RECORDING METHOD FOR SSc ORGANS INVOLVMENTS

We defined the different organ involvements of SSc using, for each organ, the 2013 Classification criteria for SSc of ACR/EULAR[1]

The cutaneous and articular involvement was evaluated clinically: skin thickening of the fingers, fingertips lesion, telangiectasia. The articular involvement was evaluated clinicated by the presence of articular pain or synovitis.

For cardiac involvement we considered an elevated troponin level (not explained otherwise) or a specific signal on cardiac MRI (confirmed by a radiologist of our centre).

Pulmonary involvement was recorded using radiographic criteria, all patients with SSc-ILD in this study had a CT demonstrating ILD, confirmed by a radiologist of our centre.

Digestive can be heterogeneous and involve all part of the digestive tract from mouth to anus, and some of its symptoms can be non-specific, to improve specificity we chose to use objective criteria such as: the presence of an oesophagitis of grade B or more according to Los Angeles classification on a gastric fibroscopy made by a gastroenterologist of our centre, the presence of oesophageal or lower digestive tract motor involvement in manometry made in the gastroenterology department of our centre, the presence of a small intestinal overgrowth (diagnosed with a respiratory test), the presence of a chronic intestinal pseudo-obstruction (confirmed by a radiologist of our centre with a CT).

Finally, kidney involvement was recorded if the patient had a history of scleroderma renal crisis in their medical record.

RECORDING METHOD FOR COMORBIDITIES

Tobacco use was recorded from the patient medical record, we considered past or present smoking to be relevant.

Obesity was recorded using the body mass index (BMI (>/= 30 kg/m2))..

Myocardial infarct was recorded in the medical record.

Cardiopathy was recorded in the patient medical record (presence of a valvar dysfunction, cardiac rhythmic disorder, hypertensive cardiopathy with altered function, specific involvement of SSc or any other cardiac disease), confirmed if possible by a cardiac echography, an EKG…

Lower limbs endarteritis disease was recorded from the medical record.

High blood pressure was recorded from the medical record

Stroke was recorded from the medical record.

Dementia was recorded from the medical record.

Chronic obstructive pulmonary disease was recorded from the medical record, if possible confirmed by functional respiratory test.

Gastric ulcer was record from the medical record, if possible confirmed by a digestive endoscopy or the prescription of a proton pump inhibitor.

Hepatopathy was recorded from the medical record, and the cause was recorded if available.

Diabetes was recorded from the medical recorded and confirmed by the prescription on an anti-diabetes drug or insulin. For each diabetic patient we also recorded the type, the presence of a BMI >25, the prescription of metformin, and the prescription of insulin.

Chronic kidney disease was recorded from the medical record.

HIV infection was recorded from the medical record, and confirmed by the presence of an anti-retroviral drug in the prescription sheet.

Depression and anxiety was recorded from the medical record, and confirmed by the presence of anxiolytics or antidepressant in the prescription sheet.

Osteoporosis was defined as a history of low-energy mechanisms causing fractures and/or treated osteoporosis, and recorded as such from the patient medical record.

Polypharmacy was recorded from the last available prescription sheet, we only took into account “chronic” prescription, excluding medication that were took for acute event (infections, for instance)

Neoplasia were defined as solid tumours, leukaemia or lymphoma, and were recorded as such from the patient medical record.

DATA ANALYSIS

The analysis was performed on a python environment using conda. It was tested on an ASUS ROG Strix G733ZX_G733ZX laptop, with linux (Ubuntu 23.04). All requirements to create the

python environment can be found in the project (conda_environment.yml, requirements.txt).

DATA PREPARATION

The data was formatted, first manually (each input variable was checked and labeled, binarynumerical values were retained), then automatically. Indeed, a python script (1-prepare_data.py) allowed correction of input errors, filling missing values, creating newvariables (e.g. ‘polymedication>n...) and formatting time values with the calculation of a newvariable, ‘duration_days’. The main outcome is ‘deceased’ status (binary ‘0’ or ‘1’).

SIMPLE STATISTICS AND SURVIVAL PLOTS

Then we calculated simple statistics (Prevalence of disease, % of death, mean, median of numerical values...), and plot Kaplan-Meier survival curves over time on subgroups. Statistics are performed using pandas and numpy. Survival curves are plotted with scikit survival library using 'kaplan_meier_estimator'.

SURVIVAL REGRESSION MODEL COMPUTING AND 'COMORBIDOME PLOT’

First, correlations between variables are checked by plotting a correlation matrix with 'seaborn' in order to assess a covariate issue.

Then a regression model is performed by fitting the formatted variables using the lifelines python library. The 'CoxPHFitter' is used for the semi-parametric model, and it implements fitting Cox’s proportional hazard model non-parametrically, using Breslow’s method (the entire model is the traditional semi-parametric Cox model). Ties are handled using Efron’s method. The 'WeibullAFTFitter' is used for parametric model, and it implements a Weibull AFT model. We used a parametric analysis to confirm the results of Cox model.

To evaluate the performances of the model, we used several metrics:

> Cross-validation: with the 'cross_val_score' of the scikit-learn library, it splits the dataset in 'n' parts, it takes apart a fraction for validation and uses the rest to train a model. It applies the model on validation part and calculates accuracy. It performs this operation 'n' times and calculates the mean value.

Concordance_index, likelihood ratio, asumptions: are attributes of lifelines model (see doc:https://lifelines.readthedocs.io/en/stable/fitters/regression/CoxPHFitter.html?highlight=concordance_index_). The model is used to perform predictions on random patients of the dataset to evaluate clinical relevance.

We plotted log(HR) values on simple graphs. And we plotted 1/HR on 'a Comorbidome plot' as seen on other studies (1). It plots each variable of the model, the size of the circle is proportional to prevalence, and the proximity to center is proportional to 1/HR. The dot circle is placed were 1/HR=1. Inside the circle, the variables are associated to mortality. The statistical significance is represented with a red star.

A fine-tuning was performed by repeating automated analysis with several parameter values. The parameters used for the paper after fine-tuning are:

- set_alive_lost: True (lost patients are not right censured)
- prevalence_min: 2.0 (if the prevalence < 2%, the variable is excluded of the model)
- prevalence_max: 95.0 (if the prevalence > 95%, the variable is excluded of the model)

DATA AVAILABILITY

- All the code used is public, and can be found at: https://gitub.u-bordeaux.fr/public_projects/02.2023-t2-comorbidome
- The data are private and kept in a local server.

REFERENCE:

(1) Hoogen, Frank van den, Dinesh Khanna, Jaap Fransen, Sindhu R. Johnson, Murray Baron, Alan Tyndall, Marco Matucci-Cerinic, et al. « 2013 Classification Criteria for Systemic Sclerosis: An American College of Rheumatology/European League against Rheumatism Collaborative Initiative ». *Annals of the Rheumatic Diseases* 72, nº 11 (novembre 2013): 1747-55. https://doi.org/10.1136/annrheumdis-2013-204424.

(2) Jean-Marie Grosbois, Axelle Détrée, Adeline Pierache, Nathalie Bautin, Thierry Pérez, Benoit Wallaert,Cécile Chenivesse, Olivier LeRouzic. (2023)Impact of Cardiovascular and Metabolic Comorbidities on Long Term Outcomes of Home-based Pulmonary Rehabilitation in COPD. International journal of chronic obstructive disease

Table 5: evaluation of the model – clinical examples of patients and correlation between their HR and comorbidity profile

Patient	HR	Sex	Age
303	0.106	F	22
252	0.228	F	44
26	0.567	F	57
140	0.863	F	69
170	1.171	F	69
131	2.157	F	75
102	9.022	M	67
152	12.912	F	79

Patient	Comorbidities										
	Tobacco use	COPD	Gastric ulcer	Liver disease	CKD	Neoplasia	D/A	Osteoporosis	Obesity	Cardiovascular	Polypharmacy
303	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	0
252	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	2
26	YES	NO	NO	NO	NO	NO	NO	NO	NO	NO	1
140	YES	NO	NO	NO	NO	YES	NO	NO	NO	NO	3
170	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	3
131	NO	NO	NO	NO	NO	NO	NO	NO	NO	NO	4
102	YES	NO	NO	NO	NO	YES	NO	NO	NO	YES	8
152	NO	NO	YES	NO	NO	NO	NO	NO	NO	YES	15

Patient	SSc involvement			
	Cardiovascular	Lung	Renal	Digestive
303	NO	NO	NO	NO
252	NO	NO	NO	YES
26	NO	YES	NO	NO
140	NO	NO	NO	NO
170	NO	YES	NO	YES
131	YES	YES	NO	NO
102	YES	NO	NO	NO
152	NO	YES	NO	NO

HR, hazard ratio
F, female
M, male
COPD, chronic obstructive pulmonary disease
CKD, chronic kidney disease
D/A, depression/anxiety
SSc, systemic sclerosis