Project Description, Scoping and Datasets

A Data Driven Tool to Assess Clinical Outcome Risk

Team 62

María Paula Álvarez, Cristian Rodríguez, Luis Daniel Chavarría,
Jeyson Guzmán, Juan Barrios and Luis Serna.

Correlation One and Colombian Ministry of Information Technologies and Communications | Data Science for All Colombia 6.0
Week 5 Project Submission
May 2022

In this document we present an approximation to our project's description, scoping, and datasets to use.

Problem overview

Medical treatment has had a great evolution throughout history, from the prevention until the treatment of diseases that previously had no cure. Nevertheless, the time it takes for doctors to identify diseases and begin the right treatment has been a race against time since a long time ago. Developed countries which have more resources have more capacity to identify these problems earlier than developing countries where a huge gap between the quality and the time of response can be seen.

Nowadays plurpathologic patient cohort management is done reactively instead of proactively. This means that people must schedule their controls themselves and doctors must usually fix damage that could've been prevented.

Problem definition

With this project we attempt to construct a data-product able to assess the risk of different clinical outcomes such as hospitalization or death risk. This product will provide reliable feedback on a cohort of patients and identify who has the highest risk of a particular clinical outcome.

Medical outcomes assessment is a current research problem in public health focused on the estimation of clinical results or effects that different treatments or interventions can have in a patient population. The early determination of these outcomes can help to bring an adequate response for each patient, better focused on its symptoms, functional status, and quality of life, as well as reduce healthcare costs by taking a preventive approach and drawing out possible risk factors.

In this context, we intend to answer the question: *what are the medical outcome predictors in a sample Colombian patient population based on previously gathered clinical information?*

Importance of the problem

Medicine has come a long way since its inception and its practices have had drastic changes throughout history. In ancient times their applications were mainly rituals or involved the usage of plants but nowadays thanks to technology, investigation, and artificial intelligence, etc. we are witnesses of the medical advance and new techniques that contribute more efficiently to the protection of health, the cure of diseases, also to improve the quality of health services and prevention of health risk.

The quality of health services is one of the main concerns of governments around the world. For the Ministry of Health in Colombia, it is a priority to guarantee access and quality care in which the minimum conditions for a service as important as this one is met. The Colombian law 1431 of 2011 states “The General System of Social Security in Health will be aimed at generating conditions that protect the health of Colombians, being the well-being of the user the central axis and articulating nucleus of health policies”.

Considering the previous description, we can develop a predictive model to measure the attention scales, hospitalization time and reduce the probability of having patients unfortunately die.

Data sets to use

We have been provided, and were allowed to use, anonymized pluripathologic patient data for a 1-year follow-up program from a healthcare provider in Colombia. The anonymized data contains 5,511 patients in the entire year 2020, 1 row per patient. There are 137 variables for each one of them, some functional, some related to the use of resources and some other about chronic pathologies. No data dictionary has been provided, and there is some missing data for some patients.

The name of the columns is self-explanatory for some of the variables; however, we present the next table to describe the dataset and generate a list of features to be aware of (size, missing data, data types and groups of variables). In addition, we attach to this document a spreadsheet with a detailed description of each variable.

Dataset	Source	Description	Topics
Pluripathologic patients features.	Colombian healthcare provider.	Anonymized pluripathologic patient data. 5511 patients in the entire year 2020. 137 variables for each patient.	Patient characteristics. Clinical billing. Use of medical resources. Chronic diseases. Clinical outcome.

Annex 1. Data Dictionary

No.	Name	Description	Type	Values
[1]	edad	The age of the patient in years.	int	18 to104
[2]	sexo	Gender	str	['F' for Female, 'M' for Male]
[3]	n_urg	Number of urgency cases during 2020	float	0 to 9
[4]	n_hosp	Number of hospitalizations during 2020.	float	0 to 39
[5]	n_cx	Number of surgeries during 2020	float	~ -0.3 to ~23 (anonimyzed)
[6]	n_ce_med_general	Number of general medicine external consultation during 2020	float	anonymized
[7]	n_ce_med_especializada	Number of external medical appointments during 2020.	int	0 to 39
[8]	estancia_sala	Time in hospitalization room in days.	int	0 to 120 days
[9]	estancia_uce	Time on Special Care Unit in days.	float	0 to 23
[10]	estancia_uci	Time on Intensive Care Unit in days.	float	0 to 45
[11]	pancreatobiliar	Number of pancreatobiliar episodes during 2020	float	0 to 2
[12]	glaucoma	Number of glaucoma episodes during 2020	int	0 to 3
[13]	hipertension	Number of episodes related to arterial hipertension during 2020.	int	0 to 12
[14]	cerebrovascular	Number of episodes of brain strokes in medical consultations during 2020	int	0 to 10
[15]	vertigo_y_alteraciones_auditivas	Number of episodes of vertigo and ear difficulties during 2020	int	0 to 4
[16]	bradicardias_y_enfermedades_de_la_conduccion	Number of bradycardias and conduction diseases episodes during 2020.	int	0 to 5
[17]	otras_genitourinarias	Number of other genitourinary diseases episodes during 2020.	int	0 to 15
[18]	depresion_y_alteraciones_del_animo	Number of Depression and Mood Disorders episodes during 2020	int	0 to 17
[19]	epoc	Number of episodes related to COPD during 2020.	int	0 to 13
[20]	enfermedad_renal_cronica	Number of episodes of Chronic Kidney Disease in medical consultations during 2020	int	0 to 16
[21]	enfermedades_de_la_tiroides	Number of episodes of thyroid disease during 2020	int	0 to 8
[22]	otras_enfermedades_digestivas	Number of other digestive diseases episodes during 2020.	int	0 to 3
[23]	hematologicas	Number of hematological diseases episodes during 2020.	int	0 to 16
[24]	alteraciones_otorrinolaringologicas	Number of Otorhinolaryngological alterations episodes during 2020	int	0 to 7
[25]	obesidad	Number of episodes related to Obesity during 2020.	int	0 to 7
[26]	enfermedades_de_la_prostata	Number of episodes of Prostate Diseases in medical consultations during 2020	int	0 to 7
[27]	esquizofrenia	Number of episodes of esquizofrenia during 2020	int	0 to 10
[28]	vascular_periferica	Number of peripheral vascular disease episodes during 2020.	int	0 to 4
[29]	alteraciones_de_la_agudeza_visual	Number of visual acuity disturbances episodes during 2020.	int	0 to 2
[30]	cromosomicas	Number of Chromosomal diseases episodes during 2020	int	0 to 7
[31]	osteoporosis	Number of episodes related to Osteoporosis during 2020.	int	0 to 8
[32]	infecciones_cronicas	Number of episodes of Chronic Infections in medical consultations during 2020	int	0 to 6
[33]	anemia	Number of episodes of anemia during 2020	int	0 to 8
[34]	somatomorfos	Number of neurotic, stress-related and somatoform diseases episodes during 2020.	int	0 to 8
[35]	otras_enfermedades_oculares	Number of other eye diseases episodes during 2020.	int	0 to 3
[36]	colitis_y_gastrointestinales_inferiores	Number of Colitis and Lower Gastrointestinal episodes during 2020	int	0 to 9
[37]	artrosis	Number of episodes related to Osteoarthritis during 2020.	int	0 to 10

[38]	enfermedad_isquemica_cardiaca	Number of episodes of Ischemic Heart Disease in medical consultations during 2020	int	0 to 13
[39]	otras_respiratorias	Number of episodes of other respiratory diseases during 2020	int	0 to 5
[40]	cataratas	Number of cataract and other lens diseases episodes during 2020.	int	0 to 1
[41]	hepatopatia_cronica	Number of chronic liver diseases episodes during 2020.	int	0 to 10
[42]	otras_neurologicas	Number of other Neurological episodes during 2020	int	0 to 12
[43]	otras_psiquiatricas	Number of episodes related to Other psychiatric disorders during 2020.	int	0 to 7
[44]	valvulares	Number of episodes of Valves in medical consultations during 2020	int	0 to 6
[45]	otras_dermatologicas	Number of episodes of other dermatology diseases during 2020	int	0 to 4
[46]	autoinmunes	Number of autoimmunes diseases episodes during 2020.	int	0 to 12
[47]	trastornos_del_sueño	Number of sleep disorders episodes during 2020.	int	0 to 10
[48]	enfermedad_venosa_y_linfatica	Number of venous and lymphatic disease episodes during 2020	int	0 to 8
[49]	migraña_y_sindromes_faciales_dolorosos	Number of episodes related to Migraine and facial pain syndromes during 2020.	int	0 to 7
[50]	dislipidemia	Number of episodes of Dyslipidemia in medical consultations during 2020	int	0 to 7
[51]	esclerosis_multiple	Number of episodes for esclerosis multiple during 2020	int	0 to 6
[52]	demencias	Number of dementia episodes during 2020.	int	0 to 11
[53]	alergia	Number of allergy episodes during 2020.	int	0 to 7
[54]	asma	Number of Asma episodes during 2020	int	0 to 6
[55]	ulceras	Number of episodes related to chronic ulcers during 2020.	int	0 to 12
[56]	enfermedad_inflamatoria_intestinal	Number of episodes of Inflammatory Bowel Disease in medical consultations during 2020	int	0 to 6
[57]	neuropatias_perifericas	Number of episodes for peripheral neuropathies during 2020	int	0 to 6
[58]	epilepsia	Number of epilepsy episodes during 2020.	int	0 to 12
[59]	artropatías_inflamatorias	Number of inflammatory arthropathies episodes during 2020.	int	0 to 7
[60]	lumbalgia_cronica	Number of Chronic Low Back Pain episodes during 2020	int	0 to 4
[61]	otras_enfermedades_cardiovasculares	Number of episodes related to Other cardiovascualar diseases during 2020.	int	0 to 9
[62]	parkinson	Number of episodes of Parkinson in medical consultations during 2020	int	0 to 8
[63]	cancer	Number of episodes for cancer disease during 2020	int	0 to 13
[64]	fibrilacion_auricular	Number of atrial fibrillation episodes during 2020.	int	0 to 14
[65]	diabetes	Number of di diabetes diseases during 2020.	int	0 to 18
[66]	falla_cardiaca	Number of Heart Failure episodes during 2020	int	0 to 19
[67]	gastrointestinales_superiores	Number of episodes related to Erophagus, stomach and duodenum diseases 2020.	int	0 to 6
[68]	fact_otros	Billing Others in Colombian Pesos (COP)	int	0 to 1.739.431 COP
[69]	fact_estancias	Billing related to hopital stay (COP)	float	0 to 39479405
[70]	fact_ayudas_dx	Total billing for diagnostic aids in colombian pesos [COP].	float	0 to 23373779.36
[71]	fact_interconsultas	Billing related to consultations during 2020.		
[72]	fact_medicamentos_insumos	Total billing for drug supplies during 2020	float	anonymized
[73]	fact_otros_conceptos	Billing related to other concepts during 2020.	float	
[74]	fact_procedimientos_paquetes_qx	Billing Surgical Procedures and Packages (COP)	float	0 to 994.705,96 COP
[75]	fact_rondas	Billing related to rondas for each patiente (COP)	float	0 to 3592219
[76]	fact_banco_sangre	Total billing for blood bank in colombian pesos [COP].	float	0 to 6648881.74
[77]	fact_planta_oxigeno	Total billing for oxygen plant in colombian pesos [COP].	float	~-0.8 to ~20
[78]	fact_consulta_externa	Total billing for external consultarion during 2020	float	anonymized

[79]	fact_salud_oral	Billing related to oral health during 2020.	float	
[80]	fact_total	Total Billing (COP)	float	0 to 9.994.257,92 COP
[81]	peso	Weight for each patient in Kg	float	0 to 157
[82]	talla	Size in centimeters [cm] (contains missing values).	float	0 to 188
[83]	saturacion_oxigeno	blood oxygen saturation percentage (contains missing values)	int	65 to 100
[84]	perímetro_muslo	Thigh perimeter (contains missing values).	float	anonymized
[85]	perímetro_cintura	Waist perimeter in cm.	float	
[86]	pliegue_triceps	triceps fold in milimiters	float	0 to 95.0 mm
[87]	pliegue_abdomen	abdomen fold in milimeters	float	0 to 5510 cm
[88]	pliegue_muslo	Thigh crease measure in centimeters [cm] (contains missing values).	float	0 to 95
[89]	presión_arterial_sistólica	Sistolic Blood Pressure in milimeters of mercury [mmHg] (contains missing values)	int	70 to 270
[90]	presión_arterial_diastólica	Diastolic Blood Pressure in milimeters of mercury [mmHg] (contains missing values)	int	70 to 270
[91]	frecuencia_cardíaca_en_reposo	Heart rate at rest.	float	
[92]	auto-calificacion_nivel_de_ejercicio	Excersice Level - Selfevaluation	int	1 to 5
[93]	constantes	Other metric	float	0 to 3.03
[94]	mets_-índice_metabólico	Metabolic equivalent of task (METS) in watts per kilogram [W · kg-1] (contains missing values).	float	-1,5 to 49.6
[95]	vo2_-_máxima_cantidad_de_oxígeno	maximum volume of oxygen that the body can process during exercise [ml/kg/min]	float	-12,7491.08 to 14.1742
[96]	índice_de_fragilidad_groningen	Groningen frailty index (contains missing values).	int	0 to 15
[97]	calificacion_indicefragilidad	Qualification in the hospital's pluripathologic fragility index. Higher is worse.	str	Normal / Frágil - No calificado
[98]	tiempo_segundos_monopodal	Monopodal time in seconds	int	0 to 210 seconds
[99]	calificacion_apoyo_monopodal	Monopodal calification in ordinal range	str	Aceptable - Bueno - Malo (fragil) - No calificado
[100]	tiempo_segundos_5metros	Time a patient takes to run a standard distance in seconds [s] (contains missing values).	int	0 to 9
[101]	calificación_velocidad	speed rating	str	['No calificado' 'Deficiente' 'Malo - Frágil' 'Aceptable' 'Bueno']
[102]	indice_tobillo_brazo	Ankle arm index, is a method for determining peripheral arterial disease	str	Not qualified, normal, mild or moderate and severe
[103]	diabetes_mellitus	Whether the patient has diabetes mellitus or not.	str	'Si' for True / 'No' for False
[104]	tipo_diabetes_mellitus	type of diabetes mellitus	str	I, II, Don´t apply
[105]	es_insulinorequiriente	Medical patiente requires insuline in case of episode	str/bool	'Si' for True / 'No' for False
[106]	tiempo_con_el_diagnóstico	Time with dyabetes diagnosis in years (contains missing values).	int	0 to 2020
[107]	glicemia	the measure of free glucose concentration in blood, serum or blood plasma.	float	0 to 993
[108]	hemoglobina_glicada	Glycated hemoglobin (contains missing values).	float	0 to 109
[109]	control_diabetes	Whether the patient's diabetes is in control or not.	str	No aplica / controlada / No controlada
[110]	tiene_hta	Has Arterial hypertension	str	Don´t apply, Controlled, Not Controlled
[111]	control_hta	Has controlled it's hypertension?	str	Controlada - No aplica - Controlada
[112]	tiempo_con_el_diagnóstico2	Time with arterial hypertension diagnosis in years (contains missing values).	int	0 to 2020
[113]	epoc_bodex	Index that gives us a prognostic approximation to chronic obstructive pulmonary disease	str	['No aplica' 'Leve' 'Moderada' 'Grave']
[114]	enfermedad_coronaria	Coronary heart disease	str	No, IAM, stable angina, inestable angina
[115]	insuficiencia_cardíaca	Whether the patient's cardiac insufficiency is under control or not.	str	No / Controlada / No controlada
[116]	valvulopatía	Has valve disease	str/bool	'Si' for True / 'No' for False
[117]	arritmia_o_paciente_con_dispositivo	Patiente has arrhythmia or device that controls it	str/bool	'Si' for True / 'No' for False
[118]	sufre_de_alguna_enfermedad_cardiovascular	Does the patient have cardiovascular diseases?	str/bool	'Si' for True / 'No' for False
[119]	tabaquismo	tobacco user	str	['No' 'Si' 'Ex-fumador']

[120]	cuantos_cigarrillos_día	Number of cigarettes per day	`int`	0 to 60
[121]	años_de_consumo	Number of years the person has been smoking.	`int`	0 to 90
[122]	lipoproteina	Levels of lipoprotein in the body (mg/dL)	`float`	0.0 - 99.96 mg/dL
[123]	hdl	Levels of High-density lipoprotein (HDL) cholesterol	`float`	0 to 9922
[124]	colesterol_total	Total cholesterol in milligrams per deciliter [mg · dL-1] (contains missing values).	`float`	0 to 1,798
[125]	trigliceridos	are measured in milligrams (mg) of triglycerides per deciliter (dl) of blood.	`float`	0 to 1299
[126]	clasificación_de_framinghan	Framinghan clasification	`str`	Unclassified, low risk, high risk
[127]	creatinina_1_consulta	Result for the creatinin exam in the first medical appointment.	`float`	
[128]	tasa_de_filtración_glomerular_tfg	glomerular_filtration_rate_tfg (amount of blood passing through the glomeruli in mL/min)	`float`	0.0 to 99,97 Ml/min
[129]	estadio_de_la_enfermedad_renal	Result of renal disease in terms of estadios clasification table	`str`	estadio (1, 2, 3A, 3B, 4, 5)
[130]	microalbuminuria	Microalbuminuria in milligrams of albumin per gram of sample [mg · g-1] (contains missing values).	`float`	0 to ~9,886.8
[131]	tsh	This is a test that measures the amount of thyroid-stimulating hormone in the blood. [mU/L]	`float`	0 to 285
[132]	clase_funcional	Heart failure classification.	`str`	Unclassified, functional class 1/2A/2B/3/4
[133]	creatinina_2_consulta	Result for the creatinin exam in the second medical appointment.	`float`	
[134]	tasa_de_filtración_glomerular_tfg3	glomerular_filtration_rate_tfg3 (amount of blood passing through the glomeruli in mL/min)	`float`	0.0 to 99,85 Ml/min
[135]	cambio_de_tfg	Change of glomerular filtration rate between each episode	`float`	-89 to 75
[136]	úlcera_de_pie_diabético	Does the patient have diabetic foot ulcer?	`str/bool`	'Si' for True / 'No' for False
[137]	estado_vital	vital state	`str`	['vivo' 'fallecido']

Acronyms	
Name	**Description**
n	Count
urg	Urgencies
hosp	Hospitalization
cx	Surgery
ce	External consultation
med	Medicine
uce	Special Care Unit
uci	Intensive Care Unit
epoc	Chronic obstructive pulmonary disease (COPD)
fact	Billing
dx	Diagnostic
qx	Surgical
mets	Metabolic equivalent of task (METS)
hta	Arterial hypertension
hdl	High-density lipoprotein (HDL) cholesterol
perímetro	Lenght measure (cm)
pliegue	Crease measure (cm)
tsh	Thyroid stimulating hormone (TSH)