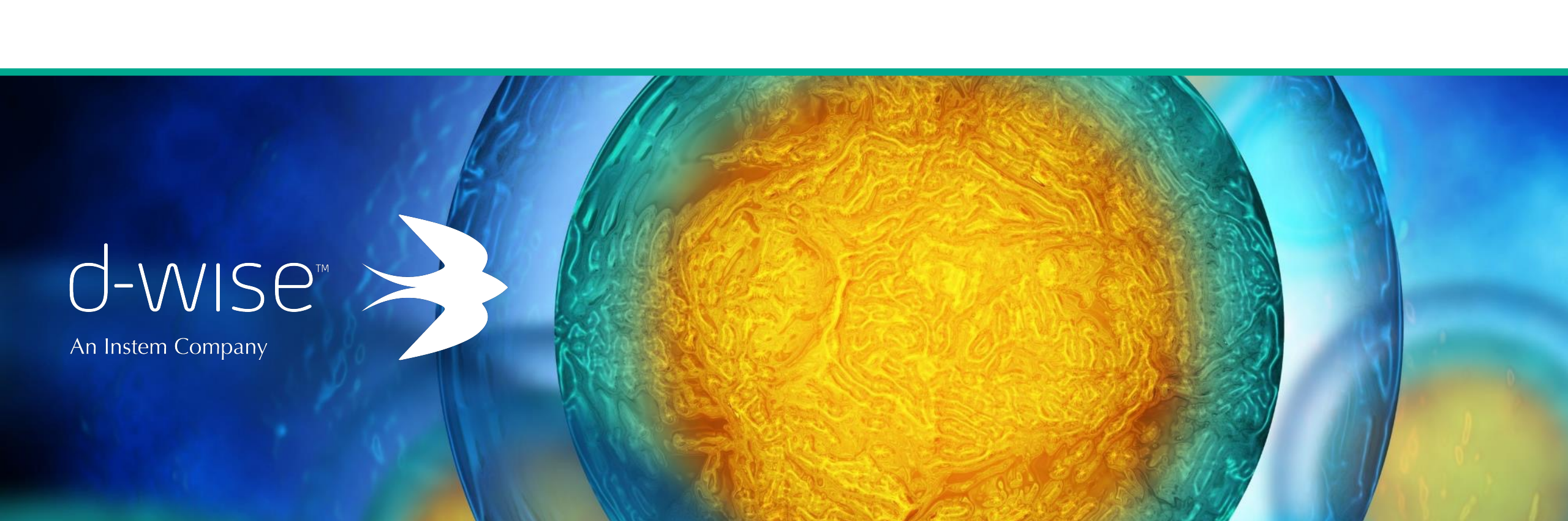




d-wise™

An Instem Company



See no data, hear no risk.

Cathal Gallagher

Potential Title

I got 99 problems, but measuring risk of reidentification of patients in public release documents when you don't have access to data... ain't one



Disclaimer

The opinions expressed in this presentation and on the following slides are solely those of the presenter and not necessarily those of d-wise/Instem.

d-wise/Instem do not guarantee the accuracy or reliability of the information provided herein.

k-anonymity

k-anonymity is a property possessed by certain anonymized data. The concept of k-anonymity was first introduced by Latanya Sweeney and Pierangela Samarati in a paper published in 1998.

A release of data is said to have the k-anonymity property if the information for each person contained in the release cannot be distinguished from at least $k - 1$ individuals whose information also appear in the release.

	Race	Birth	Gender	ZIP	Problem
t1	Black	1965	m	0214*	short breath
t2	Black	1965	m	0214*	chest pain
t3	Black	1965	f	0213*	hypertension
t4	Black	1965	f	0213*	hypertension
t5	Black	1964	f	0213*	obesity
t6	Black	1964	f	0213*	chest pain
t7	White	1964	m	0213*	chest pain
t8	White	1964	m	0213*	obesity
t9	White	1964	m	0213*	short breath
t10	White	1967	m	0213*	chest pain
t11	White	1967	m	0213*	chest pain

Quasi Identifiers



Sex



Height



Weight



Age



Country

- White
- Black
- Asian
- Pacific Islander

Race

- Hispanic
- Non-Hispanic

Ethnicity

- 17.4
- 19.9
- 27.2
- 35.4

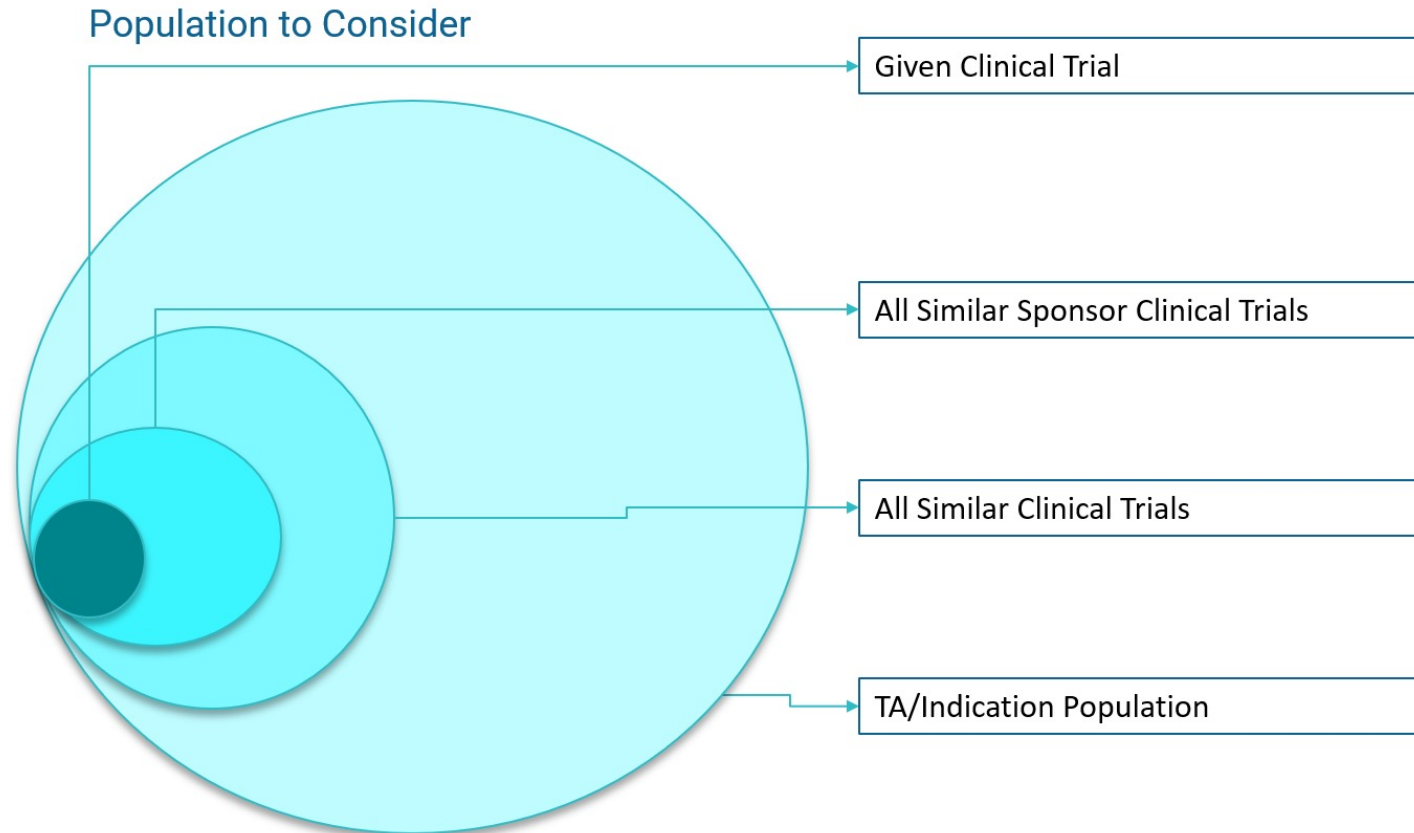
BMI

k-anonymity Rules

k=11

A public sharing threshold of 0.09 indicates that a person's information cannot be shared unless cannot be distinguished from at least 10 other individuals within the same release.

What is you Population Size?



k-anonymity Rules

Age

- No Change
- 5 Year Age band
- 10 Year Age Band
- 15 Year Age Band
- 20 Year Age Band
- 25 Year Age Band
- 30 Year Age Band
- Child 10s 65+ Band
- Child Adult 65+ Band
- Drop

Country

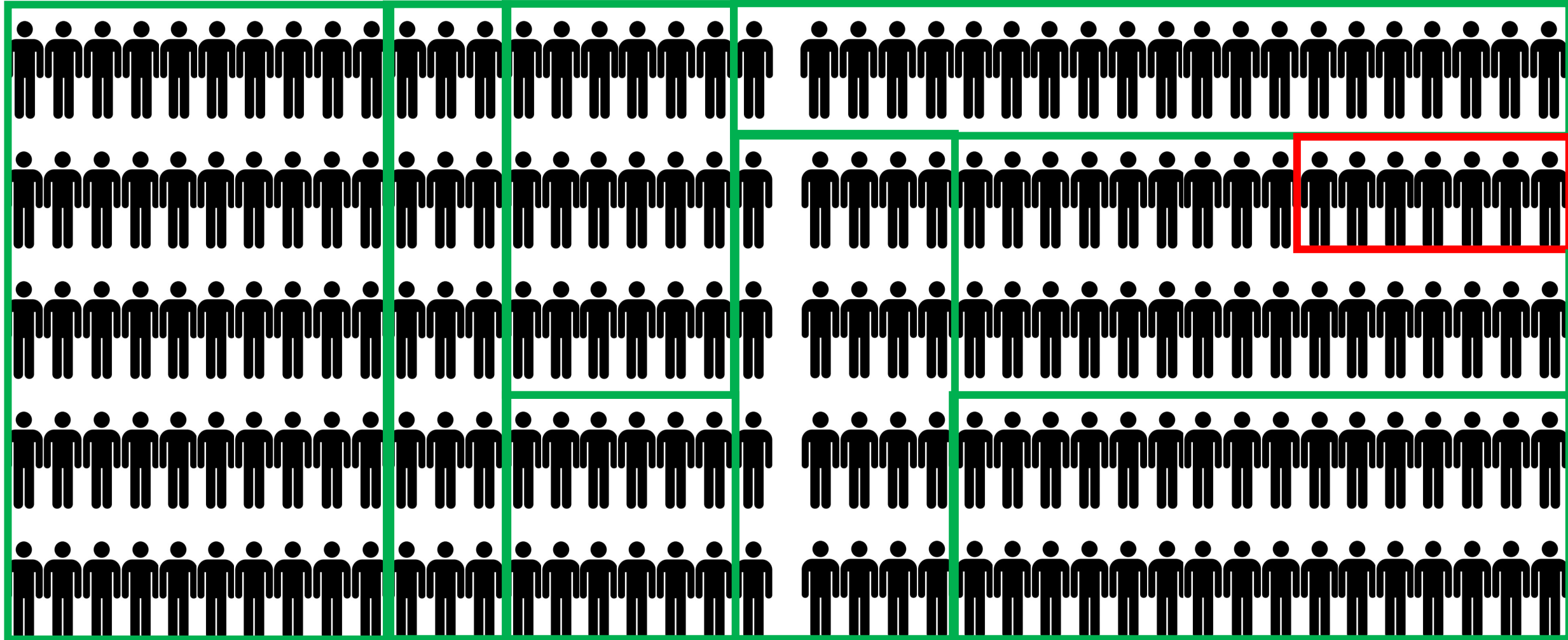
- No Change
- Low Frequency 40
- Low Frequency 50
- Low Frequency 60
- Drop

Sex

- No Change
- Drop

100 Possible combinations

k-anonymity Rules



Process



Conference Meetings



CSR.pdf



Data.sas7bdat

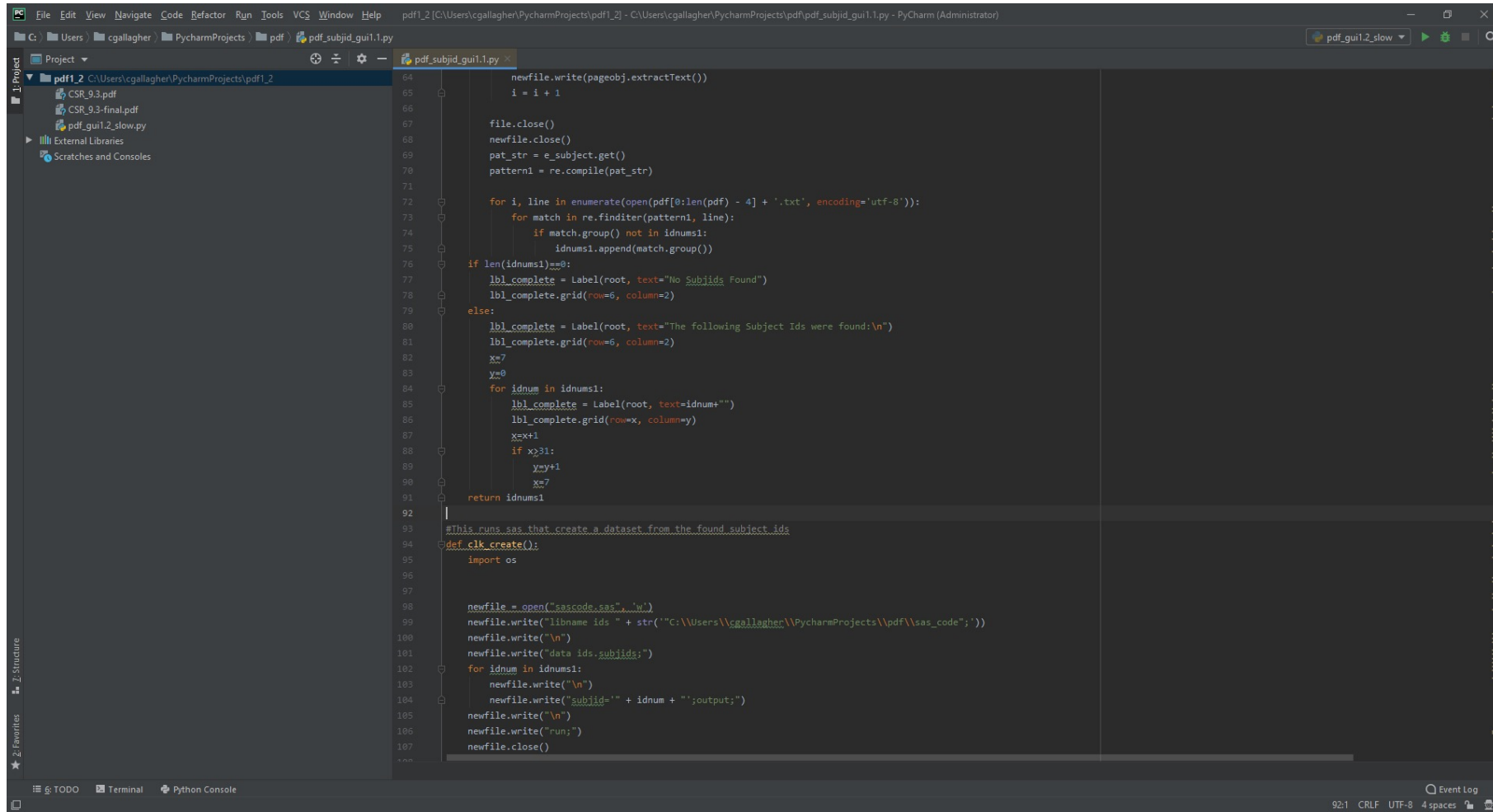
Health Canada Submission – No Data

16.2 Listings out of scope



Participant ID	Age (yrs) / Sex / Race	Visit	Date/Time Sample Taken
12345-123	41/M/W	DAY 1	25SEP2020 / 11:03

Python Script



The screenshot shows the PyCharm IDE with a project named 'pdf1_2'. The file explorer on the left shows the project structure, including 'pdf1_2', 'CSR_9.3.pdf', 'CSR_9.3-final.pdf', 'pdf_gui1.2_slow.py', 'External Libraries', and 'Scratches and Consoles'. The main editor displays the code for 'pdf_subjid_gui1.1.py'. The code is a Python script that processes a PDF file to extract subject IDs. It uses the 'pypdf' library to open the PDF and 're' for regular expressions. The script iterates through the PDF pages, extracts text, and searches for subject IDs. It then creates a dataset from the found subject IDs and writes it to a file named 'sascode.sas'. The script is written in a dark theme with syntax highlighting.

```
64 newfile.write(pageobj.extractText())
65 i = i + 1
66
67 file.close()
68 newfile.close()
69 pat_str = e.subject.get()
70 pattern1 = re.compile(pat_str)
71
72 for i, line in enumerate(open(pdf[0:len(pdf) - 4] + '.txt', encoding='utf-8')):
73     for match in re.finditer(pattern1, line):
74         if match.group() not in idnums1:
75             idnums1.append(match.group())
76
77 if len(idnums1)==0:
78     lbl_complete = Label(root, text="No Subjids Found")
79     lbl_complete.grid(row=6, column=2)
80 else:
81     lbl_complete = Label(root, text="The following Subject Ids were found:\n")
82     lbl_complete.grid(row=6, column=2)
83     x=7
84     y=0
85     for idnum in idnums1:
86         lbl_complete = Label(root, text=idnum+"\n")
87         lbl_complete.grid(row=x, column=y)
88         x=x+1
89         if x>31:
90             y=y+1
91             x=7
92     return idnums1
93
94 #This runs sas that create a dataset from the found subject ids
95 def clk_create():
96     import os
97
98     newfile = open("sascode.sas", 'w')
99     newfile.write("libname ids " + str("C:\\Users\\cgallagher\\PycharmProjects\\pdf\\sas_code;"))
100     newfile.write("\n")
101     newfile.write("data ids.subjids;")
102     for idnum in idnums1:
103         newfile.write("\n")
104         newfile.write("subjid=" + idnum + ";output;")
105     newfile.write("\n")
106     newfile.write("run;")
107     newfile.close()
```

Mined Data - Notepad

 data mining example.txt - Notepad

File Edit Format View Help

Participant

ID

Age(yrs)/

Sex/

Race Visit

Date/Time Sample

Taken

12345-123 41/M/W DAY 1 25SEP2020 / 11:03 |

Regular expression

Subjid Age/Sex/Race

"\d\d\d\d\d-\d\d\d \d\d/\w/\w"

Regular expression

Subjid

"\d\d\d\d\d-\d\d\d"

Result

Rules

- Age – Numeric Band
- Race– Low frequency
- Sex – Keep
- Dates – Subject date offset



All patients on the trial were from a single country. There were no mentions of ethnicity, height, weight or BMI.

Mining issue Problems

- REGEX
- PDFs are not always what they look like
- Getting a complete dataset
- Setting a precedent



Contact me

Cathal Gallagher
Delivery Lead, Transparency
C: (+44)7786.967.185
Cathal.Gallagher@d-wise.com



d-wise an Instem Company
www.d-wise.com

