

POLITECHNIKA ŚLĄSKA
WYDZIAŁ AUTOMATYKI, ELEKTRONIKI I INFORMATYKI



**Politechnika
Śląska**

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

Streszczenie

Wsparcie diagnostyki onkologicznej w opartym o dane omiczne procesie wielokryterialnego wnioskowania

parametryzowanego obrazem klinicznym pacjentów o podobnym profilu

Diagnostyka tarczycy w kierunku zmian nowotworowych to proces podejmowania decyzji w warunkach niepewności, wynikającej często z braku jednoznacznych objawów fenotypowych świadczących o charakterze tych zmian. Pojawiające się guzki tarczycy w zdecydowanej większości okazują się być zmianami łagodnymi, co do których wystarczające jest leczenie farmakologiczne lub za pomocą jodu radioaktywnego. W przypadkach podejrzenia złośliwości, stosuje się zwykle leczenie operacyjne, z późniejszym podawaniem leków zastępujących hormony produkowane m. in. przez komórki pęcherzykowe tarczycy. Badania kontrolne takich pacjentów należy prowadzić przez całe ich życie. Objawy kliniczne nie są wystarczającym czynnikiem potwierdzającym złośliwość zmiany. Dużym wsparciem w diagnozowaniu zmian złośliwych jest biopsja cienkoigłowa, pozwalająca na oszacowanie prawdopodobieństwa złośliwości zmiany na podstawie pobranego materiału komórkowego. Ryzyko to kategoryzuje zmianę w 6-stopniowej skali Bethesda. Niestety sposób pobrania nie pozwala na zachowanie struktur tkankowych i ocena polega jedynie na weryfikacji czy materiał wykazuje zmiany sygnalizujące zezłośliwienie tkanki.

Dogłębna analiza wytycznych sporządzonych przez polskie towarzystwa naukowe, wydające rekomendacje w zakresie diagnostyki i leczenia nowotworów tarczycy pokazuje, że poza brzegowymi kategoriami oceny cytologicznej, istnieje duże ryzyko błędnego oszacowania zmiany jedynie w oparciu o obraz kliniczny. Rozwiązaniem wspierającym może być wtedy analiza profilu ekspresji genów. Rozszerzona diagnostyka tarczycy w kierunku zmian nowotworowych może opierać się o testy molekularne, mające na celu wykrywanie zmian somatycznych (modyfikacji genowych wywołanych błędami w procesach replikacji i niedoskonałym działaniem mechanizmów naprawy DNA lub czynnikami środowiskowymi). Jednak rozwiązania te nie są w rutynowym użyciu. Fundamentem do podejmowania decyzji klinicznych pozostaje informacja kliniczna.

Mimo to, w pracach nad nowymi sposobami predykcji ryzyka złośliwości w guzkach tarczycy wykorzystuje się jedynie cechy molekularne. Ich wartość pochodzi z przetwarzania danych pobranych z materiału komórkowego. Tu również występuje niepewność wynikająca z faktu, że guzki mogą być małe i dokładne pobranie materiału może być utrudnione.

Motywacją do napisania niniejszej dysertacji, były założenia projektu o akronimie MILESTONE (STRATEGMED2/267398/4/NCBR/2015), w którym autor brał udział. Jego celem było opracowanie nowych narzędzi diagnostyki molekularnej i obrazowania w indywidualizowanej terapii raka piersi, tarczycy i gruczołu krokowego. Wyniki dodatkowej analizy retrospektywnej, wykonanej przez autora na danych gromadzonych w szpitalnej bazie danych ośrodka referencyjnego, zajmującego się diagnostyką i leczeniem nowotworów tarczycy, uwidocznily niewykorzystany potencjał danych klinicznych jako danych wspierających projektowanie testów molekularnych.

Celem tej pracy jest chęć sprawdzenia, czy brak możliwości podjęcia decyzji w oparciu o ocenę kliniczną oznacza słabość tych cech i stanowi powód do ich unikania w modelach predykcji złośliwości guza tarczycy bazujących na danych molekularnych? W pracy przeprowadzono badania eksperymentalne, w ramach których przeanalizowano wpływ jakości z jaką została wykonana ocena cytologiczna, stanowiąca kluczowy czynnik decyzyjny w procesie diagnostycznym, na jakość metod klasyfikacji wykorzystujących jednocześnie dane molekularne i kluczowe cechy kliniczne.

Opracowano model syntetycznego ryzyka klinicznego, pozwalający na uśrednienie ryzyka złośliwości nowotworu na potrzeby wsparcia cech omicznych. Zweryfikowano, czy syntetyczna zmienna opisująca obraz kliniczny pacjenta może być wykorzystana do dyskretyzacji wartości ciągłych ekspresji genów, w celu zwiększenia separowalności przypadków podejrzenia złośliwości u pacjentów.

Wszystkie podejmowane w niniejszej dysertacji starania mają na celu zaprezentowanie, że pierwotnie mało informatywne dane kliniczne mogą stanowić wartość, jeżeli zostaną odpowiednio przetworzone i we właściwy sposób wykorzystane. W pierwszej części pracy autor prezentuje potencjał, jaki drzemie w danych klinicznych gromadzonych w codziennej diagnostyce onkologicznej w szpitalnych bazach danych. Są to dane nienacechowane spełnianiem ścisłych kryteriów projektów badawczych. Autor wykorzystuje je do zbudowania modelu generalizującego ryzyko złośliwości guzka tarczycy w oparciu o subpopulację polskich chorych, u których zdiagnozowano nowotwór tarczycy. Przeprowadza badania retrospektywne w celu identyfikacji kluczowej terminologii, wykorzystania jej do kategoryzacji danych nieposiadających odpowiednich cech, a tym samym do zwiększenia siły zbioru, na którym może stworzyć sieć bayesowską i następnie wyznaczać prawdopodobieństwo a priori złośliwości klinicznej. Wartości zwracane przez ten model, wykorzystuje w dalszej kolejności jako syntetyczną cechę kliniczną i bada jej wpływ na modele łączące dane omiczne z klinicznymi, by potwierdzić lub odrzucić hipotezę o przydatności cech klinicznych w procesie analizy materiału molekularnego.

Uzyskane w ramach niniejszej rozprawy wyniki potwierdzają założenie, że sensowne i celowe jest wykorzystanie syntetycznego ryzyka klinicznego w modelach predykcyjnych złośliwości guza tarczycy, ale nie poprzez zwykłą integrację danych w formie fuzji, lecz przez wykorzystanie informacji klinicznej w procesie dyskretyzacji wartości ekspresji genów. Uzyskane wyniki pokazują, że podstawienie ryzyka budowanego w oparciu o zbiór pochodzący z codziennej diagnostyki, zamiast surowych cech, pozwala w integracji z cechami omicznymi uzyskać wyniki porównywalne z konsylium patologów – najbardziej wiarygodnym źródłem wiedzy medycznej. Wszystko to odbywa się bez nadmiernego wykorzystywania wiedzy eksperckiej w procesach przetwarzania danych, z identyfikacją kluczowych cech w ramach uczenia maszynowego. Wszystkie modele budowane w ramach tej rozprawy są budowane w oparciu o dane. Ich parametry są estymowane również na podstawie danych.

Uzyskane w ramach niniejszej rozprawy wyniki dają nadzieję, że wykorzystanie ogólnodostępnych danych klinicznych, gromadzonych przez wiele lat w systemach szpitalnych, pozwoli na zmniejszenie kosztów opracowywania nowych klasyfikatorów molekularnych oraz pozwoli na wykorzystanie informacji fenotypowej do jeszcze bardziej precyzyjnej segmentacji próbek.

Zaproponowane podejście i opracowane algorytmy mogą zostać zastosowane do bardziej szczegółowych analiz, które np: pozwoliłoby na zdefiniowanie potencjalnych genów, których ekspresja powiązana jest z widocznymi w obrazie cytologicznym zmianami. Byłby to krok w kierunku identyfikacji powiązań między genotypem i fenotypem. Zaproponowany model syntetycznego ryzyka klinicznego może być wykorzystany do estymacji samego, klinicznego, ryzyka złośliwości guzka tarczycy, w celu porównania pacjentów między sobą i selekcji tych, dla których ryzyko jest podobne. Takie podejście umożliwi zapoznanie się z istniejącą już dokumentacją medyczną i przyjętą w podobnych sytuacjach strategią leczenia. Model może być również wykorzystany do parametryzowania ekspresji genów na potrzeby opracowywania nowych testów molekularnych lub wzmacniania segmentacji dla istniejących dyskretnych klasyfikatorów, gdzie znane są różnicujące cechy molekularne.