Narodowy Instytut Onkologii
im. Marii Skłodowskiej-Curie
Państwowy Instytut Badawczy
Oddział w Gliwicach

Gliwice, 23 czerwca 2023

dr hab. n. med. Michał Jarząb
Centrum Diagnostyki i Leczenia Chorób Piersi

Recenzja rozprawy na stopień doktora pt. „Machine learning methods in support of multiomics signature identification for breast cancer patient subpopulations" przedstawionej przez mgr inż. Joannę Tobiasz

Przedstawiana przeze mnie recenzja wykonana została na zaproszenie i zlecenie Rady Dyscypliny Inżyniera Biomedyczna Politechniki Śląskiej. Przedstawiona mi do oceny praca przygotowana przez mgr inż. Joannę Tobiasz to licząca 160 stron monografia, opracowana w języku angielskim pod nadzorem promotora prof. dr hab. inż. Joanny Polańskiej i ko-promotora Dra Christosa Hatzisa. Mo ona typowy układ, składa się ze zwięzłego wprowadzenia przedstawiającego cele i zawartość pracy, a także rysującego tło merytoryczne, przedstawienia materiału i metod, a następnie omówienia wyników wraz z krótką ich dyskusją, przeplatającą prezentację uzyskanych danych. Całość kończy krótki rozdział zawierający podsumowanie i wnioski, a następnie suplement zawierający dodatkowe wyniki w rycinach i tabelach. Układ pracy jest w pełni poprawny i bardzo dogodny do śledzenia toku rozumowania Autorki. Całość kończy przedstawienie piśmiennictwa, z którego korzystano i do którego odnosi się tekst monografii, spis tabel (jest ich 30) oraz rycin (jest ich 51), wykaz skrótów, podziękowania oraz informacja o źródłach finansowania pracy.

Na pierwszych stronach Autorka błyskotliwie rysuje rozbieżność wyników badań podstawowych i i ich klinicznego wykorzystania w zakresie podtypów molekularnych raka piersi, mimo ponad 20-letniej już historii stanowiącą nadal barierę niepokonaną. Wskazuje na cele pracy, z których główny oparty jest o próbę wykorzystywania profili białek w materiale klinicznym i próbę zidentyfikowania w ten sposób podtypów biologicznych, dla lepszego scharakteryzowania podgrup raka piersi i powiązania ich biologii z obrazem klinicznym. Autorka stawia hipotezę, że zaawansowane metody analizy pozwolą wyodrębnić podtypy w sposób odporny na ograniczenia wyjściowego materiału, w którym przeprowadzona były badania.

Autorka we wstępie pracy krótko przedstawia biologiczne i kliniczne różnice pomiędzy podtypami raka piersi. Prezentuje podtypy w kontekście ich biologii i rokowania, ujmując je nieco stereotypowo (dla podtypu HER2: „oporny na chemioterapię, wrażliwy na terapię anty-HER2", dla TNBC: „najgorsze rokowanie raków potrójnie ujemnych"). Oczywiście, Autorka jako nie-lekarz korzysta przygotowując dyskusję z dostępnych źródeł; dwie wartościowe moim zdaniem publikacje mogą być pomocą (Taylor i wsp., BMJ 2023, doi: 10.1136/bmj-2022-074684, Loibl i wsp., Lancet 2021 doi: 10.1016/S0140-6736(20)32381-3). Byłbym ciekaw dyskusji ze strony Autorki kwestii wpływu wyjściowego zaawansowania nowotworu (różnego między podtypami), dostępnych metod terapii oraz interakcji tych czyników na rokowanie; jestem na nią otwarty zarówno w ramach obrony pracy, jak i w bezpośredniej spotkaniu.

W następnej części wstępu Autorka przedstawia i dyskutuje historię badań nad podtypami raka piersi i dostępne metody ich analizy, w tym badania w ramach sieci atlasu genomu raka (TCGA, The Cancer Genome Atlas Network). Dane dotyczące raka piersi zebrane w ciągu prac konsorcjum TCGA stanowią dla Niej podstawę analizy. Z tego względu proszę Autorkę o przedyskutowanie podczas obrony tego, jak sposób pobierania próbek podczas prac konsorcjum TCGA (pobranie podczas operacji versus biopsja przed leczeniem, zastosowanie leczenia przedoperacyjnego) może wpływać na potencjalną populację chorych objętych badaniem, a pośrednio na jego reprezentatywność i wyniki.

Praca oparta jest o szeroki wachlarz metod biostatystycznych i bioinformatycznych, zarówno powszechnie stosowanych, jak i opracowanych rzez zespół prof. Joanny Polańskiej, w którym Doktorantka pracuje. Metody wydają się w pełni adekwatne do stawianych celów, jednak pełną ocenę tej części pracy pozostawiam (jako lekarz) pozostałym PT Recenzentom. Uznanie budzi we mnie troska o właściwą prezentację danych czytelnikowi, zarówno w kontekście ich selekcji, jak i przedstawienia.

Podstawowy zbiór danych objętych analizą stanowi tabela ekspresji 166 białek od 407 chorych, u których znane było przypisanie molekularnego podtypu raka PAM50. Zestaw białek określono po procedurach filtracji obserwacji brakujących; ponieważ wyjściowo analizą objęto 281 białek, jest to istotne ograniczenie użytej metody. Autorka jest świadoma aspektów technicznych, właściwie opisuje je i analizuje, a także rozumie pewną wycinkowość obserwacji dokonanej tym sposobem. Przedstawia w tym kontekście także odniesienie do wykorzystania metody analizy nadreprezentacji ścieżek (Reactome ORA), która stanowi źródło danych do analizy. Proszę o rozszerzenie podczas obrony zagadnienia, na ile białka obecne w wysokim lub niskim stężeniu mogą być reprezentatywne dla procesów biologicznych o różnym charakterze (białka strukturalne versus regulacyjne); a więc na ile reprezentatywna jest analiza szlaków po znaczącej filtracji danych wyjściowych?

Przedstawienie wyników pracy podzielono na 3 części, wyodrębnione w postaci osobnych rozdziałów. W pierwszym z nich Autorka analizuje zagadnienie, jak zidentyfikować

subpopulacje pacjentów chorych na raka piersi na podstawie danych o ekspresji białek. Przedstawia wyniki analizy ścieżek (ORA), analizuje kwestie efektu serii (batch effect), przegląda wpływ różnych algorytmów grupowania/klasteryzacji oraz różnych miar różnicy i przechodzi do bezpośrednich wyników dotyczących liczby klastrów, na które rozdziela się badana grupa (3-6 podgrup). Największa z populacji, to jest chore z podtypem luminalnym A jest w opinii Autorki najbardziej heterogenna i może być podzielona na 3 podgrupy (A1-A3). Brakuje mi w tym rozdziale (choć częściowo zagadnienie to jest dyskutowane dalej) przedstawienia, na ile efekt serii może wpływać na obserwowany podział i na ile udało się go skutecznie skorygować. Na przedstawionej w pracy rycinie 4.2A i C widać wyraźnie podgrupę próbek silnie wyodrębniającą się (ujemne wartości składowej U2), która najsilniej rzutuje na wyniki analizy. Co to za próbki, wydaje się że stanowią podgrupę 2 lub 3 płytek na których wykonywano badania RPPA. Czy dane TCGA zawierają informacje o źródle próbki (ośrodek)? Czy Autorka ma wiedzę co do tego, czy próbki pochodzące z różnych ośrodków były w zrównoważony sposób rozkładane pomiędzy serie eksperymentalne? Czy ta podgrupa próbek wyodrębnia się w którymś z uzyskiwanych klastrów z przewagą?

W kolejnym rozdziale Autorka przechodzi do scharakteryzowania populacji chorych pod kątem dostępnych w repozytorium TCGA danych klinicznych, w tym danych dotyczących przeżycia. Autorka jest ograniczona naturą dostępnych danych i trybem ich gromadzenia; szczególnie rzutuje to na analizę czasu przeżycia. Czy dostępne są informacje o tym, jak aktualizowano dane TCGA po bezpośrednim zgłoszeniu próbek? Skąd bierze się relatywnie krótki czas obserwacji, brak zdarzeń w niektórych podtypach i dużego stopnia niezbalansowanie pomiędzy pacjentami z wystąpieniem zdarzenia i bez niego (vide Tab. 5.1-5.2)? Jak podkreśla Autorka, obserwowane braki mają wpływ na skuteczność wnioskowania w dalszych częściach analizy; jednak biorąc pod uwagę znaczący i publiczny charakter danych TCGA oraz ich szerokie użycie, fakt ten ma doniosłą wartość naukową, a jego szeroka analiza w monografii i przyszłych badaniach ma znaczenie dla badaczy zajmujących się zagadnieniem. Autorka bardzo kompetentnie omawia definicje różnych miar skuteczności leczenia onkologicznego, które porównywane są dalszym ciągu pracy. Kładzie nacisk na znaczenie nadreprezentacji chorych o dobrym rokowaniu (mniej agresywne podwtypy raka piersi), ale od zgromadzenia danych TCGA minął już czas na tyle długi, że nawet w podtypach agresywnych na pewno mamy do czynienia ze zdarzeniami; szwankuje zapewne (o ile istnieje) mechanizm aktualizacji tych danych. W mojej opinii lepszym sposobem przedstawienia danych byłoby rozpoczęcie od danych dotyczących podgrup zaawansowania (stage/TNM), gdyż dane te są bardziej kompletne i rzetelne, a równocześnie częściowo rzutują na parametry przeżycia niezależnie od ocenianej proteomicznie biologii guza.

Najciekawszy z punktu widzenia klinicysty element analizy w tym podrozdziale Autorka pozostawia na koniec: analizując zmienne ciągłe, bada związek wieku chorych w momencie zachorowania oraz obecność subpopulacji komórek odpornościowych w guzie

(estymowanych na podstawie algorytmu CIBERSORT). Obserwuje różnice między podtypami, w tym wyraźnie niższy wiek pacjentek w podtypie luminalnym A1 w stosunku do pozostałych raków o fenotypie hormonowrażliwym. Obserwuje też silne różnice w nacieku limfocytarnym, w tym różnice w obecności limfocytów B pomiędzy wyodrębnionymi podtypami luminalnymi, o dużej istotności statystycznej obserwowanej różnicy. Wynik ten ma charakter znaczący, biorąc pod uwagę próby zastosowania immunoterapii również w tym podtypie raka piersi oraz obserwowany brak skuteczności u dużej części chorych w tej populacji. Autorka poświęca więcej uwagi limfocytom T oraz makrofagom i mastocytom, prezentując szczegółowe dane graficzne w głównej części tekstu. Różnicę dotycząca subpopulacji makrofagów M1/M2 Autorka wybija w dyskusji jako najbardziej znaczącą; zagadnienie to warte jest dyskusji w kontekście obecności lub braku w niektórych podtypach raka piersi nacieku zapalnego. Ogólnie Autorka wnioskuje, iż klasyfikacja na podstawie danych proteomicznych nie wnosi więcej niż klasyfikacja na podstawie PAM50. Perspektywicznie wartościowe są potencjalne różnice, które dostrzega między klasyfikacjami – genomiczną i proteomiczną; na pewno zagadnienie to zasługuje na dalsze badania.

W Rozdziale 6 Autorka rozprawy charakteryzuje wyodrębnione subpopulacje pod kątem molekularnym. Tu pojawia się dyskusja dotycząca efektu serii, głównie pod kątem różnic obserwowanych w profilowaniu mRNA; rozszerzenie tej analizy, jak zaznaczono wcześniej, pod kątem związku z wykrytymi podtypami wydaje się wartościowe.

Modelowo doskonałą częścią wyników jest podrozdział 6.2, w którym Autorka analizuje markery specyficzne dla wykrytych podtypów raka, używając równocześnie klasyfikacji transkryptomicznej, proteomicznej, jak i transkryptomicznej ograniczonej do zestawu białek badanych na poziomie mRNA. W czytelny sposób zidentyfikowane są markery specyficzne dla podtypów, zarówno białkowe jak i mRNA, a na podstawie wyników Autorka podejmuje próbę opisania sygnatury specyficznej dla podtypu biologicznego raka piersi. Obserwowane różnice dotyczą genów o dużym znaczeniu biologicznym i klinicznym, w tym np. cykliny E1, kaweoliny czy GATA3.

W mojej opinii rozdział 7 – stanowiący podsumowanie – nie w pełni wyczerpuje potencjalnych wniosków praktycznych wynikających z badań przedstawionych przez Autorkę. Nie jest to jednak zarzut w stosunku do rozprawy; jako dzieło głównie w zakresie głównie biostatystyki musi ona być ograniczona w kontekście praktycznym, a zawartość monografii wskazuje na doskonałe zrozumienie przez mgr inż. Joannę Tobiasz wielu zagadnień o charakterze biologicznym i medycznym. Autorka wskazuje na skuteczność oryginalnych metod analitycznych, opracowanych przez Nią i zespół prof. Polańskiej, do analizy danych proteomicznych i transkryptomicznych raka piersi. Świadoma ograniczeń technicznych, Autorka podkreśla ich znaczenie dla pewności wniosków, które formułuje.

Autorka w pracy cytuje różnorodne i dobrze dobrane piśmiennictwo. Biorąc pod uwagę pewne spowolnienie badań nad podtypami molekularnymi raka piersi w ostatnich 5 latach, relatywnie niewiele pozycji literaturowych pochodzi z ostatniego czasu. Bardzo

proszę Doktorantkę o zwięzłe podsumowanie badań z ostatnich 2 lat w tym zakresie podczas obrony, wydaje mi się niezwykle ważne kontynuacja prac które zostały w tym zakresie przez Nią podjęte, gdyż samo zagadnienie dalekie jest jeszcze od ostatecznych wniosków praktycznych i pełnego przeniesienia zdobytej wiedzy do praktyki klinicznej.

Autorka nie uniknęła drobnych błędów redakcyjnych, które wymieniam z obowiązku recenzenta. Na stronie 8 (wiersz 19) w zdaniu odnoszącym się do liczby ozdrowieńców w roku 2020 prawdopodobnie źle przetłumaczono jedno ze słów lub doszło do zmiany szyku; sens nie jest w pełni czytelny. W tabelach 5.1 i 5.2 nie jest wskazana jednostka miary czasu, a wskazane wartości nie w pełni pasują ani do lat, ani do miesięcy. Praca przygotowana jest doskonałą angielszczyzną, a jej lektura stanowi intelektualną satysfakcję.

Reasumując chcę podkreślić, że praca mgr inż. Joanny Tobiasz jest ciekawym i wartościowym opracowaniem, które stanowi oryginalny głos w dyskusji dotyczącej podtypów biologicznych raka piersi. Zawierata wiele spostrzeżeń nowych w dyskursie naukowym, o szeokich implikacjach praktycznych. Łączy wykorzystanie oryginalnych metod z wykorzystaniem publicznie dostępnych danych, rozwiązuje problemy istotne z punktu widzenia zarówno rozwoju wiedzy, jak i praktycznej aplikacji wyników. Wysoko oceniam warsztat naukowy Doktorantki i Jej zaangażowanie w przygotowanie dysertacji. Zauważone przeze mnie niewielkie braki czy nieścisłości w niczym nie umniejszają bardzo dużej merytorycznej wartości rozprawy.

W podsumowaniu stwierdzam, że przedstawiona mi do oceny rozprawa doktorska mgr inż. Joanny Tobiasz pod tytułem „Machine learning methods in support of multiomics signature identification fro breast cancer patient subpopulations" spełnia warunki określone przez art. 187 ustawy z dnia 20 lipca 2018 – Prawo o szkolnictwie wyższym i nauce (tekst jednolity Dz. U. z 2020 r. poz. 85, z późn. zm.), stąd przedstawiam Wysokiej Radzie Dyscypliny Inżynieria Biomedyczna Politechniki Śląskiej wniosek o dopuszczenie mgr inż. Joanny Tobiasz do dalszych etapów przewodu doktorskiego. Równocześnie, podkreślając wkład Doktorantki w oryginalne opracowanie danych i dalszy potencjał uzyskanych wyników wnioskuję o wyróżnienie pracy, jeśli spełnia ona kryteria formalne w tym zakresie.