An Phríomh-Oifig Staidrimh Central Statistics Office

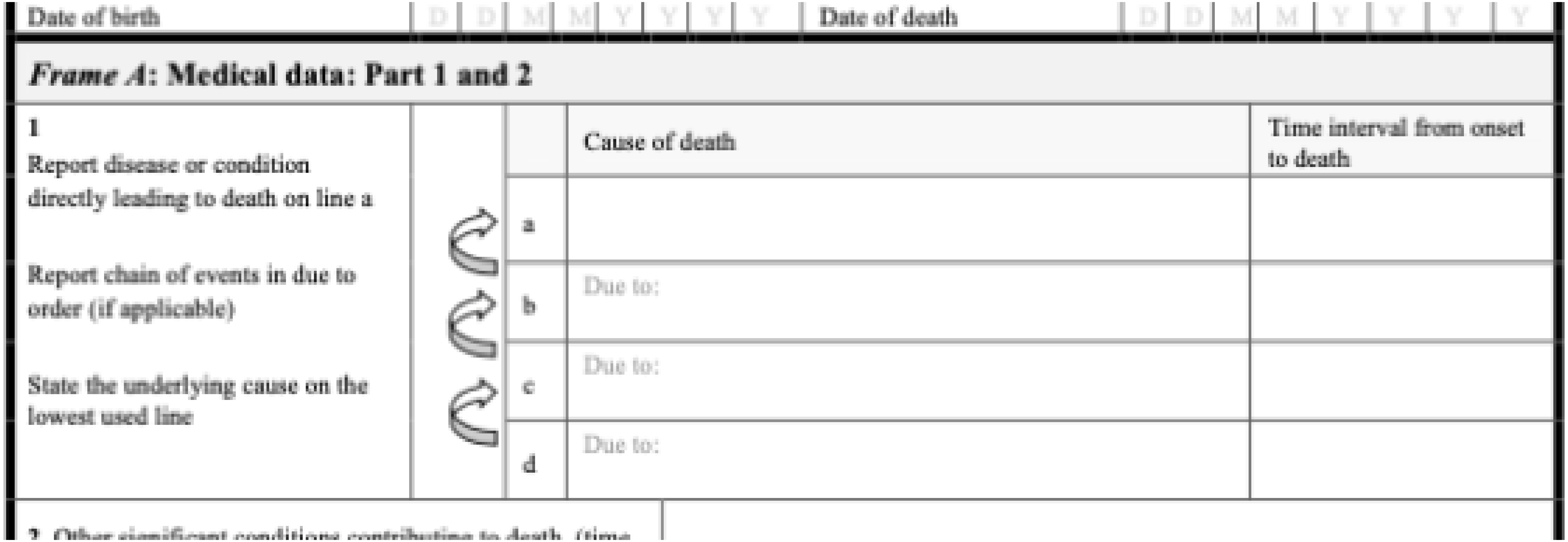
Mortality Statistics and Cause of Death Classifications

Sean O’Connor Life Events and Demography, Central Statistics Office, Ireland

Overview

- CSO produces mortality statistics on quarterly basis.
- Deaths are coded to International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10).
- Each record is assigned an Underlying cause of Death code (UCOD) - ~35k deaths a year in Ireland
- All the narrative on the death certificate are coded also as there is a causal relationship needed to be followed to get the UCOD.

ICD- 10



Date of birth	D	D	M	M	Y	Y	Y	Y	Date of death	D	D	M	M	Y	Y	Y	Y		

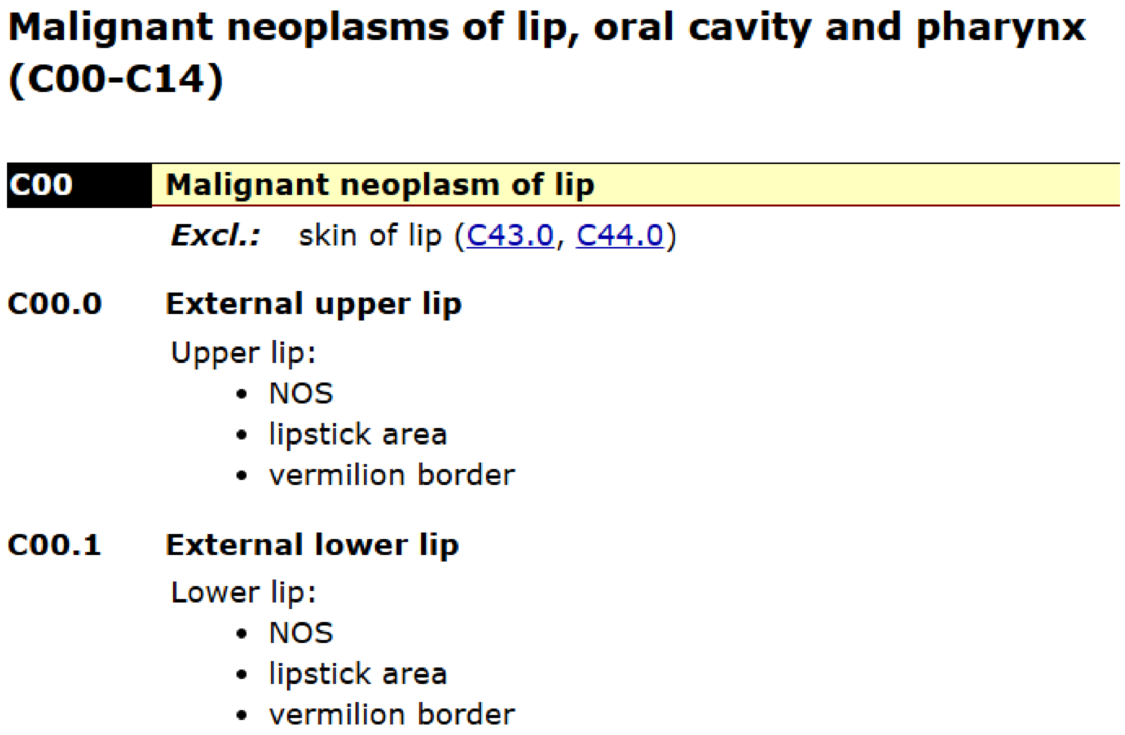
***Frame A*: Medical data: Part 1 and 2**

1 Report disease or condition directly leading to death on line a		Cause of death	Time interval from onset to death
	a		
Report chain of events in due to order (if applicable)	b	Due to:	
State the underlying cause on the lowest used line	c	Due to:	
	d	Due to:	

2 Other significant conditions contributing to death (time

ICD-10

- Close to 8k unique ICD-10 codes.



Malignant neoplasms of lip, oral cavity and pharynx (C00-C14)

C00 **Malignant neoplasm of lip**

Excl.: skin of lip (C43.0, C44.0)

C00.0 **External upper lip**

Upper lip:
- NOS
- lipstick area
- vermilion border

C00.1 **External lower lip**

Lower lip:
- NOS
- lipstick area
- vermilion border

Coding

- Coding is carried out in IRIS which is standard internationally used death coding software.
- IRIS can code around ~55% of the death certs automatically.
- The remaining certs are manually coded by CSO mortality coders.

Common Issues

- Lack of standardisation in how narratives are written.
  - Some can be overly wordy.
  - Multiple ways time intervals will be included (5 weeks, five weeks, 5 wks, 35 days etc.)
- Frequent misspellings for various medical terms.

Project

- Could we improve our rate of automation and lessen the number of certs which needed to be manually coded?
- Could we develop a tool which could assist coders with their manual classification work?

An Phríomh-Oifig Staidrimh
Central Statistics Office

ML-Assisted Cause of Death Classification

Labhaoise Barrett, Life Events and Demography, Central Statistics Office, Ireland

Project Team

This project was a collaborative effort between the Life Events and Demography Division and the Data Science Team at the Central Statistics Office (CSO). The technical development was led by:

Sarah Murphy, *Statistician*

James Kelly, *Graduate*

With thanks to additional colleagues who provided support throughout the project

Contents

1. Introduction to the problem
2. Overview of Solutions:
   - Spell Check and Time Intervals
   - Machine Learning development and Results
   - User friendly recommender

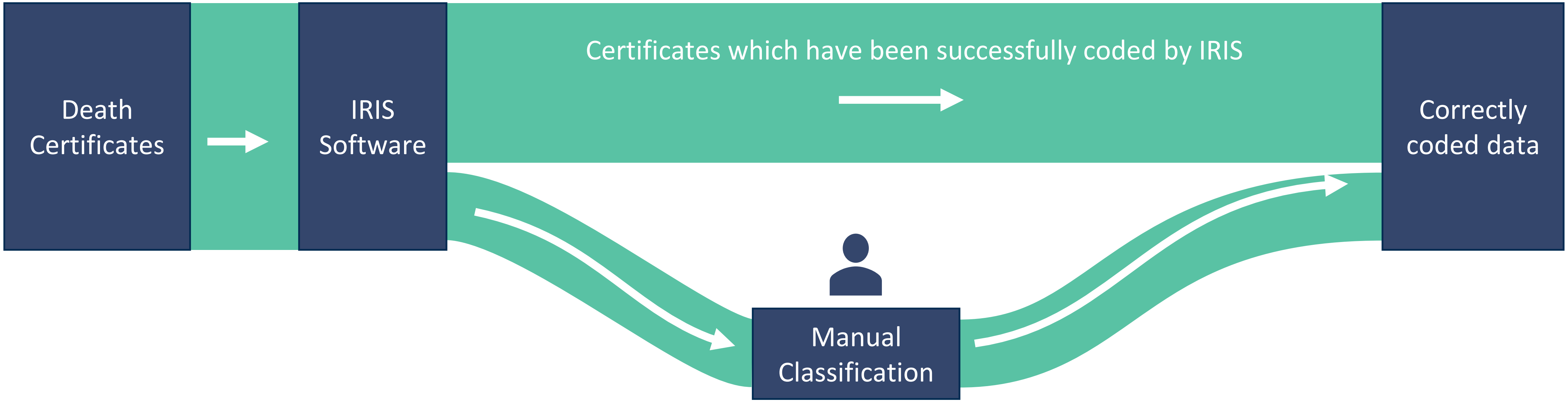
Goal

- To help the human coders with their classification of death certificates

Solution

- A predictor tool based on a supervised machine learning classification model to provide a shortlist of likely ICD-10 codes.

Existing process



Death Certificates

IRIS Software

Certificates which have been successfully coded by IRIS

Manual Classification

Correctly coded data

The Data

Challenges:

- 80,000 individual lines of natural text describing a disease a year.
- The text itself was messy, requiring a cleaning process before any model could be applied.
- There are multiple ways of describing the same disease, including standard and non-standard acronyms.

Upsides:

- There are already many previously correctly classified text lines, either by the IRIS system (55%) or by our manual coders (45%).
- This provides a great base to train and then test our classifier.

Data Cleaning and Preprocessing

Spell Check:

- Levenshtein distance for similarity
- Prefer medical dictionary matches
- Handle case, apostrophes, and duplicated characters
- Use Soundex for phonetic similarity

Time Intervals:

- Standardised natural language inputs of time intervals and dates, for example first Jan 21 becomes January 2021

Limiting the Classification Space

- The ICD-10 system includes thousands of possible causes of death, but only a limited number occur frequently in Ireland.
- Rare codes are underrepresented in our data, making them unsuitable for model training.
- Including too many codes introduced labels that were virtually unseen in training data
- This underscores the importance of expert human coders in identifying and managing rare cases

Feature Engineering — TF-IDF

- Vectorisation: TF-IDF
- Includes unigrams and bigrams
- Sublinear term frequency scaling
- Removes stopwords and rare terms

Algorithm Candidates

Evaluated Models:

– Logistic Regression

– Random Forest

– Stochastic Gradient Descent Classifier (SGDClassifier)

– XGBoost (initially considered)

Metric: Metric: F2 Score (F-beta with β = 2) to emphasise recall

Validation: Nested Cross-Validation

Evaluation Strategy

Nested Cross Validation:

- Outer loop: 3-fold cross-validation to estimate generalisation
- Inner loop: Grid search to select best parameters
- Final model refit on full data

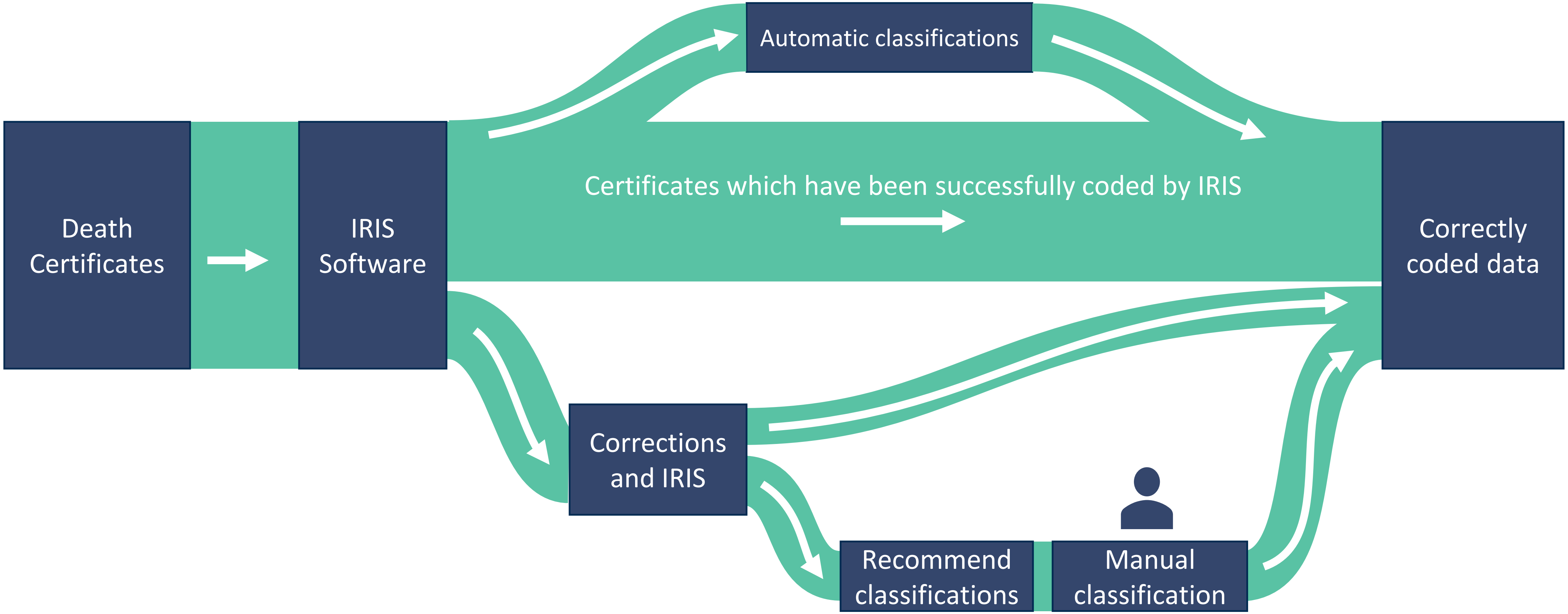
Model Results

Model	Performance	Results
Random Forest	Highest overall	Mean F2 Score: 0.86
Logistic Regression	Comparable	< Random Forest
SGDClassifier	More variable	Lower and less consistent

Pipeline Overview

1. Clean raw text
2. Vectorise with TF-IDF
3. Train with cross-validation
4. Tune hyperparameters
5. Save best model
6. Weekly inference + human review in Shiny app

The Pipeline


Automatic classifications
Death Certificates
IRIS Software
Certificates which have been successfully coded by IRIS
Correctly coded data
Corrections and IRIS
Recommend classifications
Manual classification

In Production

Previously:

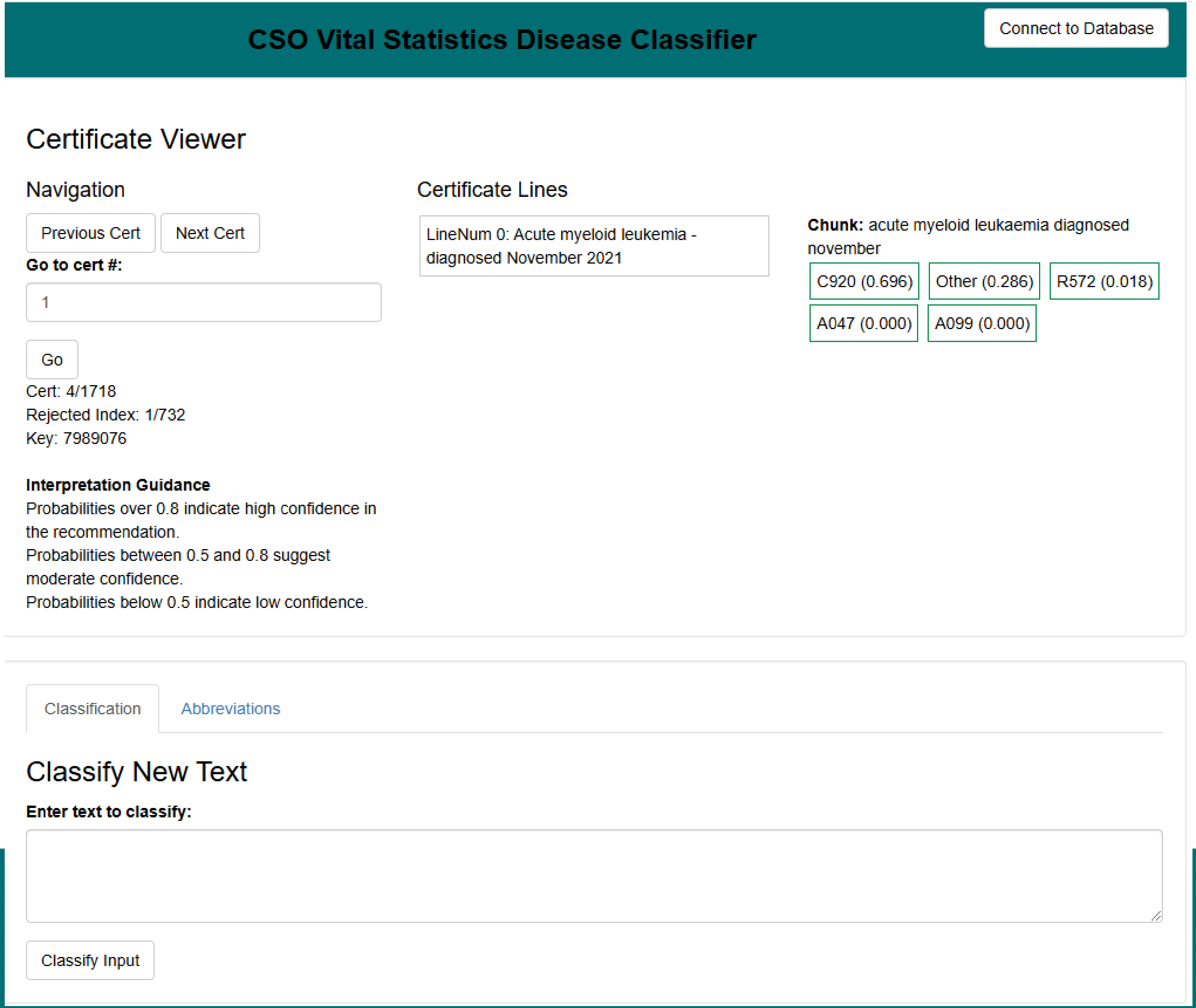
- Coders searched long PDFs or ICD databases

Now:

- Shiny app suggests likely codes
- Coders can review and override suggestions
- Interface supports their judgment, not replaces it

User Interface

Built in R Shiny



CSO Vital Statistics Disease Classifier

Certificate Viewer

Navigation

Go to cert #: 1

Cert: 4/1718
Rejected Index: 1/732
Key: 7989076

Interpretation Guidance
Probabilities over 0.8 indicate high confidence in the recommendation.
Probabilities between 0.5 and 0.8 suggest moderate confidence.
Probabilities below 0.5 indicate low confidence.

Certificate Lines

LineNum 0: Acute myeloid leukemia - diagnosed November 2021

Chunk: acute myeloid leukaemia diagnosed november

C920 (0.696) | Other (0.286) | R572 (0.018) | A047 (0.000) | A099 (0.000)

Classify New Text

Enter text to classify:

Questions?

Contact:
sean.oconnor@cso.ie
labhaoise.barrett@cso.ie
demography@cso.ie