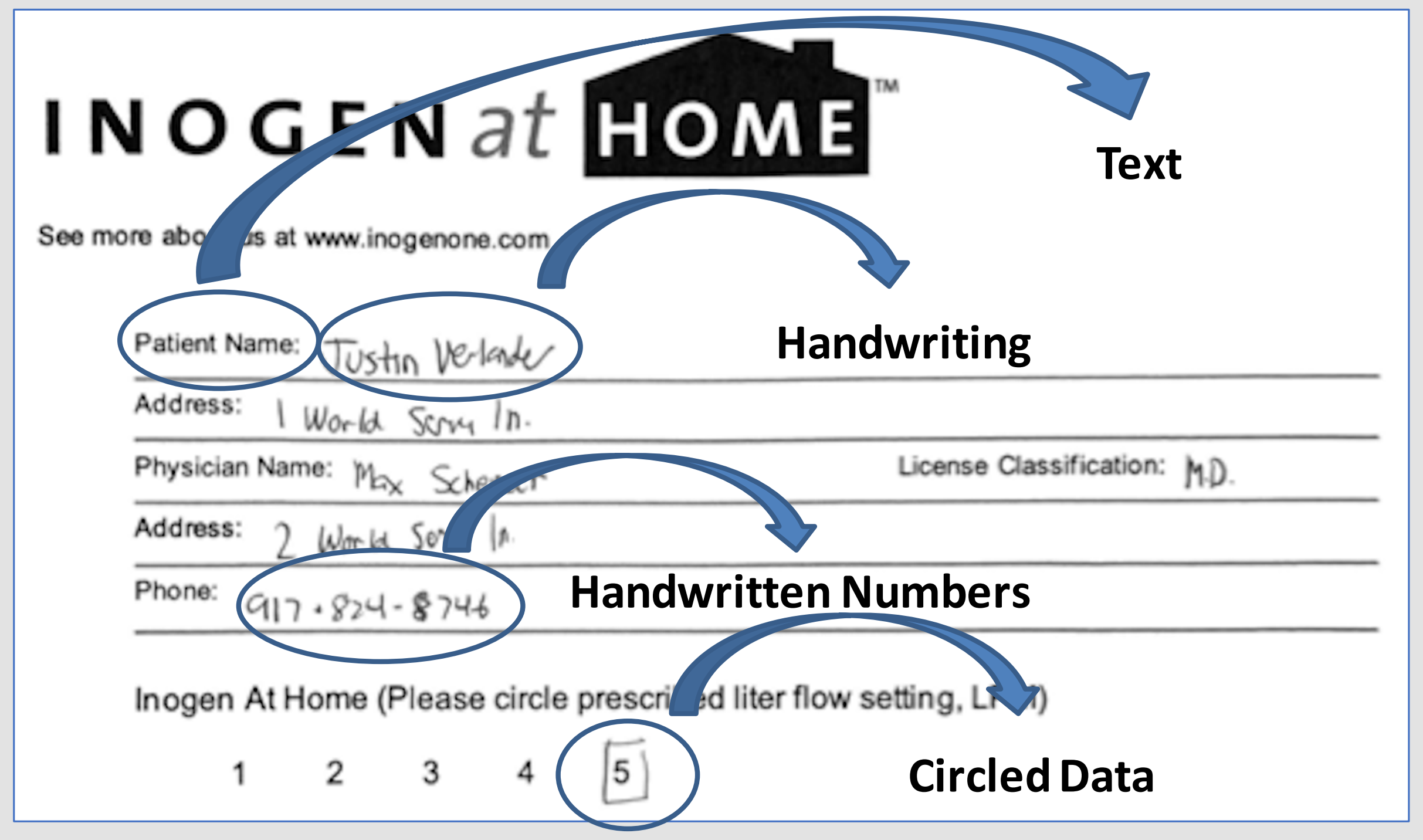
Methodology

Overview

- Creating our own mock database
- Criteria for choosing a software
- Testing the software

Three main software considerations:

1. HIPAA compliance 🔒
2. Performance
3. Price


Text
Handwriting
Handwritten Numbers
Circled Data

INOGEN at HOME™

See more abo[illegible]s at www.inogenone.com

Patient Name: Justin Velander

Address: 1 World Sorry ln.

Physician Name: Max Sche[illegible] License Classification: M.D.

Address: 2 World So[illegible] ln.

Phone: 917-824-8746

Inogen At Home (Please circle prescri[illegible]d liter flow setting, L[illegible])

1 2 3 4 [✓ 5]

No Inogen customer or patient data was used in this visual. Hypothetical patient information used for example purposes only.

- Evaluated each on a scale of 0 to 1
- 0 = failed entirely
- 1 = perfect performance

Evaluated Methods

Tesseract OCR (Python) :

- Package in python that implements Google's OCR software

Pros:

- Free (Open-Source Software)
- Highly customizable
- Very fast

Cons:

- Lower accuracy on OCR tests
- Collaboration becomes more complicated

Version 3.0

UiPath™

Version 2022.10.5

Automation platform with OCR Component (OCR scripting)

Pros:

- ~80% handwriting accuracy
- Effectively extracts to a CSV

Cons:

- Failure to read circles
- Requires a large amount of processing power

Machine Learning Optical Recognition Service (AWS)

Pros:

- Very good at handwriting
- Adaptable to numerous form types

Cons:

- Base version bad with circles
- Sometimes interprets scribbled out characters

Amazon Textract

2023 Version

Base64.ai

Version 2023

Artificial Intelligence OCR software

Pros:

- Handwriting and circle/box filling
- Easy, Support from Base64.

Cons:

- Circled Data
- Some issues with tables

OCR software

Pros:

- Business oriented model
- Trainable

Cons:

- Inferior performance compared to other software tested
- Does not limit outputs to a lexicon of English words or the Latin alphabet

ABBYY®

Fine Reader 16

Cloud Vision API

Version 2023

Document AI – machine learning, OCR platform

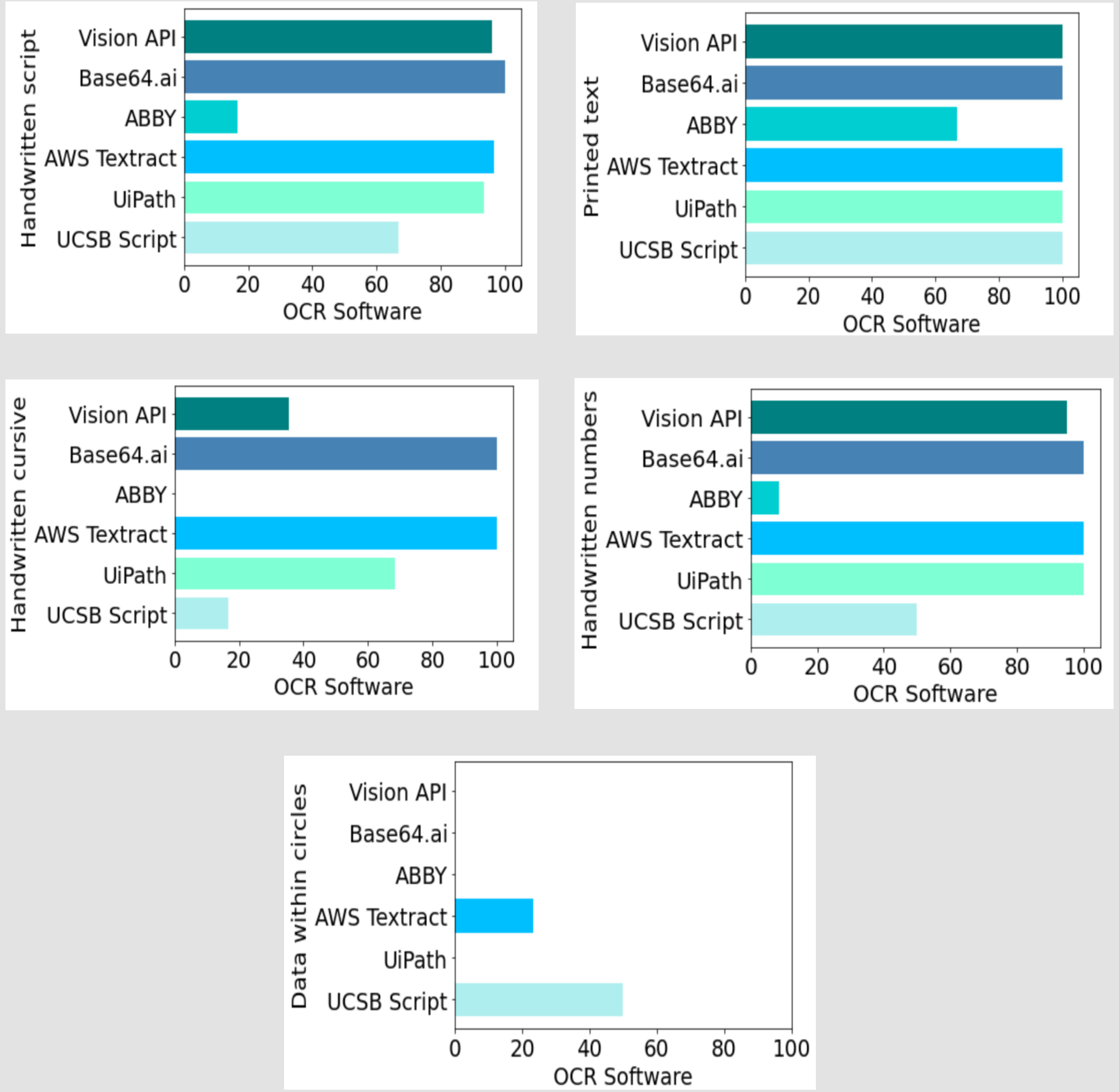
Pros:

- Confident text extraction
- Identified symbols and room for ML

Cons:

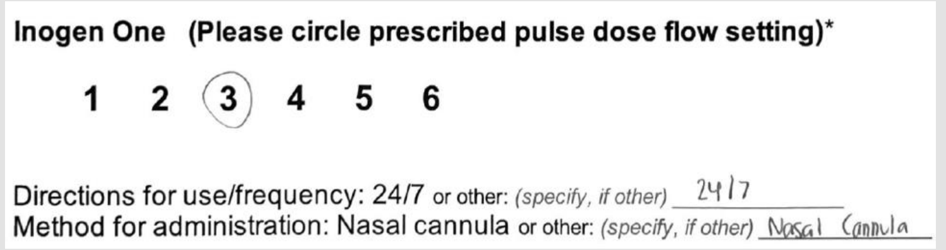
- Poor results with circle data
- Unclear data sorting

Combined Results (%Accuracy)



Current Work

Working with companies to get customized solutions:



Inogen One (Please circle prescribed pulse dose flow setting)*

1 2 [✓ 3] 4 5 6

Directions for use/frequency: 24/7 or other: *(specify, if other)* 24/7

Method for administration: Nasal cannula or other: *(specify, if other)* Nasal Cannula

-Circled values extraction

-Flexible value identification (circled or written)

Working on improving Python solution:

- Testing new packages
- Identifying regions of interest
- Improving hand-writing accuracy
- Image pre-processing

References

https://docs.aws.amazon.com/textract/index.html
https://guides.nyu.edu/tesseract/usage
https://base64.ai/features/data-extraction-api/handwritten-document

Conclusions

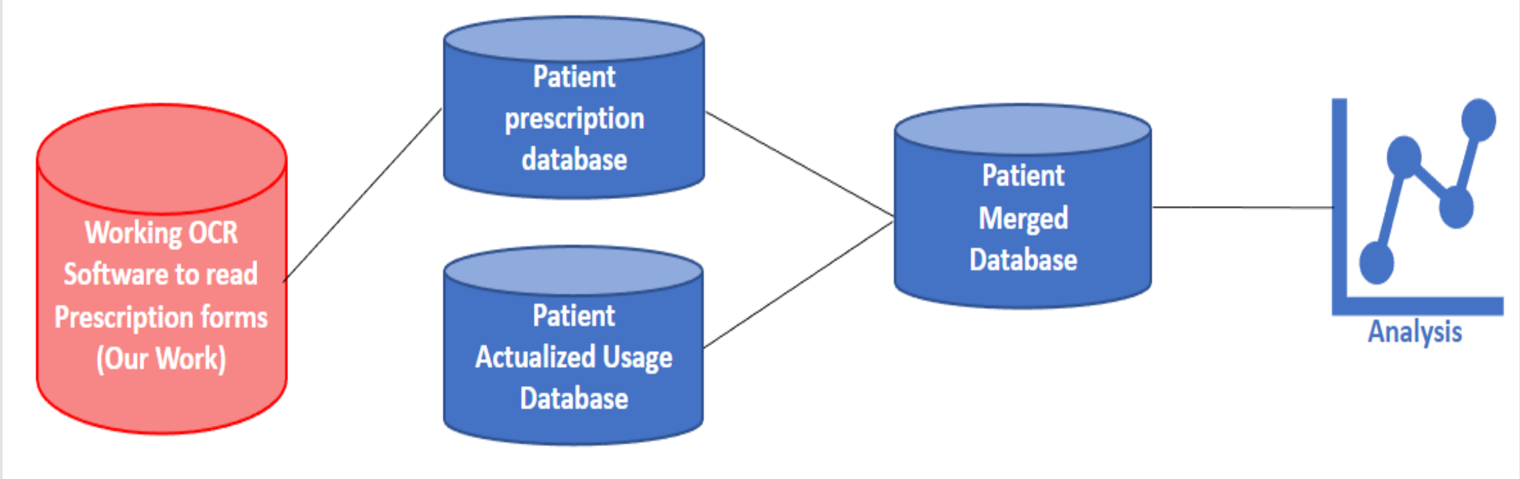
- Created and tested mock patient prescription PDFs
- AWS and Base64.ai were the most accurate and flexible software
- Tesseract (python) and Google API can be explored
- Circled values weren't recognized by any method

Future Goals

We are creating a tool that allows Inogen to generate a patient prescription database

This new database can help them answer questions like:

- What proportion of patients are adhering to their prescribed flow setting?
- What kind of patients are not adhering to their prescribed flow setting?
- In cases of inconsistency with flow setting and prescription; how is the flow setting being misused (higher or lower)?


Working OCR Software to read Prescription forms (Our Work)
Patient prescription database
Patient Actualized Usage Database
Patient Merged Database
Analysis

Acknowledgements

Purdue Datamine Team
- Paula Sarmiento M.Sc.
- Kush Malhotra
- David Glass M.Sc.
- Maggie Betz
- Nicholas Lenfestey MPH

UCSB Capstone
- Trevor Ruiz Ph.D
- Tim Robinson

Inogen
- Norbert Leinfellner Ph.D.
- Stanislav Glezer M.D. MBA
- Horst Pichler M.Sc.
- Juhi Madan M.Sc.
- Niloufar Rajabshirazi M.Sc.
- Eddy Choi