Editorial

European Society of Thoracic Surgeons big data utilization—part 1: research interest for the thoracic community

Michele Salati

Division of Thoracic Surgery, United University Hospitals of Ancona, Ancona, Italy

Correspondence to: Michele Salati, MD, PhD, Division of Thoracic Surgery, United University Hospitals of Ancona, Via Conca 71, 60126 Ancona, Italy. Email: michelesalati@hotmail.com.

Submitted Jan 13, 2018. Accepted for publication Apr 11, 2018.
doi: 10.21037/jtd.2018.04.103
View this article at: http://dx.doi.org/10.21037/jtd.2018.04.103

Background

The knowledge is delivered at different levels using several processes. In fact we can obtain information and learning through different activities: experiencing, conceptualizing, analyzing, and applying.

In medical science the process of knowledge follows a well-defined pathway, that often starts with operations of data collection. At the end of nineties, Fayyad, Piatetsky-Shapiro and Smyth theorized and described the steps of this process (1). Once obtained a consistent collection of data, it is possible to apply algorithms through the data mining and analytic phases in order to derive knowledge from them, as shown in *Figure 1*. This process defined "Knowledge Discovery in Databases" should deliver information with specific attributes: valid, new, useful and understandable.

Using this perspective, it seems clear that knowledge can be heavily influenced by the characteristics of the analyzed database, both for its architecture and for its content.

Databases could store different types of data: structured (represented by elementary attributes for an item), semi-structured (represented by schematically collected values with a certain degree of flexibility), and unstructured (represented by informations coded in natural language). Data could also present different temporal dimensions: stable data have a very low probability of changing and usually maintain valuable information over the time, data with long term variability are characterized by a low but constant rate of change and need of updating to deliver proper informations (i.e., follow up data), and data with high variability present a high rate of change that imposes strategies for real time, constant interval or irregular updating. Moreover data-flows within a database could be extremely different, varying from monolithic systems where data are repetitively uploaded using standardized methods and structured platforms, to peer to peer informative systems where data could be uploaded with the highest level of freedom without any form of central control. Finally, databases could describe the "real world" they want to represent with different degree of reliability. This could be formally assessed using standardized data quality dimension and processes of data cleaning. A rigorous knowledge extraction from databases should always be preceded by these phases of data validation.

All the abovementioned features of a database could influence the procedures of analysis, research and inference founded on it.

Characteristics of the European Society of Thoracic Surgeons (ESTS) database and its use for research

With the aim of improving quality in thoracic surgery, the ESTS created the ESTS database, that was used as the cornerstone for discovering knowledge about the thoracic surgery field in Europe.

This registry underwent several changes and refining over the years. Nowadays it presents the following general characteristics:

(I) It is an on-line database with a dedicated webpage hosted within the ESTS web site. The access is allowed through specific user identifier and password to any thoracic surgery unit having a staff surgeon member of the ESTS. Contribution to ESTS database is voluntary, even though it offers

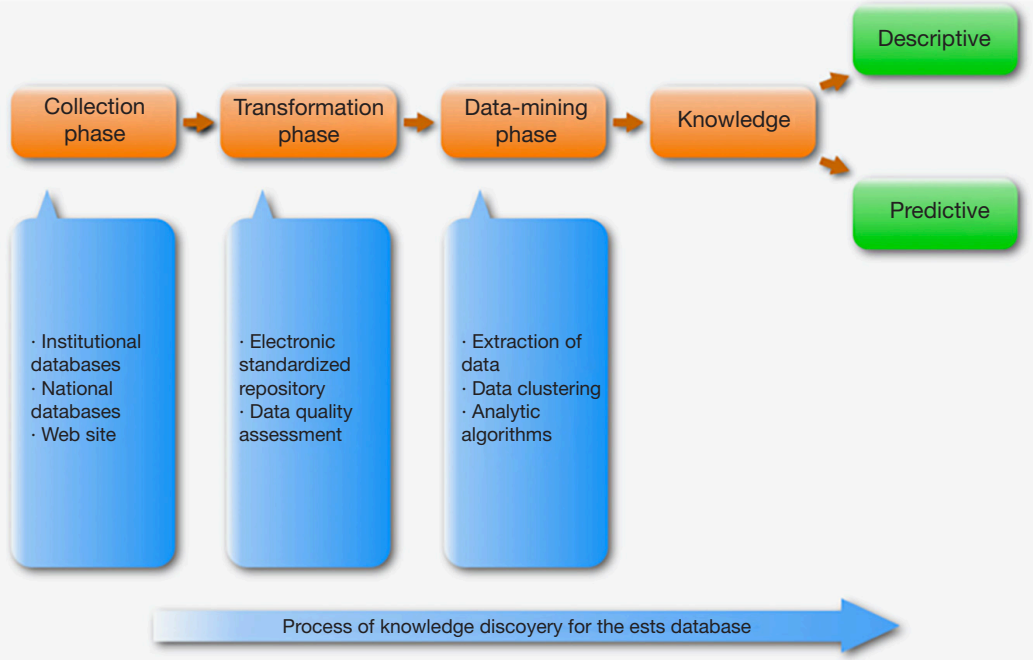

Figure 1 The process of knowledge discovery in databases. From Salati M. Reasons to participate in European Society of Thoracic Surgeons database. J Thorac Dis 2015;7(S2):S112-S117, with permission.

several benefits to the participant Units as stated in the ESTS Annual Report (http://www.ests.org/collaboration/database_ reports.aspx);

(II) Within the on-line platform, five different sections can be accessed: the lung core section (collecting data about lung resections), the satellite thymic module, the satellite NET (neuro-endocrine tumors) module, the satellite mesothelioma module and the satellite chest wall module. The implementation of these satellite modules was considered strategic by ESTS in order to systematically acquire information and to deliver general knowledge as well as clinical benchmarks about less frequent (in comparison to lung resection) procedures in thoracic surgery. Each section presents structured pages where the fields/items of interest can be filled with the help of legends and pop-ups;

(III) The ESTS database collects more than one thousand variables (lung: 155 variables, thymic: 232 variables, NET: 205 variables, mesothelioma: 276, chest wall: 270 variables). Data are reported in a structured format in order to be ready for the analytic phase as much as possible without further data cleaning processes. Often the attributes of an item can be selected from a multiple choice list, in order to minimize variability of data collection and syntactic errors (*Figure 2*);

(IV) The collection of data can follow two different flows. In the first, data can be uploaded using the on-line platform. This provides a prospective imputation of informations within the ESTS database through repetitive and standardized procedures. In the second, large amount of data can be transferred en-block from a pre-existent repository (such as National Registries) to the ESTS database on a regular basis (usually once a year), after a procedure of compatibility verification among them;

(V) The vast majority of collected data are stable. The follow up section that could increase the variability rate of information is still affected be data under-report.

Using the information collected within the ESTS database, it was possible to carry out several study projects during the last ten years.

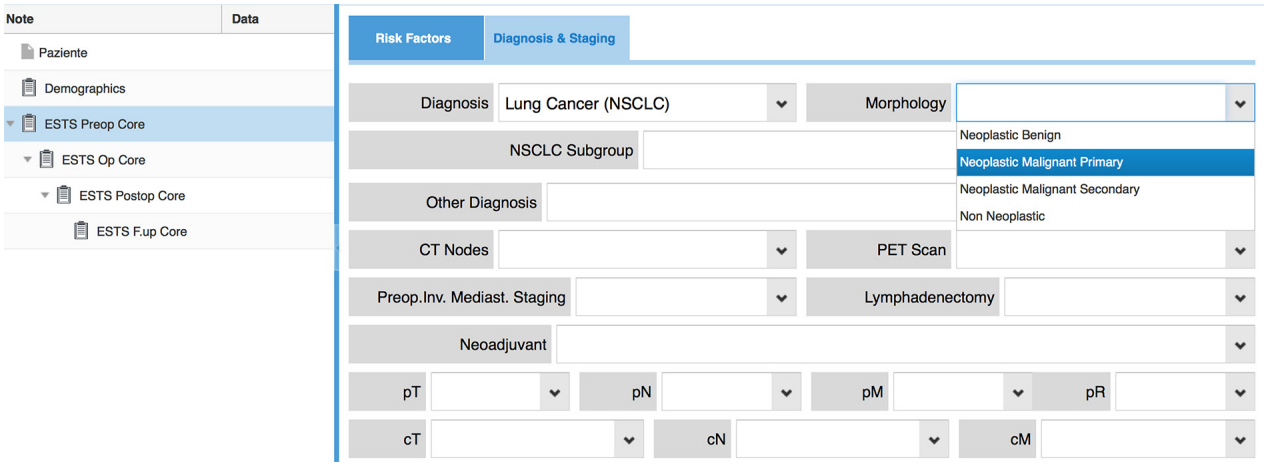

Figure 2 ESTS core section screenshot. ESTS, European Society of Thoracic Surgeons.

Some of them represent a contribution for the assessment and improvement of the quality of care in thoracic surgery. The first studies published using the ESTS database were exactly related to the definition of a risk score for patients submitted to lung resection (the European Society Objective Score) (2,3) and to the development of methodologies for evaluating the performance of thoracic surgery units taking into account the impact of preoperative, intraoperative and postoperative processes of care in a unique quality indicator as the Composite Performance Score (CPS) (4,5). More recently other papers addressed the same research topic, analyzing outcomes and risk factors for complications after video assisted lobectomy, a procedure that is increasingly widespread as reported within the ESTS database (6,7). Moreover, at the beginning of 2017, two new and updated risk scores for predicting morbidity and mortality after anatomic lung resection, using almost 48 thousand patients collected in the database until August 2015, were developed, offering a pivotal tool for monitoring and improving the quality of care in the European thoracic surgery scenario (8).

The ESTS database with its satellite modules gave also the chance to perform research about rare diseases that, due to the multi-institutional nature of this registry, were sufficiently represented within this database. This allowed to publish evidences inherent tumors of the thymus for identifying prognostic factors after surgical resection (analysis on 2,151 thymic tumors) (9) and for clarifying the role of adjuvant treatment in locally advanced thymomas (analysis on 370 Masaoka stage III thymomas) (10). In the same way, it was possible to perform analysis about uncommon lung neoplasm such as the neuroendocrine tumors (11) and the carcinoid tumors (12).

Moreover, the ESTS database represented the perfect testing ground to explore unusual research subject in medical literature, such as the quality of data management. Using the large amount of data stored within the registry, it was possible to develop indicators and methodologies for data quality assessment (13). These procedures are now available for validating the quality and reliability of data before the analytic phase in any study performed using the ESTS database.

This registry was also used to show general characteristics and results of thoracic surgery practice in Europe (14). In this regard, the ESTS Database Annual Report represents an authoritative reference published every year for benchmarking and comparing activity among European Units (http://www.ests.org/collaboration/database_reports.aspx).

These studies here reported represent just a non-exhaustive list of papers and research lines directly obtained from the exploration of data gathered within the ESTS database, and some others could be cited. Moreover, as a relapse of this body of knowledge disseminated, other investigations will be encouraged and supported in the next future. But, even though the ESTS database has proven to be an excellent instrument to perform research and increase the level of understanding in the thoracic surgery field, some limitations and pitfalls should be always taken into consideration in the process of extracting information from this big data registry.

The issue of big amount of data: start thinking about data mining procedures

The collection of data within the ESTS database, after an initial period of adjustment and refinement, is proceeding with a stable growth. Ten to twelve thousand new procedures are basically uploaded each year, leading to a linear increase of data available for analytic processes. For sure this data growth is not yet comparable to the one usually observed for "big data". At the same time information gathered within the ESTS database doesn't exactly have characteristics identifying "big data": high number, velocity and variety.

Nevertheless, considering the present status of this registry, the procedures for data cleaning and analysis are becoming more and more challenging due to the amount of cases stored. The methodologies for assessing the quality and the reliability of the entire dataset as well as of subset of data are more complex and time demanding. The analytic phase, even for the simpler models, requires greater effort and technical support in order to be solid and reproducible.

Due to the aspects above mentioned, in the near future the ESTS Database Committee should consider of implementing procedures of data mining (15) in order to extract valuable information from the large amount of procedures collected within the database. The creation of analytic models automatic and replicable over time could offer the chance of glimpsing patterns of knowledge from this large dataset. As a consequence, the obtained preliminary results will be able to discover characteristics and association of data for leading proper formal analysis and extracting evidence-based science.

Considering this point of view, some methodologies, previously defined to perform studies using the ESTS database (as mentioned before about the CPS), are nowadays integrated within the online platform of the registry providing real time information to the end-users (this is exactly a form of data mining). In this way it is possible to show updated details about procedures performed and obtained results for each Unit contributing to the database as well as to instantly calculate its own composite performance score. This metric of quality of care (that will be extensively described in the next chapter) is derived using several indicators calculated following a rigorous analytic model. The model is currently embedded within the ESTS database software platform, that displays through a specific dashboard the level of the unit composite performance score in comparison to the mean of the other contributors of the registry. This information could be useful to identify potential lack in quality of care delivered by one or a group of institutions and implement formal studies to verify this hypothesis.

In 2017, the ESTS Database Committee published an official report about patterns of care and outcomes of surgery for malignant lung neoplasm analyzing more than 62 thousand patients collected within the ESTS database in ten years (14). In this study the results were split in three sections reporting findings for different cohorts of patients. In the first section (62,774 patients) were described the general baseline characteristics of patients gathered within the European database, the type of lung resection they were submitted to and the preferred surgical approach used. The second section (51,931 patients) reported information about the stage of disease of the subgroup of patients affected by primary lung cancer and the trends of management of potential lymph node metastatic disease across Europe. Finally, in the third section (51,756 patients), were presented raw preoperative outcomes for the entire cohort of patients submitted to anatomic lung resections, and in specific subgroups of potential high-risk patients. This overview of the clinical practice in Europe, obtained using more than 170 thoracic surgery Units contributing to the ESTS database, defines benchmarks of activity and standard of reference for the future. As each single step of analysis is clearly reported within the methods paragraph, some of them could be extracted and adjusted in order to obtain a systematic and automatic overview of the thoracic surgery activity described from the ESTS database. This analytic inference could represent in the next future a fundamental data mining procedure to compare and understand the trends of activity in Europe and especially to plan and lead specific rigorous analysis to interpret them.

The problem of generalization in multi-institutional and international databases

Extracting knowledge from the ESTS database, it is necessary to critically consider which is the degree of generalization at European level of the results obtained using this registry. In fact, the number of European Units that are currently contributing to the ESTS database is around 240 (*Figure 3*). Nevertheless, the representation at data uploading is extremely heterogeneous considering the rate of Unit contribution per nation. As reported in *Figure 3*, some Countries present a solid and exhaustive participation to the ESTS database, but unfortunately others are

Contributors ESTS Database*

EUR Nation	N Contributors
Albania	1
Austria	2
Belgium	10
Bulgaria	2
Czech Republic	2
France	106
Germany	11
Greece	6
Hungary	13
Italy	22
Luxembourg	1
Netherlands	8
Polang	2
Portugal	4
Romania	7
Slovakia	1
Slovenia	1
Spain	19
Switzerland	2
Turkey	12
United Kingdom	6
Total	238

*contribution to core section

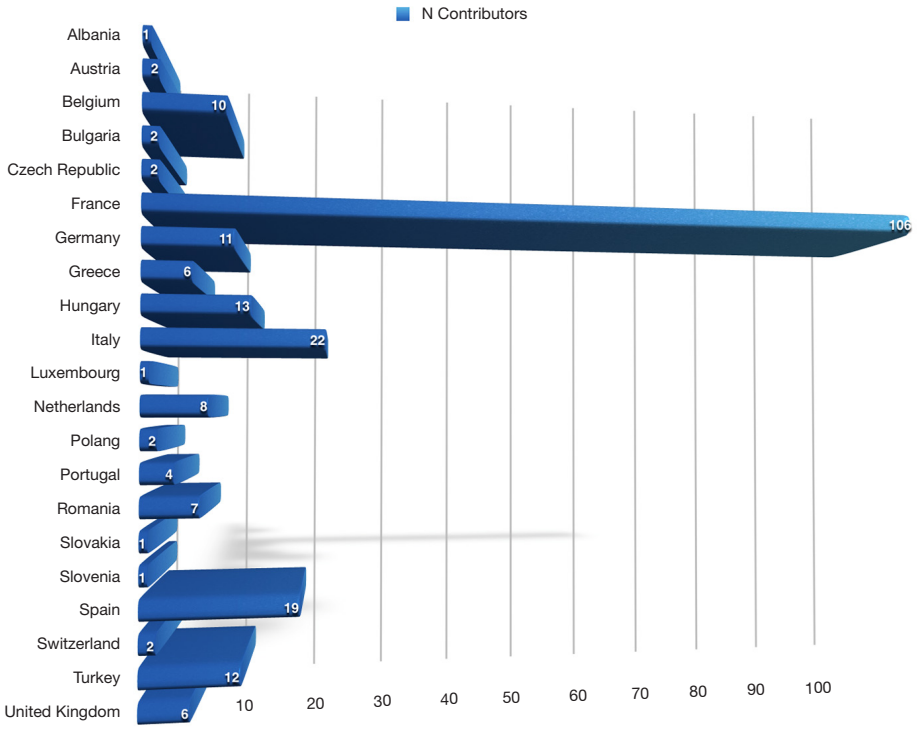

Figure 3 ESTS database contributors. ESTS, European Society of Thoracic Surgeons.

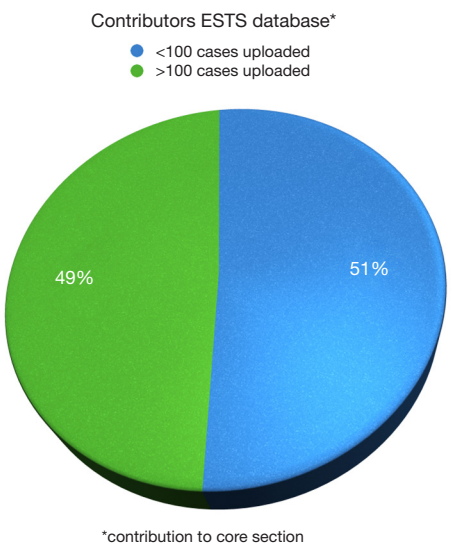

Figure 4 Proportion of low (<100 cases) and high (>100 cases) contributors. ESTS, European Society of Thoracic Surgeons.

represented by very few units. The distribution of the Units participating to the ESTS database could influence (as a consequence of different epidemiology, national clinical practice and organization of specific health care systems) the information derived by studies founded on the European registry. Moreover, as shown in *Figure 4*, half of the contributors to the ESTS database, uploaded less than 100 total cases, and the variation of contribution among Units seems very high if we consider that the mean value of uploaded cases is 376 with a standard deviation of 658 cases.

These perspectives should be always considered when we try to compare practice and outcomes at an institutional or national level to benchmarks obtained from the analysis of these data. Even more relevant is this limitation at the moment when strategies of quality of care improvement should be defined (evaluating ESTS database derived information) and then implemented in the daily clinical practice.

Do we trust our data?

Another limiting factor in the analysis and interpretation of big registries is represented by the general reliability level of data collected. Several factors that are related to the architecture of a database, to the procedures of data flows and imputation and to the update and timeliness of data could affect the correspondence between a database and the real word that it intends to describe.

In particular, due to the voluntary nature of the ESTS database, the contributors are not forced by any form of supervision for an exhaustive and constantly updated upload of their patients to the registry. This could introduce internally in the system of collection multiple bias of patients selection with obvious relapse at the moment of the data analysis. Moreover, we should consider that multi-institutional databases are often conditional on a limited number of variables composing the dataset in order to obtain a more efficient and complete data collection process. This factor, on the other hand, could impact the availability of some important information and the overall system capacity of interpreting properly the real word. In order to minimize this question, the ESTS database defined within the registry, that counts 155 variables, a specific group of items defined "core variables", which were highlighted to the end users as mandatory and most relevant for the following analytic phase. Moreover, several satellite sections of the central lung module of the ESTS database were developed with the aim of obtaining a wider exploration and collection of data on specific aspects of the thoracic surgery field.

Finally, the results of studies performed using the ESTS database are related to its data quality. Several studies were performed in the past to develop methodologies of data quality assessment of this registry. In particular metrics as completeness, accuracy and consistency were adopted to measure the quality of data collected in this registry. The studies showed an overall acceptable level of quality for the ESTS database, but at the same time revealed some weaknesses. For instance some core variables present a low level of completeness that not allows its use within analytic models. Moreover a great proportion of units were found to be characterized by a data quality level under the mean of the entire group of contributors when the quality was measured as a combination of completeness and consistency. As a consequence dashboards (*Figure 5*) reporting real time the completeness of core variables data were developed and showed to each contributor with the purpose of highlight deficiencies in the data collection process. Anyway, from this perspective to improve the use and to increase the reliability of the ESTS database and its information, specific strategies of data quality management should be reinforced in the near future, including actions of direct data quality assessment checks based on the original data source.

Conclusions

ESTS developed a medical registry that nowadays can be considered one of the most important databases for collecting data about thoracic surgery procedures worldwide. Due to a constant rate of upload, a big amount of data is stored within this registry. This implies that many information are potentially contained within the ESTS database that can be used to lead research and clinical practice about our specialty, as has happened in the past. However, at the same time, it should be taken into account that dealing with big data derived from multiple contributors will require to face with several challenges in the next future. The ESTS database should adopt methodologies for assuring a large-scale quality and reliability of data. Moreover policies for increasing the recruitment rate should be developed in order to obtain a more homogeneous national and institutional contribution. Finally, the ESTS could lead the process of defining and tailoring data mining procedures inherent to big medical databases. This perspective could represent a unique chance to found the knowledge extraction from the ESTS database on a methodology able to sustain the expected data rise of this registry.

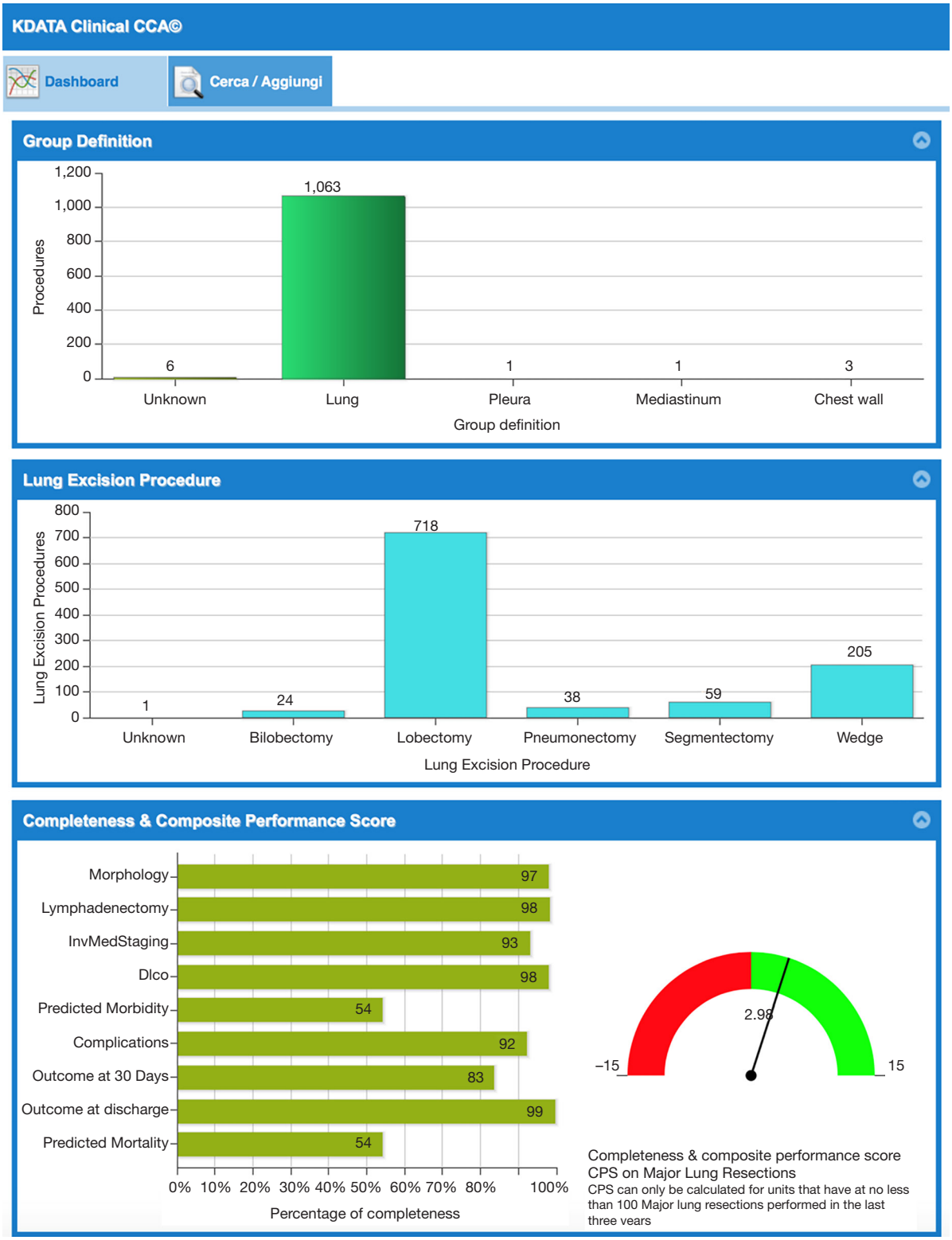

Figure 5 ESTS database dashboards screenshot. ESTS, European Society of Thoracic Surgeons.

Acknowledgements

None.

Footnote

Conflicts of Interest: The author has no conflicts of interest to declare.

References

1. Fayyad U, Piatetsky-Shapiro G, Smyth P. The KDD process for extracting useful knowledge from volumes of data. Commun ACM 1996;39:27-34.
2. Berrisford R, Brunelli A, Rocco G, et al. The European Thoracic Surgery Database project: modelling the risk of in-hospital death following lung resection. Eur J Cardiothorac Surg 2005;28:306-11.
3. Brunelli A, Varela G, Van Schil P, et al. Multicentric analysis of performance after major lung resections by using the European Society Objective Score (ESOS). Eur J Cardiothorac Surg 2008;33:284-8.
4. Brunelli A, Berrisford RG, Rocco G, et al. The European Thoracic Database project: composite performance score to measure quality of care after major lung resection. Eur J Cardiothorac Surg 2009;35:769-74.
5. Brunelli A, Rocco G, Van Raemdonck D, et al. Lessons learned from the European thoracic surgery database: the Composite Performance Score. Eur J Surg Oncol 2010;36 Suppl 1:S93-9.
6. Pompili C, Falcoz PE, Salati M, et al. A risk score to predict the incidence of prolonged air leak after video-assisted thoracoscopic lobectomy: An analysis from the European Society of Thoracic Surgeons database. J Thorac Cardiovasc Surg 2017;153:957-65.
7. Begum SS, Papagiannopoulos K, Falcoz PE, et al. Outcome after video-assisted thoracoscopic surgery and open pulmonary lobectomy in patients with low VO2 max: a case-matched analysis from the ESTS database. Eur J Cardiothorac Surg 2016;49:1054-8; discussion 1058.
8. Brunelli A, Salati A, Rocco G, et al. European risk models for morbidity (EuroLung1) and mortality (EuroLung2) to predict outcome following anatomic lung resections: an analysis from the European Society of Thoracic Surgeons database. Eur J Cardiothorac Surg 2017;51:490-7.
9. Ruffini E, Detterbeck F, Van Raemdonck D, et al. Tumours of the thymus: a cohort study of prognostic factors from the European Society of Thoracic Surgeons database. Eur J Cardiothorac Surg 2014;46:361-8.
10. Leuzzi G, Rocco G, Ruffini E, et al. Multimodality therapy for locally advanced thymomas: A propensity score-matched cohort study from the European Society of Thoracic Surgeons Database. J Thorac Cardiovasc Surg 2016;151:47-57.e1.
11. Filosso PL, Rena O, Guerrera F, et al. Clinical management of atypical carcinoid and large-cell neuroendocrine carcinoma: a multicentre study on behalf of the European Association of Thoracic Surgeons (ESTS) Neuroendocrine Tumours of the Lung Working Group. Eur J Cardiothorac Surg 2015;48:55-64.
12. Filosso PL, Guerrera F, Evangelista A, et al. Prognostic model of survival for typical bronchial carcinoid tumours: analysis of 1109 patients on behalf of the European Association of Thoracic Surgeons (ESTS) Neuroendocrine Tumours Working Group. Eur J Cardiothorac Surg 2015;48:441-7; discussion 447.
13. Salati M, Falcoz PE, Decaluwe H, et al. The European thoracic data quality project: An Aggregate Data Quality score to measure the quality of international multi-institutional databases. Eur J Cardiothorac Surg 2016;49:1470-5.
14. Salati M, Brunelli A, Decaluwe H, et al. Report from the European Society of Thoracic Surgeons Database 2017: patterns of care and perioperative outcomes of surgery for malignant lung neoplasm. Eur J Cardiothorac Surg 2017;52:1041-8.
15. Koh HC, Tan G. Data Mining Application in Healthcare. J Healthc Inf Manag 2005;19:64-72.

Cite this article as: Salati M. European Society of Thoracic Surgeons big data utilization—part 1: research interest for the thoracic community. J Thorac Dis 2018;10(Suppl 29):S3549-S3556. doi: 10.21037/jtd.2018.04.103

