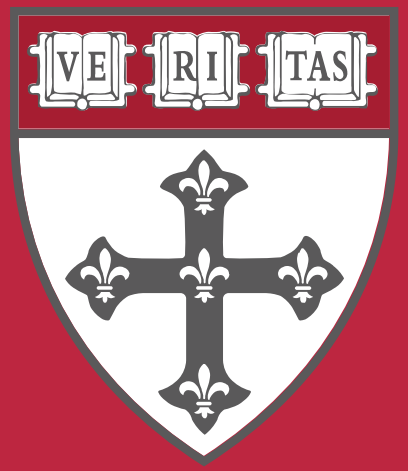


Corso di Laurea Magistrale in Data Science

24/03/2020

BIG DATA IN EPIDEMIOLOGY: A study of diabetes and gender differences in ALS risk, in a population of three million Israeli individuals.

Stefano Rola
Matr. 790383

Advisor: prof. Vittadini
Co-advisor: dr. Rotem
Co-advisor: dr. Bellavia

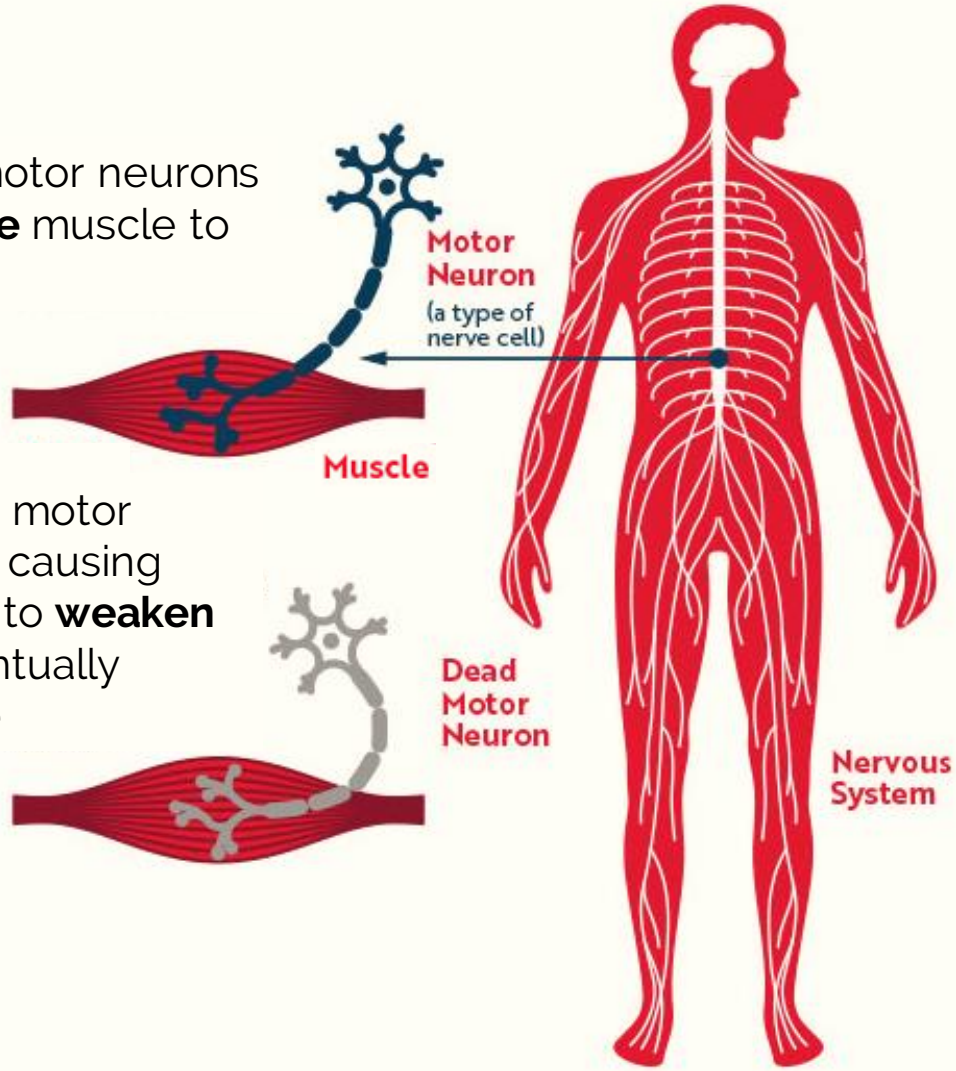
AGENDA

INTRODUCTION

ALS: AMYOTROPHIC LATERAL SCLEROSIS

Healty motor neurons **stimulate** muscle to contract

ALS **kills** motor neurons, causing muscles to **weaken** and eventually **paralyze**



- **5,000+ EVERY YEAR**
- **4.42 per 100,000**
- **2-5 YEARS**
- **DIFFICULT TO DIAGNOSE**
- **NO CURE**

ALS: RISK FACTORS

10% 'Familial ALS'

90% Non-genetic factors

ALS: RISK FACTORS

METABOLIC DISORDERS

- Weight loss
- Hypermetabolism
- Hyperlipidemia
- Insulin resistance
- Glucose intolerance
- Increased energy consumption

DIABETES MELLITUS

- 2010, Jawaid et al.
- 2015, Kioumourtzoglou et al.
- 2015, Mariosa et al.
- 2015, Sun et al.
- 2018, D'Ovidio et al.

BIG DATA IN EPIDEMIOLOGY

HOW?

- EHR/EMR
- Repurposed data
- Record linkage

VOLUME

- Monitoring
- Quick intervention
- Rapid iteration

VELOCITY

- Imaging data
- Behavioural data
- Geo-location

VARIETY

WHY?

- To monitor **drug** and device safety
- Measure **hospital quality**
- **Predict** outbreaks of epidemics
- **Avoid** preventable diseases
- Reduce the **costs** of healthcare delivery
- Reduce the amount of medical **errors**

Stefano Rola - 24/03/2020

MHS: MACCABI HEALTHCARE SERVICES

ALS

BIG DATA

MHS: MACCABI HEALTHCARE SERVICES

ALS

BIG DATA

Israel's second largest integrated healthcare organization, serving **25%** of the Israeli population

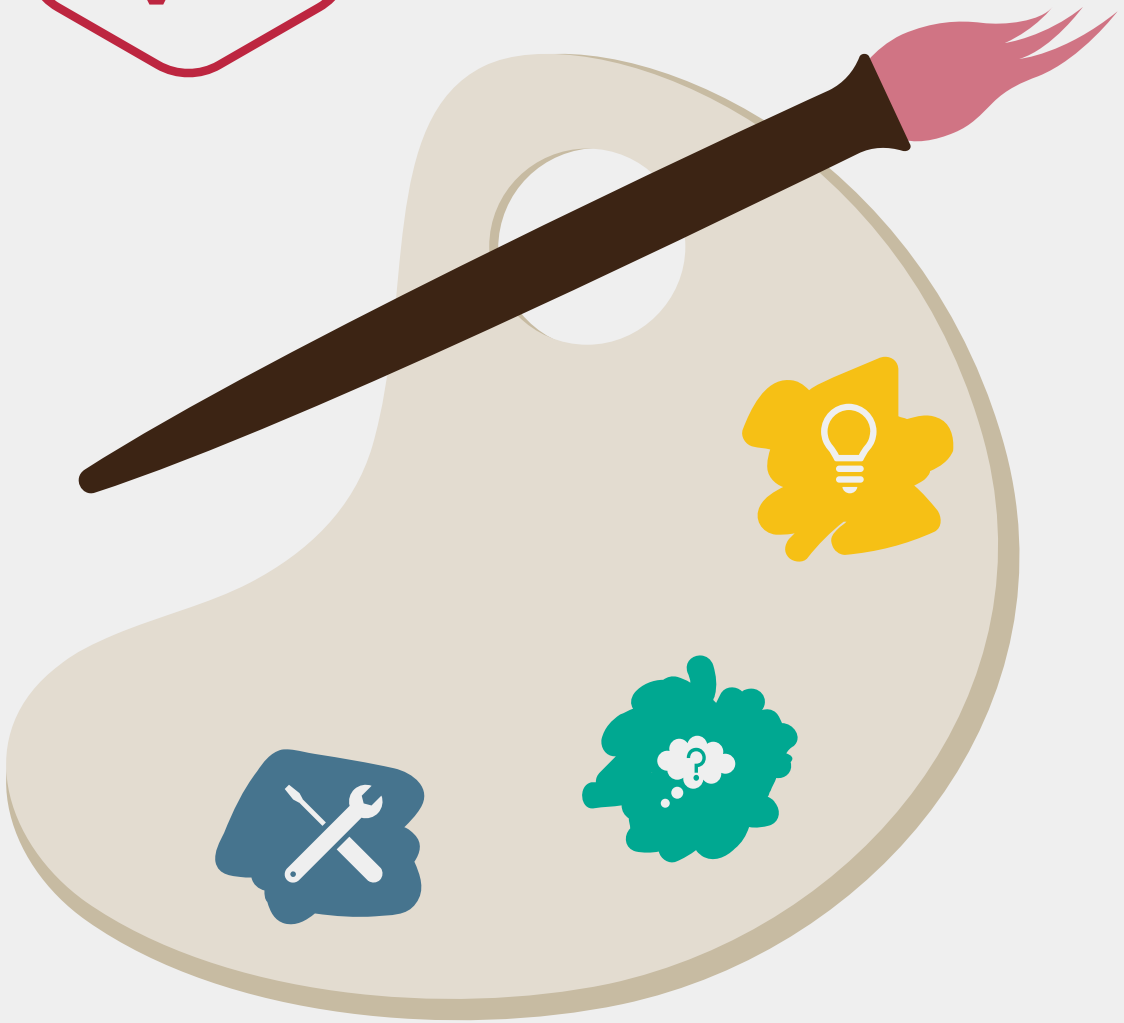
Linked to records from Israel's Central Bureau of Statistics, allowing **linkage** to additional information

All medical history of ~**3 million** individuals

Recorded computerized medical data since **1998**

METHODS

Study design



WHAT?

Nested Case-Control Study

WHY?

- Rarity of the outcome (unbalanced dataset)
- Large number of individuals
- OR estimates the underlying HR in cohort

HOW?

1. Selection of all cases
2. **Randomly** sample **100** controls per case, matched on the **year of birth** and on **being alive** at the time of ALS diagnosis

Outcome & Exposures

OUTCOME

- **DEFINITION**:
  First occurrence ever of **ALS** according to ICD-9 diagnosis code: 335.20

- **TIME RESTRICTIONS:**
  Cases in the system less than 1 year before ALS diagnosis removed;
  Controls in the system less than 1 year before ALS diagnosis of matched case removed;
  ALS diagnoses **prior 2001** removed.

EXPOSURE

- **DEFINITION**:
  Diabetes cases identified with ICD-9 diagnosis code: 250.00;

- **TIME RESTRICTIONS:**
  3 Years Lag applied between exposure's diagnoses and ALS

CONFOUNDERS

- **DEFINITION**:
  Obesity cases identified with ICD-9 diagnosis code: 278.00-278.02;
  Last **BMI** measurement before ALS diagnosis;
  Country of birth, District, SES.

- **TIME RESTRICTIONS:**
  3 Years Lag applied between confounding information and ALS

TOTAL: 547 cases, 51,707 controls

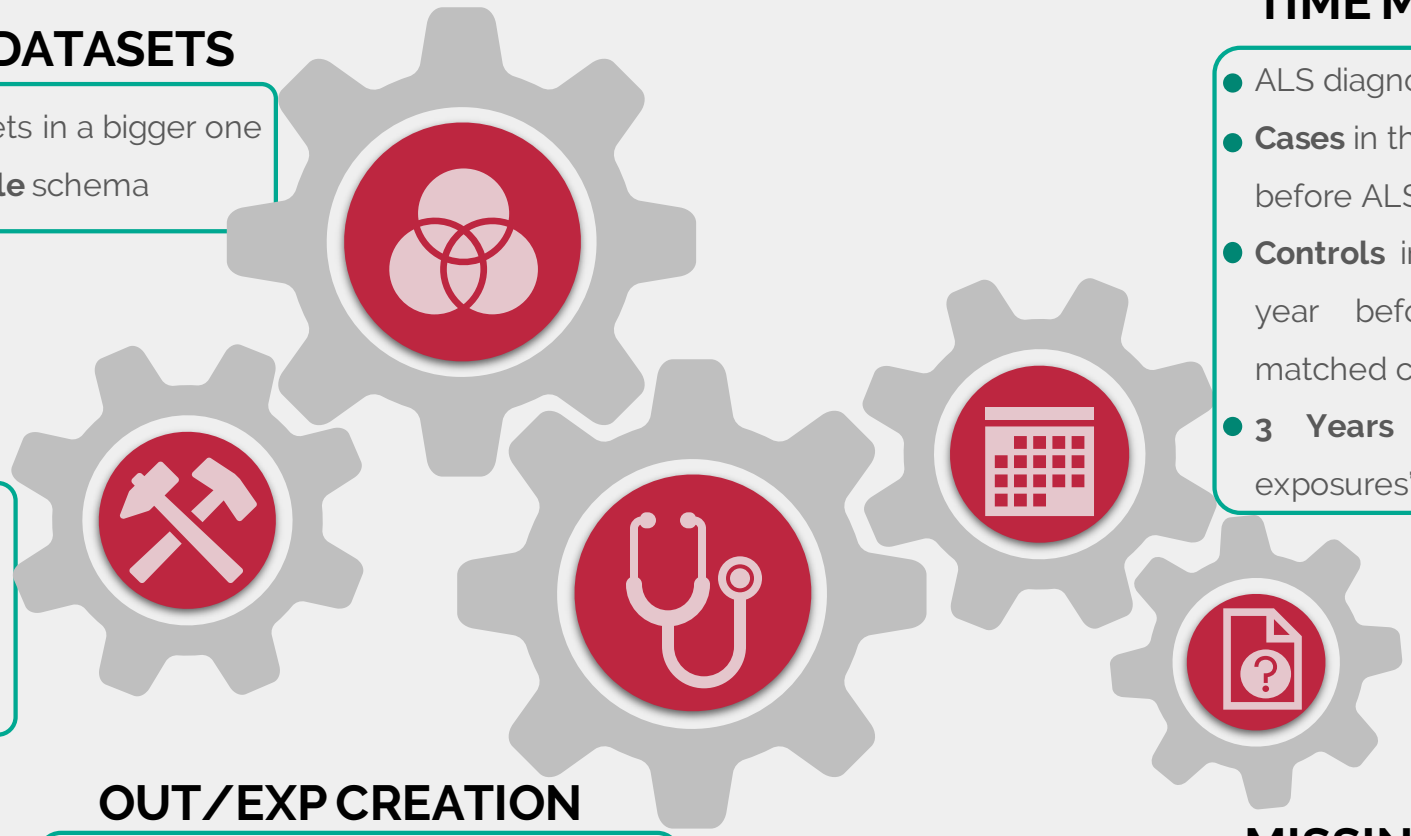
Pre-Processing

MERGING DATASETS

- **Join** small datasets in a bigger one
- Keeping a **flexible** schema

TIME MANAGEMENT

- ALS diagnosis **prior 2001** removed
- **Cases** in the system less than 1 year before ALS diagnosis removed
- **Controls** in the system less than 1 year before ALS diagnosis of matched case removed
- **3 Years Lag** applied between exposures' diagnoses and ALS



DATA MINING

- MDClone **Platform**
- **Multiple datasets** extracted, each regarding a different diagnosis or feature

OUT/EXP CREATION

- **ALS**, **Diabetes** and **Obesity** coded as a binary variable (presence/absence)
- **Overweight** if patient obese or having BMI > 25

MISSING VALUES

- Merging operation **caused** missing
- Not all **real** missing values
- Use of a **standard** format

Statistical Analysis

$$logit(p) = \beta_0 + \sum_{j=1}^{m} \beta_j X_{ij} + \sum_{j=1}^{m} y_k C_{ik}$$

1. ALS ~ Diabetes
2. ALS ~ Diabetes + Overweight
3. ALS ~ Diabetes * Overweight
4. ALS ~ Diabetes + Overweight + Country of birth + District + SES

1. ALS ~ Diabetes
2. ALS ~ Diabetes + Overweight
3. ALS ~ Diabetes * Overweight
4. ALS ~ Diabetes + Overweight + Country of birth + District + SES

1. ALS ~ Diabetes
2. ALS ~ Diabetes + Overweight
3. ALS ~ Diabetes * Overweight
4. ALS ~ Diabetes + Overweight + Country of birth + District + SES

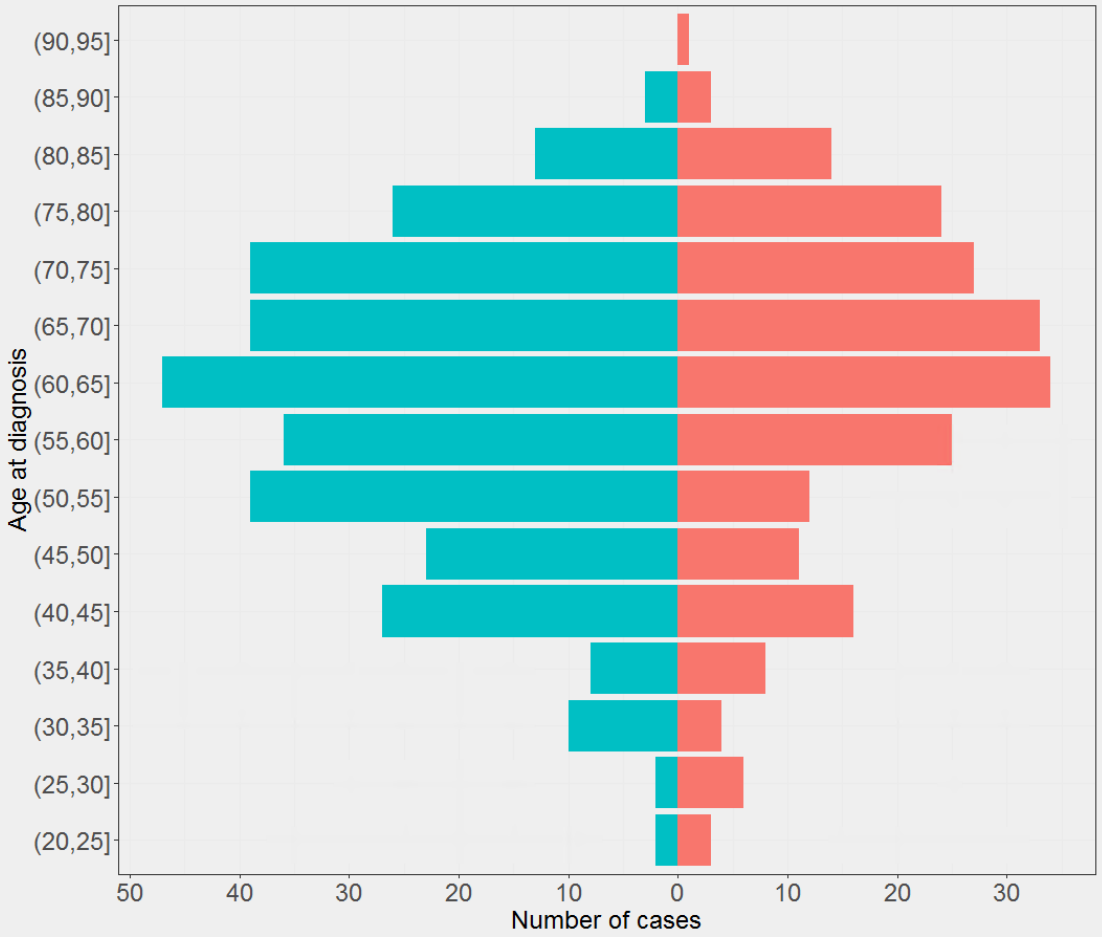
RESULTS

Descriptive Statistics

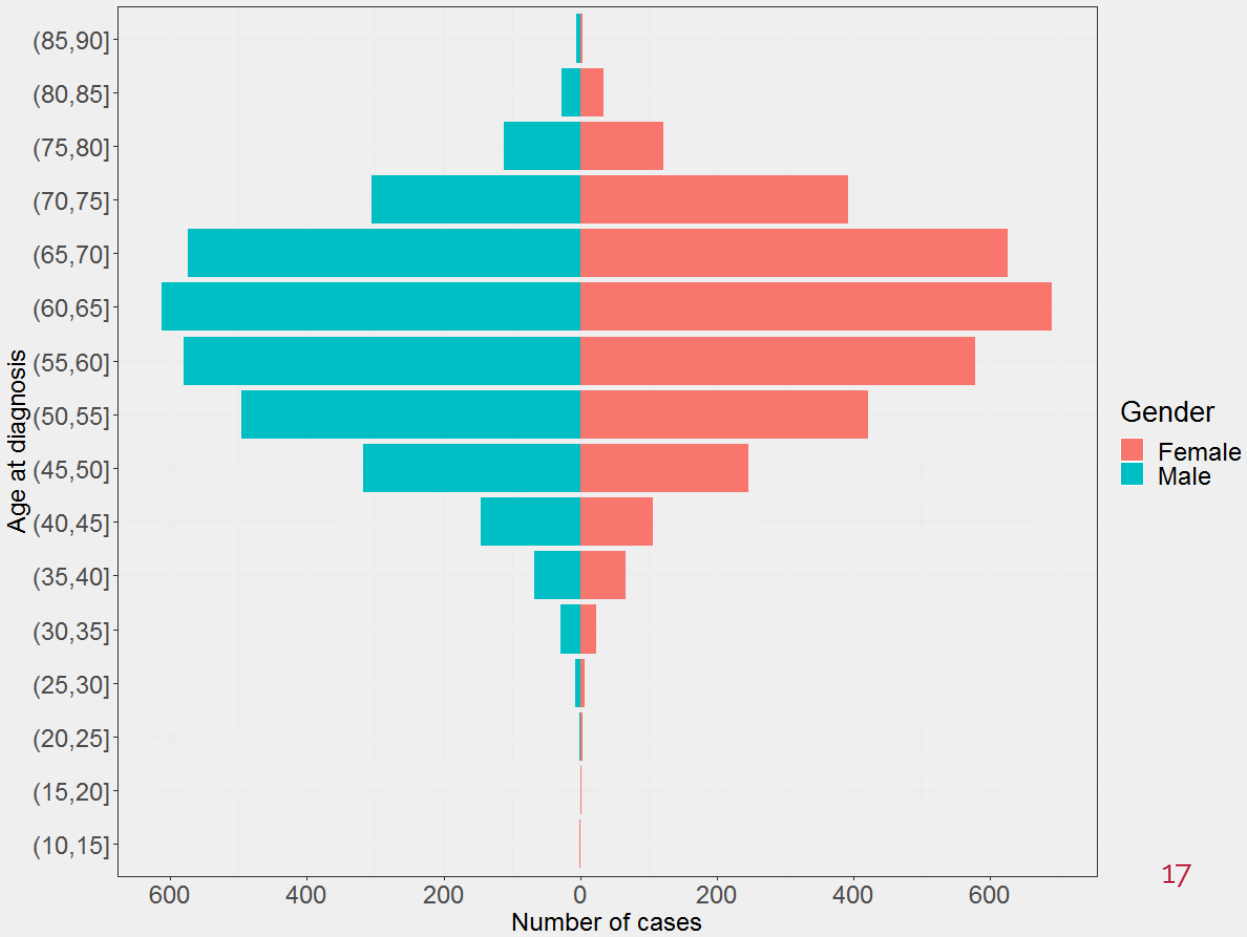
	Overall	ALS: Yes	ALS: No
N	52,254	547	51,707
Diabetes	6,602(13%)	78(14%)	6,524(13%)
Overweight	13,290(25%)	168(31%)	13,122(25%)
Age		59.9(15.7)	
Birth date	1949(15.7)	1950(16)	1949(15.7)
Gender			
Male	24,986(48%)	322(59%)	24,664(48%)
Female	27,268(52%)	225(41%)	27,043(52%)
SES (1-10)	6.1(1.8)	6.2(1.8)	6.1(1.8)
Country of birth			
Israel	27,697(53%)	284(52%)	27,692(53%)
USSR	12,341(24%)	124(23%)	12,217(24%)
Other	11,937(23%)	139(25%)	11,798(23%)
District			
Jerusalem&Shfela	12,535(24%)	116(21%)	12,419(24%)
Center	11,257(22%)	124(23%)	11,133(21%)
North	10,159(19%)	88(16%)	10,071(19%)
Sharon	9,934(19%)	122(22%)	9,812(19%)
South	8,065(16%)	72(13%)	7,993(16%)
Obesity	1,791(3%)	16(3%)	1,775(3%)
BMI	27.9(5.3)	27.6(5)	27.9(5.3)

Descriptive Statistics

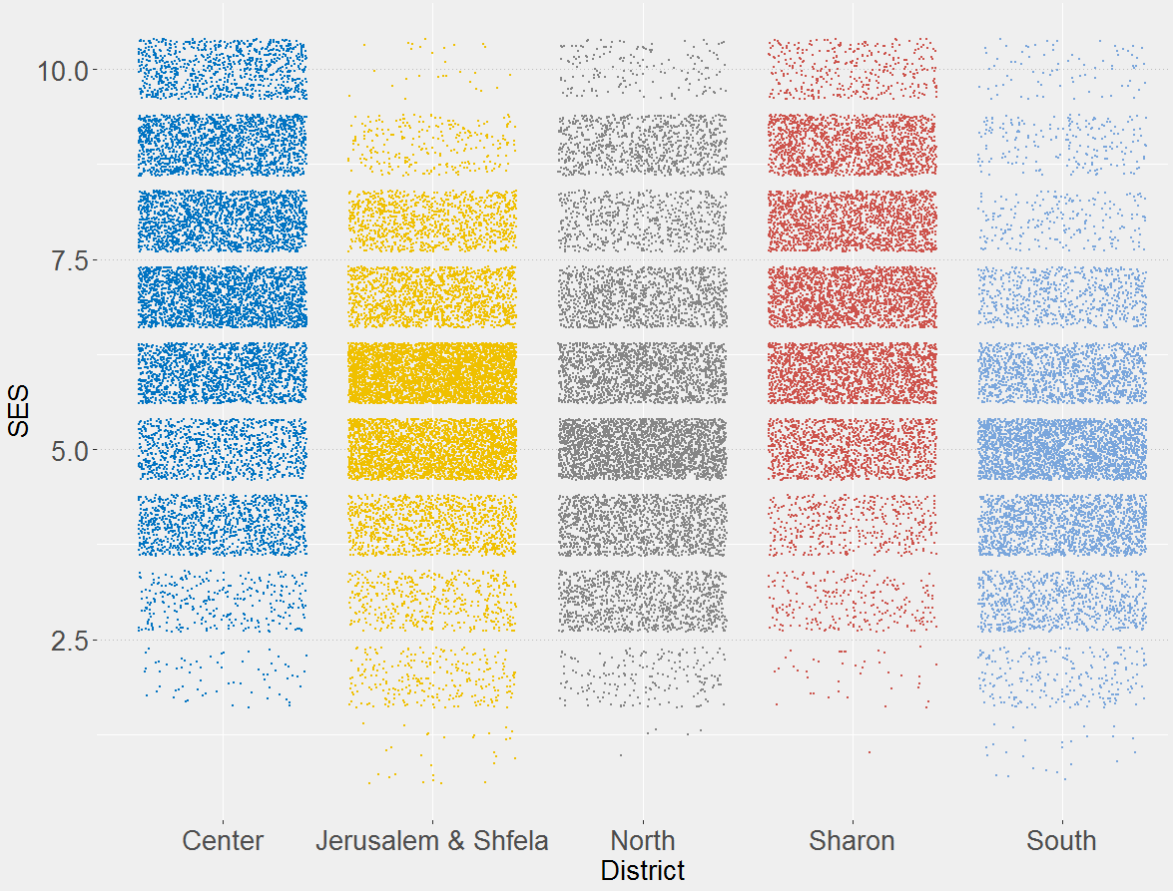
ALS

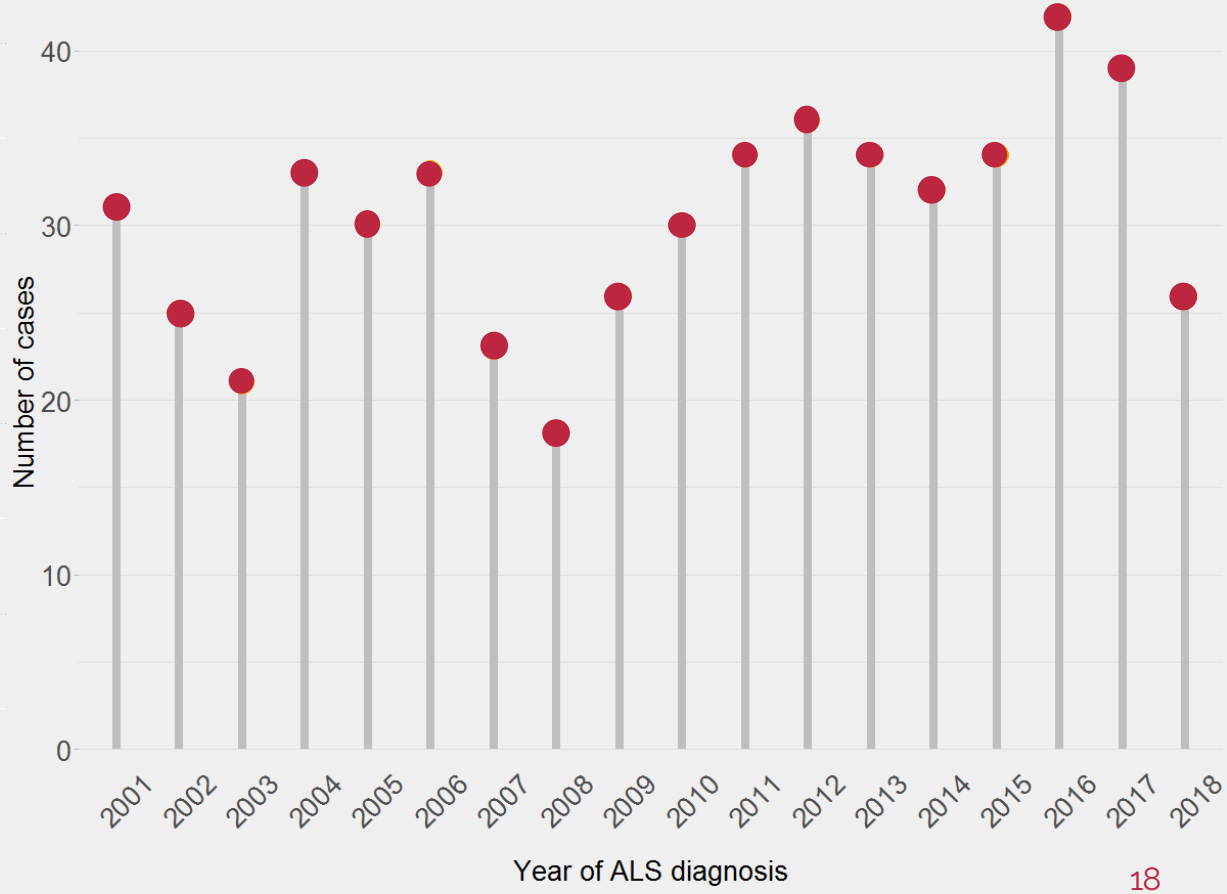


DIABETES



Descriptive Statistics





Main Analysis

TOTAL

Models	Diabetes	Overweight	Interaction
Model 1	1.17 (0.91 1.49)		
Model 2	1.07 (0.83 1.38)	1.29 (1.06 1.56)	
Model 3	1.24 (0.87 1.78)	1.35 (1.10 1.66)	0.75 (0.46 1.23)
Model 4	1.14 (0.88 1.48)	1.37 (1.13 1.66)	

MALE

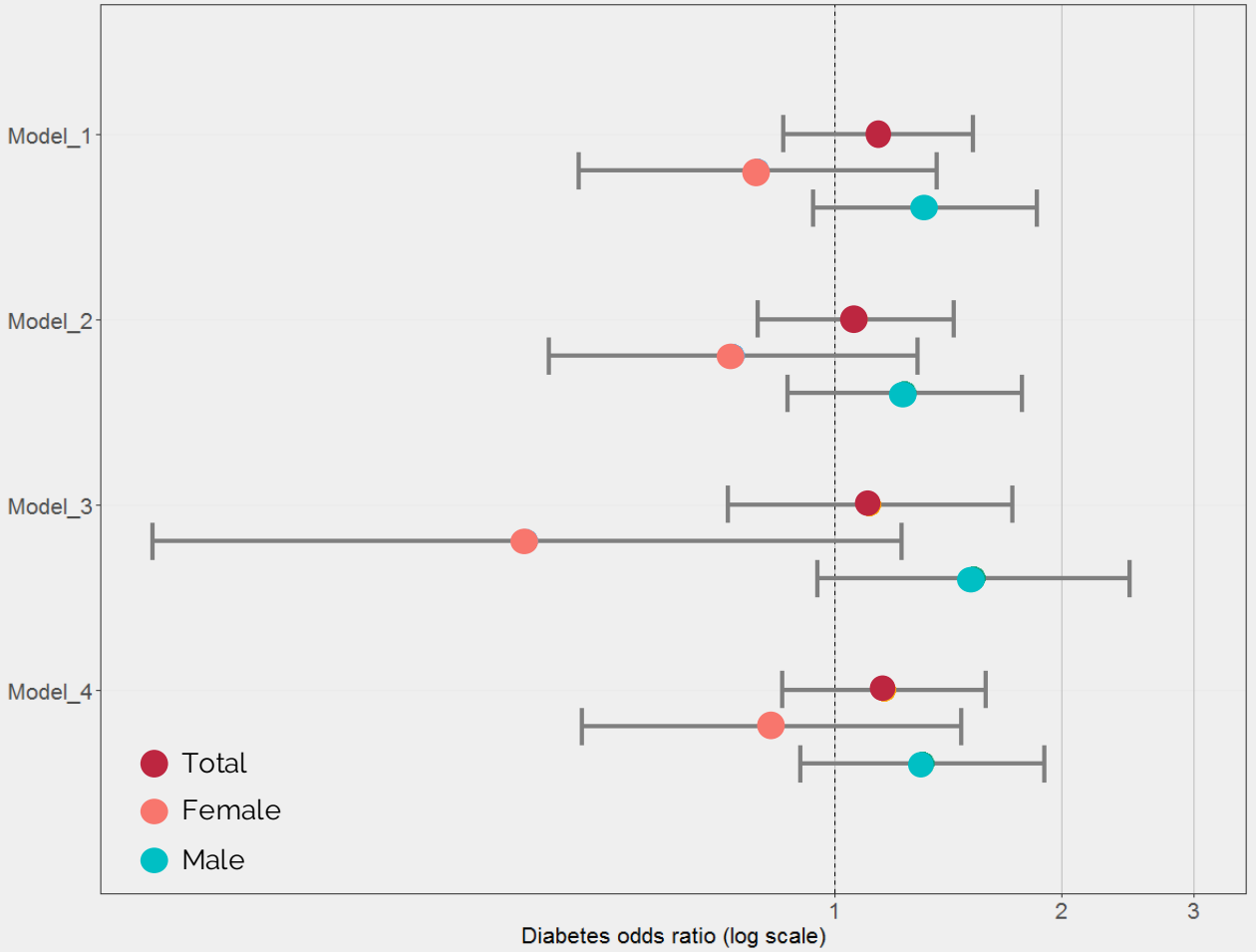
Models	Diabetes	Overweight	Interaction
Model 1	1.32 (0.97 1.80)		
Model 2	1.23 (0.89 1.69)	1.24 (0.97 1.59)	
Model 3	1.59 (1.04 2.43)	1.37 (1.05 1.80)	0.59 (0.32 1.10)
Model 4	1.28 (0.93 1.77)	1.25 (0.97 1.60)	

FEMALE

Models	Diabetes	Overweight	Interaction
Model 1	0.92 (0.60 1.39)		
Model 2	0.83 (0.54 1.27)	1.34 (1.00 1.81)	
Model 3	0.73 (0.37 1.45)	1.31 (0.95 1.80)	1.22 (0.51 2.93)
Model 4	0.89 (0.57 1.40)	1.57 (1.15 2.13)	

Additional Analysis

	Overall	ALS: Yes	ALS: No
N	52,254	547	51,707
Low risk pre-diabetes	7,418(14%)	109(20%)	7,309(14%)
High risk pre-diabetes	311(0.6%)	4(0.7%)	307(0.6%)
Diabetes	4,843(9%)	53(10%)	4,790(9%)



CONCLUSIONS

Strengths and limitations

- First ever ALS analysis in Maccabi → Consistency
- Large population → Robustness

- Outcome and exposures ascertainment → Inaccuracy
- Information since 1998 → Misclassification

What's next

1 ALS and DIABETES

- Diabetes type 1/2 distinction
- More confounding factors

2 TIME FACTOR

- Survival analysis
- Trajectories and trends

3 MEDICATIONS

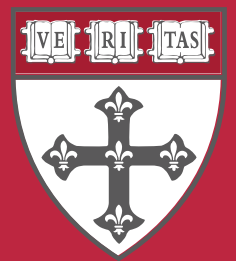
- Identify candidates for ALS
- More accurate results
- Taking time and dosage

4 MACHINE LEARNING

- Evaluate the relationship between complex and multi-factorial elements

QUESTIONS?

> Where is the wisdom we have lost in knowledge?
> Where is the knowledge we have lost in information?
>
> -- T.S. Eliot



s.rola@campus.unimib.it

stefano-rola