Natural Language Processing Methods to Parse Birth and Death Records

Matthew Jin[1], Justina Zurauskiene[2], and Zeynep Madak-Erdogan[3]

[1]Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, [2]Institute of Cancer and Genomic Sciences, University of Birmingham, [3]Department of Food Science and Human Nutrition, University of Illinois at Urbana-Champaign

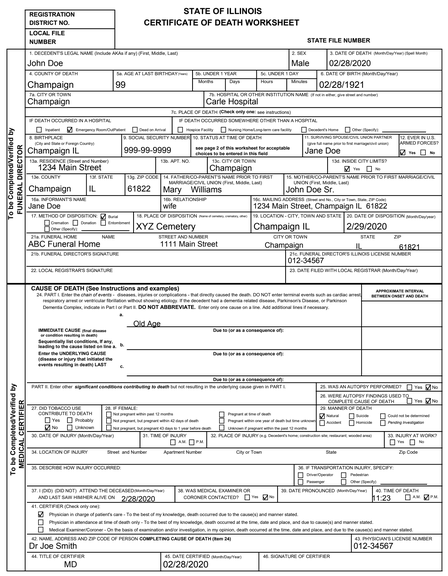
INTRODUCTION & AIMS

The goal of this project is to investigate health disparities arising from birth and death outcomes that exist across the residents of Champaign County through the parsing of electronic birth and death records provided by the Champaign Urbana Public Health District (CUPHD) from the last decade. This project is currently focused on identifying maternal health disparities and how these disparities are associated with various birth outcomes. To accomplish this task, we develop a pipeline for digitizing birth and death records by using natural language processing (NLP), which allows our program to read and understand the language in the birth and death records.

This research will provide an autonomous tool for future projects, which require the parsing of not only electronic birth and death records but also other significant records in medicine.

A sample birth certificate file, which is fake filled for training purposes, is shown below. Note the complexity of the document. The data for the fields vary in type, shape, and length. For example, data may come in the form of checks crossed off in small boxes.

SAMPLE BIRTH CERTIFICATE FILE



REGISTRATION DISTRICT NO. / LOCAL FILE NUMBER

STATE OF ILLINOIS
CERTIFICATE OF DEATH WORKSHEET

STATE FILE NUMBER

To Be Completed/Verified by FUNERAL DIRECTOR

1. DECEDENT'S LEGAL NAME (Include AKAs if any) (First, Middle, Last)			2. SEX	3. DATE OF DEATH (Month/Day/Year) (Spell Month)	
John Doe			Male	02/28/2020	
4. COUNTY OF DEATH	5a. AGE AT LAST BIRTHDAY (Years)	5b. UNDER 1 YEAR (Months / Days / Hours / Minutes)	5c. UNDER 1 DAY	6. DATE OF BIRTH (Month/Day/Year)	
Champaign	99			02/28/1921	
7a. CITY OR TOWN	7b. HOSPITAL OR OTHER INSTITUTION NAME (If not in either, give street and number)				
Champaign	Carle Hospital				
7c. PLACE OF DEATH (Check only one; see instructions)					
8. BIRTHPLACE (City and State or Foreign Country)	9. SOCIAL SECURITY NUMBER	10. STATUS AT TIME OF DEATH		11. SURVIVING SPOUSE/CIVIL UNION PARTNER (give full name prior to first marriage/civil union)	12. EVER IN U.S. ARMED FORCES?
Champaign IL	999-99-9999	see page 2 of this worksheet for acceptable choices to be entered in this field		Jane Doe	Yes / No
13a. RESIDENCE (Street and Number)	13b. APT. NO.	13c. CITY OR TOWN		13d. INSIDE CITY LIMITS?	
1234 Main Street		Champaign		Yes / No	
13e. COUNTY	13f. STATE	13g. ZIP CODE	14. FATHER/CO-PARENT'S NAME PRIOR TO FIRST MARRIAGE/CIVIL UNION (First, Middle, Last)	15. MOTHER/CO-PARENT'S NAME PRIOR TO FIRST MARRIAGE/CIVIL UNION (First, Middle, Last)	
Champaign	IL	61822	Mary Williams	John Doe Sr.	
16a. INFORMANT'S NAME		16b. RELATIONSHIP	16c. MAILING ADDRESS (Street and No., City or Town, State, ZIP Code)		
Jane Doe		wife	1234 Main Street, Champaign IL 61822		
17. METHOD OF DISPOSITION: Burial / Cremation / Donation / Entombment / Other (Specify)	18. PLACE OF DISPOSITION (Name of cemetery, crematory, other)		19. LOCATION - CITY, TOWN AND STATE	20. DATE OF DISPOSITION (Month/Day/Year)	
	XYZ Cemetery		Champaign IL	2/29/2020	
21a. FUNERAL HOME NAME	STREET AND NUMBER	CITY OR TOWN	STATE	ZIP	
ABC Funeral Home	1111 Main Street	Champaign	IL	61821	
21b. FUNERAL DIRECTOR'S SIGNATURE			21c. FUNERAL DIRECTOR'S ILLINOIS LICENSE NUMBER		
			012-34567		
22. LOCAL REGISTRAR'S SIGNATURE			23. DATE FILED WITH LOCAL REGISTRAR (Month/Day/Year)		

To Be Completed/Verified by MEDICAL EXAMINER

CAUSE OF DEATH (See Instructions and examples)

24. PART I. Enter the chain of events - diseases, injuries or complications - that directly caused the death. DO NOT enter terminal events such as cardiac arrest, respiratory arrest or ventricular fibrillation without showing etiology. If the decedent had a dementia-related disease, Parkinson's Disease, or Parkinson Dementia Complex, indicate in Part I or Part II. **DO NOT ABBREVIATE.** Enter only one cause on a line. Add additional lines if necessary.

		Approximate interval between onset and death
a. Old Age		
Immediate cause (final disease or condition resulting in death)	Due to (or as a consequence of):	
Sequentially list conditions, if any, leading to the cause listed on line a. Enter the UNDERLYING CAUSE (disease or injury that initiated the events resulting in death) LAST	b.	
	Due to (or as a consequence of):	
	c.	
	Due to (or as a consequence of):	

PART II. Enter other **significant conditions contributing to death** but not resulting in the underlying cause given in PART I.

25. WAS AN AUTOPSY PERFORMED? Yes / No
26. WERE AUTOPSY FINDINGS USED TO COMPLETE CAUSE OF DEATH? Yes / No
29. MANNER OF DEATH: Natural / Accident / Suicide / Homicide / Could not be determined / Pending Investigation

27. DID TOBACCO USE CONTRIBUTE TO DEATH: Yes / Probably / No / Unknown
28. IF FEMALE: Not pregnant within past 12 months / Not pregnant, but pregnant within 42 days of death / Not pregnant, but pregnant 43 days to 1 year before death / Pregnant at time of death / Pregnant within one year of death but time unknown / Unknown if pregnant within the past 12 months
30. DATE OF INJURY (Month/Day/Year)
31. TIME OF INJURY: A.M. / P.M.
32. PLACE OF INJURY (e.g. Decedent's home, construction site, restaurant, wooded area)
33. INJURY AT WORK? Yes / No
34. LOCATION OF INJURY: Street and Number / Apartment Number / City or Town / State / Zip Code
35. DESCRIBE HOW INJURY OCCURRED:
36. IF TRANSPORTATION INJURY, SPECIFY: Driver/Operator / Passenger / Pedestrian / Other (Specify)
37. I (DID) (DID NOT) ATTEND THE DECEASED (Month/Day/Year) AND LAST SAW HIM/HER ALIVE ON: 2/28/2020
38. WAS MEDICAL EXAMINER OR CORONER CONTACTED? Yes / No
39. DATE PRONOUNCED (Month/Day/Year)
40. TIME OF DEATH: 1:23 A.M. / P.M.

41. CERTIFIER (Check only one)
- Physician in charge of patient's care - To the best of my knowledge, death occurred due to the cause(s) and manner stated.
- Physician in attendance at time of death only - To the best of my knowledge, death occurred at the time, date and place, and due to cause(s) and manner stated.
- Medical Examiner/Coroner - On the basis of examination and/or investigation, in my opinion, death occurred at the time, date and place, and due to the cause(s) and manner stated.

42. NAME, ADDRESS AND ZIP CODE OF PERSON **COMPLETING CAUSE OF DEATH (Item 24)**: Dr Joe Smith
43. PHYSICIAN'S LICENSE NUMBER: 012-34567
44. TITLE OF CERTIFIER: MD
45. DATE CERTIFIED (Month/Day/Year): 02/28/2020
46. SIGNATURE OF CERTIFIER

METHOD

The pipeline will receive the path to a folder of scanned birth certificate files in PDF format. These files are two pages each. First, we convert these PDF files into image files (.jpeg).

Then, we remove the horizontal and vertical lines in the image files by finding all horizontal and vertical contours (cv2.findContours()) in the image and "erasing" them by drawing the same contours (cv2.drawContours()) with the color white, effectively blurring the horizontal and vertical lines.

With the processed image files, the pipeline calculates the size of the document of each page and zooms in on different parts of the birth certificate image to allow for more accurate character recognition.

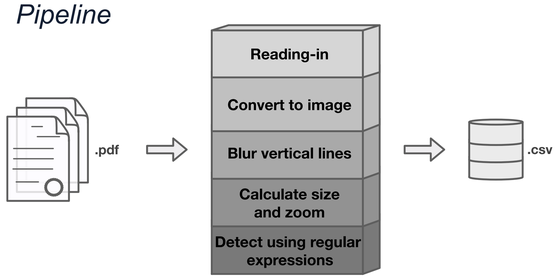
We use Tesseract, an Optical Character Recognition (OCR) software, to read in the zoomed-in portion of the birth certificate. At this step, the pipeline has a string of text describing this zoomed in portion.

Finally, the pipeline uses regular expressions pattern matching to extract desired patterns such as dates, city names, and more.

Regular Expression Pattern

```
re.compile(r'[A-Za-z]+.\d{2},.\d{4}')
```

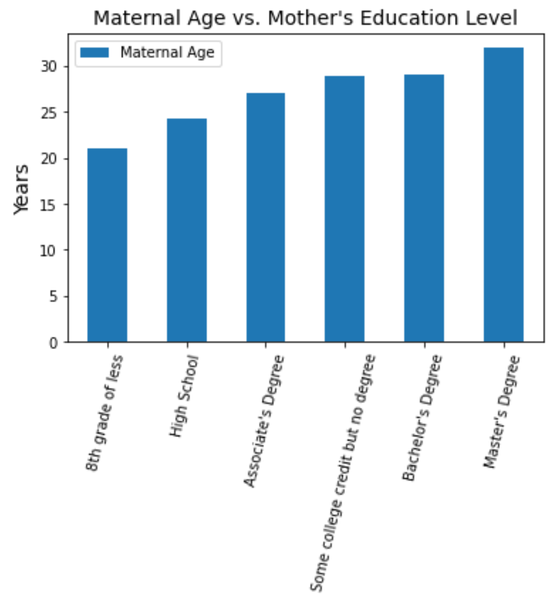
Diagram of Pipeline

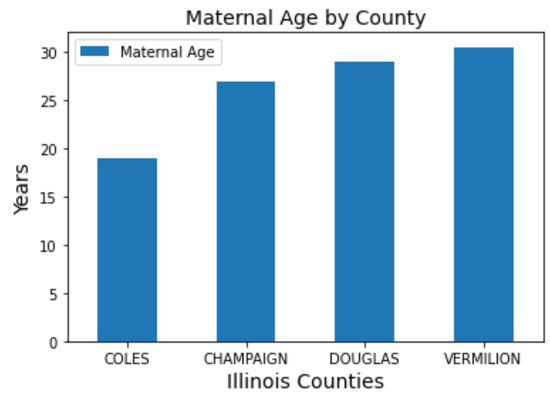


RESULTS

The pipeline is currently able to extract the first page of birth certificate files with over 98% accuracy. The first page contains fields with mostly socioeconomic data. Currently, our CSV database reveals patient data using a subset of 21 scanned birth certificate records provided by the CUPHD.

On the first page of the birth certificate file, there is information about the date of birth of the child and the mother. With this information, we can plot how the maternal age varies depending on different fields of interest.





DATABASE

The data pipeline will ultimately output a CSV database of patient information. This table is stored as a Pandas DataFrame during the execution of the pipeline.

The row names of the DataFrame are filenames with unique numbers for patient identification. The column names of the DataFrame are the field names within the birth certificate and death certificate files. A snippet of the birth data table is show below.

	21. BIRTHING PARENT'S EDUCATION
Birth cert 12B.pdf	Some college credit but no degree
Birth cert 3B.pdf	Master's Degree
Birth cert 2B.pdf	High School
Birth cert 13B.pdf	Some college credit but no degree
Birth cert 5B.pdf	Bachelor's Degree

FUTURE DIRECTION

- Continue to extract data from birth certificate files, including the second page of these files which contains the medical history of the mother and infant
- Analyze and relate different fields of interest with natural language processing and deep learning
- Start parsing data from death records to create a new CSV database

ACKNOWLEDGEMENTS

NCSA SPIN Program
- Dr. Zurauskiene
- Dr. Madak-Erdogan

Champaign Urbana Public Health District

ILLINOIS