Politechnika Śląska
Wydział Automatyki, Elektroniki i Informatyki

Wsparcie diagnostyki onkologicznej w opartym o dane omiczne procesie wielokryterialnego wnioskowania parametryzowanego obrazem klinicznym pacjentów o podobnym profilu

mgr inż. Aleksander Płaczek

Rozprawa doktorska

przygotowana pod kierunkiem:
Dr hab. inż. Dariusza Mrozka, prof. Pol. Śl.
wspierana od strony medycznej przez:
Dr hab. n. med. Michała Jarząba z PIB-NIO, Oddział w Gliwicach

Gliwice, czerwiec 2022

mgr inż. Aleksander Płaczek

Abstract

Cancer diagnosis support in multi-criteria inference process based on omics data parametrized by clinical profile of similar patients

Diagnostics for neoplastic thyroid lesions is a decision-making process against a background of uncertainty due to a lack of clear phenotypic symptoms indicating the nature of such lesions. Emergent thyroid nodules are in the vast majority diagnosed as benign lesions, treatable pharmacologically or with radioactive iodine. Surgical treatment is usually applied if malignancy is suspected, followed by administration of drugs to replace the hormones produced by, i.a., thyroid follicular cells. Affected patients must be subject to monitoring throughout their lives. Clinical signs are not a factor sufficient to confirm that the lesion is malignant. Fine-needle biopsy provides substantial support for the diagnostics of malignant lesions. It allows estimating the probability of a lesion being malignant, based on the collected cellular material. Such risk puts a lesion in one of 6 categories of the Bethesda system. Unfortunately, the method of collecting the samples does not allow for retention of tissue structural integrity and the evaluation is made only as far as to verify if the material shows lesions indicating tissue malignancy.

An in-depth analysis of the guidelines prepared by the Polish research association, which provides recommendations on diagnostics of thyroid neoplasms and their treatment, shows that except for the boundary categories of cytologic evaluation, there is a huge risk of an incorrect estimation of the lesion based on a clinical picture alone. In such a case, a gene expression profile analysis may be used as a complementary solution. An extended diagnostics for neoplastic thyroid lesions may be based on molecular tests for the detection of somatic mutations (genetic alterations caused by errors in replication processes and imperfect functioning of the DNA repair mechanisms or by environmental factors). However, such solutions are not in regular use. Clinical information remains the basis for clinical decision-making.

Even so, only molecular characteristics are used in research concerning new methods of malignancy risk prediction in thyroid nodules. Their value originates from processed data derived from the cellular material. Here, further uncertainty arises since the nodules may be small, and accurate collection of the material may be more difficult.

Assumptions of the project under the acronym MILESTONE (STRATEGMED2/267398/4/NCBR/2015), in which the author participated, were the motivation to write this dissertation. Its aim was to develop new molecular diagnostic and imaging tools for individualised treatment of breast, thyroid, and prostate cancer. The results of an additional retrospective analysis carried out by the author on the data gathered from the hospital database of reference site carrying out diagnostics and treatment of thyroid neoplasms exposed the untapped potential of clinical data as a support for molecular test modelling.

The aim of this work comes with the intention to verify if the inability to make a decision based on clinical evaluation implies limitations of such characteristics and if it is the reason for their avoidance in thyroid cancer prediction models based on the molecular data. Experimental research was carried out within the scope of the work, including an analysis of cytologic evaluation quality

impact as a key decisive factor in the diagnostic process, on the quality of classification methods employing both molecular data and key clinical features. A model of synthetic clinical risk was developed, enabling averaging of the neoplasm malignancy risk to back up the omic features. It was verified whether the synthetic variable describing the patient's clinical picture may be used for discretisation of gene expression continuous variables to increase the separability of cases of suspected malignancy in patients.

All efforts made in this dissertation are aimed to show that the clinical data, originally not particularly informative, may be valuable if processed and used correctly. In the first part, the author shows the potential hidden in the clinical data from day-to-day oncological diagnostics collected in the hospital databases. These are the data that do not follow the strict criteria of research projects. The author utilizes them to create a model generalizing thyroid nodule malignancy risk based on a subpopulation of Polish patients diagnosed with a thyroid neoplasm. He carries out retrospective research to identify key terminology and its application in the categorization of data devoid of appropriate features, and at the same time to increase the cardinality suitable to build a Bayesian Network and determine the a priori probability of the clinical malignancy. The values returned by this model are then used by the author as a synthetic clinical feature. Its impact on the omic and clinical models is subsequently examined to either confirm or reject the hypothesis on the usefulness of clinical features in the molecular material analysis process.

Results obtained within the framework of this thesis confirm the assumption that employing the synthetic clinical risk in thyroid tumour malignancy risk prediction models is rational and advisable, however not by the means of simple data integration as a fusion but by using the clinical information in the process of discretisation of gene expression values. The obtained results show that substitution of raw features with a risk designed on the grounds of a set derived from day-to-day diagnostics allows, in connection with omic features, to obtain results comparable with those of a pathology case conference – the most reliable source of medical knowledge. All this takes place without the excessive use of expert knowledge in the data processing operations, with the identification of key features within the framework of machine learning. All models constructed within the framework of this thesis are data-based. Their parameters are also estimated on the basis of data.

The results obtained within the framework of this thesis bring hope that the use of generally available clinical data collected over the years in hospital systems will facilitate cost reduction for the development of new molecular classifiers and enable the use of phenotypic information for even more precise segmentation of samples.

The suggested approach and developed algorithms may be used for more detailed analyses, which could, for example, allow defining potential genes with expression linked to lesions visible upon cytologic evaluation. It would be a step towards the identification of links between genotype and phenotype. The suggested model of synthetic clinical risk may be used for the clinical estimation of thyroid nodule malignancy risk itself in patient-to-patient comparison and selection of those with similar risk. Such an approach allows consultation with the existing medical documentation and treatment strategy applied in similar situations. The model may also be used to parametrise gene expression for the needs arising during the development of new molecular tests or for increasing the segmentation of the existing discrete classifiers, where discriminatory molecular characteristics are known.