APPLYING DEEP LEARNING TO HOSPITALIZATION DATA

Springboard: Capstone 2

Ivette M Tapia

TABLE OF CONTENTS

INTRODUCTION

This report summarizes the overall process taken to build a deep learning network to predict a patient's procedure based on the hospitalization authorization data. Deep learning networks is a class of machine learning algorithms. However, deep learning differs from other machine learning algorithms in that:

1. Deep learning uses multiple layers for feature extraction and transformation. Each layer uses the output of the other layer as input. As such, it can handle fairly well non-linear processing and patterns.
2. It can learn in supervised or unsupervised manner.
3. Can learn multiple levels of representation, concepts, and abstraction.
4. Given all the layers and abstraction, it is often difficult to interpret and can be 'black-box'.

The data used to train the neural network is from the Authorization for Hospital Admission. This dataset is part of Brazil's SIHSUS Hospital Information System. This system manages the coordination and payment by Brazil's public healthcare system. The data is publically available on the web. In this application, I will be using data from 2015 – 2018. This represents 3.5 years' of data. A record in the AIH database is created when a hospital or healthcare unit generates a request for hospitalization. This dataset is large and highly dimensional.

The report starts by outlining the motivation for the project. After the motivation and background section, file conversation, extraction and initial data wrangling are discussed. In between, a summary of all the features available and action taken on the feature. After this, exploratory data analysis and further data wrangling on the features are discussed. Afterwards, feature engineering performed to enhance usefulness of the features and further reduce dimensionality is discussed. Finally, the deep neural network is described and predictive performance results are discussed.

The coding language used throughout this project is R, and python. The coding interface is Jupyter notebooks. The deliverables for the project is python code, a detailed report, and presentation slides. Throughout the report references to specific notebooks will be provided.

MOTIVATION & BACKGROUND

Healthcare records have become increasingly more digitized. This is an open an opportunity to analyze and obtain patient data at an unprecedented detail and scale. While there is potential to gain greater insights, cost reduction and efficiencies in the healthcare space exist, great challenges remain. Some of these include data availability and complexity of services. Healthcare is particularly complex due to overlapping systems, diversity of providers, services and health issues.

This project would use hospitalization data from Brazil to make predictions about key features and outcomes of an hospitalization request. Specifically, a deep learning will be used to predict: (1) procedure(s) performed, (2) hospitalization costs, and (3) hospitalization length given the information available in the approval provided in the hospitalization

Ability to accurately predict these three features of hospitalization can yield significant benefits. For example, knowing how many days a patient can be expected to be in the hospital will help hospital managers manage their capacity (especially in areas where beds are scarce). Length of stay and likely procedures can inform service charges and help all parties involved navigate the healthcare charge system better so likely costs are known in the front end.

Moreover, predicting healthcare expenditures can be tricky for insurers, providers and particularly consumers. One of the main factors that have been cited as a cause of rising healthcare expenditures is the inability of consumers to know in advance the cost of the healthcare services they consume.

The data that will be used is from the Authorization for Hospital Admission. This dataset is part of Brazil's SIHSUS Hospital Information System. This system manages the coordination and payment by Brazil's public healthcare system (covers around 34% of Brazil's population and pays for 80% of all hospitalizations). The data is publically available as .dbc files on the web. In this application, I will be using data from 2015 – 2018. This represents 3.5 years' of data.

A record in the AIH database is created when a hospital or healthcare unit generates a request for hospitalization. Providers submit demographic and health information about the patient. This request is approved, reduced, rejected, or rejected due to an error. While the patient is in the hospital, the record is updated to also contain information about procedures performed and discharge. Each row of information represents an hospitalization. If a patient is hospitalized more than 30 days, a new authorization is needed and a new record (i.e. row is created).

DATASET EXTRACTION, CONVERSION & SAMPLING

The dataset extraction was a fairly complex process with multi-steps. This was due to the fact that the data originally in .dbc format, was distributed in hundreds of files, spans multiple years, has millions of observations and is high-dimensional.

This characteristics presents challenges and opportunities. The challenges mainly stem from the file format and large size of the dataset. The .dbc format is propertary to the Brazilian Department of Health. It is basically a compressed version of a dbf files. The opportunities is that there are many features and lots of data to work with in this dataset. I will describe the extraction, conversion and sampling process below.

Step 1: Extraction

The hospitalization data was extracted from the DataSUS servers through the web. The website is as follows: http://www2.datasus.gov.br/DATASUS/index.php?area=0901&item=1&acao=25.

There are several type of datasets choose from:

- Rejected hospitalization requests
- Professional Services
- Reduced

For this project, I extracted the reduced option. This dataset contains the hospitalization authorization request data. I extracted all the hospitalization data available for the years of 2015, 2016, 2017, and up to July 2018. As highlighted above this data roughly represents 80% of all hospitalizations in Brazil.

The 2015 year has 324 files, the 2016 has 324 files, the 2017 year has 317 files and the 2018 year has 176 files. This is for a total of 1,141 files that need to be converted to make them usable.

Step 2: Conversion from .dbc files to R dataframes to CSV

While it is possible to convert these files using python, R already has a package call **read.dbc** developed by Brazilian researchers specifically written to convert these files. Since a solution already existed in R, I implemented the file conversion in R.

Once all the files were converted to R data frames, I concatenated the data frames by their respective year. Four large data frames were created, one for each year represented. These four data frames were outputted as CSV files for further use. For more details refer to R code here.

Step 3: Create random sample

The entire dataset extracted and converted takes 35GB of memory in pandas and has 113 columns. Given, that the project involves creating a neural network, the computational resources are not available to use the entire dataset.

To solve this problem, I drew a random sample for each year. The goal was to extract 40% of the entire dataset randomly. For reproducibility the random seed throughout was 42.

Each year has different number of observations. To account for this, I forced the sampling process to take the same proportion that year represents of the total observations. The calculation and results were as follows:

- Total Observations 2016 – 2018: 41,537,081
- 40% of total observations: 16,614,832
    - 2015 (28%): 4,655,541
    - 2016 (28%): 4,611,084
    - 2017 (26%): 4,624,383
    - 2018 (16%): 2,723,822
    - **Total:** 16,614,830

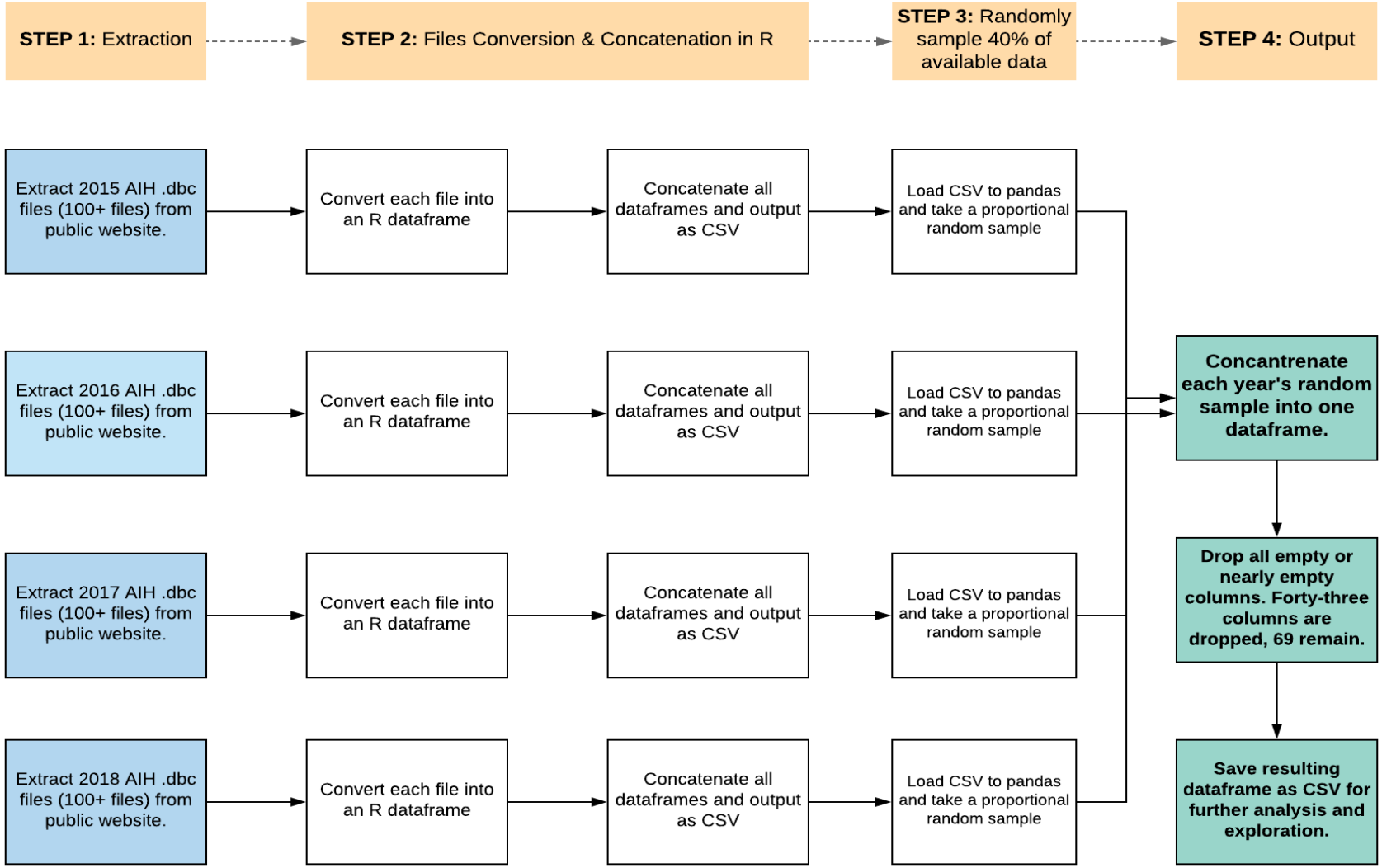
Once the samples has been extracted, I saved the results to as CSV files for further use. For more details refer to the python code here.

Step 4: Data Wrangling – *First Pass*

After the random sample was created for each year, I concatenated the four resulting yearly random sample data frames into one larger data frame. From this data frame, I dropped columns that were either completely empty of had more than 20% missing values. As a result 44 columns dropped, 69 columns / features remained. Please see section *'Dataset Features & Actions'* for a complete listing of features and actions taken on the features.

For more details refer to the python code here.

Conversion, Sampling & Wrangling



EXPLORATORY DATA ANALYSIS & WRANGLING

The remaining columns can be grouped into four themes: (1) patient demographics, (2) patient diagnosis, (3) hospitalization services and, (4) financial features, and (5) auditor metadata. Given the large number of features remaining and keeping with the main interests of this project, I decided to not use the financial features and auditing metadata and focus on the patient demographics, diagnosis and hospitalization services.

Patient Demographics

The code for all the analysis and wrangling that will be described below can be found here.

a. Findings
b. Wrangling
c. Exported as CSV

Diagnosis

The code for all the analysis and wrangling that will be described below can be found here.

d. Findings
e. Wrangling
f. Exported as CSV

Hospitalization

The code for all the analysis and wrangling that will be described below can be found here.

g. Findings
h. Wrangling
i. Exported as CSV

DATASET FEATURES & ACTIONS

Field_Name	Type of Field	Description	Action
UF_ZI	char(6)	Municipality Manager	Auditor metadata - Not Used
ANO_CMPT	char(4)	Year of AIH processing, in yyyy format.	**Part of all Groups**
MÊS_CMPT	char(2)	Month of AIH processing, in mm format.	**Part of all Groups**
ESPEC	char(2)	Specialty of Bed	**Hospitalization Group**
CGC_HOSP	char(14)	CNPJ of the Establishment	**Hospitalization Group**
N_AIH	char(13)	Number of AIH	Auditor metadata - Not Used
IDENT	char(1)	Identification of the type of AIH	Auditor metadata - Not Used
CEP	char(8)	CEP of the patient	Auditor metadata - Not Used
MUNIC_RES	char(6)	Municipality of Patient's Residence	**Demographic Group**
NASC	char(8)	Date of birth of the patient (yyyammdd)	Demographic Group – Not Used
SEXO	char(1)	Sex of patient	**Demographic Group**
UTI_MES_IN	numeric(2)	Reset	Dropped – First Pass
UTI_MES_AN	numeric(2)	Reset	Dropped – First Pass
UTI_MES_AL	numeric(2)	Reset	Dropped – First Pass
UTI_MES_TO	numeric(3)	Number of ICU days in the month	**Hospitalization Group**
MARCA_UTI	char(2)	Indicates the type of ICU used by the patient	**Hospitalization Group**
UTI_INT_IN	numeric(2)	Reset	Dropped – First Pass
UTI_INT_AN	numeric(2)	Reset	Dropped – First Pass
UTI_INT_AL	numeric(2)	Reset	Dropped – First Pass
UTI_INT_TO	numeric(3)	Number of nights in intermediate unit	**Hospitalization Group**
DIAR_ACOM	numeric(3)	Number of companion nights	**Hospitalization Group**
QT_DIARIAS	numeric(3)	Number of nights	Hospitalization Group – Not Used
PROC_SOLIC	char(10)	Procedure requested	Hospitalization Group – Not Used
PROC_REA	char(10)	Procedure performed	**Hospitalization Group**
VAL_SH	numeric(13,2)	Value of hospital services	Financial Group – Not Used
VAL_SP	numeric(13,2)	Value of professional services	Financial Group – Not Used
VAL_SADT	numeric(13,2)	Reset	Dropped – First Pass
VAL_RN	numeric(13,2)	Reset	Dropped – First Pass
VAL_ACOMP	numeric(13,2)	Reset	Dropped – First Pass
VAL_ORTP	numeric(13,2)	Reset	Dropped – First Pass
VAL_SANGUE	numeric(13,2)	Reset	Dropped – First Pass
VAL_SADTSR	numeric(11,2)	Reset	Dropped – First Pass
VAL_TRANSP	numeric(13,2)	Reset	Dropped – First Pass
VAL_OBSANG	numeric(11,2)	Reset	Dropped – First Pass
VAL_PED1AC	numeric(11,2)	Reset	Dropped – First Pass
VAL_TOT	numeric(14,2)	Total value of the AIH	Financial Group – Not Used
VAL_UTI	numeric(8,2)	Value of ICU	Financial Group – Not Used
US_TOT	numeric(10,2)	Total value, in US dollars	Financial Group – Not Used
DI_INTER	char(8)	Date of hospitalization in aaammdd format	Hospitalization Group – Not Used

DT_SAIDA	char(8)	Exit date in yyyammdd format	Hospitalization Group – Not Used
DIAG_PRINC	char(4)	Code of the main diagnosis (CID10)	**Diagnosis Group**
DIAG_SECUN	char(4)	Secondary diagnosis code (ICD10). Filled with zeros from 201501.	Diagnosis Group – Not Used
COBRANCA	char(2)	Reason for Exit / Stay	**Hospitalization Group**
NATUREZA	char(2)	Legal nature of the hospital (with content until May / 12). It was used the classification of Regime and Nature.	Auditor metadata - Not Used
NAT_JUR	char(4)	Legal nature of the establishment, as the Commission National classification - CONCLA	Auditor metadata - Not Used
DESTAO	char(1)	Type of hospital management	Auditor metadata - Not Used
RUBRICA	numeric(5)	Reset	Dropped – First Pass
IND_VDRL	char(1)	Indicates VDRL exam	**Hospitalization Group**
MUNIC_MOV	char(6)	Municipality of the Establishment	Auditor metadata - Not Used
COD_IDADE	char(1)	Unit of measure of age	Demographic Group – Not Used
IDADE	numeric(2)	Age	**Demographic Group**
DIAS_PERM	numeric(5)	Days of Stay	**Hospitalization Group**
MORTE	numeric(1)	Indicates Death	**Demographic Group**
NACIONAL	char(2)	Code of nationality of the patient	Demographic Group – Not Used
NUM_PROC	char(4)	Reset	Dropped – First Pass
CAR_INT	char(2)	Character of hospitalization	**Hospitalization Group**
TOT_PT_SP	numeric(6)	Reset	Dropped – First Pass
CPF_AUT	char(11)	Reset	Dropped – First Pass
HOMONIMO	char(1)	Indicator if the patient of the AIH is homonymous with the another AIH.	Auditor metadata - Not Used
NUM_FILHOS	numeric(2)	Number of children of the patient	Demographic Group – Not Used
INSTRU	char(1)	Degree of instruction of the patient	Demographic Group – Not Used
CID_NOTIF	char(4)	CID of Notification	Auditor metadata - Not Used
CONTRACEP1	char(2)	Type of contraceptive used	**Hospitalization Group**
CONTRACEP2	char(2)	Second type of contraceptive used	**Hospitalization Group**
GESTRISCO	char(1)	Indicator if pregnant at risk	Demographic Group – Not Used
INSC_PN	char(12)	Number of the pregnant woman in prenatal care	Hospitalization Group – Not Used
SEQ_AIH5	char(3)	Long-stay sequential (AIH type 5)	Auditor metadata - Not Used
CBOR	char(3)	Occupancy of the patient, according to the Brazilian Occupations - CBO.	Demographic Group – Not Used
CNAER	char(3)	Work accident code	Auditor metadata - Not Used
GESTOR_COD	char(3)	Reason for authorization of the AIH by the Manager	Auditor metadata - Not Used
GESTOR_TP	char(1)	Type of manager	Auditor metadata - Not Used
GESTOR_CPF	char(11)	Manager's CPF number	Auditor metadata - Not Used
GESTOR_DT	char(8)	Date of authorization given by the Manager (yyyammdd)	Dropped – First Pass
CNES	char(7)	CNES code of the hospital	Auditor metadata - Not Used
CNPJ_MANT	char(14)	CNPJ of the maintainer	Auditor metadata - Not Used
INFEHOSP	char(1)	Hospital infection status	Auditor metadata - Not Used
CID_ASSO	char(4)	CID causes	Hospitalization Group – Not Used
CID_MORTE	char(4)	CID of death	Hospitalization Group – Not Used

COMPLEX	char(2)	Complexity	**Hospitalization Group**
FINANC	char(2)	Type of financing	Financial Group – Not Used
FAEC_TP	char(6)	Financing subtype FAEC	Financial Group – Not Used
REGCT	char(4)	Contract rule	Auditor metadata - Not Used
RACA_COR	char(4)	Race / Color of the patient	**Demographics Group**
ETNIA	char(4)	Ethnicity of patient, if race color is indigenous	**Demographics Group**
SEQUENCIA	numeric(9)	Sequential of the AIH in the consignment	Auditor metadata - Not Used
REMESSA	char(21)	Shipping number	Auditor metadata - Not Used
AUD_JUST	char (50)	Auditor's justification for acceptance of the IAI without the National Health Card.	Auditor metadata - Not Used
SIS_JUST	char (50)	Rationale of the establishment for acceptance of the AIH without number of the National Health Card	Auditor metadata - Not Used
VAL_SH_FED	numeric (10, 2)	Value of the federal complement of hospital services. It is included in the total value of the AIH.	Financial Group – Not Used
VAL_SP_FED	numeric (10, 2)	Value of the federal complement of professional services. It is included in the total value of the AIH.	Financial Group – Not Used
VAL_SH_GES	numeric (10, 2)	Value of the complement of the manager (state or municipal) of hospital services. It is included in the total value of the AIH.	Financial Group – Not Used
VAL_SP_GES	numeric (10, 2)	Value of the complement of the manager (state or municipal) of profesional services.It is included in the total value of the AIH.	Financial Group – Not Used
VAL_UCI	numeric (10, 2)	Value of ICU.	Financial Group – Not Used
MARCA_UCI	char (2)	Type of ICU used by the patient.	Hospitalization Group – Not Used
DIAGSEC1	char (4)	Secondary diagnosis1	Dropped – First Pass
DIAGSEC2	char (4)	Secondary diagnosis2	Dropped – First Pass
DIAGSEC3	char (4)	Secondary diagnosis3	Dropped – First Pass
DIAGSEC4	char (4)	Secondary diagnosis4	Dropped – First Pass
DIAGSEC5	char (4)	Secondary diagnosis5	Dropped – First Pass
DIAGSEC6	char (4)	Secondary diagnosis6	Dropped – First Pass
DIAGSEC7	char (4)	Secondary diagnosis7	Dropped – First Pass
DIAGSEC8	char (4)	Secondary diagnosis8	Dropped – First Pass
DIAGSEC9	char (4)	Secondary diagnosis9	Dropped – First Pass
TPDISEC1	char(1)	Type of secondary diagnosis 1	Dropped – First Pass
TPDISEC2	char(1)	Type of secondary diagnosis 2	Dropped – First Pass
TPDISEC3	char(1)	Type of secondary diagnosis 3	Dropped – First Pass
TPDISEC4	char(1)	Type of secondary diagnosis 4	Dropped – First Pass
TPDISEC5	char(1)	Type of secondary diagnosis 5	Dropped – First Pass
TPDISEC6	char(1)	Type of secondary diagnosis 6	Dropped – First Pass
TPDISEC7	char(1)	Type of secondary diagnosis 7	Dropped – First Pass
TPDISEC8	char(1)	Type of secondary diagnosis 8	Dropped – First Pass
TPDISEC9	char(1)	Type of secondary diagnosis 9	Dropped – First Pass

FEATURE ENGENEERING

DEEP LEARNING NEURAL NETWORK

SOURCES

1. Choi et al. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. 2016. https://arxiv.org/abs/1511.05942
2. de Araujo Lima et al. Successful Brazilian Experiences in the Field of Health Information – Final Report. 2016. https://www.measureevaluation.org/our-work/health-information-systems/health-information-system-strengthening-in-lac-region-2005-2010/his-brazil-english-august2007.pdf
3. Malhotra et al. Long Short Term Memory Networks for Anomaly Detection in Time Series. 2015. Long Short Term Memory Networks for Anomaly Detection in Time Series. https://www.elen.ucl.ac.be/Proceedings/esann/esannpdf/es2015-56.pdf
4. Ralf Carvalho & Alex Paino. Deep Learning for Fraud Detection. https://engineering.siftscience.com/deeplearning-fraud-detection/
5. Sapronova et al. Deep learning for wind power production forecast. 2017. http://ceur-ws.org/Vol-1818/paper3.pdf
6. Shipmon et al. Time Series Anomaly Detection. https://ai.google/research/pubs/pub46283
7. Tang and Mendis. Unsupervised fraud detection in Medicare Australia. 2011. https://pdfs.semanticscholar.org/e7b6/9c1ba648de68d30aacb84b47ab43fcdf75e9.pdf
8. Teodoro et al. ORBDA: An openEHR benchmark dataset for performance assessment of electronic health record servers. 2017. https://www.ncbi.nlm.nih.gov/pubmed/29293556
9. Tianwei Yue & Haohan Wang. Deep Learning for Genomics: A Concise Overview. 2018. https://arxiv.org/abs/1802.00810
10. Viscondi et al. Cost Description of Prevention and Treatment of Cervical Cancer. https://www.researchgate.net/profile/Juliana_Viscondi/publication/299483604_COST_DESCRIPTION_OF_PREVENTION_AND_TREATMENT_OF_CERVICAL_CANCER_IN_BRAZIL_IN_2012/links/59b6786ea6fdcc7415bd3940/COST-

DESCRIPTION-OF-PREVENTION-AND-TREATMENT-OF-CERVICAL-CANCER-IN-BRAZIL-IN-2012.pdf