MI² DataLab Seminar

João Malato

January 31st, 2022

Wydział Matematyki i Nauk Informacyjnych
POLITECHNIKA WARSZAWSKA
LISBOA UNIVERSIDADE DE LISBOA
FACULDADE DE MEDICINA LISBOA
CENTRO ACADÉMICO DE MEDICINA DE LISBOA
Instituto de Medicina Molecular | João Lobo Antunes
FCT Fundação para a Ciência e a Tecnologia

Impact of imperfect diagnosis in ME/CFS association analyses

João Malato

January 31st, 2022

Wydział Matematyki i Nauk Informacyjnych
POLITECHNIKA WARSZAWSKA
LISBOA UNIVERSIDADE DE LISBOA
FACULDADE DE MEDICINA LISBOA
CENTRO ACADÉMICO DE MEDICINA DE LISBOA
Instituto de Medicina Molecular | João Lobo Antunes
FCT Fundação para a Ciência e a Tecnologia

Quick presentation

João Malato

- BSc Biology
- MSc Biostatistics
- PhD Biomedicine & Computational biology

- Immune-Stats Group
- LGraca Group

Background

Modern diseases

- Cancer
- Autoimmune diseases

Chronic illnesses

- Complex
- Multifactorial diseases
- Dysregulation and degeneration at base

Diseases with uncertainty in their diagnosis

- Overlapping symptoms
- Same phenotype for different diseases
- Diagnosis of exclusion

Modern problems require modern solutions
Cartoon by Gary Larson

ME/CFS

- Medically unexplained (persistent) fatigue
- ≥ 6 months
- Post-exertional malaise

- Chronic pain
- Sleep disturbances
- Cognitive difficulties

- Estimated prevalence of 0.1-2.2%
- 6 woman to 1 man

> *People with ME/CFS are often not able to do their usual activities. (…) ME/CFS may confine them to bed. (…) [O]verwhelming fatigue that is not improved by rest. (…) [M]ay get worse after any activity, whether it's physical or mental [post-exertional malaise].*

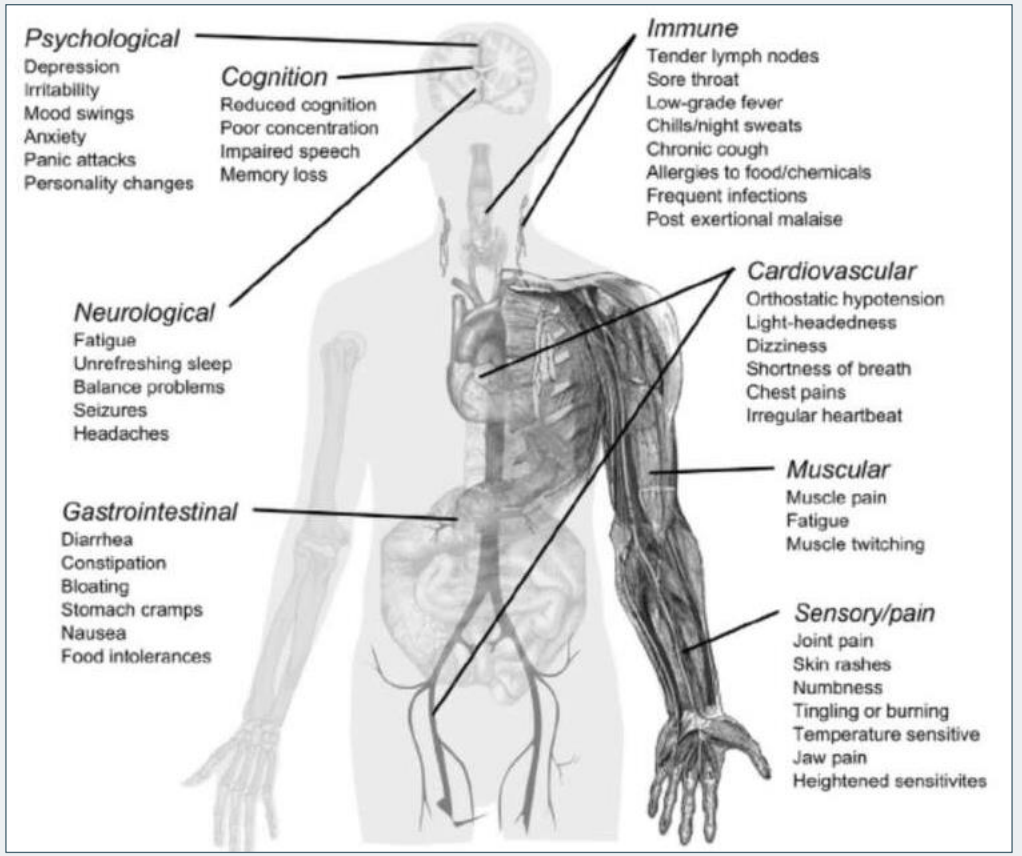
Broad definition given by the CDC

Images by The Scientist

ME/CFS is a complex disease

- Multifactorial disorder
  - Genetic variations
  - Epigenetic variations
  - Gene expression variations
  - Phenotypic variations

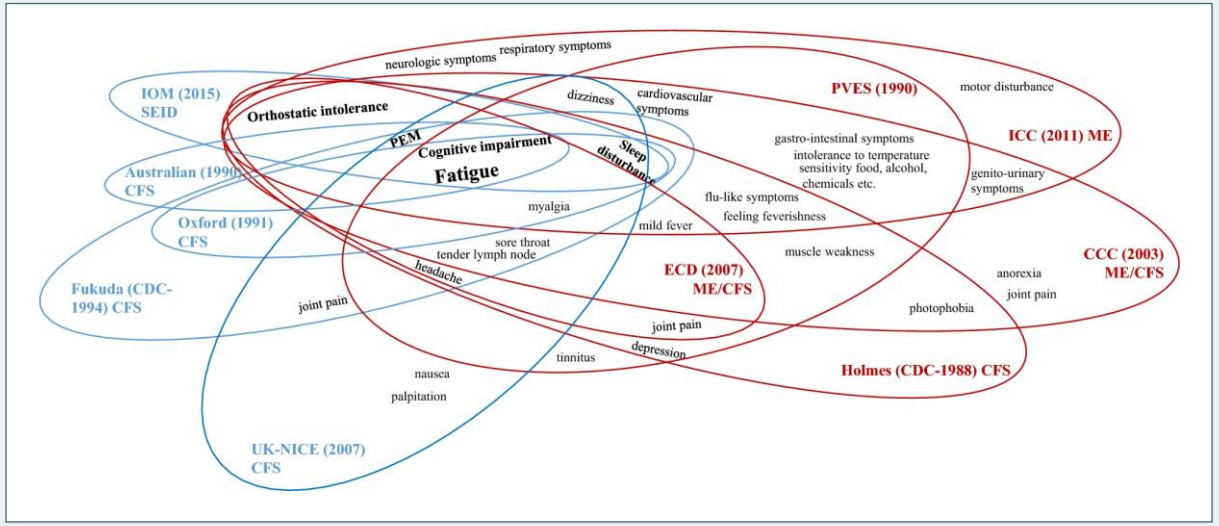
- Similarities to some autoimmune diseases



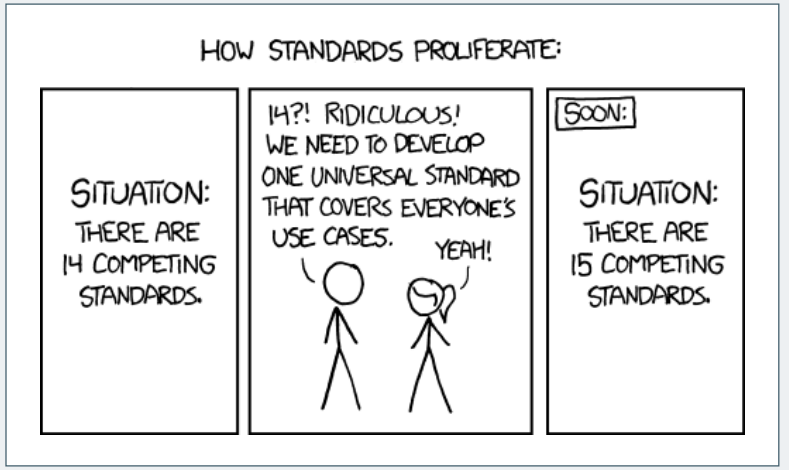
Overview of ME/CFS symptoms

ME/CFS diagnosis

- No known biomarker
- No laboratory diagnostic test
- Clinical diagnosis based on
  - Case history
  - Physical exams
  - Exclusion of other diseases
- There are more than 20 symptoms-based criteria



Key symptoms of ME/CFS by different case definitions; Lim et al. (2020)



Cartoon adapted from xkcd #927, by Randall Munroe

Study objectives

1. Diagnostic agreement and symptoms similarity

- Different case definitions for ME/CFS
- How similar are ME/CFS case criteria?
- ME/CFS patients symptom characteristics
- Can patients be further stratified?

How uncertainty around disease definition can

- Potentially disguise further subgroups of patients
- Affect the reproducibility of its research

2. Impact of misclassification

- What are the impacts of not considering potential misclassified or inherent subtypes of ME/CFS diagnosed patients?

- Common case-control association studies
- Candidate gene studies
- Serology studies (with impact of false positive tests)

Data available

CureME Group

- Established in 2006
- Advance clinical research into ME/CFS

UK ME/CFS Biobank

- Initiated in 2013
- One of the few worldwide dedicated to study ME/CFS

- Data on 600+ donors
- Healthy controls with no history of fatigue
- Multiple sclerosis patients (controls)
- ME/CFS patients

Data available

CureME Group

- Established in 2006
- Advance clinical research into ME/CFS

UK ME/CFS Biobank

- Initiated in 2013
- One of the few worldwide dedicated to study ME/CFS

- Data on 600+ donors
- Healthy controls with no history of fatigue
- Multiple sclerosis patients (controls)
- ME/CFS patients

Type of data

- Diagnostics criteria
- Symptom assessment questionnaires
- Immunological data
- Gene expression data

Analysis of diagnostic agreement

Suspected patients evaluated according to four case definitions

- CDC-1994
- IOM-2005
- CCC-2003
- ICC-2011

- Information on 275 suspected cases
- Diagnosed by at least one case definition

Analysis of diagnostic agreement

Case definition				N	% of total suspected cases
CDC-1994	IOM-2005	CCC-2003	ICC-2011		
+	+	+	+	173	62.9
+	+	+	–	32	11.6
+	–	+	+	16	5.8
+	+	–	+	16	5.8
+	–	–	–	14	5.1
+	+	–	–	10	3.6
+	–	+	–	5	1.8
+	–	–	+	3	1.1
–	–	+	+	3	1.1
–	–	–	+	1	0.4
–	+	–	–	1	0.4
–	+	–	+	1	0.4
97.8%	84.7%	83.3%	77.5%	275	100%

Frequency of suspected cases of ME/CFS according to their diagnostic outcomes using different case definitions

	CDC-1994	IOM-2005	CCC-2003	ICC-2011
CDC-1994	1.000	0.876	0.876	0.760
IOM-2005	0.876	1.000	0.840	0.752
CCC-2003	0.876	0.840	1.000	0.753
ICC-2011	0.760	0.752	0.753	1.000

Estimates if the Jaccard's similarity index for the four case definitions

Analysis of diagnostic agreement

Case definition				N	% of total suspected cases
CDC-1994	IOM-2005	CCC-2003	ICC-2011		
+	+	+	+	173	62.9
+	+	+	–	32	11.6
+	–	+	+	16	5.8
+	+	–	+	16	5.8
+	–	–	–	14	5.1
+	+	–	–	10	3.6
+	–	+	–	5	1.8
+	–	–	+	3	1.1
–	–	+	+	3	1.1
–	–	–	+	1	0.4
–	+	–	–	1	0.4
–	+	–	+	1	0.4
97.8%	84.7%	83.3%	77.5%	275	100%

Frequency of suspected cases of ME/CFS according to their diagnostic outcomes using different case definitions

	CDC-1994	IOM-2005	CCC-2003	ICC-2011
CDC-1994	1.000	0.876	0.876	0.760
IOM-2005	0.876	1.000	0.840	0.752
CCC-2003	0.876	0.840	1.000	0.753
ICC-2011	0.760	0.752	0.753	1.000

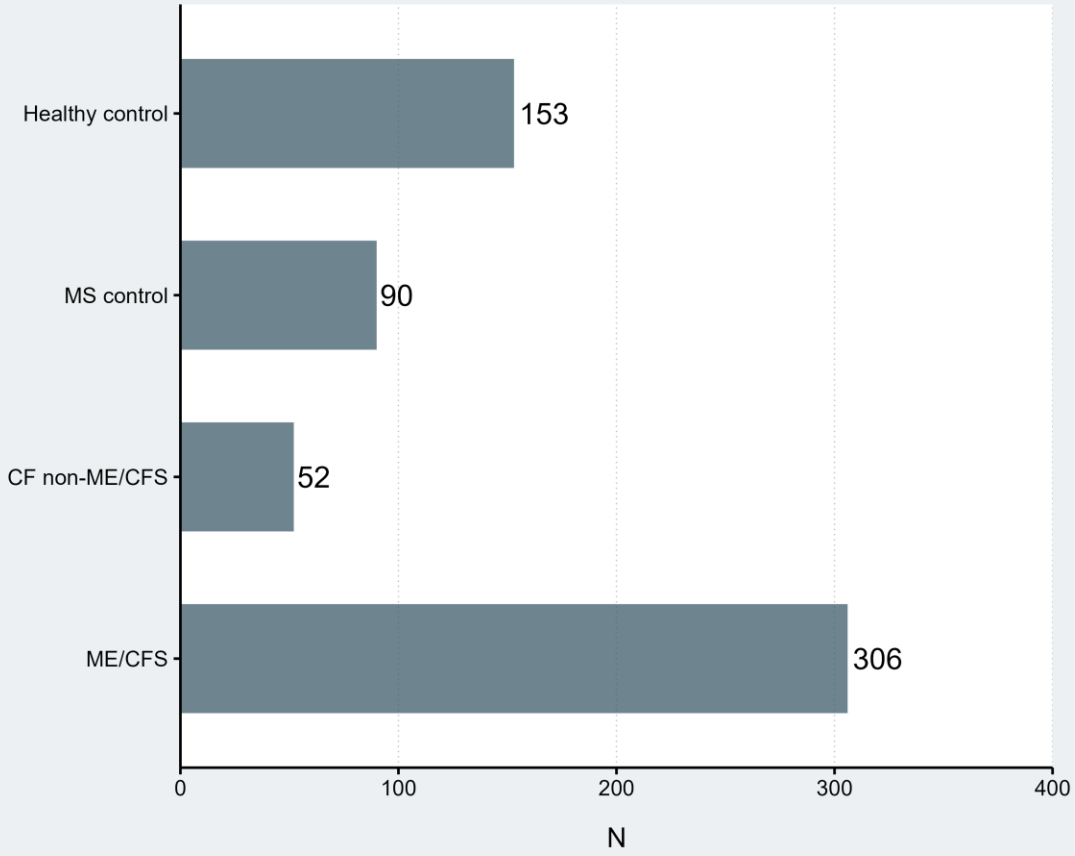
Estimates if the Jaccard's similarity index for the four case definitions

- Fraction of cases that do not match
- Misdiagnosed individuals included as patients

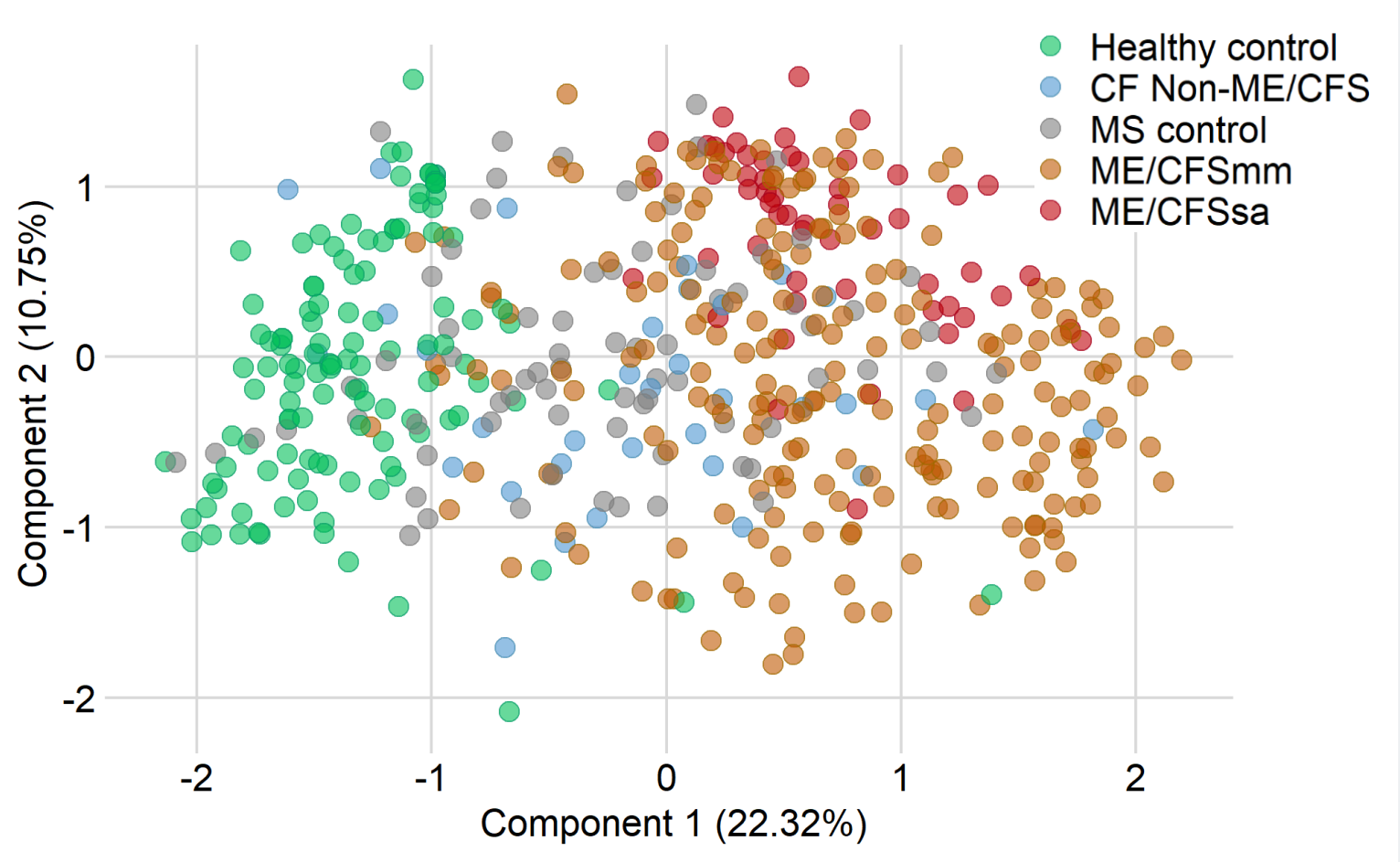
Analysis of symptoms' similarity

- 57 symptoms assessment questions
- Severity of symptoms
- Different clinical domains

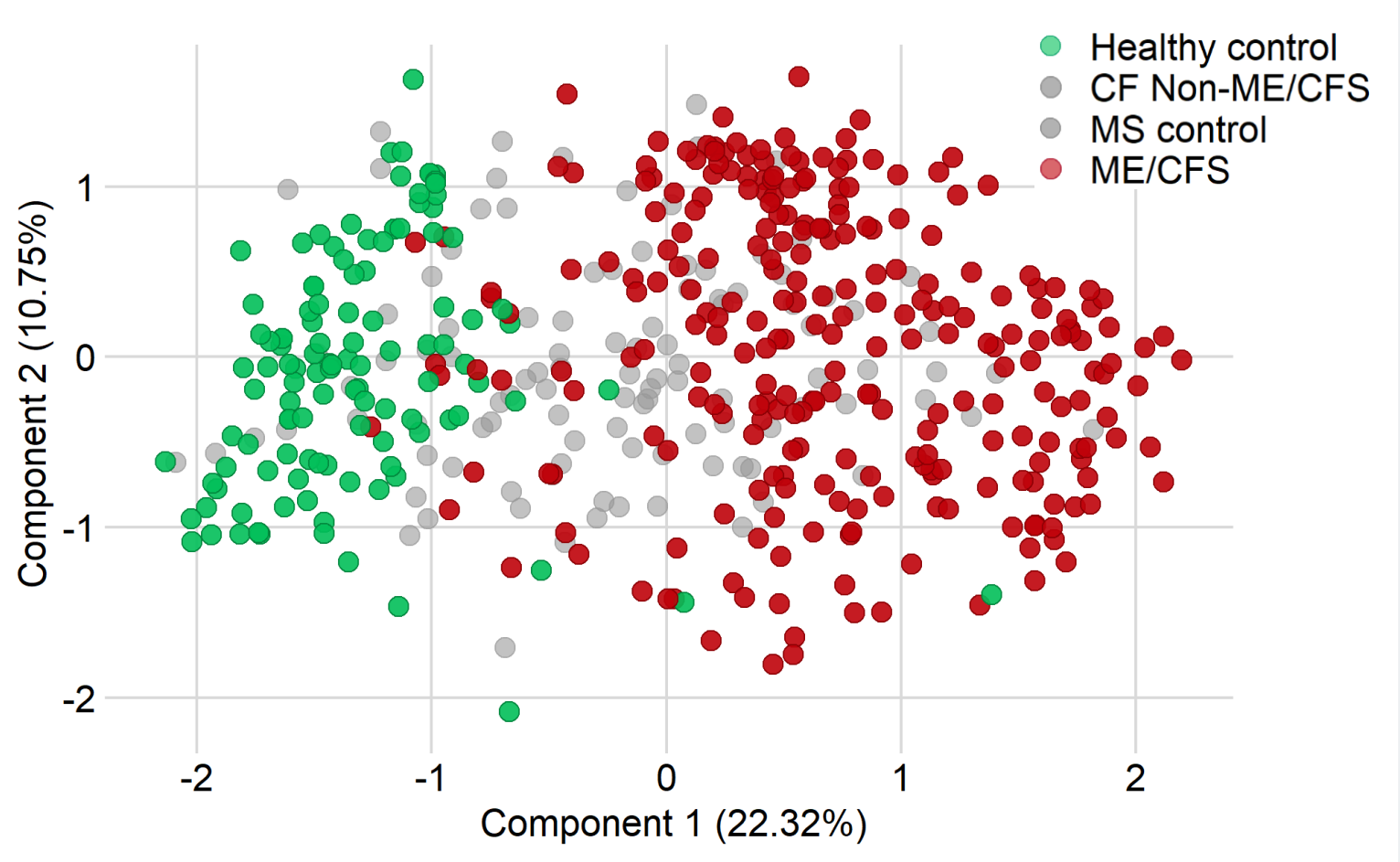
- Measure combination of answers between participants
- Distances through Cohen's K coefficient
- Measure similarities between set of answers
- Multidimensional scaling (MDS)



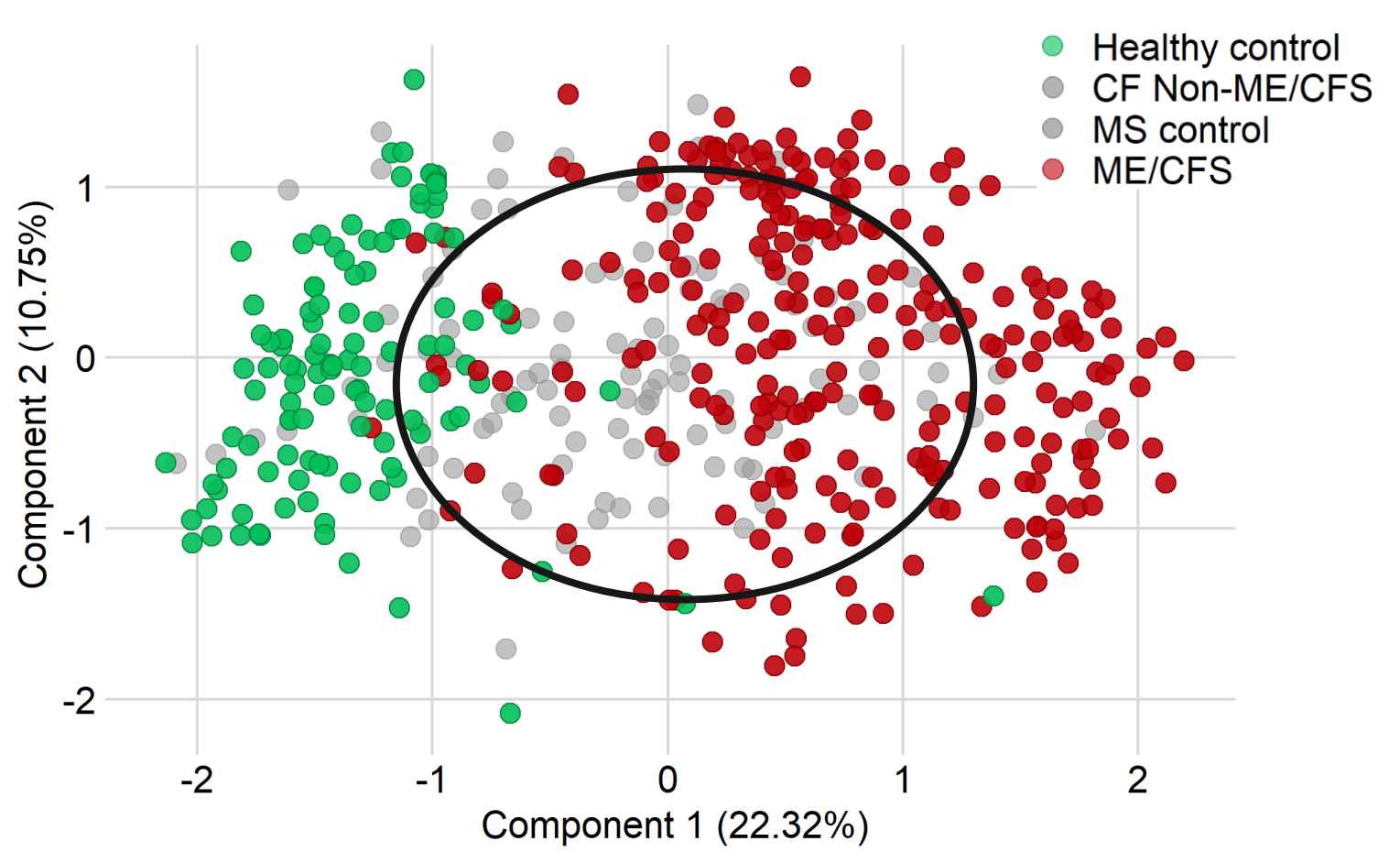
Analysis of symptoms' similarity



Analysis of symptoms' similarity

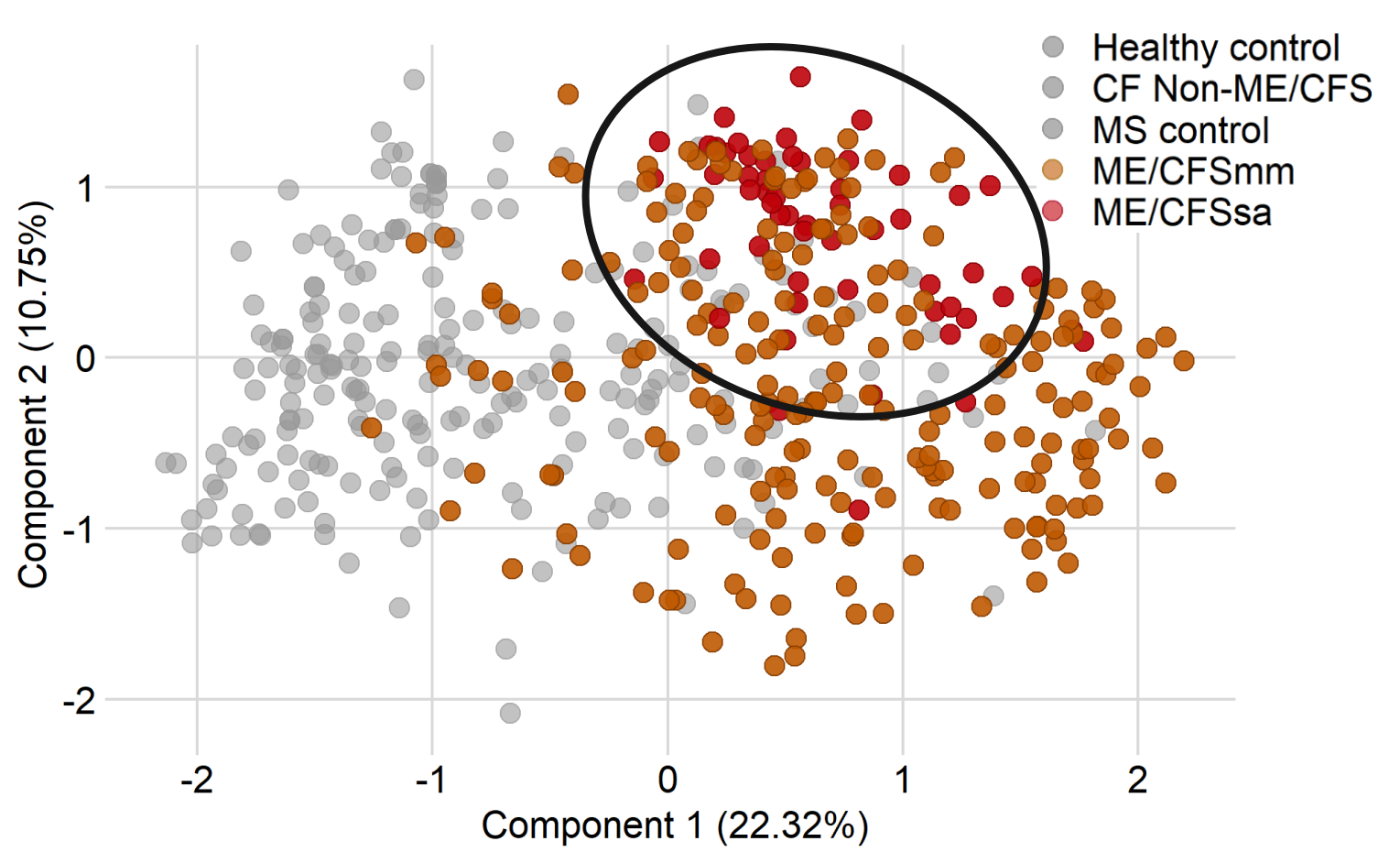


Analysis of symptoms' similarity



- Overlap between symptoms of different diseases
- Misclassification associated to diagnosis

Analysis of symptoms' similarity



- Overlap between symptoms of different diseases
- Misclassification associated to diagnosis

- Subtyping of ME/CFS patients
- (based on severity)

Uncertainty in ME/CFS

- Not a single set of diagnostic criteria for ME/CFS
- Different criteria will select distinct case cohorts
  - Misclassification associated to diagnosis of patients
- Evidence for subgroups of patients
  - Potential for patient stratification and subtyping

Impact of misclassification on association studies

Hypothetical study

- Case-control
- Gene signature
- Effects of exposure to viruses

- Studying lack of association to ME/CFS

Impact of misclassification on association studies

Hypothetical study

- Case-control
- Gene signature
- Effects of exposure to viruses

- Studying lack of association to ME/CFS
  - 2×2 frequency table
  - $H_0: \theta_0 = \theta_1$
  - Assessing the power of the study

$$f(x_i \mid n_i; \theta_i) = \prod_{\mathrm{i}=0,1} \binom{n_i}{x_i} \theta_i^{x_i} (1-\theta_i)^{n_i - x_i}$$

Risk factor	Healthy controls	Suspected cases
1	θ_0	θ_1
0	$1-\theta_0$	$1-\theta_1$

Assumptions

1. ME/CFS cohort can be divided into apparent and true positives
2. Poorly diagnosed cases are considered to be healthy controls (θ_0)
3. Misclassification rate (γ) defined false and true positives
4. γ is only dependent on the true clinical status of each suspected case

Risk factor	Healthy controls	Suspected cases
1	θ_0	θ_1
0	$1-\theta_0$	$1-\theta_1$

Assumptions

1. ME/CFS cohort can be divided into apparent and true positives
2. Poorly diagnosed cases are considered to be healthy controls (θ_0)
3. Misclassification rate (γ) defined false and true positives
4. γ is only dependent on the true clinical status of each suspected case

$$\theta_1 = \gamma\theta_0 + (1-\gamma)\,\theta_1^*$$

		Suspected cases	
Risk factor	Healthy controls	(Apparent)	(True)
1	θ_0	$\gamma\theta_0$	$(1-\gamma)\theta_1^*$
0	$1-\theta_0$	$\gamma(1-\theta_0)$	$(1-\gamma)(1-\theta_1^*)$

Assumptions

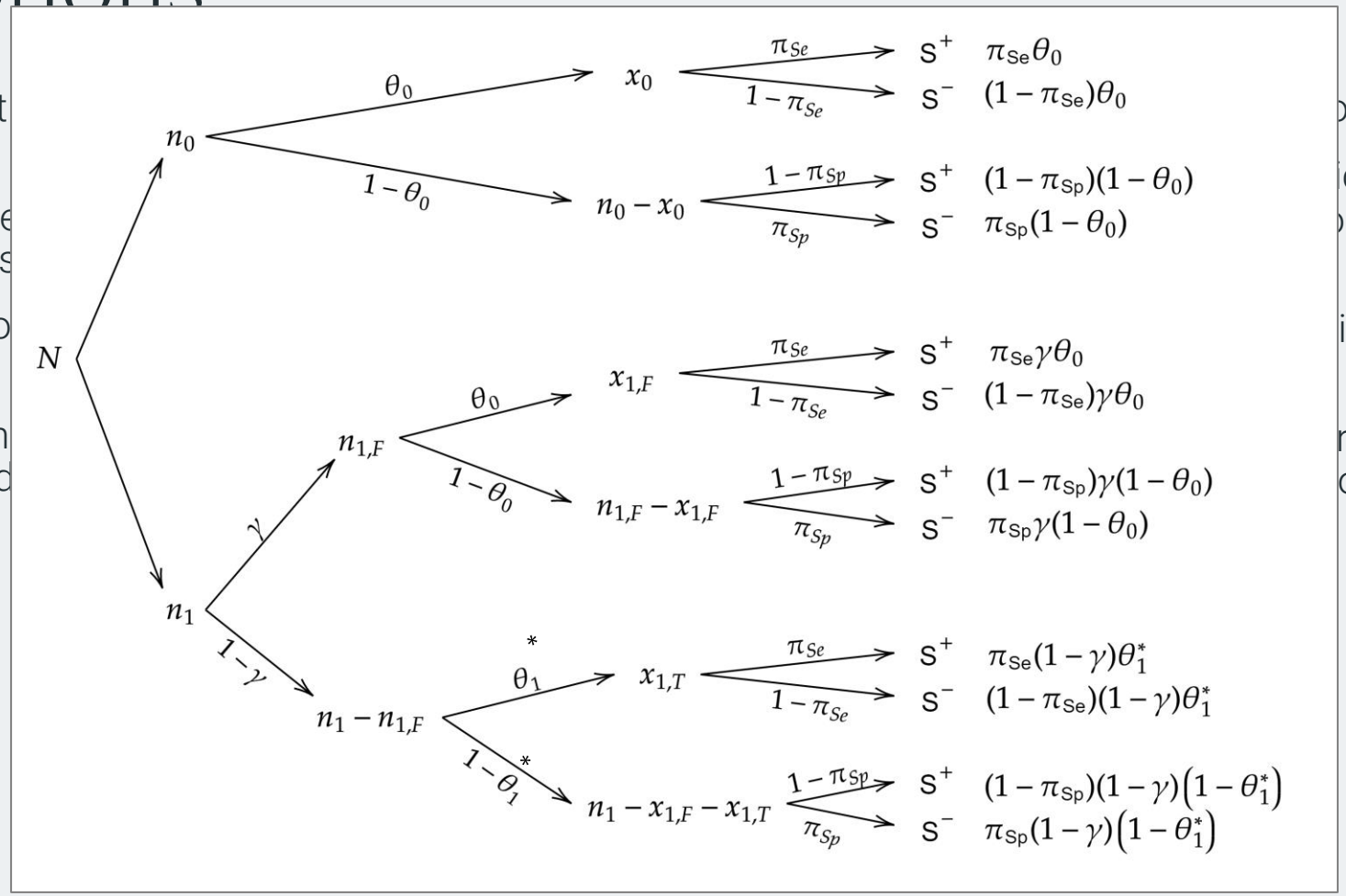
1. ME/CFS cohort can be divided into apparent and true positives
2. Poorly diagnosed cases are considered to be healthy controls (θ_0)
3. Misclassification rate (γ) defined false and true positives
4. γ is only dependent on the true clinical status of each suspected case
5. Participants are classified as seropos or seroneg
6. Tests' sensitivity (π_{Se}) and specificity (π_{Sp}) to determine accuracy for truly exposed and non-exposed individuals
7. Category of undetected false positives and false negatives
8. Binary exposure outcomes for serologic tests are not dependent on the assessed cohorts

$$\theta_1 = \gamma\theta_0 + (1-\gamma)\,\theta_1^*$$

		Suspected cases	
Risk factor	Healthy controls	(Apparent)	(True)
1	θ_0	$\gamma\theta_0$	$(1-\gamma)\theta_1^*$
0	$1-\theta_0$	$\gamma(1-\theta_0)$	$(1-\gamma)(1-\theta_1^*)$

Assumptions

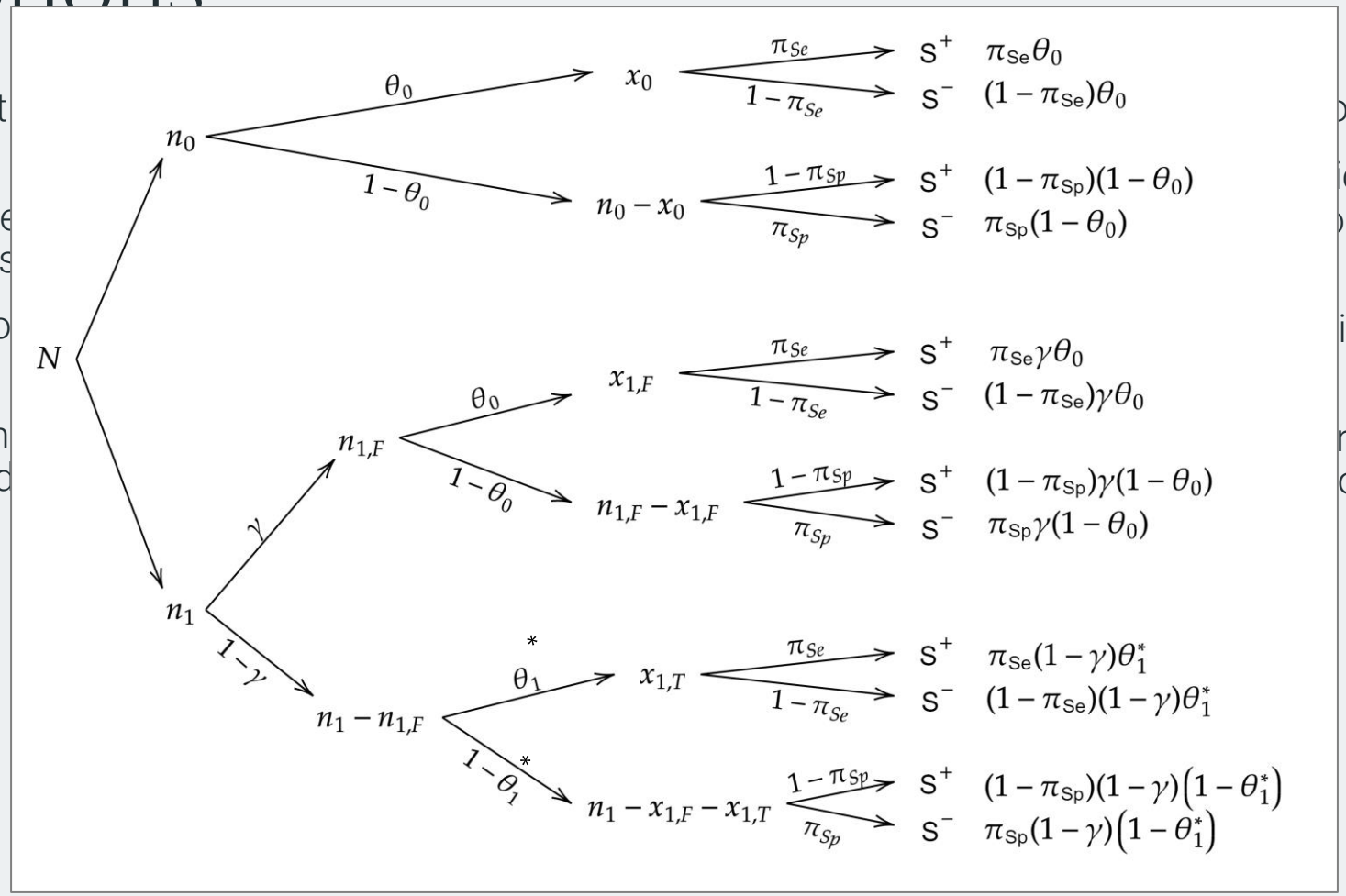
1. ME/CFS cohort … pos or seroneg true positives
2. Poorly diagnose… city (π_{Sp}) to healthy controls… sed and non-
3. Misclassificatio… itives and false positives
4. γ is only depen… rologic tests are each suspected… ohorts



$$\theta_1 = \gamma\theta_0 + (1-\gamma)\,\theta_1^*$$

Assumptions

1. ME/CFS cohort … pos or seroneg true positives
2. Poorly diagnose… city (π_{Sp}) to healthy controls… sed and non-
3. Misclassificatio… itives and false positives
4. γ is only depen… rologic tests are each suspected… ohorts



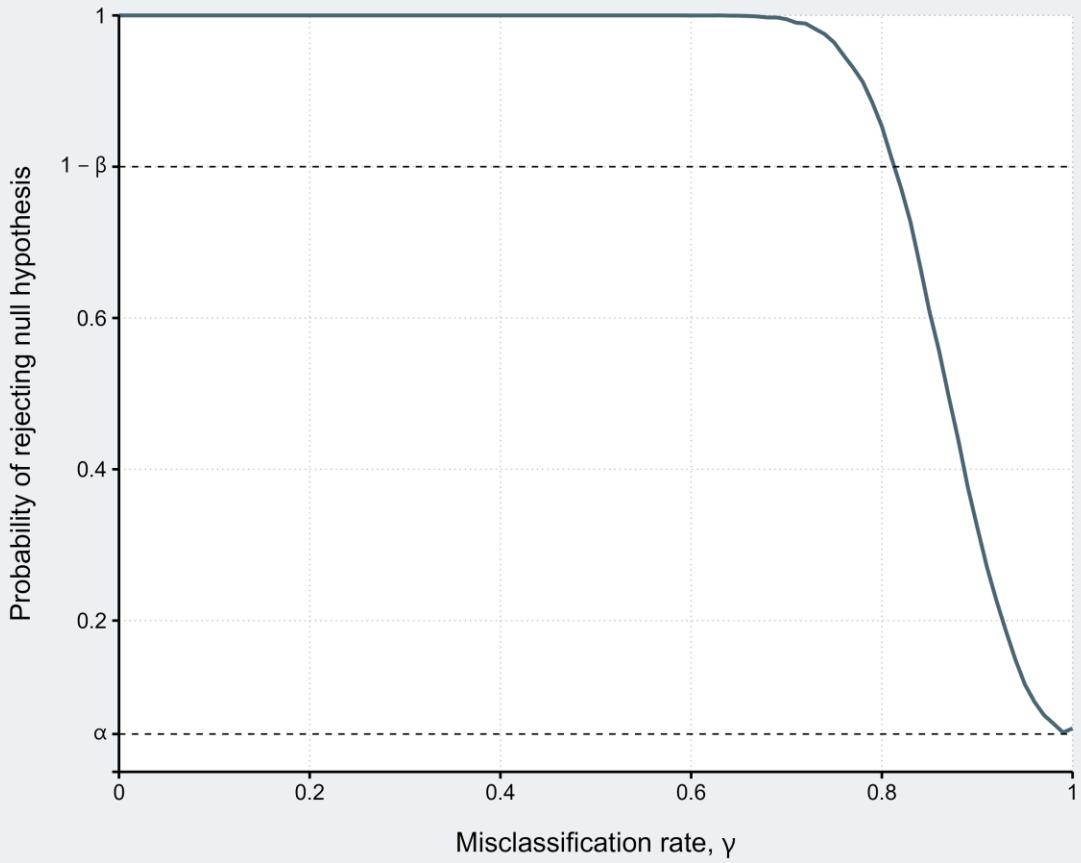
$$\theta_1 = \gamma\theta_0 + (1-\gamma)\,\theta_1^*$$

$$\theta_1 = (\pi_{Se}\gamma\theta_0) + \left[\left(1-\pi_{Sp}\right)\gamma(1-\theta_0)\right] + \left[\pi_{Se}(1-\gamma)\theta_1^*\right] + \left[(1-\pi_{Sp})(1-\gamma)(1-\theta_1^*)\right]$$

Simulation results

A ‘perfect’ study

- $n_0 = n_1 = 1{,}000$ individuals
- $\Delta_T = 5$
- $\theta_0 = 0.5$

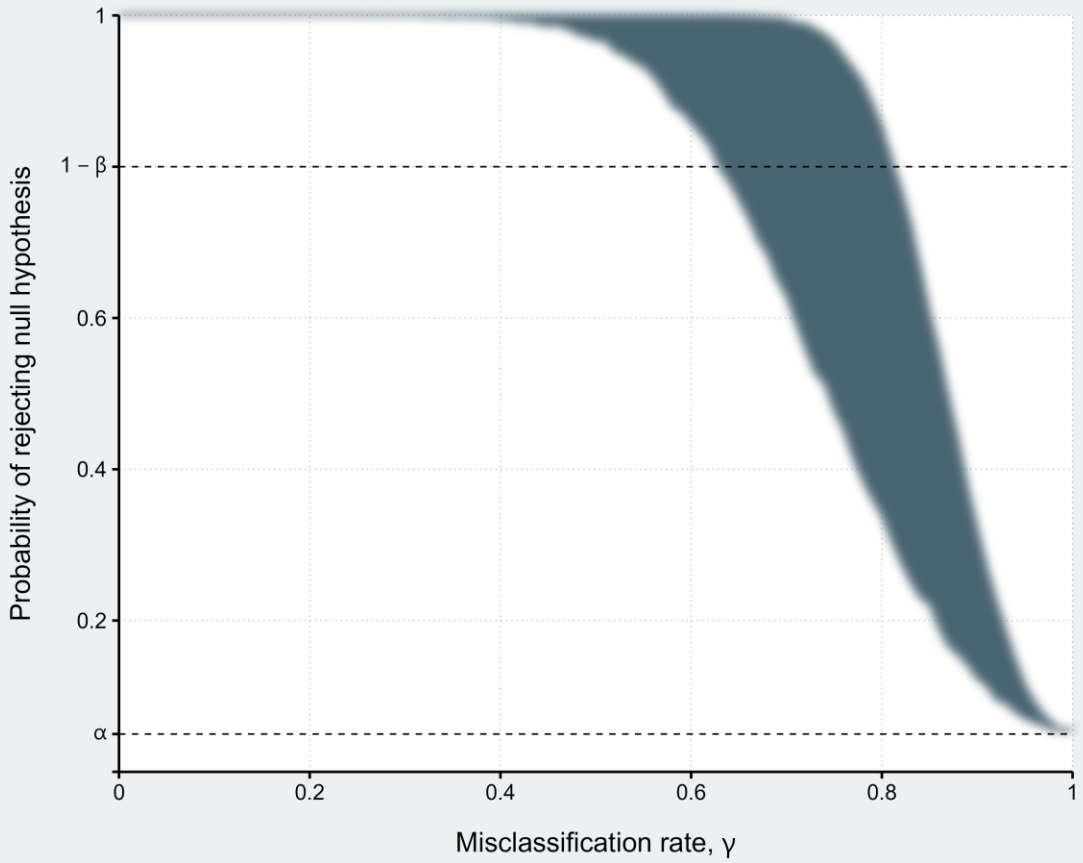


Simulation results

A ‘not-so-perfect’ study

- ~~$n_0 = n_1 =$ 1,000 individuals~~
- $n_0 = n_1 =$ 250 individuals
- $\Delta_T = 5$

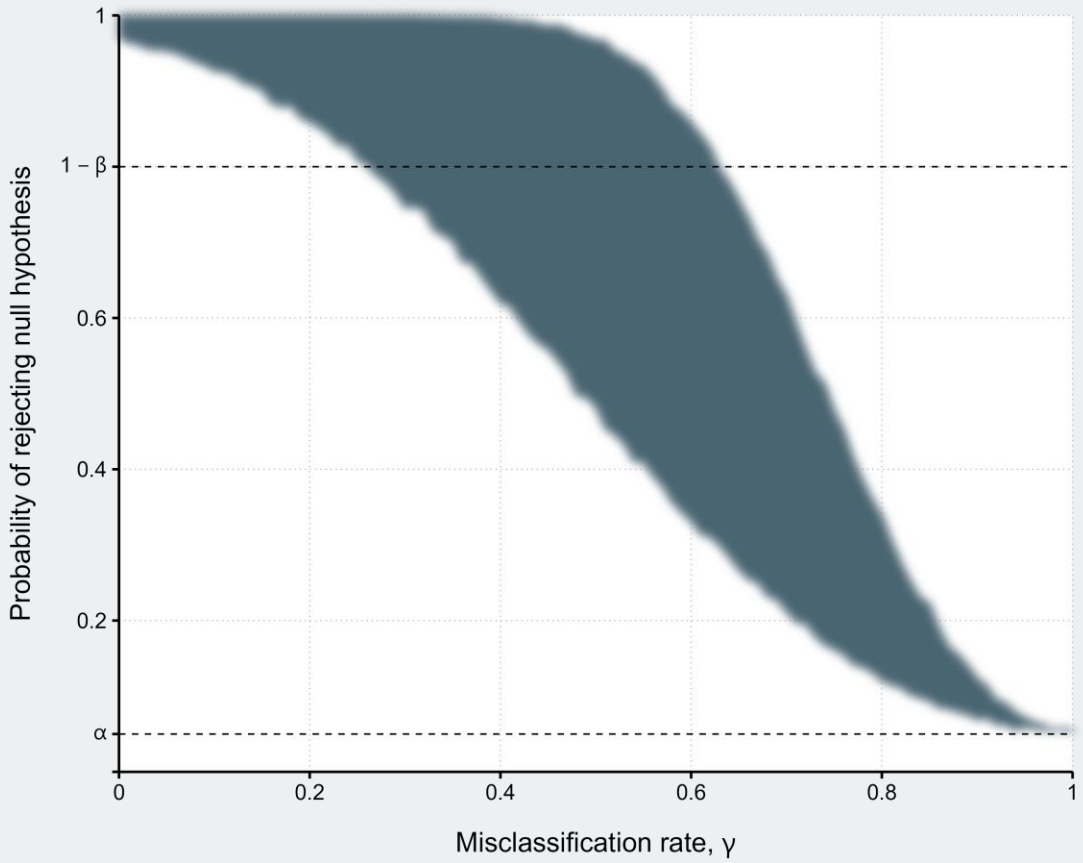
- $\theta_0 = 0.5$



Simulation results

An even ‘not-so-perfect’ study

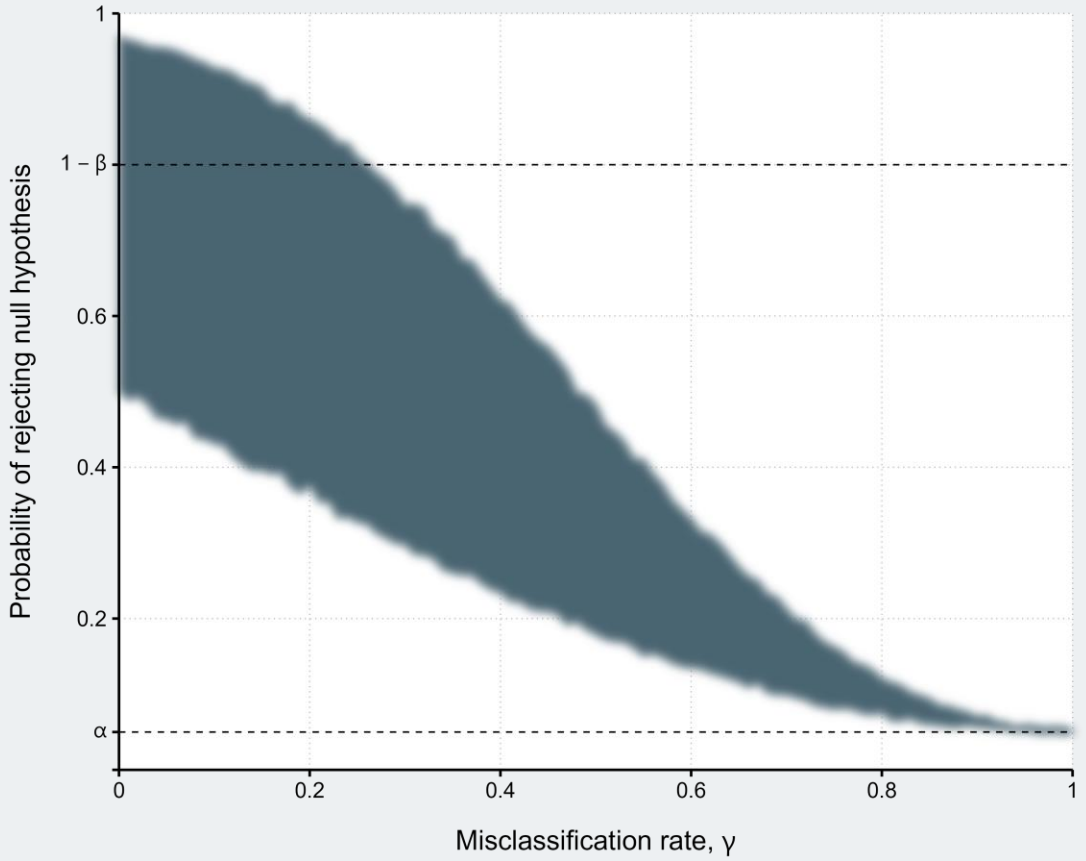
- ~~$n_0 = n_1 =$ 1,000 individuals~~
- $n_0 = n_1 =$ 250 individuals
- ~~$\Delta_T = 5$~~
- $\Delta_T = 2$
- $\theta_0 = 0.5$



Simulation results

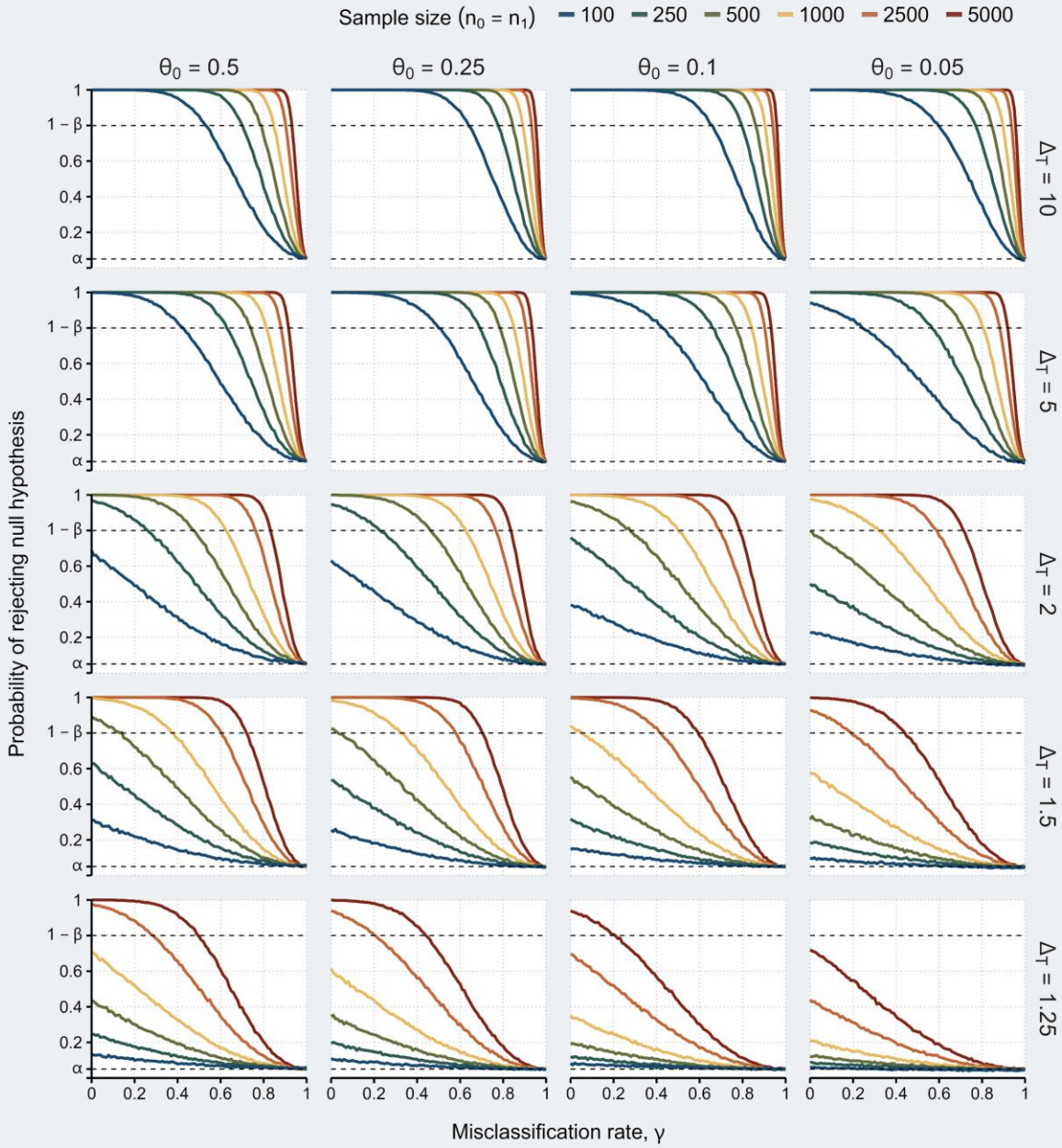
A ‘real’ study

- ~~$n_0 = n_1 = 1{,}000$ individuals~~
- $n_0 = n_1 = 250$ individuals
- ~~$\Delta_T = 5$~~
- $\Delta_T = 2$
- ~~$\theta_0 = 0.5$~~
- $\theta_0 = 0.05$



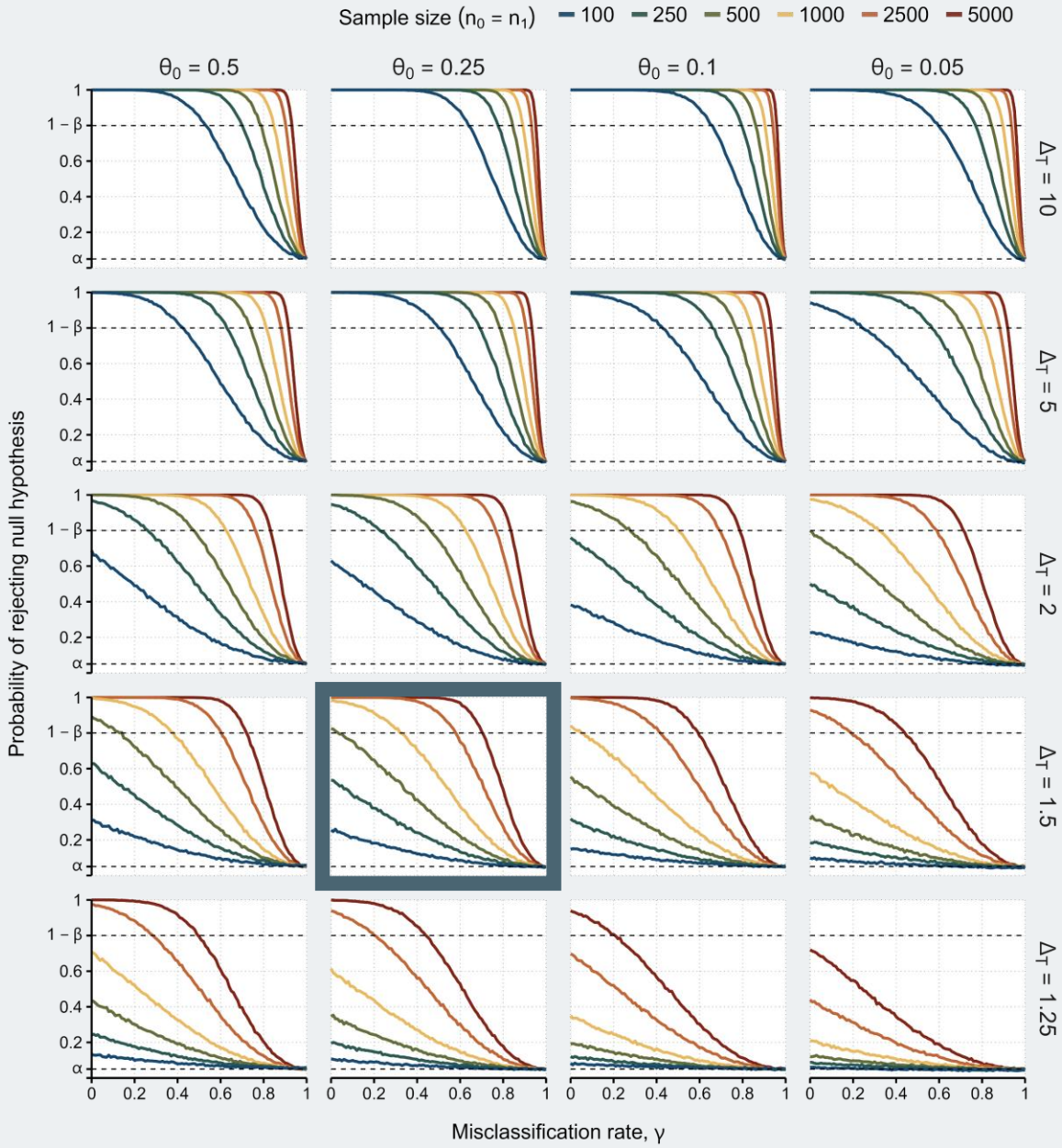
Simulation results

Δ_T \ θ_0	0.05	0.1	0.25	0.5	n_i
10	0.59	0.65	0.63	0.53	100
5	0.24	0.42	0.50	0.42	
2	–	–	–	–	
1.5	–	–	–	–	
1.25	–	–	–	–	
10	0.76	0.79	0.77	0.70	250
5	0.56	0.66	0.69	0.62	
2	–	–	0.22	0.26	
1.5	–	–	–	–	
1.25	–	–	–	–	
10	0.84	0.86	0.84	0.78	500
5	0.71	0.76	0.78	0.74	
2	–	0.27	0.46	0.47	
1.5	–	–	0.03	0.14	
1.25	–	–	–	–	
10	0.89	0.90	0.89	0.84	1000
5	0.80	0.84	0.85	0.81	
2	0.31	0.50	0.62	0.62	
1.5	–	0.05	0.32	0.38	
1.25	–	–	–	–	
10	0.93	0.94	0.93	0.90	2500
5	0.88	0.90	0.90	0.88	
2	0.59	0.69	0.76	0.76	
1.5	0.19	0.41	0.57	0.60	
1.25	–	–	0.20	0.28	
10	0.95	0.95	0.95	0.93	5000
5	0.91	0.93	0.93	0.91	
2	0.71	0.78	0.83	0.83	
1.5	0.44	0.59	0.70	0.72	
1.25	–	0.20	0.44	0.49	



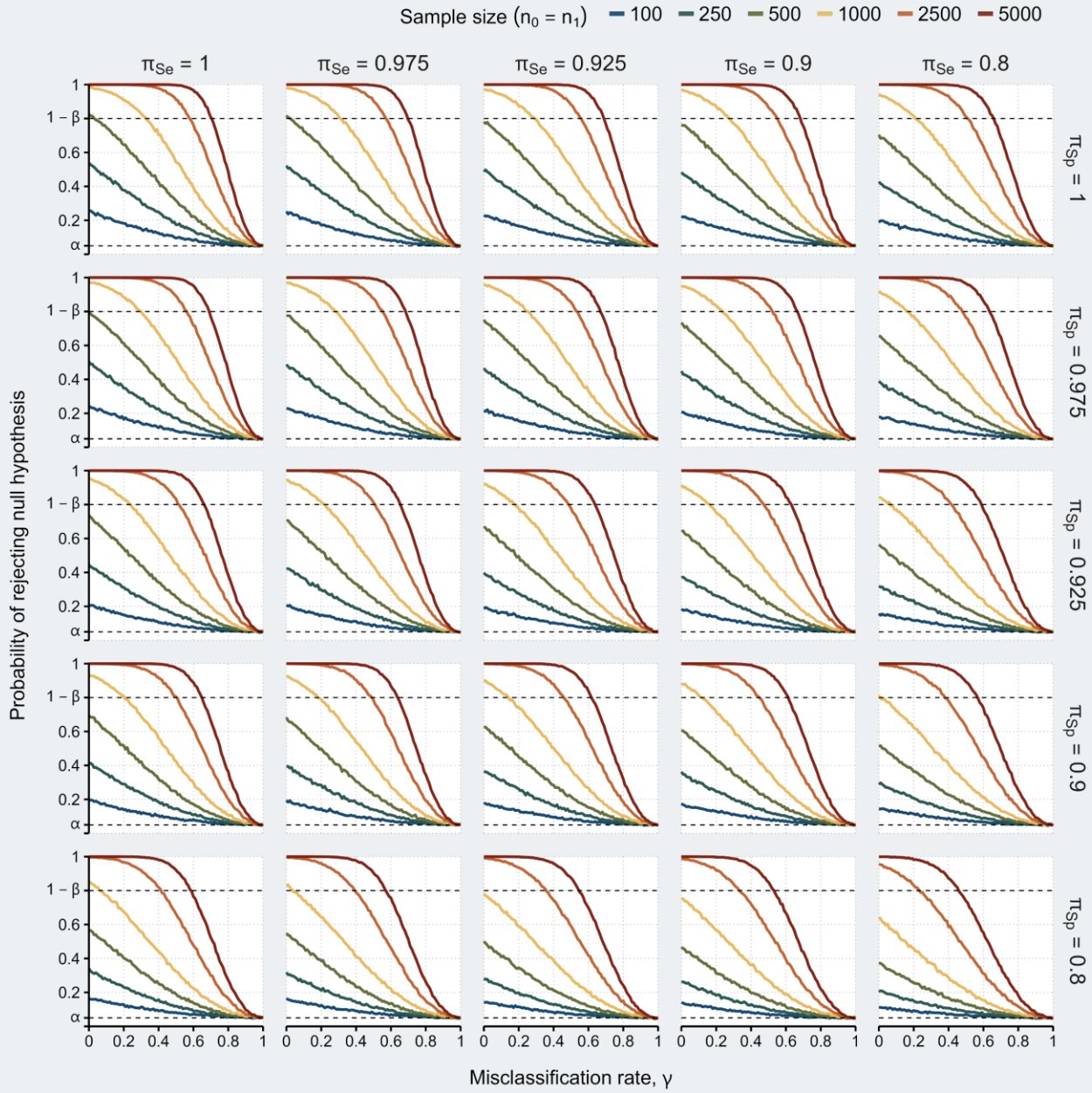
Simulation results

Δ_T \ θ_0	0.05	0.1	0.25	0.5	n_i
10	0.59	0.65	0.63	0.53	100
5	0.24	0.42	0.50	0.42	
2	–	–	–	–	
1.5	–	–	–	–	
1.25	–	–	–	–	
10	0.76	0.79	0.77	0.70	250
5	0.56	0.66	0.69	0.62	
2	–	–	0.22	0.26	
1.5	–	–	–	–	
1.25	–	–	–	–	
10	0.84	0.86	0.84	0.78	500
5	0.71	0.76	0.78	0.74	
2	–	0.27	0.46	0.47	
1.5	–	–	0.03	0.14	
1.25	–	–	–	–	
10	0.89	0.90	0.89	0.84	1000
5	0.80	0.84	0.85	0.81	
2	0.31	0.50	0.62	0.62	
1.5	–	0.05	0.32	0.38	
1.25	–	–	–	–	
10	0.93	0.94	0.93	0.90	2500
5	0.88	0.90	0.90	0.88	
2	0.59	0.69	0.76	0.76	
1.5	0.19	0.41	0.57	0.60	
1.25	–	–	0.20	0.28	
10	0.95	0.95	0.95	0.93	5000
5	0.91	0.93	0.93	0.91	
2	0.71	0.78	0.83	0.83	
1.5	0.44	0.59	0.70	0.72	
1.25	–	0.20	0.44	0.49	



Simulation results

π_{Sp} \ π_{Se}	1	0.975	0.925	0.9	0.8	n
1	–	–	–	–	–	100
0.975	–	–	–	–	–	
0.925	–	–	–	–	–	
0.9	–	–	–	–	–	
0.8	–	–	–	–	–	
1	–	–	–	–	–	250
0.975	–	–	–	–	–	
0.925	–	–	–	–	–	
0.9	–	–	–	–	–	
0.8	–	–	–	–	–	
1	0.03	0.02	–	–	–	500
0.975	–	–	–	–	–	
0.925	–	–	–	–	–	
0.9	–	–	–	–	–	
0.8	–	–	–	–	–	
1	0.32	0.30	0.29	0.27	0.20	1000
0.975	0.29	0.28	0.25	0.23	0.16	
0.925	0.23	0.21	0.17	0.16	0.06	
0.9	0.20	0.17	0.14	0.11	–	
0.8	0.06	0.03	–	–	–	
1	0.57	0.57	0.55	0.55	0.50	2500
0.975	0.56	0.55	0.53	0.52	0.47	
0.925	0.51	0.50	0.48	0.47	0.41	
0.9	0.50	0.49	0.45	0.44	0.37	
0.8	0.41	0.39	0.35	0.33	0.23	
1	0.70	0.69	0.68	0.68	0.65	5000
0.975	0.69	0.68	0.66	0.66	0.63	
0.925	0.66	0.65	0.63	0.63	0.58	
0.9	0.64	0.63	0.61	0.61	0.56	
0.8	0.58	0.56	0.54	0.53	0.45	



Application to real-life data

Example 1: *Steiner et al. (2020)*

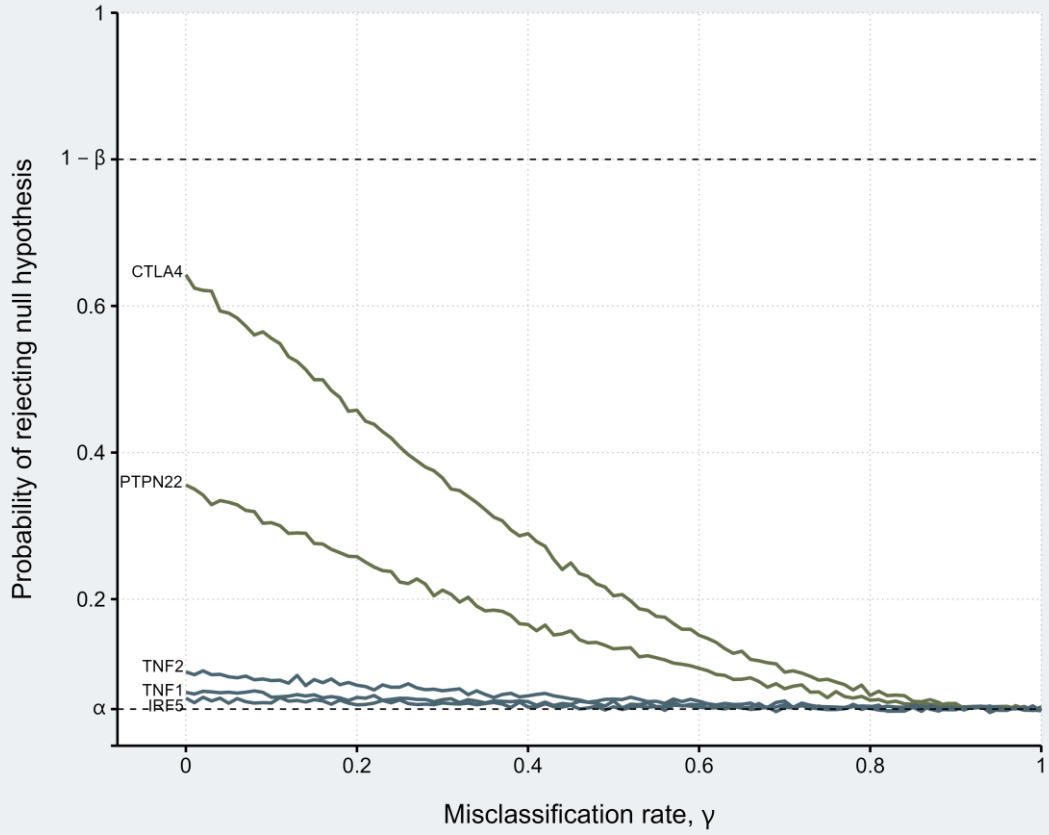
Autoimmunity-Related Risk Variants in PTPN22 and CTLA4 Are Associated With ME/CFS With Infectious Onset

Sophie Steiner[1†], Sonya C. Becker[1†], Jelka Hartwig[1], Franziska Sotzny[1], Sebastian Lorenz[1], Sandra Bauer[1], Madlen Löbel[2], Anna B. Stittrich[3,4], Patricia Grabowski[1] and Carmen Scheibenbogen[1,3]*

[1] Institute of Medical Immunology, Charité-Universitätsmedizin Berlin, Corporate Member of Freie Universität (FU) Berlin, Humboldt-Universität zu Berlin and Berlin Institute of Health (BIH), Berlin, Germany, [2] Carl-Thiem-Klinikum Cottbus gGmbH, Research Center, Cottbus, Germany, [3] BIH Center for Regenerative Therapies, Charité-Universitätsmedizin Berlin, Berlin, Germany, [4] Labor Berlin—Charité Vivantes GmbH, Berlin, Germany

- Genetic markers associated with autoimmunity
- γ = 0.24

	snp	theta0	or_t	theta1	theta1_true
1:	PTPN22	0.08	1.63	0.12	0.13
2:	CTLA4	0.56	1.54	0.63	0.66
3:	IRF5	0.51	0.94	0.50	0.49
4:	TNF1	0.16	0.89	0.15	0.14
5:	TNF2	0.13	0.84	0.11	0.11



Application to real-life data

Example 2: *Cliff et al. (2019)*

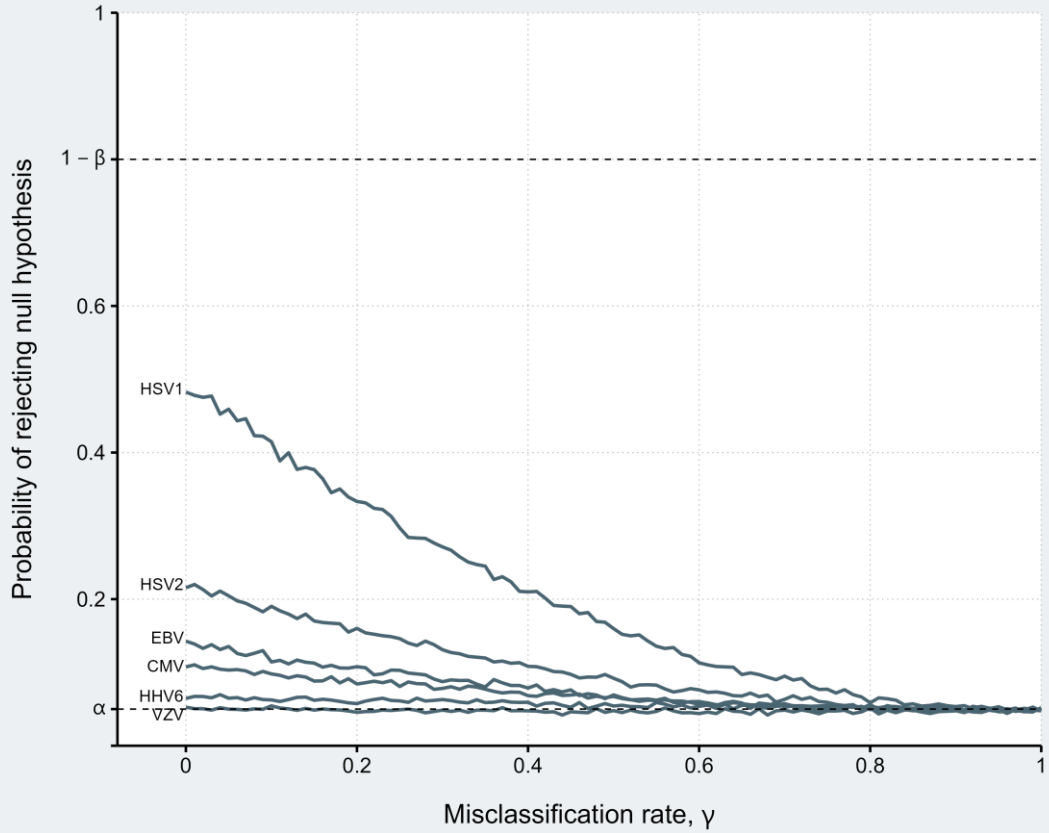
Cellular Immune Function in Myalgic Encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS)

Jacqueline M. Cliff[1], Elizabeth C. King[1], Ji-Sook Lee[1], Nuno Sepúlveda[1,2], Asia-Sophia Wolf[1], Caroline Kingdon[3], Erinna Bowman[3], Hazel M. Dockrell[1], Luis Nacul[3], Eliana Lacerda[3†] and Eleanor M. Riley[1†‡]*

[1] Department of Immunology and Infection, Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, London, United Kingdom, [2] Centre of Statistics and Applications, University of Lisbon, Lisbon, Portugal , [3] Department of Clinical Research, Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, London, United Kingdom

- Seroprevalence of six human herpes viruses
- γ = 0.22

	virus	theta0	or_t	theta1	theta1_true
1:	CMV	0.37	0.84	0.35	0.33
2:	EBV	0.93	0.65	0.88	0.89
3:	HSV1	0.42	1.60	0.51	0.54
4:	HSV2	0.34	1.36	0.40	0.41
5:	VZV	0.97	0.75	0.94	0.96
6:	HHV6	0.95	1.27	0.94	0.96



Application to real-life data

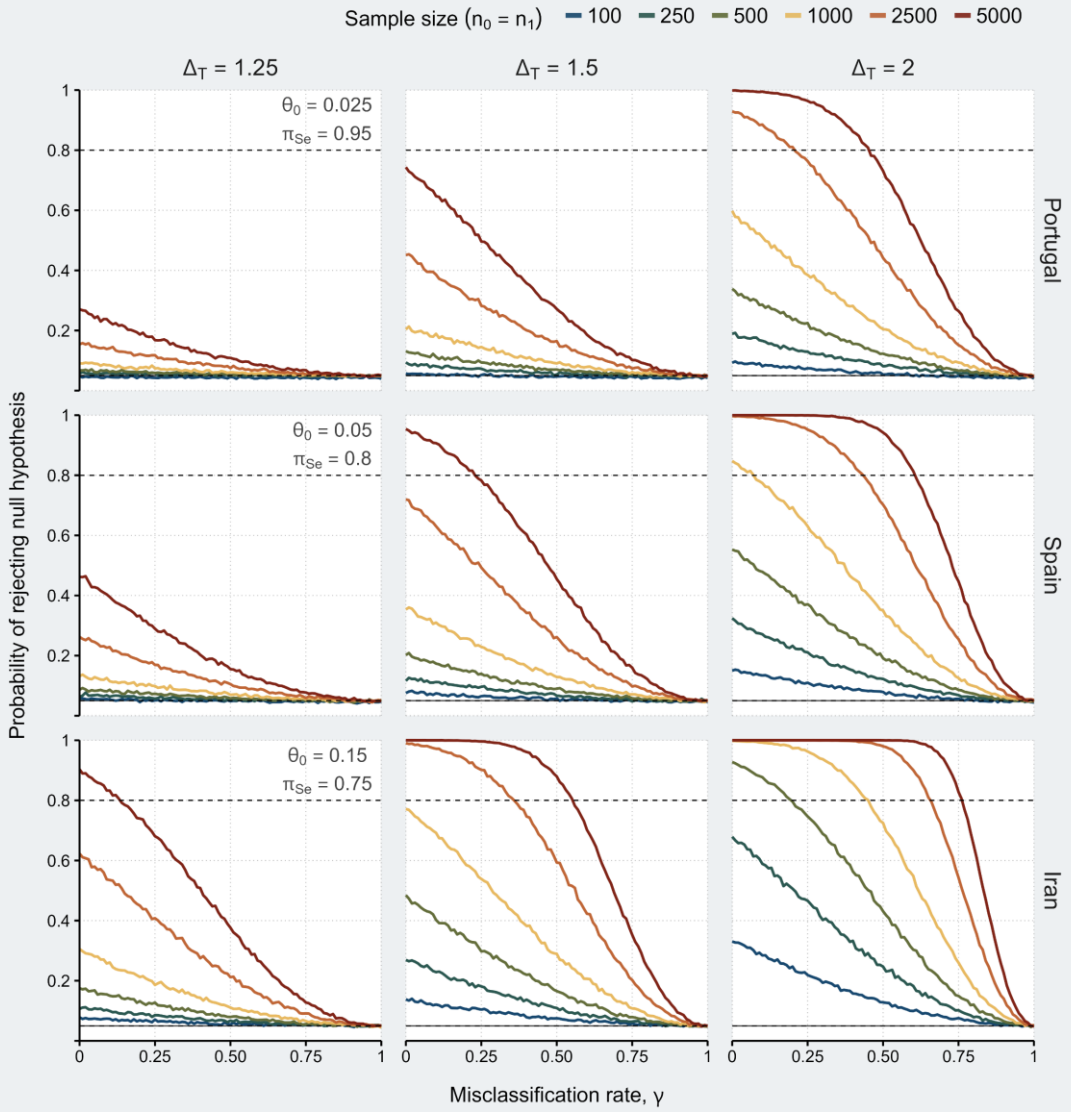
Example 3

Impact of Misclassification and Imperfect Serological Tests in Association Analyses of ME/CFS Applied to COVID-19 Data

João Malato[1], Luís Graça[1], and Nuno Sepúlveda[2,3]

[1] Instituto de Medicina Molecular João Lobo Antunes, Faculdade de Medicina, Universidade de Lisboa, Portugal
jmalato@medicina.ulisboa.pt,
[2] CEAUL – Centro de Estatística e Aplicações da Universidade de Lisboa, Portugal
[3] Faculty of Mathematics and Information Science, Warsaw University of Technology, Poland

- Links between ME/CFS and past viral infections
- COVID-19
- Serological testing



Discussion and future work

- Power studies
- Sample size estimation

- Sensitivity analysis
- Account for potential misclassification of patients
- Transversal to complex diseases

- Improving diagnosis
- Improve consistency across results

Going back to study UKMEB data

- Symptom assessment questionnaires
- Immunological data
- Gene expression data

Thank you

Impact of imperfect diagnosis in ME/CFS association analyses

João Malato

January 31st, 2022

Wydział Matematyki i Nauk Informacyjnych
POLITECHNIKA WARSZAWSKA
LISBOA UNIVERSIDADE DE LISBOA
FACULDADE DE MEDICINA LISBOA
CENTRO ACADÉMICO DE MEDICINA DE LISBOA
Instituto de Medicina Molecular | João Lobo Antunes
FCT Fundação para a Ciência e a Tecnologia