International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

An Effective Data Analysis for Lymphomas Outcomes and Risk Levels using Combined Advanced Data Mining Algorithms

[1]Manu M R, [2]T Poongodi

Submitted: 12/02/2023 **Revised:** 10/04/2023 **Accepted:** 11/05/2023

Abstract: Nowadays cancer has become a vitasonr measurable global death., because insufficiency of proper treatment and also late diagnosis treatment. with the help of various data mining techniques can recognize or predict cancer diseases threat level with different dimensions covered by medical date. The data which give space to predict risk of cancer including different factors (both risk and investigation factors). Our proposed multidimensional analysis upgraded with cognitive experts (data management and data analysis) Also getting support with oncology domain experts for huge amounts of data and excellent analytical skills. Our proposed data mining analysis evaluated lymphomas of the entire body data extract through analytical methods in ontology. In general, Neural networks (NN), Clustering, Regression, Decision Tree (DT) performing with high accuracy and precision rate than association based and Naive Bayes Classifier. Analyzing with various data mining algorithms are ready for combination or disjoint to provide an effective medical data analysis with high accuracy. Data prediction analysis of caCORE, SEER dataset system and systemic experimental integration proved that distributive system is extremely complex. Providing some betterments made to simplify and an effective generic caCORE data model. Use divide and conquer basis split huge volume it into smaller portions and expose as combinational functionality as data hiding or web-oriented interface classes are decreasing the complexity of the model. Combined Data analysis models with multidimensional analysis (MDA) performed better than other models. Our proposed methodology gives a better results and diagnosis prediction for Lymphoma analysis for different age group starting from age 25 to higher . It was calculated and analyzed combined classifiers gave better results with accuracy of 88.17%, result of independent or structured data accuracy was 90.10%.

Keywords: *structured, combinational, independent, precision*

1. Introduction

1.1. Understanding the need of data warehousing

When studying literature introducing readers to the topic of data warehousing, either a lot of features of these suites is provided or the description is spread across so many pages that it takes too much time and effort to spot the new value a data warehouse introduces and reasons why it is built. Moreover, typical and common buzzwords define advantages of deploying them like: “gaining competitive advantage” or “accelerating business growth”. A typical database contains a vast number of detailed information and is possibly the best solution one can choose for quick and reliable storage of vast amount of data and for acquisition of details. A typical accompanying application built on top of a data source is designed for dealing with rather single instances of data entities, which can be of arbitrary form. A record representing a person, or a car or whatever can be very efficiently manipulated and combined with other records it is bound with by a relationship. Even when filling a sort of tabular displaying structure with data, the number of rows retrieved is in most cases relatively small to what databases in business can store. There is a price to pay for this very “close-up” view. Namely, data consumers are almost separated from any sort of “big picture”. Assume a database application business area supports shopping center trading - they can tell what sort of products Svenson or Kowalski bought on a specific day in a specific shop, but there is actually no way for them to see the information from wider perspective. Therefore, they would struggle to analyses the data already stored and to draw any conclusions. Following that line, they are incapable of making business decisions being provided with* details only. What mentioned users actually see, can be expressed by a metaphor-alike table below.

Research scholar, Professor, SCSE Galgotias University, Greater Noida, Uttar Pradesh, India manuramachandran18@gmail.com

t.poongodi@galgotiasuniversity.edu.in

Table 1: Small scope view of data (Browse Data: Sales_fact, 2021)

	product_id	time_id	customer_id	promotion_id	store_id	store_sales	store_cost	unit_sales
1	173	748	2094	54	1	4.29	1.8447	3
2	1119	748	2094	54	1	9.51	3.5187	3
3	1242	748	2094	54	1	7.92	2.8512	4
4	460	748	2094	54	1	6.44	2.7048	4
5	104	748	2094	54	1	11.67	3.9678	3
6	27	748	2094	54	1	7.95	3.816	3
7	67	748	1277	54	1	7.44	2.9016	4
8	217	748	1277	54	1	2.72	0.8432	4

The above table describes the data's in the warehouse with varying data unit . A bird's eye perspective a data warehouse delivers allows its users to understand the relations and spot the facts in data that spread on a much wider number of instances (records) or database structures that reflect company framework. All the details visible in the previous approach are sacrificed (still though, DW tools allow users to magnify selected areas of data and pull original records from source databases before lossy aggregation and summarizing transformed them – the feature, called most often a "drill – through" will be covered more precisely later on) Therefore, we can think about a data warehouse not only its technical definitions and architectural demand terms but also as a different approach to the data. There is nothing new or particularly innovative in this – all sort of charts used to summarize the data even in the times when no database technologies were known share the attitude. DW tools evolved from them in order to challenge situations when huge amount of stored data from heterogeneous, mutually incompatible in terms of standards sources, reporting versatility, easiness of use and quick report delivery have to be combined.

1.2. Multidimensional perspective of the data, cubes, measures and customizable calculation.

The way how one accomplishes a need of aggregated, multidimensional view is based on a very aboriginal ingredient of actually any database application – a relation. Relations join database entities reflecting the real – world way they interact. A quick returning instance: a database for shopping center business would possibly store information about products offered, types of products, customers, shop assistants, promotions and shops themselves, to name just a few. A case of selling a product would be reflected in a transaction that is also made persistent. Connecting each and every transaction with information like what was bought, who bought, in the scope of which promotion, by whom she or he was assisted, when and where is the responsibility of relations. Therefore, they provide excellent traces for observing data from different angles. If we know (because of the data stored) about all the transactions in a for instance previous fiscal year, we can ask questions like what properties are most common among customers who generate highest revenue for the company or at what time sales of what sort of products were the highest. In fact, number of these types of questions is probably boundless and depends on the analyst imagination and curiosity only. To answer them, we have to look at the data from a perspective of a customer, product, place, time etc.

The next key implementation factor of a data warehouse are measures. They are often single numerical columns pulled directly from the database, but can also be composed of many fields put together in a formula. What's most relevant, values to be observed in data are delivered by them. In the mentioned example of shopping centers, an income, number of items sold or cost incurred in single transaction could be good candidates for measures. As pointed out, measures could also be more complicated formulas, making use of values of more than one column. In order to be prepared for situations if this is the case, tools facilitating processes of creation of a data warehouse application shall allow developers to make use of predefined logical (returning Boolean results), mathematical (like standard deviation, regressions) or data navigation functions (like the ones that form an array or set from records provided or can navigate through relations) to operate on database-alike structure. If the development is done from scratch, without using any available warehouse builders, it's almost sure functionalities listed above will have to be implemented.

Multiple perspectives sharing the same data are combined into a structure called a cube. The name is frankly derived from the very first way we image a multidimensional structure. The most advanced one most of us can think of in the term of the shape is a 3D cube and this was used to reflect the multiple viewpoints of the data a data warehouse provides. In real data warehouses, their data storages layer for aggregated and summarized data differs substantially form the way an

operational database is designed. In details, much more redundancy is allowed and the general schema is star-alike, with a centric table storing events users need to analyses and related entity tables surrounding it. This can be multiplied as many times as the count of observed events. Common name for mentioned event storage is “fact table”. In operational databases, schemas get much more complicated than a start-alike one. Moreover, established dimensions can be divided into levels on which data can be analyzed. An exemplary case could be a perspective (dimension) of a location, that could be separated into countries, states, cities, districts levels etc. according to user needs and depth of location data.

2. Traces of Similarity to Data Warehousing in Data Mining

When putting data warehousing and data mining techniques side by side, only partial, little resemblance can be spotted. They both share the similar goal and benefit of data owners which is making them more aware of valuables facts and information contained in huge amount of data they have been collecting. This is partially done thanks to summaries of data these techniques deliver and abilities to discover some trends in data their owners didn't know anything about, or at least weren't sure about their existence. Both techniques are designed to make large amount of data more comprehensive. At that point though, the similarities end.

In the literature this chapter is based on ([2]), the definition of data mining in use, that spans over the whole book, is (to quote it rather directly) that it is analysis of (very often large amount of) data to find unsuspected relationships. Still a great difference between these techniques and data warehousing cannot be seen. The reason for that is because it is hidden in the way analysis is done. In a data warehouse, instances (or better to say samples- which in most cases a derived directly from records, single ones or some of them joined) are summarized in a rather straightforward way and viewed from separate perspectives. Let sums of integer fields (like money paid for goods) of records belonging to certain groups (this implements perspectives) that are calculated be an example of that. In data mining techniques this is a much more sophisticated process with some stages that can be separated in it. Moreover, there are a few focused data mining tasks and methods used to fulfil them are adjusted for the job as well.

2.1. Creation or selection of a model

At the very beginning, we need to find out what kind of model describes the data best. It is a tricky task in most cases because of the different structure of available data. At the moment we distinguish the cases in our observation (these can be database records for instance) and their attributes (represented by column attributes), many challenges may arise. Some variables are continuous, other discrete like the ones that allows only categorized or enumerated values, the other terms of the same variables which implies that a standardized description of every case would require adding new variables to it for the sake of compatibility, with empty observation value as the only possible. And this matrix-alike form of very general model still does not support the way observation of cases changes in time. Anyway, researchers are not left alone at any of the data mining process stages, at this one in particular. Some heritage of already developed process components lends a helping hand and there are some predefined models that describe certain sorts of input well. A score function allows to verify whether a chosen representation of the data describes data properly and to adjust model parameters in order to reach appropriate accuracy. Functions like misclassification rate, sum of squared errors or like hood [3] are most commonly used because of their general character elasticity in model verification and fine-tuning. Score function formulas shall also be reasonably priced in terms of computational effort required to calculate their values. Being prone to surprisingly high changes of results for some specific input from the available data is not desired as well and such score functions could lead to inappropriate model verification.

Optimization of the score function

On the way for adjusting model parameters an extreme value of the score function is to be found, for instance a minimum if it measures the error in representing data with selected model. Because value of this function can be measured for arbitrary sample from the data repository and the possibilities are usually numerous, the problem is usually transformed to an optimization task. Depending of its exact nature, this can be tacked using algebraic transformations, differential calculus or heuristic search. One has to pay attention though to avoid the model over-fitting the data, which means representing them too precisely. Such cases decrease efficiency of prediction new cases values.

2.2. Managing the data

The origin of the problem is, as one could expect, large amount of data, but in fact not the possible storage shortages cause it but delays in accessing it. Some optimization algorithms simply do not take into account the fact that some input data they use when finding extreme values may not be accessible as quickly as other ones. If the data set used for optimum search is relatively small and it can be stored in RAM memory of the computer, any difficulties related to that disappear. Large

datasets however, which are stored in relational databases on much slower disk-based media, can slow the algorithm execution substantially.

A remedy for that is a boost in fetching data provided by state-of-the-art database engine, with usage of indexing and query optimization techniques playing relevant role. Still though, it cannot be compared with a situation where optimization input is located in random-access memory and freely accessible.

2.3. Data mining specific tasks

Data exploration: There is no main goal defined for this sort of data mining activities and it focuses rather on graphical representation of data and simplification techniques in case of its high dimensionality (like principal component analysis) and projections to make the input more visual and human-readable. Researcher effort in this area has been directed into proposing solutions for mapping very complicated relations into simplified structures. When this is achieved, analysts can conclude from the data more easily or this gets feasible at all.

2.3.1. Descriptive modelling

The area of application embraces approaches to apply mathematical or structural models to input data and by saying that, estimations of probability distributions in data, its segmentation or clustering in groups of entities sharing selected characteristics is meant.

2.3.2. Modelling for prediction, classification and regression

While previous task aims rather to make data more explanatory and to acquire as much information of it as possible, current one's target is to forecast a value of selected attributes basing on their (and additional ones) values from the past. According to [3], these predictions shall rather be considered in limited, defined scope of time, not for future in general; predictions of both categorical and continuous attributes are handled, by classification and regression, respectively. What's distinguishes predicting techniques from explanatory ones (mentioned in the last paragraph) is that they do not focus on the whole attribute space.

2.3.3. Discovery of patterns and rules

Although previous tasks can help detecting patterns and relations in data analysts weren't aware about as well, this task is strictly about that and mainly applies association rules in the effort of spotting them. Trying to detect boundary sections in input data – regions when standard properties end and abnormalities begin can lead to successful rules discovery. Matching a proper model to the data in this case is of lesser importance.

2.3.4. Retrieval by content

Explaining that the possibly very best example of this task is the most popular web search engine allows to image what it is about. The challenge is to find best matches in the data (these can be of every type – relational table entries, media, textual documents or whatever) when being provided with a small trial sample of what's desired. According to literature, the distance, or similarity between sampled entry and stored data it to be evaluated, generally speaking, still though, techniques to achieve that vary a lot, especially at the stage of transforming general input “signal” (text, graphics, sound, categorical entries for instance) to its mathematical representation.

2.4. Analysis of National Cancer Institute USA cancer cases data

The goal of this chapter is to draw conclusions from analysis and observation performed on large cancer diseases dataset (1975-2021) over USA population. Using data warehousing analysis and data dimensions like sex, race, stage of disease and is morphology and primary site, residence area, age at and year of diagnosis and others and data mining techniques, questions like

- which are the most frequent and most growing (rates) types of cancer
- what is the age structure of diseases occurrence?
- what is the relation of ratios of some specific cancers (including the most frequent ones)
- what are the trends over years in selected as well as general cancer diseases prevalence and how they relate to selected factors (like mentioned data dimensions)
- in general, how data warehousing and mining techniques perform in drawing meaningful

conclusions from the data and how well are they capable of visualizing trends and regularities

2.5. Introduction into SEER Stat system

The system can be considered as a medical data warehouse with its front-end available to all who sign an agreement limiting distribution of database content that is going to be made available to them and download client-side software from NCI website.

The data repository itself can be accessed on-line, using light client only which allows creation and execution of queries than are then being sent to the SEER Stat for servers-side processing and response generation. The response is then retrieved by the client and displayed in a tabular form. Elasticity of queries formulation allows it to simulate Excel and Analysis Services pivot tables

(which dimensions are locked in rows or columns, also overlapping, is fully adjustable). Processing queries generating even large datasets generated by the server-side scripts took rarely more than a minute. Further options for data usage are downloading it in a binary form (around 500 MB) or as textual flat files (~1,5 GB) with specification of fields to use it in third party analytical tools. Working with SEER Stat is organized into sessions, in which users can analyses aspects of data with separated concerns. Sessions share almost all dimensions of the data and allow narrowing the resulting dataset using conditional statements very similar in form to standard SQL dialect WHERE clauses. Available sessions deal with:

- frequencies of particular diseases or sets of diseases, containing counts of cases
- rations of diseases (normalized results, containing all counts, rates as well population serving as computational basis)
- survival chances when developing specific types of cancers
- case listing
- limited duration - prevalence of selected diseases (a history of diseases frequencies over an extended period of time, especially useful when performing estimations for medical funds and resources management)
- MP-SIR (Multiple Primary - Standardized Incidence Ratios) – for analysis of multiple instances of cancer appearing subsequently

At every session, data source needs to be selected, sessions aspects adjusted (for instance, whether crude of age adjusted rates are demanded), result set desired dimensions chosen (including their location in pivot tables) and selection/projection conditions put in. After processing the query, tabular result set is returned with the granularity level corresponding with input parameters. It's worth noting that SEER Stat itself does not visualize data so an external tool for that purpose is needed. Fraction of possible analyses is available via interactive web pages hosted by NCI (these web applications tend to have better visualization features, with graphs and charts generated on the fly), still though, SEER Stat potential is far higher when compared with quick access web layer of the system.

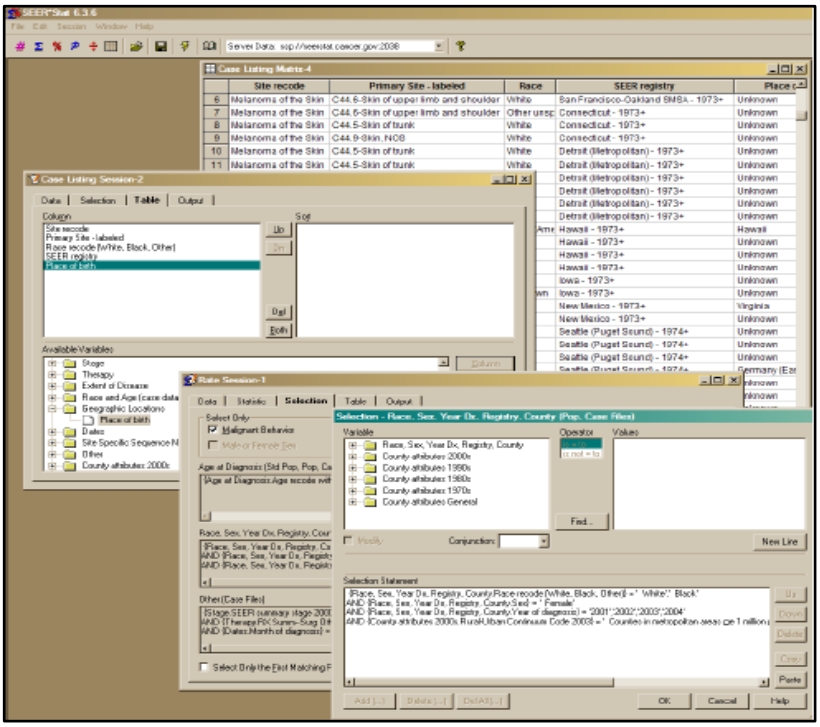

Fig 1: SEER Stat screenshot taken during an analysis task.

Visible are the result matrix, selection expression designer, session window and dimensions explorer. The above table represent the SEER statics in which variable SITE record denotes the type of the record, the field site label denotes the name given for the record ,where RACE denotes whether it is white or black men, SEER registry denotes the which record at which indexing and place denotes the location

2.6. Aspect of data analysis and eventual ETL processes

In the first stage of data analysis, when SEER Stat software was used only, no transformation and loading

of data into any database was required since this was entirely handled by the tool and hidden from the researcher. The only aspect necessary to notice was the distribution of values of parameters that were used for plotting. While SEER database makes large number of parameters of data available, one has to take care whether they are filled with data for sections that are researched. For instance, disease stage parameters vary over the time with only one being actual through all 30 years represented and other being limited to some period of time. The newest, most precise TNM cancer staging classification is made available for 2020+ records only. Similar characteristics were displayed by other parameters. Because of that, necessary checks of parameters availability were performed before they were used for multidimensional analysis, otherwise a risk of false results would be significant. Checks were performed by consulting parameters dictionary for possible lacks of data for given time period as well as by executing case listing sessions limited to fractions of records/timeline under investigation. By verifying presence of desired attributes in one-to-one data snapshot, a certainty was achieved that high-granularity plotting will provide dependable views.

In the second stage of analysis, where large number of records had to be extracted from SEET database and exported to data mining tools, the check described above was also necessary to be sure data are complete. It turned out not transformation and normalization of records was necessary when forcing them into data mining tool tables since SEER data source was fully normalized and standardized, without any need of manipulation and transformation of column types and their values. Such smooth operation was partly possible sue to elasticity of MS SQL Data Mining software which accepted data as they were, without complaining. All that was needed was to set columns of mining data tables to types that matched data extracted from SEER.

3. Data Warehousing Analysis

Following analysis of data provided by the SEER system has been performed using its built-in data warehousing features, then visualized with a spreadsheet tool. Fig: 2 All cancer sites rate and year of diagnosis

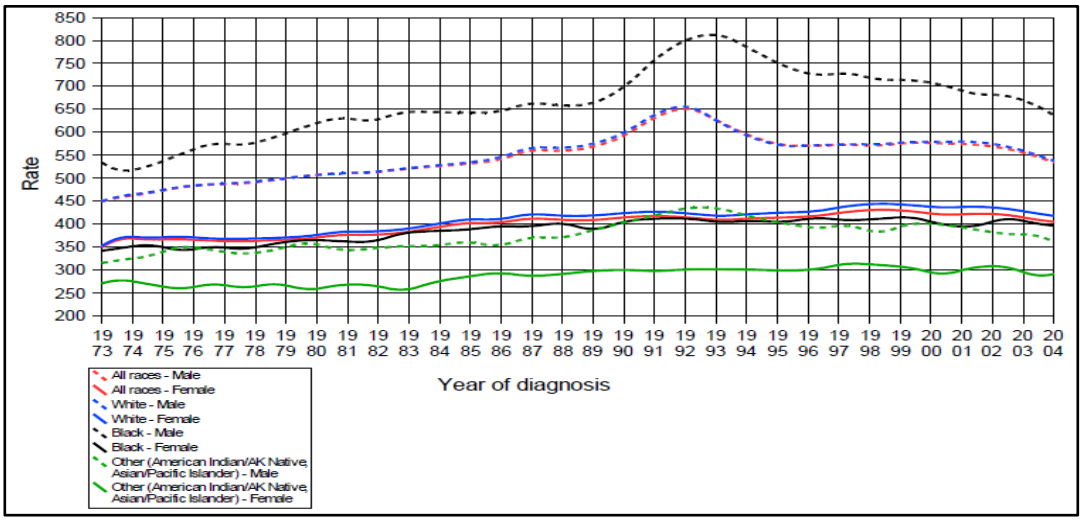

Fig 2: All cancer sites rate and year of diagnosis.

The chart has been readjusted (shifting from discrete into continuous and y-axis relocation) for the sake of visualization improvement.

3.1. System Analysis of All cancer sites rate and year of diagnosis

- Most promising observation is that overall rate of cancer diseases over many years (1973 – 2021, as NCI statistics are available) does not grow intensively and continuously and tended to fall in last years, returning to the level observed in middle eighties.
- Women tend to suffer much less than men from cancerous diseases in general, with general rate difference nowadays equaling ~150 cases less per 100 000 of population, comparing to general male rates its almost 40% less. Naturally, these trends summarize all types of cancer so particular morphology types of it may occur with much higher rate in women than in men.
- In cases of both sexes, races other than black and white proved to be much more resistant to cancer diseases in general, accounting to the rate a 100 (in case of females) to even 150 cases less (males) per 100 000, which amounts to around 30% and 27% less, respectively. This fact could be used in planning distribution of resources for medical screening and early detection clinical investigations

to ensure equal treating of all races, taking into consideration how likely which of them are to develop a cancerous disease.

- Reasons for extraordinary peak in developing cancers in general for the years 2012-2013 haven't been found, however, traces of it were spotted in some medical publications [8].

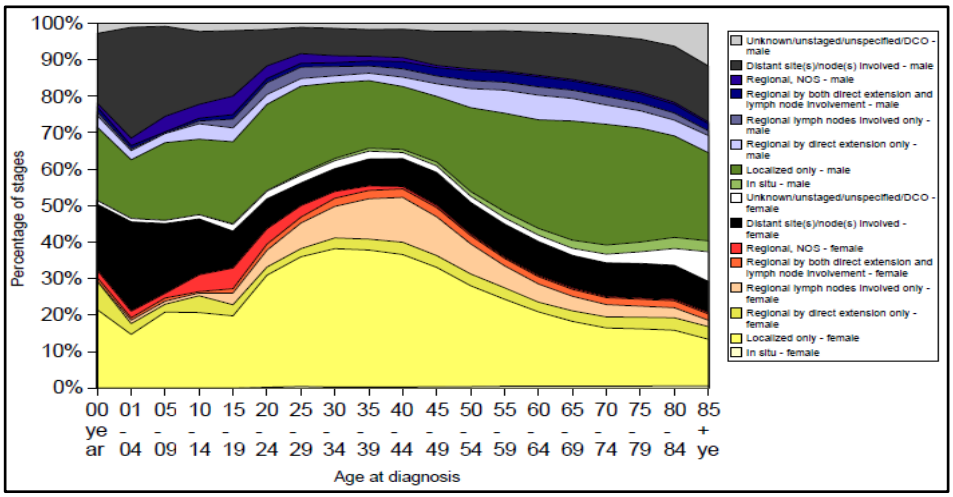

Fig 3: Percentage of cancer stages by age at diagnosis

System Analysis of Percentage of cancer stages by age at diagnosis

- The most alarming fact observed is a very high share of distant metastasis cancers in the overall balance of all cancers detected. Since these stages are very often considered terminal and incurable, with palliative care to increase quality of life, conclusions can be drawn that, when seen overall, cancer detection and screening is still very poor
- The fact supporting such hypothesis is an extremely low rate of in situ cancers, which is their earliest development stage where treatment is in most cases very successful, with full recovery being a normal situation
- Since selected cancers, like breast, prostate and melanoma do not share this distribution of stages, a conclusion can be drawn that, while some most frequent or most growing in rates cancer attracted attention of media and society, resulting in earlier detection, the high number of less frequent cancers remain undetected until it is actually too late. Distantly metastasized cancers, when summed up with extensive lymph node involvement cancers, which also tend to have not a good prognosis, this chart emphasizes how serious threats cancers are nowadays and a fact that, as a counterbalance, many stages remain local only does not improve situation by much. Extended warehousing analysis using SEER Stat

Following suggestions from opponents raised at the thesis defense, a research part of the document has been extended, among others, mainly minor improvement. The extension is focused on exploiting more of the SEER Stat software analysis features as well as applying data mining algorithms on the data extracted from SEER database (which its software itself does not offer as it is a data warehouse only). First part of the extension is related to key functionality of the SEER Stat previously not used, namely providing various survival probabilities for specified diseases and populations and other, additional conditions. The second part makes use of case listing feature in the mentioned software, which allows to browse and copy detailed cases data with any attributes available, in contrary to generalized, summarized and grouped calculations that founded data warehousing analysis. Case data were input for data mining algorithms to search patterns and regularities among them which may not have been expected to be there.

It is worth noting that, after using SEER Stat for the frequencies and rates analysis (as in the previous sub-chapter) and for survival probabilities analysis (current chapter), most important values of this tool have been applied. Naturally, further extensions of analysis still exist, these though are rather more detailed and precise follow-ups of attempts present in the thesis and are not completely different kind of approaches to the data. The last remaining key feature of SEER Stat is the analysis of prevalence of diseases, which hasn't been incorporated in the document. Reason for this is because prevalence is somewhat similar to rates and frequencies analysis, only focused on entire number of patient suffering for specific diseases at given time, not only on new occurrences.

Therefore, it is of particular value for staff responsible for management of medical treatment financing as it provided view of entire numbers describing precedence of specific diseases.

Most of the charts in this chapter based on time series have been modified to represent the data as continuous functions although in fact there are discrete. It has been done so to better (in author's opinion) represent changes in the measured values over time. Standard errors were also available, however, since these virtually never exceeded 5%, were omitted when plotting result for the sake of better and more comprehensive visualization.

3.2. Observed 5-year survival of all site's cancers by stage, age year of diagnosis in whites

Observed 5-year survival rates in white males by cancer (all types) stage, age at diagnosis and year of diagnosis

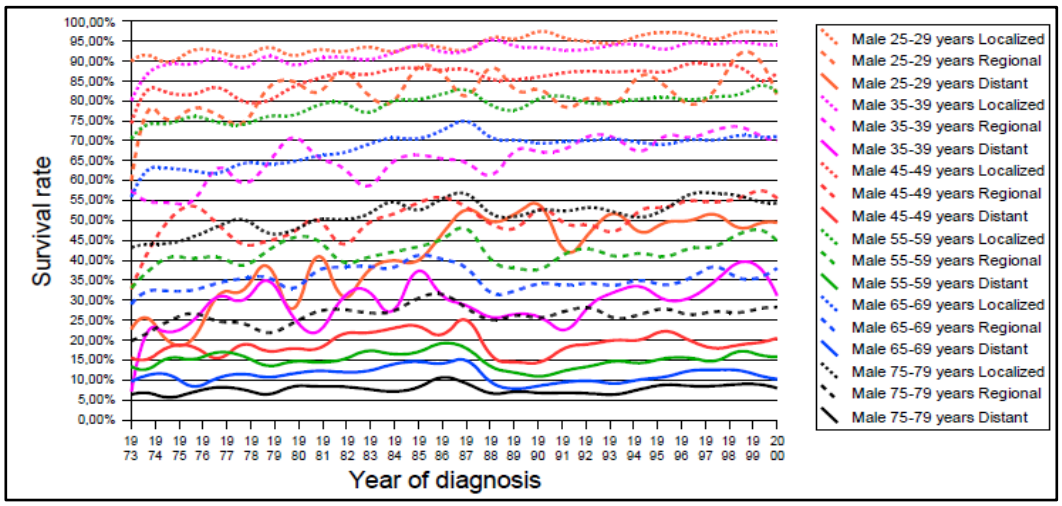

Fig 4: 5-y survival of all-sites cancers in males over time

Observed 5-year survival rates in white females by cancer (all types) stage, age at diagnosis and year of diagnosis

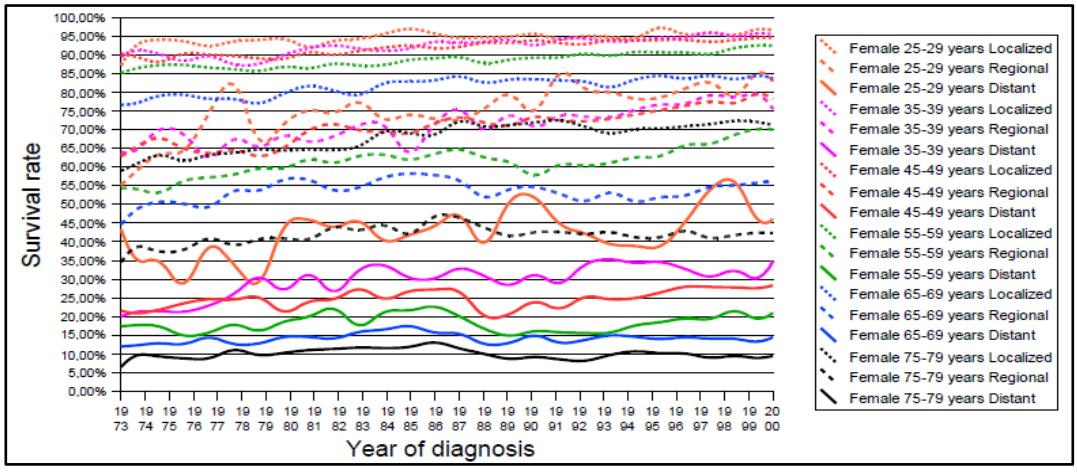

Fig 5: 5-y survival of all-sites cancers in females over time

3.3. Data Analysis of 5-years survival of all sited cancers in females over time

- These charts are considered as one of the most surprising as well as significant across the document since it reveals that, although more than thirty years passed since the point in time its time-scale begins, in general, in most cases no improvement in 5-year survival of diagnosed cancer diseases has been observed. The survival rates stay flat which is not what one could expect observing developments in other aspects of daily life. It's impossible in the scope of the thesis to predict reasons staying behind such a constant (and very low, particularly in case of cancers that metastasized) survival in so many years. Only speculative causes can be proposed like cancers evolving to become more aggressive and invasive, more resistant to drugs in use (possibly due to increased environment pollution or widely adopted unhealthy lifestyle which diminish any improvement in medical practices). It is noticeable that the chart is generally independent from factors like cancers incidence of how early these have been diagnosed. It's also worth noting that the charts sum all sites cancers in general which means that it does

not deny improvement in treating particular (the ones most frequent for instance) diseases. This is investigated later on when focus is moved on most prevalent types of cancers and improvement in their treatment over time.

- Females generally tend to have better prognosis than males and the difference varies from 5% to even 20% in particular age and stage compilations. In case of cancers that metastasized distantly, this subtle difference may sometimes mean the survival is twice higher, although still very poor. This observation is present in many publications [17][18][9] concerning more specific cancer sites. It turns out the pattern spreads to generalized all-sites view as well.
- Especially noticeable is immense difference in observed survival depending on the stage the disease was diagnosed at and this fact is in charge for all ages at any time. The chances of survival differ so much they mostly do not overlap for three general stages of cancer and this happens despite generally poorer prognosis for older patients.
- Noticeable are the periodical variations on survival percentage. While an exact reason for this cannot be provided, one could suggest these may depend on variety of factors like economic situation of the nation over time or environment related aspects.

3.4. Recent changes in curability of most impacting types of cancers

As previous experiment proved, overall curability of cancers has improved very little and, in fact, these improvements are related to small group of stages and ages when diagnosed with the disease. While this is a very grim finding in general, more detailed view of how medical practices perform in specific types of cancer is considered to be necessary, mainly because of the findings described in previous chapters. It turned out that impact on society (measured as an estimated, observed number of patients that succumbed to the disease) different types of cancerous diseases result in completely different loss of life. This is due to the product of varying incidence of them and different prognosis as well as diagnosing them at different stages. Therefore, a more detailed observation has been carried through, taking into account only these cancers that proved to be most devastating. As in other parts of the research, relative 5-year relative survival has been chosen as a measure of efficiency of treatments applied to cancer patients. This measure, due to its characteristics [13], has been considered as most informative since it compares the survival of cancer patients to the survival of the expected survival among cancer-free individuals. Periods chosen were 5-year periods starting from 2020 and 2021 (the newest available) and patients who were diagnosed with selected cancers in 2019- 2020 and 2018-2019 were observed. Observations of results allow to spot minor improvement of treatment efficiency, however still examples of worse survival nowadays than 10 years ago can be seen. Overall, the best gain exists in case of stages of cancer when it has spread to local lymph nodes. The most significant increase in survival is present in case of lymphoma where it varies from 15% to 20%, while in other areas of improvement the gain is limited to not more than 10% mostly.

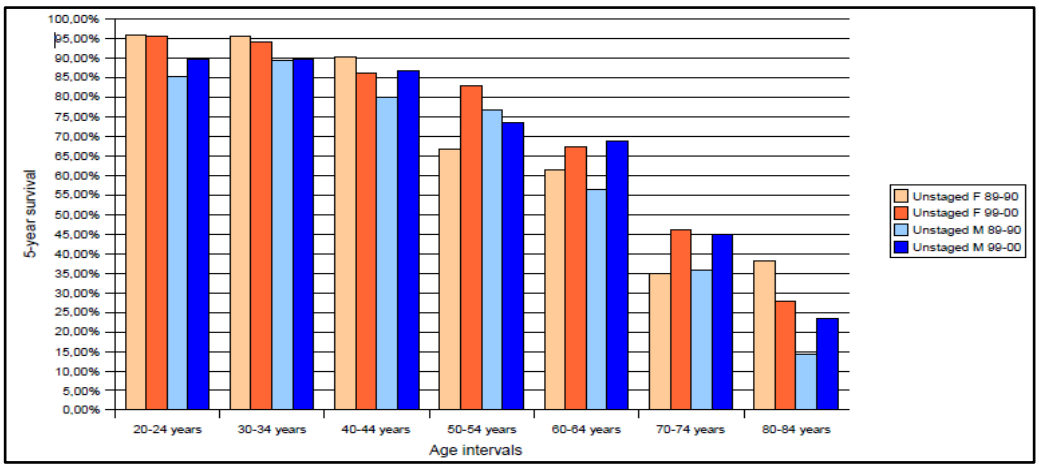

Fig 6: Comparison of 5-year survival of non- hodgkin lymphoma cancers in whites diagnosed in 89-90 and 99-00 by sex, stage and age

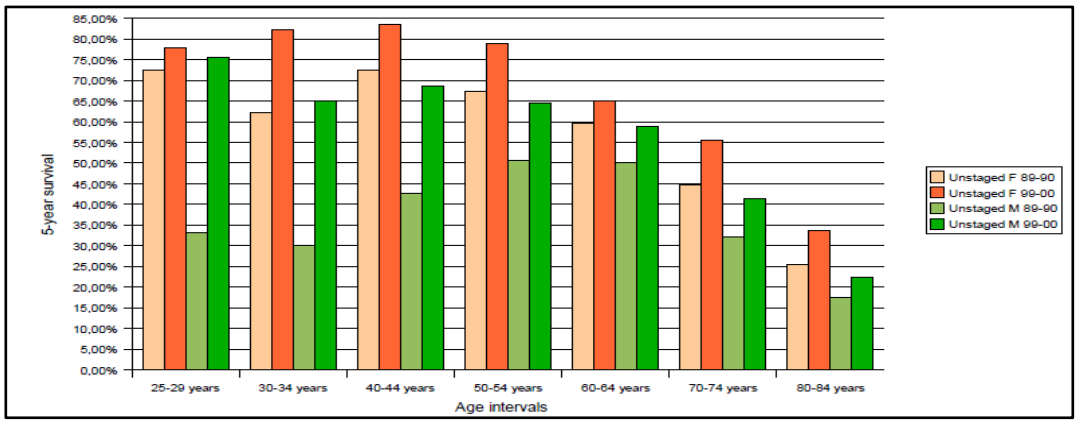

Fig 7: 5-year survival of non-hodkin lymphoma cancer comparison

Comparison of 5-year survival of colon cancers in whites diagnosed in 89-90 and 99-00 by sex, stage and age

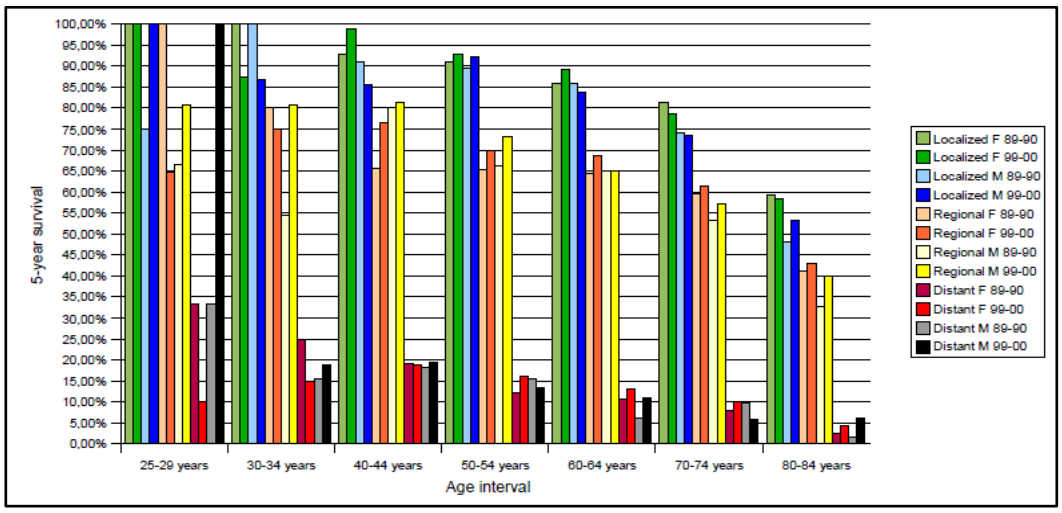

Fig 8: 5-year survival of colon cancer comparison

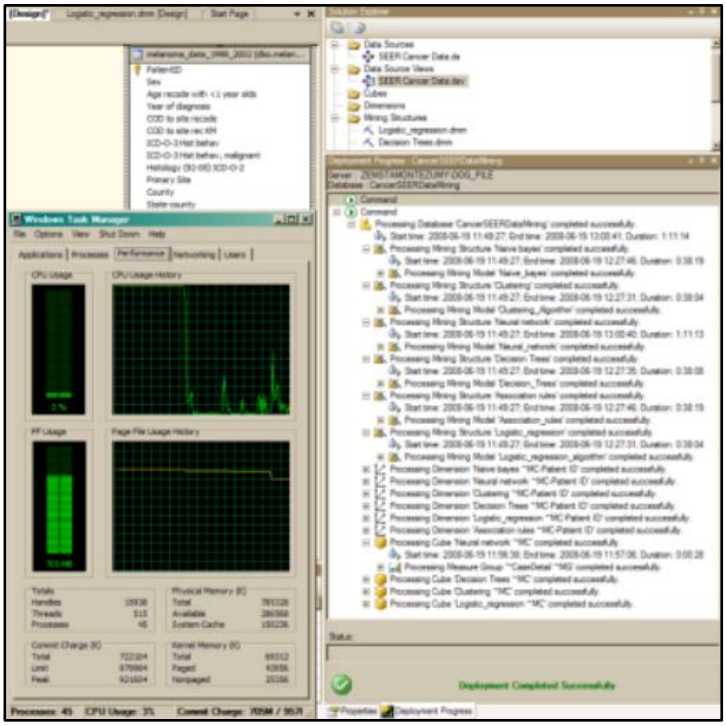

Fig 9: Processing of 40 000 records by data mining algorithms.

As an aspect to be predicted from all attributes of cases, a patient's disease result (either alive or deceased due to the illness, which may be the cancer patient was diagnosed with or any other disease) was chosen. This was because the final result has been considered as a most valuable measure whether treatment succeeds or not. The data records count was about 40 000.

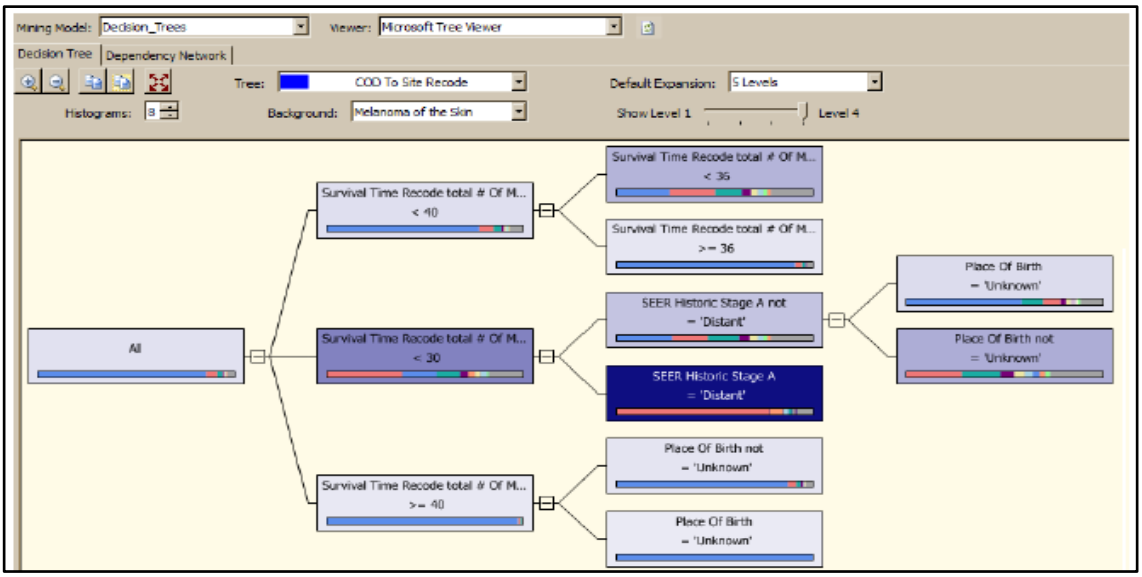

Fig 10: Poorly grown decision tree with too high inhibiting factor, dominated by strongest attributes

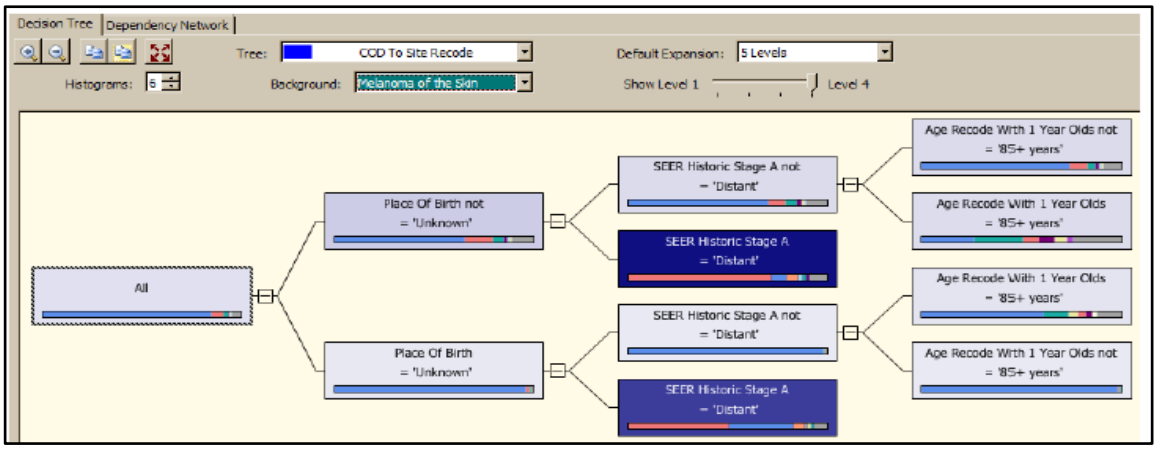

Fig 11: Poorly grown decision tree with too high inhibiting factor, dominated by strongest attributes

Each parameter in the record denoting decision tree which have different sub nodes which represents survival time record which means the recovery time for patience and historical analysis by place of birth and age records . First attempt resulted in malfunctioned tree being created by the algorithm, which was due to standard growth-inhibiting parameter being chosen. This caused most influencing attributes (which were also the most obvious as such) to outmatch any other as tree nodes. After minimizing the inhibiting factor, more detailed tree was gained, which was then pruned as, from certain node, it didn't contain any valuable differentiation of attributes. The tree is depicted below, being separated among three figures with dots of the same color marking cutting point.

Table 2: Comparison of SEER Stat and caCORE

System aspect	**SEER Stat**	**caCORE networked integrated system**
Type of data source/data model and its content	A set of fixed-structure databases of cancer diseases cases (which may be seen as data warehouse facts), available on-line, as a download for off-line use and as flat files for custom	An extensible global ontology covering as many aspects of oncology knowledge, not only the structure of cancer diseases themselves connecting an extensible number of data sources

	processing. The database structure cannot be changed and users cannot contribute to it by extending models and supplying new data sources. Any operation like this would require very intense cooperation with NCI system responsibles. The content is statistical data about cancer diseases, very rich in its form, but does not cover other oncology concepts, like drugs in use, extensive therapy descriptions, symptom mappings etc.	that supply the system with real data having to comply with data structures it enforces (by providing the mapping between their structure and following the system structure as much as possible). Every research group can expand the system by following a well-described integration procedure. The caCORE database can then be considered as a peer-to-peer network of standalone database engines.
Analytical methods in use	Data warehousing methods already supplied with the system, like multidimensional analysis, aggregations, drill-through views, calculated columns, trends, statistics, probabilities and domain-specific calculations. No data mining functionality like decision trees, neural networks, classification and clustering etc. algorithms are present. No visualization is present (web front-end visualizations exploit little system potential. However, exporting the data for visualization or further analysis proved to be rather easy (time consuming though).	At the moment, no analytical tools are to be found in the system, which, being still under development, serves only small partition of data it could possibly make accessible. Planned are analysis services which can be exposed through the system just like data sources are, using common interfaces. This could lead to implementing any type of analysis (DW, DM, custom) in the system, at the moment though it's in the too early stage, the documentation on how to deploy analytical services in caCORE also is poor at the moment (unlike the integration documentation).
Qualifications required to use the system/easiness of use	The system is very easy to use and little experience in database-alike application and Excel pivot tables is fully sufficient to make a lot of use of it. Even statistical adjustments are presented in comprehensive manner. All in all, very little IT qualifications are necessary to handle the SEER Stat and related tools. Availability of online mode makes usage even easier.	A differentiation needs to be made at this point. Browsing the caCORE dictionaries and CDEs (common data elements, repositories of ontology entities) is quite easy due to availability of intuitive, browser-based tools. Consuming the data is more difficult, but performing and integration process can be considered very hard, time-consuming, requiring professional skills in both oncology as well as IT, thus enforcing teams of researchers from both disciplines.
How fast does it take to start using the system	The system is ready to use in terms of both finished installation as well as easiness to operate. Naturally, referring to the documentation is necessary to perform more refined analysis, but the	Just understanding of the system concept (which is by far more complicated than SEER Stat) takes a lot of time for research papers skimming (if not detailed reading). navigating through the web—enabled structures is straightforward but using

	manuals are also very straightforward.	the data becomes difficult and certainly cannot be compared with SEER Stats quick start (a matter of minutes, not more than an hour compared to a couple of days). The integration process would certainly take more than week (in fact it could take much more than a week) if performed by inexperienced staff of both IT and oncology specialists.
What is the system development stage	Next in a row stable release	The system core is stable and operating (second stable release) but since the P2P based system needs to be fed with peer content, it can take very long until it becomes fully functional.
Potential of the system	Very high but limited since data source is not extensible (at least not as extensible as caCORE)	The limit is only set by the advancement of ontology and coverage of it by peer data sources and exposed data services.
Documentation/user manual quality	Very good, content-rich and straightforward, one uses it quickly and efficiently.	Good only in some aspects and large in terms of volume. Getting system bird view at the beginning is difficult (due to its complexity and)

3.5. Numerical comparison of algorithms

In order to algorithmically compare predictive efficiency of selected algorithms, a second dataset was created describing cancer patients that were diagnosed a year in advance before patients described in previous, training and analysis set. This time, the count of records was about 9000. Additionally, to the algorithms described, two others, which were not used for analysis (since they didn't discover data characteristics considered significant), were also evaluated. Evaluation method used was the lift chart generator, built into the tool used for the analysis in general, which visualizes how well the created model describes data, therefore how accurate it is.

4. Conclusions

Evaluation of data mining techniques in analysis of cancer data resulted in a conclusion that while neural networks and Bayesian algorithm proved more efficient in cross-valuation of predictive power, observing models that were created by worse-performing algorithms also allowed extraction of useful and valuable patterns that weren't present in the output of best performers. This suggests as many algorithms as possible should be used for analysis since, summarily, such pack can provide better results than picking most efficient ones only. Evaluation of how future cancer data systems for analytical purposes could be build was carried by extensive usage of SEERStat, currently possibly most powerful but enclosed system and an experiment conducted over the caCORE, a system currently in development. The idea of it is to make a common data model (called ontology) available and allow interested parties to integrate with the system by means of publishing their own data, provided they are rendered compatible with the model. This compatibility is to be achieved by following an integration procedure, including data transformation and exposition. An experiment included in the thesis was basically a simulation of such integration.

This research concluded that systems with an ambition of being fully extensible by adding more and more peers supplying their own data (like for instance medical institutes from any part of the world) are very difficult to integrate with. The procedure is complicated, requires in depth knowledge of huge ontology presented by caCORE, takes a lot of effort and time and is virtually infeasible without strict cooperation of oncology and IT professionals. Despite that, a potential of such system, when fully functional, is superior over an enclosed

system like SEERStat. One has to remember though that while caCORE remains dysfunctional as for now, SEERStat is at its full steam, very robust and reliable. And to be precise, it also provides some means of adding data to it, however this requires a careful analysis of SEERStat data format and serious programming knowledge. In order then to address problems of future cancer data systems, a suggestion was made these should be basically designed like caCORE but with as much encapsulation known from object-oriented programming and as much usage of simple web services interfaces as possible. This would make integration procedures simpler and therefore tempting for researchers which are willing to participate in creating possibly the richest sources of cancer data. Additionally, scope of the caCORE-alike system could be slightly limited which would result in making its data model simpler and therefore more comprehensive. Provided these conditions are met, a system like caCORE could succeed as much as other "social network" projects did, being driven by the power of community.

Reference:

[1] Ries Lag, Melbert D et all. SEER Cancer Statistics Review, 1975-2004, National Cancer Institute, partially summarized on http://seer.cancer.gov/statfacts/html/melan.html

[2] Bibault, J.E., Giraud, P., Housset, M., Durdux, C., Taieb, J., Berger, A., Coriat, R., Chaussade, S., Dousset, B., Nordlinger, B. and Burgun, A 2018, 'Deep learning and radiomics predict complete response after neo-adjuvant chemoradiation for locally advanced rectal cancer', Scientific reports, vol. 8 no. 1, pp.1-8.

[3] Blanes-Vidal, V., Baatrup, G. and Nadimi, E.S 2019, 'Addressing priority challenges in the detection and assessment of colorectal polyps from capsule endoscopy and colonoscopy in colorectal cancer screening using machine learning'. Acta Oncologica, vol. 58 sup1, pp.S29-S36.

[4] Borkowski, A.A., Wilson, C.P., Borkowski, S.A., Thomas, L.B., Deland, L.A. and Mastorides, S.M 2018, 'Apple machine learning algorithms successfully detect colon cancer but fail to predict KRAS mutation status'. arXiv preprint arXiv, vol. 1812.04660.

[5] Bychkov, D., Linder, N., Turkki, R., Nordling, S., Kovanen, P.E., Verrill, C., Walliander, M., Lundin, M., Haglund, C. and Lundin, J 2018, 'Deep learning based tissue analysis predicts outcome in colorectal cancer', Scientific reports, vol. 8, no. 1, pp.1-11.

[6] Bychkov, D., Turkki, R., Haglund, C., Linder, N. and Lundin, J 2016, 'Deep learning for tissue microarray image-based outcome prediction in patients with colorectal cancer', International Society for Optics and Photonics, vol. 9791, p. 979115.

[7] Cârțână, E.T., Gheonea, D.I. and Săftoiu, A 2016. 'Advances in endoscopic ultrasound imaging of colorectal diseases', World journal of gastroenterology, vol. 22, no. 5, p.1756.

[8] Caruso, S., Bazan, V., Rolfo, C., Insalaco, L., Fanale, D., Bronte, G., Corsini, L.R., Rizzo, S., Cicero, G. and Russo, A 2012. 'MicroRNAs in colorectal cancer stem cells: new regulators of cancer stemness?'. Oncogenesis, vol. 1, no. 11, pp.e32-e32.

[9] Chen, L.D., Liang, J.Y., Wu, H., Wang, Z., Li, S.R., Li, W., Zhang, X.H., Chen, J.H., Ye, J.N., Li, X. and Xie, X.Y 2018. 'Multiparametric radiomics improve prediction of lymph node metastasis of rectal cancer compared with conventional radiomics', Life sciences, vol. 208, pp.55-63.

[10] Cho, S.B. and Won, H.H 2003, 'Machine learning in DNA microarray analysis for cancer classification', In Proceedings of the First Asia-Pacific Bioinformatics Conference on Bioinformatics 2003, vol. 19, pp. 189-198.

[11] Choi, Y.R., Kim, J.H., Park, S.J., Hur, B.Y. and Han, J.K 2017. 'Therapeutic response assessment using 3D ultrasound for hepatic metastasis from colorectal cancer: application of a personalized', 3D-printed tumor model using CT images, Plos one, vol. 12, no 8, p.e0182596.

[12] CACORE SOFTWARE DEVELOPMENT KIT 3.2.1 Programmer's Guide, U.S. Government work, Revised July 16, 2007

[13] L. Breiman, J. Friedman, and C. J. Stone, and R. A. Olshen, "Classification and regression trees," Routledge, 1st ed., 2017.

[14] T. Chen, and C. Guestrin, "Xgboost: A scalable tree boosting system," in Conf. Proc. ACM SIGKDD knowledge discovery and data mining, pp. 785-794, Mar. 2016.

[15] G. A. Rao, G. Srinivas, K. V. Rao, and P. P. Reddy, "Characteristic mining of Mathematical Formulas from Document-A Comparative Study on Sequence Matcher and Levenshtein Distance procedure", J. Comp. Sci. Eng., vol. 6, no. 4, pp. 400-404, Apr. 2018.

[16] D. Morina and A. Navarro. The R package survsim for the simulation of simple and complex survival data. Journal of Statistical Software, 59(2):1–20, 2014.

[17] Y. Li, K. S. Xu, and C. K. Reddy. Regularized parametric regression for high-dimensional survival analysis. In Proceedings of SIAM International Conference on Data Mining, pages 765–773, 2016.

[18] A. Ezzat, M. Wu, X. L. Li, and C. K. Kwoh, "Computational prediction of drug-target interactions using chemogenomic approaches: an empirical survey," Briefings in Bioinformatics, no. 8, 2018.

[19] Z. Shen, Y.-H. Zhang, K. Han, A. K. Nandi, B. Honig, and D.-S. Huang, "miRNA-Disease Association Prediction with Collaborative Matrix Factorization," Complexity, vol. 2017, pp. 9, 2017

[20] M.-M. Gao, Z. Cui, Y.-L. Gao, J.-X. Liu, and C.-H. Zheng, "Dual-network sparse graph regularized matrix factorization for predicting miRNA–disease associations," Molecular Omics, 2019.

[21] Junyi Gao, Cao Xiao, Lucas M. Glass, Jimeng Sun, "COMPOSE: Cross-Modal Pseudo-Siamese Network for Patient Trial Matching," ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp.803 – 812, 2020.

[22] Chi Yuan, Patrick B Ryan, Casey Ta, Yixuan Guo, Ziran Li, Jill Hardin, Rupa Makadia, Peng Jin, Ning Shang, Tian Kang, et al. 2019. Criteria2Query: a natural language interface to clinical databases for cohort definition. Journal of the American Medical Informatics Association 26, 4 (2019), 294–305

[23] Chunhua Weng, Samson W Tu, Ida Sim, and Rachel Richesson. 2010. Formal representation of eligibility criteria: a literature review. Journal of biomedical informatics 43, 3 (2010), 451–467.

[24] David Soutar, Gerry Robertson, "CANCER SCENARIOS: An aid to planning cancer services in Scotland in the next decade", Canniesburn Hospital, Beatson Oncology Centre, Glasgow.

[25] Nicholas Huba, Yan Zhang, "Designing Patient-Centered Personal Health Records (PHRs): Health Care Professionals' Perspective on Patient-Generated Data," Journal of Medical Systems, vol 36, issue 6, dec 2012.

[26] Xingyao Zhang, Cao Xiao, Lucas M. Glass, Jimeng Sun, " DeepEnroll: Patient-Trial Matching with Deep Embedding and Entailment Prediction", WWW '20: Proceedings of The Web Conference 2020, pp 1029 – 1037, 2020.

[27] Antonio Boffa, Paolo Ferragina, Giorgio Vinciguerra, "A Learned Approach to Design Compressed Rank/Select Data Structures", ACM Transactions on Algorithms (TALG), article accepted, mar 2022

[28] Fears et al, Cancer Res. 2002

[29] Ries Lag, et all, "SEER Cancer Statistics Review 1975-2000", NCI; 2003

[30] Besusparis, J., Laurinavicius, A. and Ilyas, M 2019, 'A cascade-learning approach for automated segmentation of tumour epithelium in colorectal cancer', Expert Systems with Applications, vol.118, pp.539-552.

[31] Altini, N., Marvulli, T.M., Caputo, M., Mattioli, E., Prencipe, B., Cascarano, G.D., Brunetti, A., Tommasi, S., Bevilacqua, V., Summa, S.D. and Zito, F.A 2021, 'Multi-class Tissue Classification in Colorectal Cancer with Handcrafted and Deep Features', In International Conference on Intelligent Computing, Springer, pp. 512-525).

[32] Babu, T., Singh, T., Gupta, D. and Hameed, S 2021, 'Colon cancer prediction on histological images using deep learning features and Bayesian optimized SVM', Journal of Intelligent & Fuzzy Systems, Preprint, pp.1-12.

[33] Berger-Kulemann, V., Schima, W., Baroud, S., Koelblinger, C., Kaczirek, K., Gruenberger, T., Schindl, M., Maresch, J., Weber, M. and Ba-Ssalamah, A 2012. 'Gadoxetic acid-enhanced 3.0 T MR imaging versus multidetector-row CT in the detection of colorectal metastases in fatty liver using intraoperative ultrasound and histopathology as a standard of reference', European Journal of Surgical Oncology EJSO, vol. 38 no.8, pp.670-676.

[34] Coenegrachts, K., Bols, A., Haspeslagh, M. and Rigauts, H 2012, 'Prediction and monitoring of treatment effect using T1-weighted dynamic contrast-enhanced magnetic resonance imaging in colorectal liver metastases: potential of whole tumour ROI and selective ROI analysis', European journal of radiology, vol. 81, no. 12, pp.3870-3876.

[35] Dawson, I.M.P., Cornes, J.S. and Morson, B.C 1961, 'Primary malignant lymphoid tumours of the intestinal tract', Report of 37 cases with a study of factors influencing prognosis, British Journal of Surgery, vol. 49, no. 213, pp.80-89.

[36] de Wit, M., Kant, H., Piersma, S.R., Pham, T.V., Mongera, S., van Berkel, M.P., Boven, E., Pontén, F., Meijer, G.A., Jimenez, C.R. and Fijneman, R.J 2014, 'Colorectal cancer candidate biomarkers

identified by tissue secretome proteome profiling', Journal of proteomics, vol. 99, pp.26-39.

[37] Ding, L., Liu, G., Zhang, X., Liu, S., Li, S., Zhang, Z., Guo, Y. and Lu, Y 2020. 'A deep learning nomogram kit for predicting metastatic lymph nodes in rectal cancer', Cancer Medicine, vol. 9, no. 23, pp.8809-8820.

[38] Fan, N.J., Kang, R., Ge, X.Y., Li, M., Liu, Y., Chen, H.M. and Gao, C.F 2014, 'Identification alpha-2-HS-glycoprotein precursor and tubulin beta chain as serology diagnosis biomarker of colorectal cancer', Diagnostic pathology, vol. 9, no. 1, pp.1-11.

[39] Fan, X.J., Wan, X.B., Huang, Y., Cai, H.M., Fu, X.H., Yang, Z.L., Chen, D.K., Song, S.X., Wu, P.H., Liu, Q. and Wang, L 2012. 'Epithelial–mesenchymal transition biomarkers and support vector machine guided model in preoperatively predicting regional lymph node metastasis for rectal cancer', British journal of cancer, vol. 106, no. 11, pp.1735-1741.