STATISTICS IN MEDICINE: FROM *IN SILICO* MODELING FOR DRUG DESIGN TO MEDICAL STATISTICS AND EDUCATION

Sorana D. BOLBOACĂ

A collection of publications

Submitted as **Habilitation Thesis**

by
Sorana D. BOLBOACĂ
Faculty of Medicine
"Iuliu Haţieganu" University of Medicine and Pharmacy Cluj-Napoca
13 Emil Isac, 400023 Cluj-Napoca, Cluj, Romania
E-mail: sbolboaca@umfcluj.ro
Web-page: http://sorana.academicdirect.ro/

March 8, 2012

REZUMAT

Teza 'Statistică în Medicină: de la modelarea *in silico* pentru designul de compuşi activi la statistica şi educaţia medicală' prezintă şi sumarizează o serie de cercetări originale realizate pe parcursul ultimilor şapte ani. Cercetările prezentate în această teză aparţin la trei domenii diferite dar într-un anumit fel legate, respectiv modelarea *in silico* pentru designul a noi compuşi activi, statistică medicală şi sprijinirea formării şi educaţiei. Statistica poate fi considetară punctul central al acestei teze de abilitare deoarece nu putem vorbi nici despre modelare moleculară fără a considera evaluarea statistică a modelelor dar nici de educaţie fără a vorbi de analiza statistică a datelor experimentale.

Prezenta teză a fost scrisă pe baza mai multor publicaţii reprezentate de articole publicare în reviste cotate ISI (cu sau fără factor de impact) sau în jurnale naţioanle sau internaţionale recenzate, în volumele unor conferinţe, în carte sau capitol de carte. Teza are trei părţi principale reprezentate de: realizări ştiinţifice şi profesionale, perspective şi referinţe.

Prima parte a tezei este structurată în trei capitole: primul capitol prezintă viziunea proprie asupra domeniilor cercetare şi motivaţia cercetării, capitolul doi prezintă lista publicaţiilor incluse în această teză de abilitare iar capitolul trei prezintă cele trei direcţii enunţate anterior şi principalele contribuţii. Secţiunile celui de al treilea capitol prezintă stadiul actual al cunoaşterii şi punctează contribuţiile personale în lumina cunoştinţelor de specialitate.

În prezent, cercetarea de noi medicamente a devenit o cercetare interdisciplinară, în care cunoştinţele şi aptitudinile din diferite domenii de cercetare sunt puse în slujba identificării a noi compuşi activi, viitori candidaţi ai medicamentelor noi.

În acest context, primul capitol din contribuţiile la stadiul actual al cunoaşterii referă modelarea moleculară şi pune în evidenţă principalele contribuţii în lumina noţiunilor cunoscute deja. Capitol este împărţit în cinci secţiuni, prima secţiune fiind dedicată definiţiilor, parcursului istoric şi al tendinţele. A doua secţiune prezintă extragerea de informaţii din structura chimică prin intermediul descriptorilor moleculari. Contribuţile în ceea ce priveşte descriptorii moleculari pot fi rezumate astfel: utilizarea polinoamelor caracteristice şi a celor de numărare ca descriptori moleculari; utilitatea familiilor de descriptori moleculari ca descriptori; dezvoltarea, implementarea şi evaluarea unei noi familii de descriptori moleculari (MDFV); utilizarea şi implementarea unei noi familii de descriptori (SAPF, necesită validarea). Conceptul de familie de descriptori moleculari aparţine echipei de cercetare din care fac şi eu parte. Secţiunea a treia prezintă metodele utilizate pentru identificarea şi evaluarea modelelor structură-activitate şi prezintă algoritmul utilizat în calcularea şi validarea descriptorilor din cadrul familiilor de descriptori implementate; modalitatea de selectare a acelor descriptori care au capacitatea de a descrie legătura dintre structura chimică a compusului şi activitatea/proprietatea de interes; un model standradizat de raportare a modelelor obţinute în prin aplicarea metodei de regresie; şi prezentarea modelelor obţinute prin metodologiile aplicate. Secţiunea a patra descris strategiile şi metodele utilizate pentru a valida modelele: 9 parametrii statistice pentru a evalua modelele, 6 estimatori pentru a evalua puterea de predicţie a modelelor, 12 estimatori diagnosticul unui model. Stategiile şi metodele utilizate sunt prezentate în contextul publicaţiilor personale. Secţiunea cinci prezintă un set de instrumente pentru calcularea de familie moleculare descriptori prin tăiere de vârfuri (MDFV-Toolkit), şi o aplicaţie software creată pentru a ajuta cercetătorii în calcularea estimatorilor puterii de predicţie a unui model structură-activitate (aplicaţia CPP).

O analiză statistică necorespunzătoare a datelor experimentale rezultate dintr-un studiu bine proiectat poate fi salvată de o re-analiză a datelor. Dar, un studiu prost conceput nu poate fi salvat chiar de o analiză statistică sofisticate.

În acest punct de vedere, al doilea capitol prezintă din punct de vedere statistic încredere în rezultatele cercetărilor medicale. Capitol este împărțit în cinci secțiuni: metode statistice şi aplicații ale acestora, intervale de încredere pentru diferiți estimatori statistici, utilizarea şi aplicațiile de analiză a asocierii, respectiv software-ul dezvoltat. Principalele contribuții în ceea ce priveşte metodele statistice şi aplicațiile lor sunt în următoarele domenii: analiza erorilor în aranjamente aleatorii, blocuri randomizate, şi designul de tip pătrat Latin; traducerea unui experiment într-un model matematic, identificarea planului experimental capabil să producă rezultatul optim folosind o tehnică de optimizare bazată pe matricea ortogonală; protocolul de cercetare şi analiza statistică a datelor experimentale în cercetările de pre-clinice şi clinice. A doua secțiune prezintă în contextul utilitații metodele folosite în calcularea intervalelor de încredere asociate diferiților estimatori statistici. Mai multe metode noi de calcul au fost introduse şi evaluare comparativ cu metodele publicate în literatura de specialitate. A treia secțiune a acestui capitol se ocupă analiza asocierii şi corelației ca metode statistice de analiză a datelor medicale. Contribuțiile în analiza de regresie liniară şi neliniară, precum şi în analiza de corelație, pe lângă analiza a comportamentului algoritmului genetic realizat pentru identificarea modelelor structură-activitate sunt evidențiate. Ultima secțiune a acestui capitol prezintă patru program dezvoltate şi implementate pentru sprijinirea cercetătorilor în raportarea rezultatelor statistice.

Învățământul medical şi evaluarea acestuia se mută spre identificarea şi punerea în aplicare a acelor strategii capabile să asigure calitatea, validitatea şi fiabilitatea educație şi evaluare. Globalizarea şi libera circulație a persoanelor duce la schimbări culturale, personale şi economice. Toate aceste schimbări trebuie să se reflecte în strategiile de educație şi evaluare.

În acest sens, al treilea capitol prezintă informații cu privire la instruire, educație şi evaluare. Capitol este împărțit în trei secțiuni: prima secțiune prezintă principalele linii istorice în educația medicală; a doua secțiune prezintă strategii educaționale şi de instruire şi evidențiază contribuțiile personale; şi a treia parte prezintă principalele tehnici şi metode utilizate în evaluare cu sublinierea contribuției la acest domeniu.

Parte a doua a tezei prezintă perspectivele ştiințifice şi profesionale. Această parte este divizată în trei secțiuni: prima secțiune prezintă succint cercetările în curs de desfăşurare; secțiunea a doua descrie direcțiile de cercetare şi modurile probabile de punere în practică a acestora; secțiunea a treia prezintă succint planurile de evoluție şi dezvoltare a propriei cariere profesionale, ştiințifice şi academice.

ABSTRACT

The habilitation thesis entitled 'Statistics in Medicine: From *in silico* Modeling for Drug Design to Medical Statistics and Education' collects and summarizes several original researches carried out over the last seven years. The researches presented in this thesis belong to three quite different but intertwined fields, *in silico* modeling for drug design, medical statistics and supporting training and education. Statistics could be seen as the central point of this habilitation thesis since we cannot talk neither about molecular modeling without taking about models assessment and evaluation nor about education without taking about statistical analysis of data. The basic principle of this monograph is the translation of theory to applications as well as translation of scientific research knowledge towards theoretical works.

This thesis consist of several publications, some of them published in ISI journals (with or without impact factor), others published in peer-reviewed journals or in conference proceedings, and consists of three main parts: scientific and professional realization, perspectives and references.

First part is structured on three parts: first chapter presents the overview and motivation, second chapter list the papers of this habilitation and third chapter present the main contributions. The main body of this thesis is represented by the chapter that presents the main contributions and is organized according with the defined fields. Each chapter from this section starts by presenting the main contribution in the view of the already known information/knowledge (state-of-art in the field).

Nowadays, drug research become an interdisciplinary approach, all forces bringing together knowledge and skills to integrate and translate the structural information with/in biological systems in order to identify active compounds, drug candidates and even new drugs.

In this context, the first chapter of the main contributions in expressing the therapeutic effect as a function of chemical structure presents the state-of-art in the field and highlights the main contributions. The chapter is divided in five sections, starting with definition, history and trends. The second section presents the extraction of information from chemical structure through molecular descriptors. Contributions to molecular descriptors could be summarized as: the use of characteristics and counting polynomials as descriptors; the use of entropies and energies as descriptors; the usefulness of molecular descriptors family as molecular descriptors; development, implementation and validation of a new family of descriptors, called molecular descriptors family on vertices; and development and implementation of another new family of descriptors, called structural atomic property family, family that need validation. Note that, the concept of molecular descriptors family belongs to our research team. The third section presents the approaches used to identify and evaluate the therapeutic models, describing from the point of view of contributions how the values of molecular descriptors families are calculated and validated, approaches used to select those descriptors able to link the structure of the chemical compounds with activity/property, how normality of experimental data (condition of applying the simple or multiple regression analysis approach) could be tested, chart model for reporting regression results, and presents several new expressions. The fourth section described the strategies and methods used for validate the models: 9 statistical parameters to assess the models, 6 estimators to assess the models predictive power, 12 estimators for models diagnostic. The reflection of these strategies and estimators in personal publication are highlighted. The fifth section presents one toolkit for calculation of molecular descriptors family on vertices (MDFV-toolkit) and one software application created to assist researchers in computing estimators for model classification power (Classification of Prediction Power).

An improper statistical analysis of the experimental data resulted from a well-designed study can be rescued by a re-analysis of data. Nevertheless, a poorly designed study could not be rescue even sophisticated statistical analysis is carrying out.

In this context, the second chapter of the main contributions deals with confidence in medical research results. The chapter is divided in five sections: statistical methods and their applications, confidence in the results from diverse medical studies, the use and applications of association analysis, and the developed software. The main contributions regarding statistical methods and their applications are in field of: analysis of errors in random arrangements, randomized blocks and Latin squares designs; translation of an experiment into a model; identification of the experimental design able to produce the optimal outcome using an optimization technique based on orthogonal array; development of research protocol and statistical analysis of experimental data in pre-clinical and clinical researches. The second section presents the methods and usefulness of confidence intervals in reporting statistical estimators. Several new computational methods were introduced and assessment compared with methods previously published. The third section deal with association and correlation analysis as statistical method for analysis of medical data. Contributions in linear and non-linear regression analyses as well as in correlation analysis, beside the analysis of the behavior of genetic algorithm in identification of models are highlighted. The last section of this chapter presents four developed and implemented program for supporting the researchers in reporting statistical results.

Medical education and assessment constantly move towards identification and implementation of those strategies able to assure quality, validity and reliability in both education and assessment. Globalization and free movement of people lead to changes of culture, people and economic activity. All these changes must reflect in medical education and assessment strategies.

In this regards, the third chapter of the main contribution is related with training and education. The chapter is divided in three sections: first section presents main lines of history and trends in medical education; second section presents the education part and highlighted personal contributions; and third part presents the main techniques and methods used in assessment with highlighting the contribution to this field.

The second part presents the perspectives being divided in three sections. The first section summarizes the ongoing research. The second section of this part describes the intended works in the light of both research and education. The last section briefly presents the professional, scientific and academic carrier plans.