UNIVERSIDAD DE CHILE
FACULTAD DE MEDICINA
ESCUELA DE POSTGRADO

ASSESSMENT OF EXPORTED CLINICAL DATA FROM HOSPITAL INFORMATION SYSTEM FOR CLINICAL REUSE IN MULTIPLE MYELOMA RESEARCH

VIVIANA MARIBEL TORRES MANQUELAFQUÉN

TESIS PARA OPTAR AL GRADO DE MAGISTER EN INFORMÁTICA MÉDICA

Director de Tesis: Prof. Dr. Mauricio Cerda
Co- Director de Tesis: Prof. Dra. Petra Knaup

(2016)

UNIVERSIDAD DE CHILE
FACULTAD DE MEDICINA
ESCUELA DE POSTGRADO

ASSESSMENT OF EXPORTED CLINICAL DATA FROM HOSPITAL INFORMATION SYSTEM FOR CLINICAL REUSE IN MULTIPLE MYELOMA RESEARCH

VIVIANA MARIBEL TORRES MANQUELAFQUÉN

TESIS PARA OPTAR AL GRADO DE MAGISTER EN INFORMÁTICA MÉDICA

Director de Tesis: Prof. Dr. Mauricio Cerda
Co- Director de Tesis: Prof. Dra. Petra Knaup

(2016)

UNIVERSIDAD DE CHILE
FACULTAD DE MEDICINA
ESCUELA DE POSTGRADO

INFORME DE APROBACION TESIS DE MAGISTER

Se informa a la Comisión de Grados Académicos de la Facultad de Medicina, que la Tesis de Magister presentada por la candidata

VIVIANA MARIBEL TORRES MANQUELAFQUÉN

ha sido aprobada por la Comisión Informante de Tesis como requisito para optar al Grado de Magister en Ciencias en Informática Médica en el Examen de Defensa de Tesis rendido el día 31 Marzo 2016.

.....
Prof. Dr. Mauricio Cerda
Director de Tesis
(Santiago, Chile)

COMISION INFORMANTE DE TESIS

.....	.....
Prof. Dr. Rodrigo Assar	**Prof. Dr. Sergio Bozzo**
.....	.....
Prof. Dr. Rodrigo Martínez	**Prof. Dr. Steffen Hartel**
	Presidente Comisión

UNIVERSIDAD DE CHILE
FACULTAD DE MEDICINA
ESCUELA DE POSTGRADO

ASSESSMENT OF EXPORTED CLINICAL DATA FROM HOSPITAL INFORMATION SYSTEM FOR CLINICAL REUSE IN MULTIPLE MYELOMA RESEARCH

VIVIANA MARIBEL TORRES MANQUELAFQUÉN

TESIS PARA OPTAR AL GRADO DE MAGISTER EN INFORMÁTICA MÉDICA

Director de Tesis: Prof. Dr. Mauricio Cerda
Co- Director de Tesis: Prof. Dra. Petra Knaup

(2016)

UNIVERSIDAD DE CHILE
FACULTAD DE MEDICINA
ESCUELA DE POSTGRADO

ASSESSMENT OF EXPORTED CLINICAL DATA FROM HOSPITAL INFORMATION SYSTEM FOR CLINICAL REUSE IN MULTIPLE MYELOMA RESEARCH

VIVIANA MARIBEL TORRES MANQUELAFQUÉN

TESIS PARA OPTAR AL GRADO DE MAGISTER EN INFORMÁTICA MÉDICA

Director de Tesis: Prof. Dr. Mauricio Cerda
Co- Director de Tesis: Prof. Dra. Petra Knaup

(2016)

This thesis is dedicated to my Parents, Fernando Torres Castillo and Gladys Manquelafquen Chávez, who always gave me their support and encouragement in my entire career and in achievements. They have taught me that whatever I propose myself in life can be achieved with effort and perseverance.

I thank them for their endless love and faith reposed in me to achieve my goals and dreams.

ACKNOWLEDGEMENTS

This thesis work was supported by a DAAD (Deutscher Akademischer Austausch Dienst) grant at Heidelberg University, Germany.

I gratefully acknowledge to the staff of the Institute of Medical Biometry and Informatics, Heidelberg University, Germany. In special I would like to thank to Professor Petra Knaup for give me the opportunity to participate in her work team, and to Mr. Martin Löpprich who guided me all time to accomplish my thesis objectives.

Also I thank Professor Harmut Dickhaus and Dr. Roland Metzner, for their support during my stay in Heidelberg.

I would like to express also my gratitude to Dr. Mauricio Cerda for his collaboration and permanent support in this thesis work and during my stay in Germany.

Finally many thanks to my family, who always had encouragement words reach the target of this work.

CONTENTS

ABSTRACT

The Department of Multiple Myeloma of the University Hospital Heidelberg uses a scientific database to enter clinical data for research purposes. Entered data is about demographic, diagnostic, laboratory and treatment data of patients with multiple myeloma. Currently the transfer of data from the Hospital Information System (HIS) to the scientific database must be done manually by trained specialists, which is time consuming and prone to transcription errors. Recently it became possible to export the data from the HIS as Comma Separated Value (CSV) files to import it into the database. However there is no evaluation of this reusing process to verify if the quality of the data is suitable for scientific purposes.

This thesis assessed the quality of exported laboratory data from the HIS, as CSV files, for secondary use in Multiple Myeloma research, in two types of processes: (1) in a manual process of data transcription, and (2) in the automatic process of data transference. The automatic transference was implemented as an Extract, Transform and Load (ETL) process. A comparison was carried out between manual and automatic data collection methods. The criteria to measure data quality were correctness and completeness. As results of this analysis, the manual process had a general error rate of 2.6% to 7.1%, obtaining the lowest error rate if data fields with a not clear definition were removed from the analysis ($p<0.000$). In the case of automatic process, the general error rate was 1.9% to 12.1%, where lowest error rate is obtained when excluding information missing in the HIS but transcribed to the scientific database from other physical sources. When the sample is adjusted to only data that follow a standardized protocol and present in the HIS, manual process presented a higher error rate of 1,8% in comparison with automatic process 0,18% ($p<0.000$). In addition to the speed of automatic process compared to the manual one the implemented ETL process simplifies data collection identifying, through alerts, data incompleteness and incorrectness at the point of entry, and it expanded the range of captured data.

The automatic ETL process can be used to collect laboratory data for clinical research with quality assurance if data in the HIS as well as physical documentation not included in HIS, are identified previously and follows a standardized data collection protocol.

RESUMEN

El Departamento de Mieloma Múltiple del Hospital Universitario de Heidelberg utiliza una base de datos científica para ingresar datos clínicos con fines de investigación. Dentro de los datos introducidos se encuentran datos demográficos, diagnósticos, laboratorio y tratamiento de pacientes con mieloma múltiple. Actualmente la transferencia de datos desde el Sistema de Información Hospitalaria (HIS) a la base de datos científica debe hacerse manualmente por especialistas entrenados, proceso que consume mucho tiempo y es propenso a errores de transcripción. Recientemente es posible exportar los datos del HIS como archivos de valores separados por comas (CSV) para importar estos en la base de datos. Sin embargo no existe evaluación de este proceso de reutilización que verifique si la calidad de los datos es apropiada para fines científicos.

Esta tesis evaluó la calidad de los datos de laboratorio exportados desde el HIS como archivos CSV, para uso secundario en investigación del mieloma múltiple, en dos tipos de procesos: (1) en un proceso manual de transcripción de datos y (2) en el proceso automático de transferencia de datos. Este último fue implementado como un proceso de extracción, transformación y carga (ETL). Luego, una comparación entre ambos métodos de recolección fue llevada a cabo. Los criterios para medir la calidad fueron correctitud y completitud. Como resultados de este análisis, el proceso manual presentó una tasa de error general de 2,6% a 7.1%, obteniendo la más baja tasa si los campos de datos con una definición no clara son excluidos del análisis ($p < 0.000$). En el caso de proceso automático, la tasa de error general fue 1,9% a 12.1%, donde la menor tasa se obtiene al excluir la información que falta en el HIS, pero son transcritos a la base de datos científica desde otras fuentes físicas. Cuando la muestra es ajustada con solo datos que siguen un protocolo estandarizado e incluidos en el HIS, el proceso manual presenta una tasa de error más alta de 1,8% en comparación con el automático, 0,18% ($p<0.000$). Además de las mejora en velocidad el proceso ETL implementado permite simplificar la recogida de los datos, identificar, a través de alertas, el estado incompleto e incorrecto en el punto de entrada y ampliar la gama de los datos capturados.

El proceso automático ETL puede utilizarse para recopilar datos de laboratorio para la investigación clínica, con calidad asegurada, si los datos en el HIS así como documentación física no incluidos en él, se identifican previamente y siguen un protocolo de recogida de datos estandarizados.

1. INTRODUCTION

High data quality is critical in clinical research and it is ensured by compliance of established standards. Low data quality may bring a negative result on the overall research, more over to waste resources [1,2].

Good Clinical Practice (GCP) is the most recognized standard to ensure the scientific quality and international acceptance of a study. One of the objectives of GCP is to document step by step the complete cycle of a clinical study [3]. The documentation is based on a protocol defined for each study. This protocol contains the design and logical plan to conduct the whole process in a clinical trial. Also, each study defines its own case report forms (CRF) to collect all relevant data described in the protocol [4].

The principal stages of a clinical study are: design, data collection, data entry, data verification and analysis [5]. Data collection and data entry are a main clinical stage and the quality assessment in these stages is critical. Data entry and data verification processes for scientific research has been changing in the last decade [4-6].

Traditionally specialists in the area of clinical research have used paper-based Case Report Forms (CRF) to collect relevant information to carry out scientific studies [4,6]. Nowadays, with the wide spread of information technologies, efficiency, accuracy and speed of data collection is improving by the use of Electronic Data Capture (EDC) software [7]. Currently EDC systems are the preferred system to collect, enter, and process data to carry out scientific research and it is replacing paper-based CRFs with electronic CRFs [5, 7].

On the other hand, the significant increase of clinic information registered in electronic format in hospital information systems (HIS) or Electronic Health Records (EHR) has highlighted the opportunity to directly reuse this data for scientific research.

According to GCP [3], utilizing an EHR to prepopulate an EDC can increase research data collection efficiency; eliminate transcription process and the need for source data verification when the procedure is well documented to ensure the integrity of the data. The researcher can have the complete control of the data and data that already is in the EDC could be extracted to populate a clinical data management to perform edit checks [3]. Also the data to be analysed could be available in less time [8]. In addition Weiskopf et al [9], mentions that the reuse of data for secondary purpose is a promising step to decrease research cost, increase patient-centred research and speedup the rate of new medical discoveries [3]. However, in the daily practice to reuse data from EHR and use it to populate an EDC system remains a challenge [10]. These two entities: clinical studies and HIS are in general independent of each other [11].

The study done by Dentrel et al. [2] identified barriers at the data level to reuse the data from electronic health records in terms of quality. Those barriers are: incompleteness at database level, incompleteness at data element level, incorrectness, lack of linking of data in various sources, missing provenance of data and lack of inside-knowledge of "meaning of data". In addition, certain limitation exists at organizational level that restricts the reuse of the data, for instance heterogeneous data in syntax and semantics, limited data logistic (use of different codes) and legal data privacy limitations [12]. The lack of integration in both system results in unnecessary work duplication in data collection and data entry processes due to manual transcription of data from the EHR to EDC system, and the long time required to have the complete data for analysis [13].

As it has been explained, to connect HIS and EDC systems has the potential to reuse data from the HIS, assuring a high data quality, in clinical studies. However, in this context to assess the quality and suitability of the data that can be extracted from HIS for scientific purpose is an important problem that needs to be studied.

1.1 Problem Description

The Department of Medical Informatics of the Institute of Medical Biometry and Informatics of University of Heidelberg, together with the department of internal medicine, and section Multiple Myeloma of the University Heidelberg Hospital and the National Center for Tumor diseases (NCT) carry out the project of Multiple Myeloma (MM) registry for scientific purposes.

The main objective of the Multiple Myeloma registry is to perform scientific research in order to optimize the diagnosis, treatment and quality of life of patients with Multiple Myeloma, including patients with a monoclonal gammopathy of undetermined significance (MGUS), solitary myeloma and plasma cell leukaemia [14].

In 2009 a scientific database was implemented to collect and integrate data of scientific interest related to multiple myeloma. This scientific record is based on a MySQL database. The management, planning and organization of hardware and software of this database for scientific purposes is in charge of the Department of Medical Informatics whose main purpose is to improve the quality of the recorded data for scientific research [15]. Currently, this department is working on the implementation of an EDC system based on OpenClinica, an open source software for EDC, to collect all the data for MM research. This implementation process is not concluding yet. For this reason, the data is still collected in the scientific database based on MySQL.

The MySQL database includes all data of scientific interest for the Multiple Myeloma disease; mainly patients belong to Transplantation (TRLP) section and NCT department. It includes specific data of the diagnosis and demographic characteristics of patients, laboratory parameters, the type of treatment, including new drugs and its outcome [14].

The information is collected from the HIS based in i.s.h. med (Industry Solution for Healthcare, a product of SAP) of the University Hospital of Heidelberg to the scientific database. This procedure is done by manual data transcription (Figure 1) in seven steps. (i)The

specialists open the case files within the HIS and manually select all the interest data documented in the medical discharge letter of the patients that belongs to the Multiple Myeloma section, including laboratory data. (ii) Due to the amount of the laboratory data of a patient, the specialist classifies this laboratory reports as part of one event (diagnosis, mobilisation, transplantation or 100 days after transplantation). (iii) The trained specialists select the respective laboratory value of interest and manually transcribe this data into scientific database. Transcription errors can occur during this process and limits the available time for quality assurance. In addition, researchers have to wait a long time to have data in the scientific database to perform scientific research. (iv) Data quality is retrospectively assessed by a Source Data Verification (SDV) process. (v) The data is saved in a Microsoft Excel file to make it available for the researches. (vi) Researchers can perform the respective data analysis and (vii) reports.

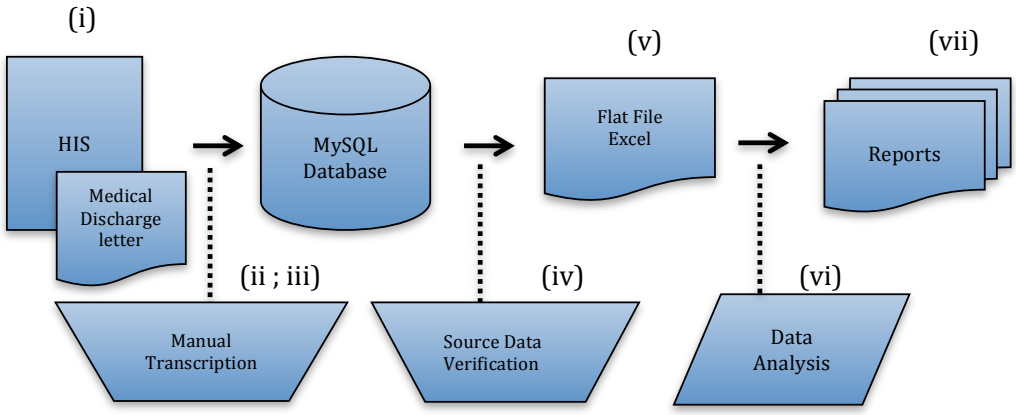

Figure 1. Manual data collection process

A recent implemented feature within the i.s.h med HIS enables the export of demographic, diagnostic procedure and laboratory data as CSV files. Clinicians could reuse this data to populate the scientific database, avoiding in this way the transcription of the data. Nevertheless the level of data quality in the transference of the data from CSV to scientific database is not clear and further efforts needs to be done to ensure it.

1.2 Background

1.2.1 Multiple Myeloma

Multiple Myeloma (MM) is the second most common blood cancer. It has a frequency of 1% among all cancers in general [16] and the exact cause has not been identified. It has an annual incidence of 4-5 cases per 100,000 people [15]. The mean age at diagnosis is presented is 65 years [16]. In the MM disease only a few patients present a remission of the illness, the rest of the patients is only able to improve their quality of life. Scientific studies are currently looking for early diagnosis, and to develop new treatments to relieve symptoms [15].

When there is a clinical suspicion that a patient has MM several tests should confirm the diagnosis [16] including blood, urine and specific tests of the bone marrow [17]. The MM is classified in three categories [17]: (i) Monoclonal gammopathy of undetermined significance (MGUS) , (ii) Asymptomatic myeloma (further subdivided into smouldering myeloma or indolent myeloma) and (iii) Symptomatic myeloma. Some patients can receive as treatment chemotherapy drugs and/or stem cell transplantation, depending of the disease stage and patient condition.

The patients that are recruited for scientific study are evaluated through laboratory parameters among others test, in four events of the disease: at the time of diagnosis, before start the chemotherapy, immediately after transplantation and after 100 days of transplantation.

1.2.2 Data Quality

Data Quality is a concept that is difficult to assess and define. In this thesis it will be referred to Data Quality, according to ISO 14005, as the *"characteristic of data that bears on their ability to satisfy stated requirements"*. Data quality is the main stone required in clinical research for scientific results [3].

In the process of quality evaluation it is important to know what kind of error can be found at different stages. Kirkwooda et al. [18] identified 4 types of errors according to different moments during a clinical trial (design, procedural, recording/data and analytical error). This thesis will be focused on errors that are present in recording/data moments. This type of error occurs when investigators enter inaccurate (including falsified) data into the CRFs (data collection process) or in the transcription of the data to the database (data entry process) [18].

Currently the literature describes 15 dimensions (Figure 2) to assess the quality of the data [19]. In this work we will assess the completeness and correctness (free of error) dimensions of data quality.

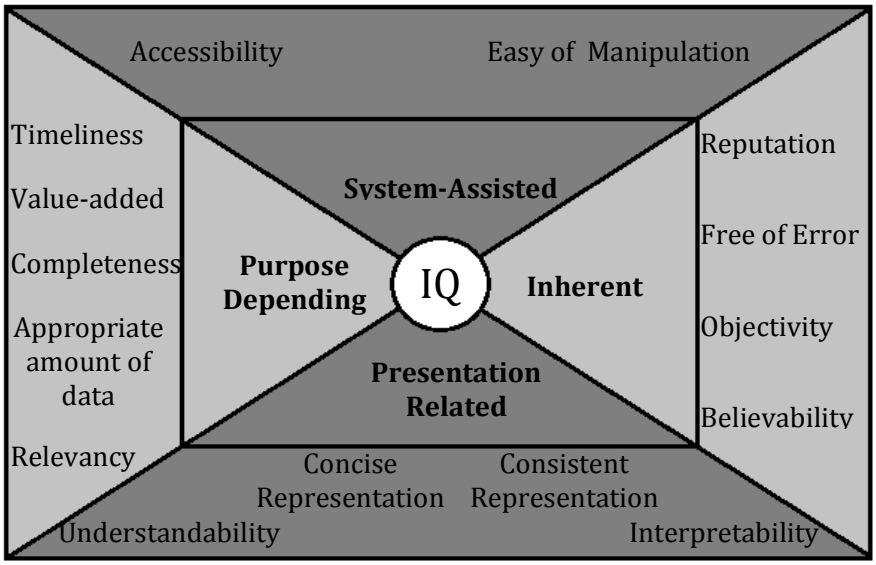

Figure 2. Quality Dimensions in 4 Quality Categories [19].

Completeness criteria refer to the question to what extent collected records have valid values in all fields. It will be defined as incomplete when missing values are found in fields where the presence of a data is expected. To assess data completeness a gold standard to compare with is required [12].

Figure 3 shows the formula for completeness ($Q_{completeness}$) on the attribute value level, where w represents an attribute value in the information system [19]. Completeness has

minimum value of zero if the corresponding attribute is not filled or contains value to null (semantically) equivalent (default). Otherwise the resulting value of the metric is one [19].

Regarding to correctness (free of errors) data is correct when the information present in the database match reality. To assess this dimension, a gold standard is required to compare the data [13]. A data value will be considered incorrect when it does not match with the gold standard. To evaluate the correctness, the formula is expressed in Figure 3. Where $w_1$represents an attribute value in the information system and w_r the respective attribute value in the real world. If the attribute value in the information system corresponds to the expression of the corresponding real world entity (distance is null) then correctness is zero otherwise the deviation is set with a maximum value of one. The metric for correctness, based on previous formula $[d_1(w_1, w_r)]$ is shown in Figure 3. [19]

$$Q_{\text{Completeness}}(w) := \begin{cases} 0 & \text{if } w = \text{Null or } w \text{ NULL (semantic) equivalent} \\ 1 & \text{else} \end{cases} \quad (1)$$

$$d_1(w_1, w_r) := \begin{cases} 0 & \text{if } w_1 \neq w_r \\ 1 & \text{else} \end{cases} \quad (2)$$

Figure 3. (1) Completeness and (2) Correctness formula to evaluate the errors., according to Heidelbrand et al. 2015.

1.2.3 ETL Tool

The integration of the data from one electronic source to another is an important task. Nowadays information systems have grown in complexity and need to be connected exchanging information to unify data view [20].

Data integration process can be carried out by an Extract Transform and Load (ETL) process. This process allows extracting data from different data sources, transform the data according to the need and load the information into a database. ETL software is generally

expensive. However in the market free open source software exists to perform this task. One of the most used open source software tools for data integration is Talend Open Studio [20].

Talend Open Studio allows customization to suit needs. This tool presents a unified platform that simplifies the development process, is easy to learn, and free. Talend Open Studio consists of a metadata repository and graphical designer with more than 400 components for connectivity, extraction and transformation. A component is a connector used to perform a specific data integration operation, e.g. databases, applications, flat files, web services. The component minimizes the amount of coding required to work on data from several different sources. This enables easy data profiling and modelling, drag and drop design and allow the reuse of the work across projects and modules [20]

Talend Open Studio can be connected to: Package applications (e.g Enterprise Resource Planning), databases, mainframes, files and web services (e.g. SOAP); data warehouse, data marts; built in advance components for ETL, including string manipulation, management of slowly changing dimensions and automatic lookup [20].

2. HYPOTHESIS

Laboratory data from HIS can be automatically verified and exported into a database for secondary use in Multiple Myeloma research, through an ETL process.

3. GENERAL OBJECTIVE

Implement and assess the quality of the exported laboratory data from the HIS by an ETL process for secondary use in scientific research.

4. SPECIFIC OBJECTIVES

4.1. - Evaluate the error of the laboratory data in the manual process of transcription from HIS to scientific database.

4.2. - Design and implement an ETL process to transfer the laboratory data as CSV files automatically into a database.

4.3. - Compare the quality of the laboratory data in terms of completeness and correctness in both processes of collecting data, manual transcription and by ETL process.

5. METHODOLOGY

A retrospective analysis was carried out to assess the quality of laboratory data. A subset of 8 laboratory parameters (Calcium, C-reactive protein, Haemoglobin, Lactate dehydrogenase, Creatinine, Thrombocytes, Albumin, Beta-2 microglobulin) was taken into account in 4 events of the disease (diagnosis, chemotherapy, transplantation and after 100 days of transplantation) from 162 patients belonging to the scientific database. This data was selected because it was already verified through a manual SDV process, defined as gold standard in this thesis. All the variables analysed were type numeric. The current manual process of transcription and the new automatic process were evaluated. The criteria to measure data quality were correctness and completeness [9].

5.1 Quality Evaluation in Manual Process

To carry out the assessment in the manual process of transcription, laboratory results from data pre-SDV with the gold standard were compared. Registered values were dichotomized (0 1) as data with error and without error respectively. As reference it has been taken into account the definition of correctness and completeness described before (Figure 3).
It was identified as incorrect if data pre-SDV has a different numerical value as data post-SDV (gold standard), and it was considered incomplete when data pre-SDV has a missing value in comparison with post-SDV data (Table 1). All patients where the date of the event did not match with the date of the event in the gold standard were excluded from the sample.

Clasification of Error			
pre-SDV		**post-SDV**	**Error**
A	==	A	Correct
A	=!	B	Incorrect
A	=!	.	Incorrect
.	=!	A	Incomplete

Table 1. Criteria to classify errors in manual process. (A, B: possible numeric results of the laboratory parameter.; "." Missing value).

5.2 ETL process design and implementation

A script was developed in Talend Open Studio software to perform the entire ETL process collecting data from CSV files from the HIS. This script was developed to work with the structure provided by the HIS of i.s.h.med and with a specific MySQL database as destination. A test MySQL database was implemented to check the entire process and to avoid interfering with the original database.

With the purpose to have all the needed information, first it was selected the sample of 162 patients through the ID of the patients already identified in the database. These IDs were introduced in the HIS one by one manually.

To obtain the laboratory values from the HIS, 3 tables were required: Demographic, Case and Laboratory values. The demographic table contains the ID of the patients and the number ID of cases; the case table contains the ID number of laboratory value (Figure 4). In this way, it is possible to do the match between the tables to identify to which patient ID the laboratory parameters belong. These tables were extracted as CSV files and saved into a folder.

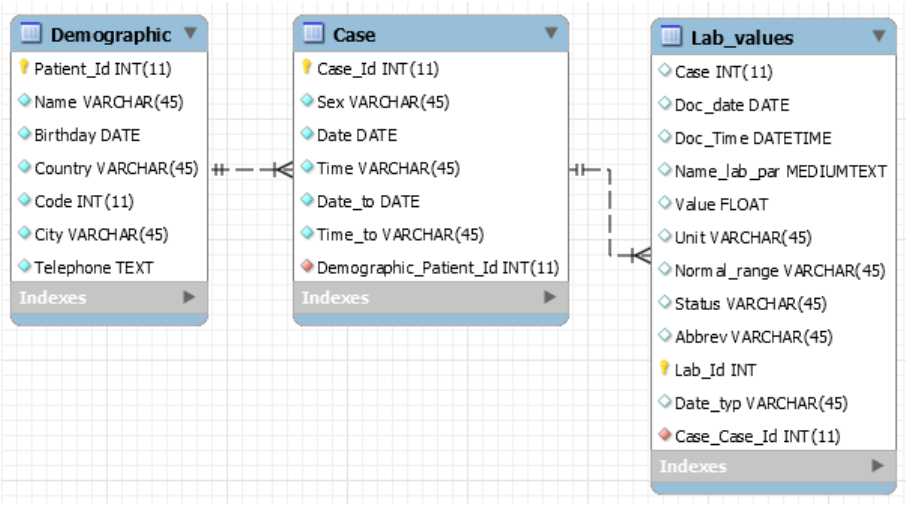

Figure 4. Selected data table from the HIS database.

Extract: The data from CSV files was integrated into the ETL software. The tables demographic, cases and laboratory with its respective data were included to have the specific

fields to be used to match data (patient ID, case ID and laboratory ID), and perform the respective analysis. These tables contain all the available information provided by the HIS. However only relevant information was taken (8 laboratory parameters with its respective date and time, patient ID and case ID) to carry out the automatic import of the data into the MySQL database.

Transform: The available information was filtered to keep columns of interest for this thesis. The provided information from the HIS had duplicate fields, hence it was used a component to eliminate fields that are duplicated through the laboratory ID parameter. Date and time data present different format, so it was transform the format into the correct one.
A mapping component was design to integrate the information provided by the 3 imported tables. As result it was only one table with the parameters to be taken into account at the moment of integration. Chronological order of the laboratory data was kept in the transform process.

Load: The information was loaded into the test MySQL database using a component of integration. A rejection error to catch all the errors found in the process was implemented. In addition an alarm system was designed for data without the appropriated integration. With the alarm, incompleteness and incorrectness data was detected.

5.3 Quality evaluation in Automatic Process

All data in the automatic process come from the results of the ETL tool. It was created a script in MySQL database to select the same data of the manual gold standard defined. The same criteria, correctness and completeness based in Figure 3 were used to compare data automatically transferred, with the gold standard. Later on it was compared both, the automatic process of data transference with the manual process of transcription.

Statistical analysis was performed using a chi square test as 2x2 table to analyse if processes are dependent of the error rate found. Test of given proportion was used to evaluate the statistical significance in the difference of error rate found in both processes of collecting data. The significance statistical level was fixed to 0.05.

6. RESULTS

6.1 Evaluate the error of the manual transcription process of the laboratory data from HIS to scientific database.

The results were evaluated through a comparison between gold standard defined and pre-SDV data. The sample was chosen by convenience where the SDV process was already performed and available. Because each event (diagnosis, chemotherapy, transplantation and after 100 days of transplantation) had a specific date where the laboratory parameters were taken, all patients were excluded where the date of the event did not match with the date of the event in the gold standard. The final total of observations included for the analysis was 3984. All analysed variables were numeric.

Principal Findings:

The result of the analysis was an overall error rate of 7.1% (283/3984) in the manual transcription process. From that percentage 81.6% (231) of the errors are due to incompleteness errors and 18.4% (52) due to incorrectness as shown in Figure 5.

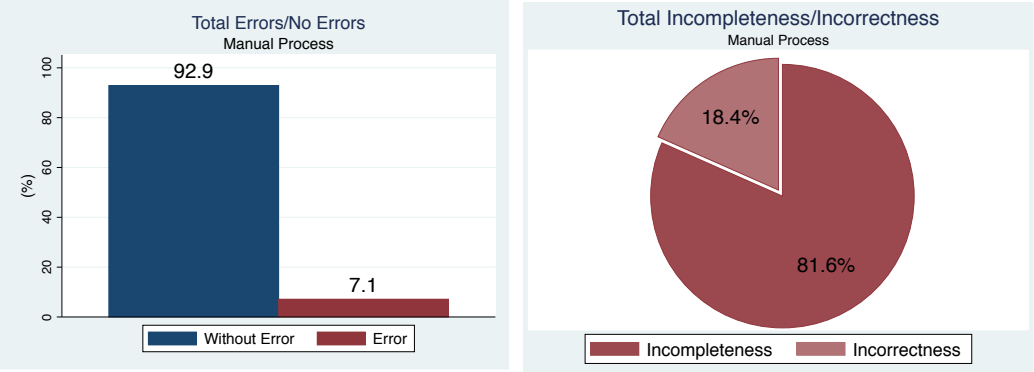

Figure 5. Total of Errors and its classification in manual transcription process.

By event the errors rates range from 5.4% to 8.9%. The difference of error rate in each event was not statistically significant. The incompleteness error was the most common type of error present in overall in all events (from 4.5% to 6.9%) in comparison with incorrectness error (0.8% to 2%), as shown in Figure 6.

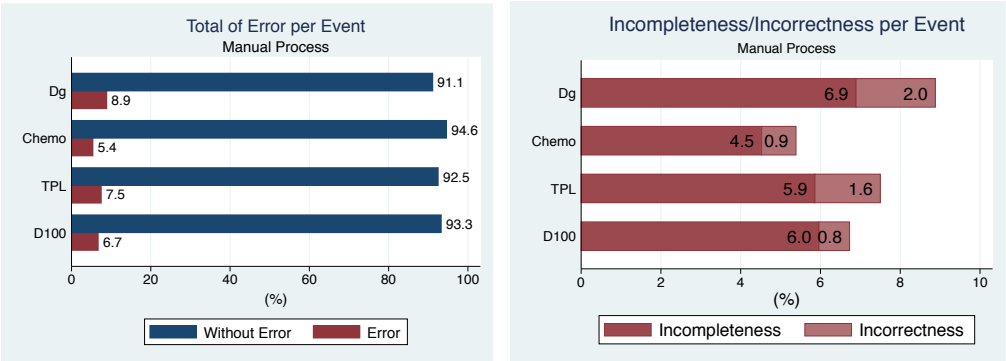

Figure 6. Total errors by event and its distribution according to incompleteness and incorrectness type of errors

Regarding the total errors analyzed per parameter, the parameter that presented highest error rate was CRP with 38.8% (from this percentage 36.8% were due to incompleteness and only 2% due to incorrectness error). For the rest of the parameters the rates of errors were similar, with a rate of error from 1% (Creatine) to 5% (Beta2 microglobulin) (Figure 7).

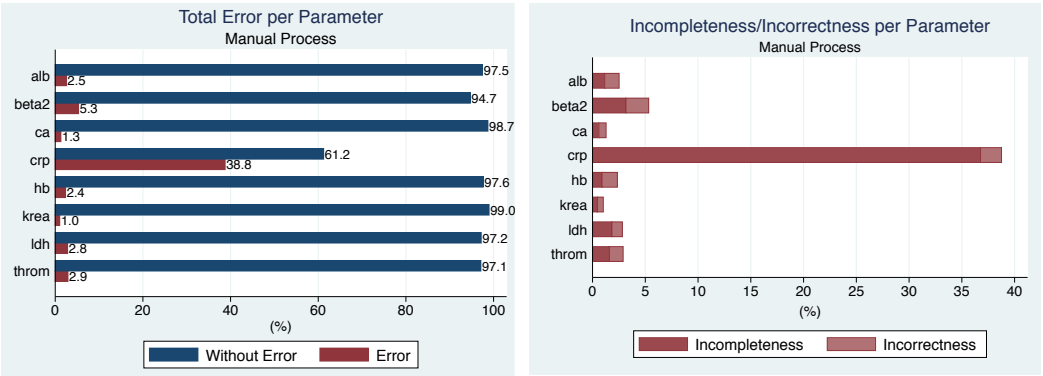

Figure 7. Total of error by parameter and its distribution between incompleteness and incorrectness errors.

The error rate of CRP dominates the error rate in the manual process. The results change if the sample is analyzed without taking into consideration the CRP parameter. It is showed that the general error rate decreases until 2.6% (87/3486) being the difference statistically significant with a $p<0.000$. Furthermore it is observed that, the incompleteness errors decrease in all events with an error rate from 0.5% to 3.2%, see Figure 8.

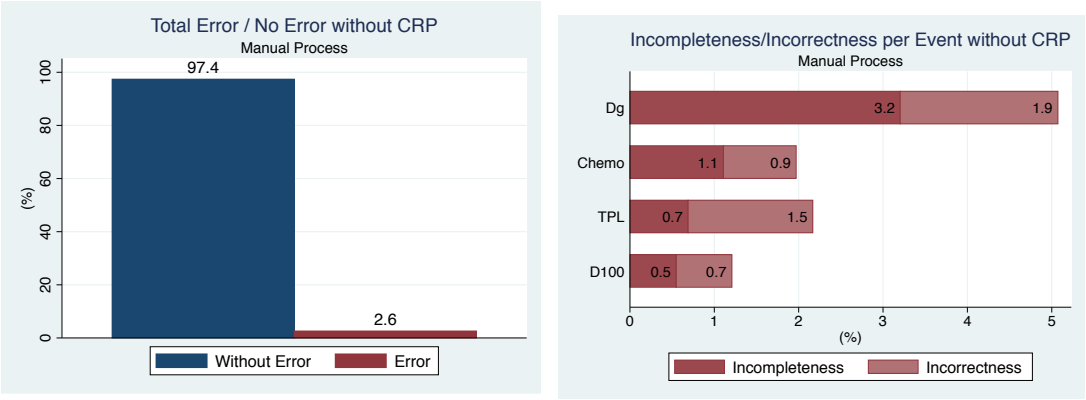

Figure 8. General Error Rate and Incompleteness and Incorrectness Error Rate by event without including the CRP parameter.

6.2 Design and implementation of an ETL process to transfer the laboratory data as CSV files automatically into a database.

A script was developed in Talend Open Studio software to perform the entire ETL process to collect data from CSV files for secondary use with scientific purpose. It was modelled an UML process to describe the activity of the program developed, see Figure 9.

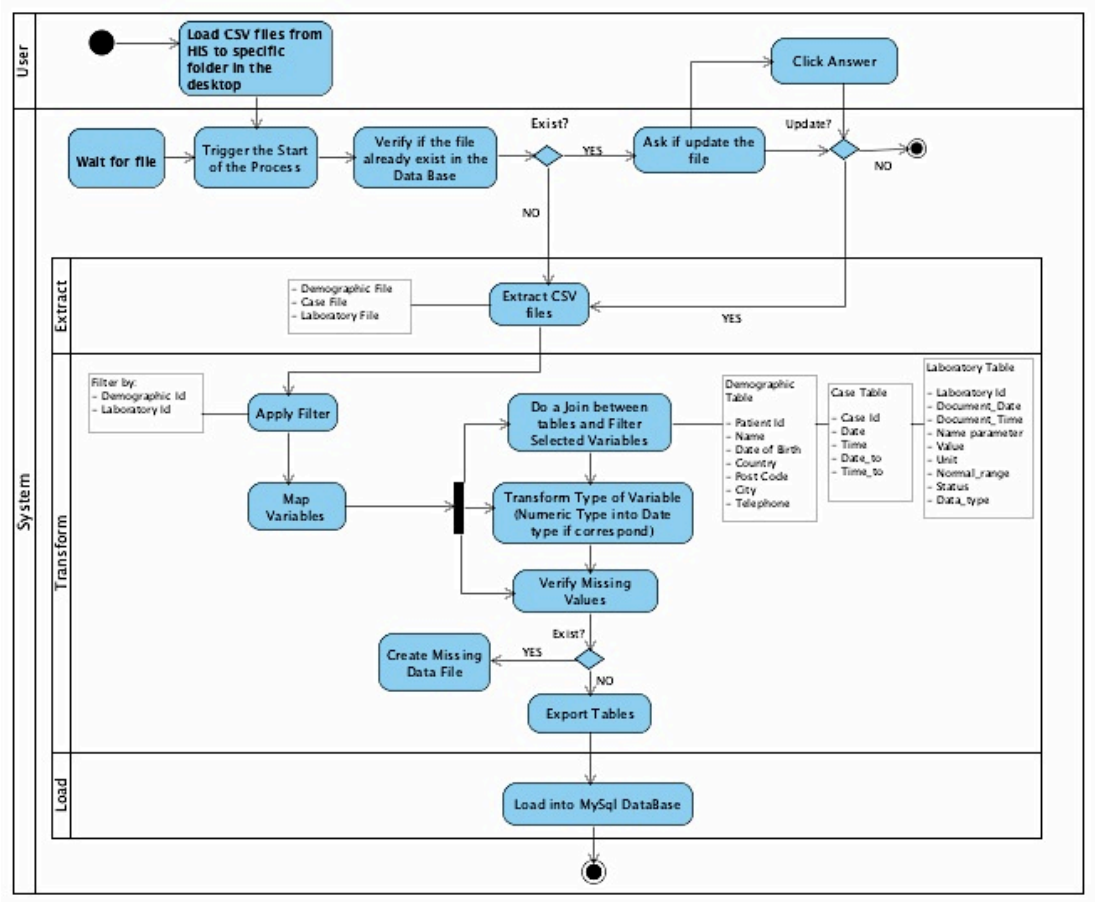

Figure 9. Activity diagram showing the entire process of ETL. It is described by a sequence where the user and the system participate. The subdivision of the system in Extract, Transform and Load clarifies the actions in each stage of the process.

Automatization

The principal function is the automatization of the process, i.e. the participation of the user was reduced to begin the process. The task of the software is to wait until the files are located in the directory defined by the user. The user must load the CSV files: Demographics, Case and Laboratory files into a defined folder that is located on the desktop. The component was developed to identify firstly if the files already exist in the database. If the files exist, the process concludes and a message appears: "the files already exist". If the files do not exist, the process continues and the ETL process begins.

Extraction

An extraction function to transfer the CSV files into the ETL software was modelled. The extraction process begins when the specific component identifies the CSV files located in a defined directory. The extraction comprises the reading of 3 files (Demographic file, Case file and Laboratory file) with the entire information, i.e. without changes in the schema and without exclusion of the variables. This activity marks the beginning of the main flow of the ETL process.

Transformation

Once the files have been integrated into ETL software, the transformation process can start. Because the original CSV files from HIS present replicated data in the files, it was applied a filter at the beginning of transformation process to have unique rows in demographics and laboratory files according to the ID in each case. The main component used was the mapping component allowing synchronizing the input tables with the output tables already defined.

When the mapping component receives a file, this does a join between the tables, it filters the variables that are not useful for the study, converts numeric data types into date data types, for instance the date of the laboratory parameter when it was taken. Also, in this part of the process, the validation of the data takes place as follows. 1) The system verifies the presence of missing values in laboratory parameters: if missing values exist, these are rejected and a file is created with all missing values found. In this way, only complete data pass to the output table. 2) The system checks if the laboratory parameters are out of the expected range, if so,

the values are rejected and another specific file is created, with incorrect results of the laboratory parameter out of range.

Load

The ETL software was linked to a MySQL database. When the ETL process completes the transformation process, data is loaded into the MySQL database and the process ends.

It is shown that the laboratory parameters that pass through this ETL tool per time, reach a maximum number of 809 data captured.

6.3 Comparison of the quality of the laboratory data in terms of completeness and correctness in both processes of collecting data, manual transcription and by the ETL process.

The automatic process developed as and ETL tool was first evaluated. The resulting of this process was compared against the gold standard. The total of analyzed data was 5184 fields. Next a comparison was carried out to compare the results between the manual and automatic process.

Principal Findings:

The general error rate found in the data collection through an automatic process was 12.1% (625/5184), from this percentage 94.9% (593) is due to incompleteness error and 5.1% (32) is due to incorrectness error (Figure 10).

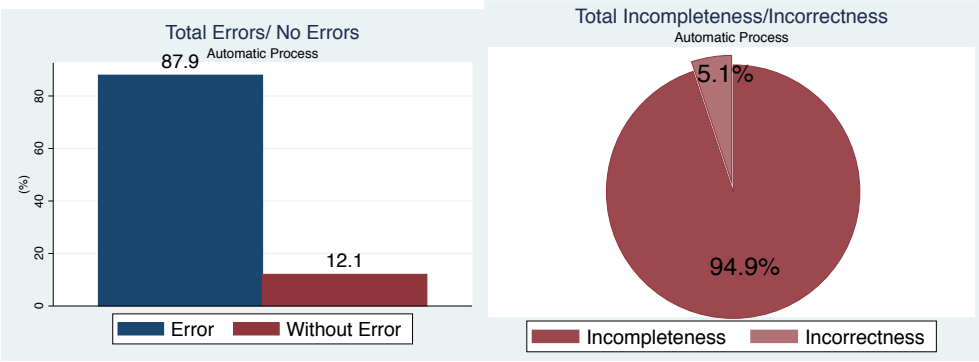

Figure 10. Total of error and its type of errors in the automatic process

The analysis of the error by event shows that diagnosis and chemotherapy event presented higher error rate with 30.3% and 13.4% respectively. Most of the error in these two events was due to incompleteness error, reaching the diagnosis event a percentage of 29.9% of incompleteness error (Figure 11).

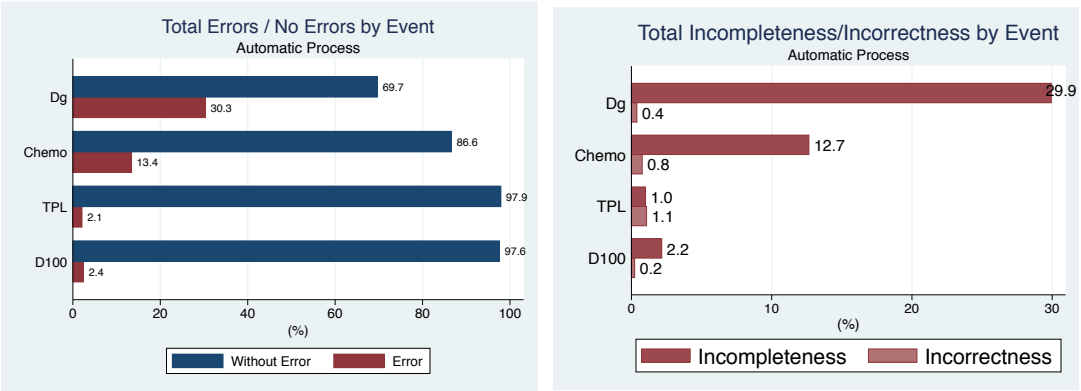

Figure 11. Total of error and without error by event and total of incompleteness and incorrectness by event.

Regarding the analysis by parameter, all laboratory parameter presented a similar error rate from 9% to 15%. Incompleteness error was observed as the most common type of error in general in all laboratory parameters, see Figure 12.

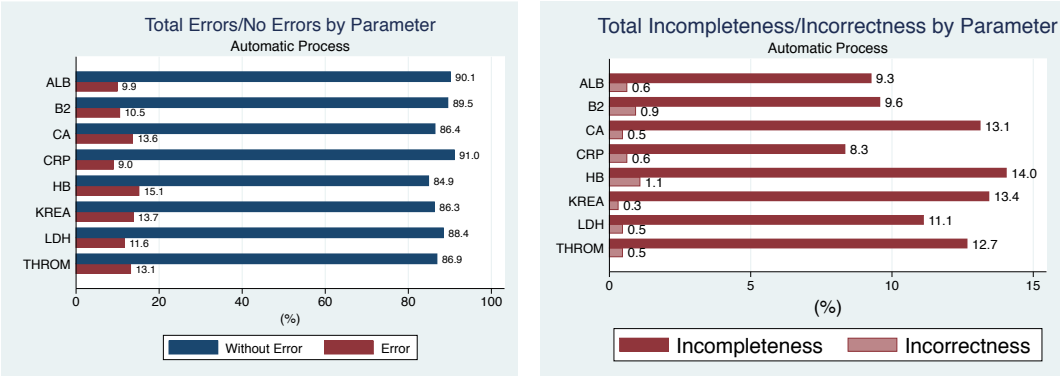

Figure 12. Total of error and without error by parameter and total of incompleteness and incorrectness by parameter.

The general error rate in the automatic process decreases to 1.9 % if the data missing in the HIS is excluded from the analysis (Figure 13). These records correspond to physical documents where laboratory data are registered. When the comparison is made only with information present in the HIS there is no difference at all, with an error rate of 0%. Nevertheless the percentage of 1.9% is maintained when is compared with the gold standard defined.

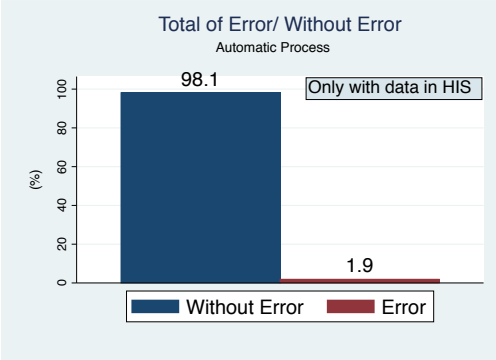

Figure 13. General error rate taking into account only data that are present in the HIS.

Comparison

To compare manual and automatic process of collecting data, the CRP laboratory parameter and excluding data not present in the HIS were excluded from the complete analysis.

The evaluation shows that the manual process presented higher error rate than automatic process (1.86% over 0.18%, respectively), in terms of incompleteness and incorrectness. The evaluation of chi square test shows that the error rate is dependent of the type of process, e.g, the manual process influences in the higher error rate found ($p<0.000$) (Table 2).

Analysis without CRP and Physical Documents				
Type of Process	**Error (n)**	**Without Error (n)**	**Total (n)**	**Proportion Exposed**
Manual	62	3263	3325	0.0186
Automatic	6	3319	3325	0.0018
chi2(1)= 46.59 Pr>chi2=0.0000				

Table 2. Analysis of Chi2 statistical test, without taking into account the CRP laboratory parameter, and excluding all physical documents not included in HIS.

If the given proportion test is applied, the automatic process presents a significant less error rate in comparison with manual process ($p<0.000$) (Table 3).

Analysis without CRP and Physical Documents			
Type of Process	**Error (n)**	**Without Error (n)**	**Total (n)**
Manual	62	3263	3325
Automatic	6	3319	3325
Proportions Test			
p-value = 0.000 (1.013e-11)			
95% CI: -1.00000000 -0.01249717			

Table 3. Analysis of given proportion statistical test, without taking into account the CRP laboratory parameter, and excluding all physical documents not included in HIS.

7. DISCUSSION

In a database for Multiple Myeloma patients, 8 laboratory parameters in 4 different events of the disease corresponding to 162 patients were analysed. The evaluation was applied to the current manual process of collecting data and also in the new automatic process of data transference created. A general error rate due to manual data entry of 7.1% was observed. The total of errors per event were similar varying in rates between 5.4% and 8.9%. When the analysis is done considering only the laboratory parameters, most error present was in the CRP (C Reactive Protein) field with 38.8%. Across all events and all laboratory parameters the most common type of error was incompleteness.

The general error rate in the manual process of transcription of the data was relatively high when compared to the average of 2.3% described in the literature for double entry method in a clinical research database [22]. The CRP parameter presented the higher error rate; this could be explained by the fact that there was no consensus and no clear protocol when the CRP value is below the detection limit. In this case, in Heidelberg laboratory the limit is 1mg/L. Some trained specialists entered the value of 1 when it is below the detection limit and others used the rule if CRP is below of this limit, they leave the item blank (‘missing value’). When the CRP parameter is excluded from the analysis, the general error rate decreases until 2.6% in the manual process of collecting data, which is similar to the rate described in the literature before.

Regarding the analysis done in the automatic process of data transference through the ETL tool implemented, it shows a general error rate of 12.1%. The events that presented higher rate of incompleteness were diagnosis and chemotherapy (29.9% and 12.7% respectively), being also this type of error the most typical error found. The highest rate found in these both events could be explained because at the time when the patient is included into the registry, most of them have already laboratory results from a physical document generated externally. This result is not included in the automatic export of the HIS and leads to an incompleteness error in the automatic process. In addition, for the event of chemotherapy,

some patients start with this phase of the treatment immediately after diagnosis. Therefore, the specialists may take the laboratory results from the diagnosis event and classify the same result as well for the diagnosis as for the chemotherapy event. The error rate of 12.1% decreased to 1.9% if physical documentation not included in the hospital information system is excluded. As expected, the error rate of 1.9% decreases to 0% when the data in the automatic process is compared directly with the HIS and not with the gold standard.

When the sample is adjusted by keeping only data that follows a standardized data collection protocol and included in HIS, the error rate is higher in manual process than automatic process (1.8% and 0.18% respectively). The error found in the manual process were due to mismatches between data Pre-SDV and Post-SDV, that could be interpreted as transcription errors. In the case of automatic process the errors found (6 in total) were due to mismatches between the gold standard (post-SDV data) and the data process by the ETL tool. This is explained because these fields defined as missing in the gold standard but the automatic process retrieve a valid value from the HIS. The mismatch could be explained by a manual exclusion of data (out of range, outdated).

The main limitations of this thesis were the reduced kinds of HIS fields chosen to compare and to have the manual evaluation already performed. This is because the HIS fields were chosen by convenience, i.e. only data with a manual SDV process already available, limiting the number of laboratory parameters to be analyzed. These fields were all numeric variables excluding other fields possible to be assessed, for instance string variables. Also, no measurement of the time of data collection by each process was available, however the tool developed has the potential to reduce time in the complete process.

The gold standard defined has extra information, where clinicians or documentaries decide if they take the result from the HIS or from an other resource for scientific purpose and also, they decide the best laboratory results from one routine day to be incorporate. This generates a bias in the analysis of the automatic process; because the nature of this process is just to take the data from HIS that follow a standardized protocol, without allowing changing it.

This study was developed taking into account only laboratory data from a specific database that contains data from Multiple Myeloma patients. However the creation of an adaptable ETL process could be expand to use with other data, such as demographic, diagnostic, and treatment data for patients with multiple myeloma, and other settings where standardized protocol of collecting data exist and the primary information is captured in electronic format. This could be replicable in Chile, in researches that present structured data, for instance obligatory notification disease, or reports of adverse events.

8. CONCLUSION

Based on the presented results of this thesis work it is possible to conclude that laboratory data as CSV file from the HIS can be automatic imported into a database for secondary use with scientific purpose by an ETL process, if exported data fields follow a standardized data collection protocol and only it is considered data included in HIS. Specifically when manual and automatic processes are compared under those conditions the manual process error rate is 10 times higher than the automatic one (1,86% and 0,18% respectively) for the case of laboratory data as Multiple Myeloma.

Automatic ETL processes, as the one implemented in this work, avoid careless errors like transcription mistakes, by automatic validation of the data at the point of entry in the scientific database. Also transcription times have the potential to be drastically reduced by ETL processes directly connecting data exported from HIS with EDC systems. Future work to should evaluate this kind of automated tools in other clinical domains and type of variables, using as starting point the presented thesis.

9. BIBLIOGRAPHY

1. Köpcke F, Kraus S, Scholler A, Nau C, Schüttler J, Prokosch H-U, et al. Secondary use of routinely collected patient data in a clinical trial: an evaluation of the effects on patient recruitment and data acquisition. Int J Med Inform. 2013 Mar;82(3):185–92.

2. Dentler K,ten Teige A, de Keise N, Cornet R. Barriers to the Reuse of Routinely Recorded Clinical Data: A Field Report. Studies in Health Technology and Informatics [Internet]. 2013 [cited 2015 Jul 28];313–7. Available from: http://www.medra.org/servlet/aliasResolver?alias=iospressISSNISBN&issn=0926-9630&volume=192&spage=313

3. CPMP/ICH. Good Clinical Practice. European Community. [Internet] 1996 [cited 2015 Jun15]. Available from: http://ec.europa.eu/health/files/eudralex/vol-10/3cc1aen_en.pdf

4. Wagner G, Leiner F, Gaus W, Haux R, Knaup-Gregori P. Medical Data Management: A Practical Guide [Internet]. Springer New York; 2006.

5. Krishnankutty B, Naveen Kumar B, Moodahadu L, Bellary S. Data management in clinical research: An overview. Indian Journal of Pharmacology [Internet]. 2012 [cited 2015 Aug 4];44(2):168. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3326906/

6. Nahm ML, Pieper CF, Cunningham MM. Quantifying Data Quality for Clinical Trials Using Electronic Data Capture. PLoS ONE [Internet]. 2008 Aug 25 [cited 2015 May 28];3(8). Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2516178/

7. El Fadly A, Rance B, Lucas N, Mead C, Chatellier G, Lastic P-Y, et al. Integrating clinical research with the Healthcare Enterprise: From the RE-USE project to the

EHR4CR platform. Journal of Biomedical Informatics [Internet]. 2011 Dec [cited 2015 May 7];44:S94–S102. Available from: http://linkinghub.elsevier.com/retrieve/pii/S153204641100125

8. Laird-Maddox M, Mitchell SB, Hoffman M. Integrating research data capture into the electronic health record workflow: real-world experience to advance innovation. Perspect Health Inf Manag. 2014;11:1e.

9. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. Journal of the American Medical Informatics Association [Internet]. 2013 Jan 1 [cited 2015 May 28];20(1):144–51. Available from: http://jamia.oxfordjournals.org/content/20/1/144

10. Mitchel JT, Kim YJ, Choi J, Park G, Cappi S, Horn D, et al. Evaluation of Data Entry Errors and Data Changes to an Electronic Data Capture Clinical Trial Database. Drug Inf J. 2011 Jul;45(4):421–30.

11. Olsen IC, Haavardsholm EA, Moholt E, Kvien TK, Lie E. NOR-DMARD data management: implementation of data capture from electronic health records. Clin Exp Rheumatol. 2014 Oct;32(5 Suppl 85):S–158–162

12. Benson T. Principles of Health Interoperability HL7 and SNOMED [Internet]. Springer; 2012. Available from: https://books.google.cl/books?id=CNcCcdUsoOYC

13. Vawdrey DK, Weng C, Herion D, Cimino JJ. Enhancing Electronic Health Records to Support Clinical Research. AMIA Summits on Translational Science Proceedings. 2014;2014:102-108.

14. Wissenschaftliche Dokumentation in der Sektion Multiples Myelom [Internet]. Heidelberg: UniversitätsKlinikum; c2015 [cited 2015 Jul 21]. Available from:

http://www.klinikum.uni-heidelberg.de/Wissenschaftliche-Dokumentation.131688.0.html

15. Klinisches Myelomregister des UniversitätsKlinikums Heidelberg[Internet]. Heidelberg: UniversitätsKlinikum; c2015 [cited 2015 Jul 21]. Available from: http://www.klinikum.uni-heidelberg.de/Myelomregister.116298.0.html

16. Vincent Rajkumar S. Multiple myeloma: 2014 Update on diagnosis, risk-stratification, and management. Am J Hematol. 2014 Oct;89(10):999–1009.

17. About Multiple Myeloma. [Internet]. United States of America: Multiple Myeloma Research Foundation; c2015 [cited 2015 Jul 15]. Available from: http://www.themmrf.org/multiple-myeloma/

18. Kirkwood AA, Cox T, Hackshaw A. Application of methods for central statistical monitoring in clinical trials. Clinical Trials [Internet]. 2013 Oct 1 [cited 2015 Aug 11];10(5):783–806. Available from: http://ctj.sagepub.com/cgi/doi/10.1177/1740774513494504

19. Hildebrand K, Gebauer M, Hinrichs H, Mielke M. Daten- und Informationsqualität: Auf dem Weg zur Information Excellence. 3a Ed. Germany: Springer-Verlag; 2015.

20. Talend Open Studio for Data Integration. User Guide 5.6.1 [Internet]. 2014. [cited 2015 Apr 21]. Available from: http://www.talend.com/download/talend-open-studio#s6

21. Goldberg S, Niemierko A, Turchin A. Analysis of Data Errors in Clinical Research Databases. AMIA . Annual Symposium proceedings / AMIA Symposium AMIA Symposium;2008:242–6