POLITECHNIKA ŚLĄSKA

KATEDRA INŻYNIERII I BIOLOGII SYSTEMÓW

Rozprawa doktorska

Opracowanie nowych algorytmów uczenia maszynowego dla heterogenicznych danych biomedycznych

Autor: mgr inż. Agata Wilk

Promotor: prof. dr hab. inż. Krzysztof Fujarewicz

Gliwice, czerwiec 2024

.

Niniejszy doktorat zarówno rozpoczął się jak i zakończył z „winy" mojego Promotora, Prof. dr hab. inż. Krzysztofa Fujarewicza, któremu serdecznie dziękuję za przekazaną wiedzę, wsparcie, i motywację (i stawianie deadlinów kiedy trzeba).

Dziękuję również Panu Prof. dr hab. inż. Andrzejowi Świerniakowi za przekazywanie mądrości naukowej i życiowej, za zaangażowanie i możliwość rozwoju pod skrzydłami autorytetu, którym dla mnie jest.

Dziękuję Prof. dr hab. inż. Damianowi Borysowi, który wciągnął mnie w „bagno" radiomiki, ale nie zostawił mnie w nim samej.

Dziękuję Pracownikom i Doktorantom z Katedry Inżynierii i Biologii Systemów oraz z Centrum Biotechnologii za wsparcie merytoryczne, emocjonalne, wspólną pracę i rozmowy przy kawie.

Dziękuję moim Współpracownikom z Narodowego Instytutu Onkologii w Gliwicach, w szczególności Kierownikowi i Koleżankom z Działu Analiz Bioinformatyczno-Biostatystycznych za motywację, wyrozumiałość, cierpliwość i wiarę w mój sukces.

Dziękuję Wszystkim, z którymi miałam zaszczyt i okazję współpracować, i przyczynili się do mojego rozwoju, a którzy nie zostali wymienieni.

Dziękuję także mojej Mamie, która nigdy we mnie nie zwątpiła.

Niniejsza praca doktorska była wspierana finansowo w ramach grantu Narodowego Centrum Nauki UMO-2020/37/B/ST6/01959.

Trzy powtórzenia biologiczne w sekwencjonowaniu,
Siedem podtypów tkanki tarczycy do klasyfikacji,
Dziewięć liczb wektorów cech w obrazowaniu,
I jeden problem danych agregacji,
W uczeniu maszynowym przez heterogenię.
Jeden, by dane połączyć, jeden, by cechy zgromadzić,
Jeden by próbki odróżnić i w modelu związać
W uczeniu maszynowym przez heterogenię.

Spis treści

Streszczenie

Hodowle komórek zdesynchronizowanych w fazie cyklu komórkowego, tkanki nowotworowe złożone z komórek różnych typów i o różnych profilach molekularnych, kohorty pacjentów różniących się stanem zaawansowania choroby, czynnikami genetycznymi i środowiskowymi, zróżnicowane populacje regionów czy krajów. Dane kliniczne, genomiczne, transkryptomiczne, proteomiczne i obrazowe, pochodzące z różnych źródeł, o różnych typach, strukturze i wymiarowości. Heterogeniczność jest nieodłącznym aspektem badań biomedycznych, przyczyną wielu zjawisk biologicznych i podstawą personalizowanej terapii. Analiza heterogenicznych danych prezentuje jednak szereg wyzwań — od złego uwarunkowania numerycznego i utrudnionej estymacji parametrów modelu, do obecności różnej liczby wektorów cech dla poszczególnych obiektów.

Kluczowym problemem naukowym, którego rozwiązanie jest przedmiotem niniejszej rozprawy, jest wykorzystanie danych dla heterogenicznych obiektów w uczeniu maszynowym. Pracę doktorską stanowi cykl siedmiu artykułów naukowych, obrazujących własne doświadczenia autorki.

Pierwsza część poświęcona jest heterogeniczności występującej na różnych poziomach. W szczególności omówiony jest dylemat pomiędzy budowaniem wspólnego modelu dla heterogenicznych grup, a niezależną analizą podgrup czy pojedynczych obiektów. Jako kompromis łączący te strategie, zaproponowano autorskie, indywidualizowane podejście do estymacji parametrów (na przykładzie modelu epidemiologicznego), polegające na estymacji części parametrów jako wspólnych dla wszystkich obiektów, a części jako niezależnych. Zastosowanie indywidualizowanego modelowania pomogło przezwyciężyć problem złego uwarunkowania numerycznego, pozwalając jednocześnie zachować elementy indywidualnej charakterystyki, co znalazło odzwierciedlenie w rozkładzie błędów dopasowania.

Druga część rozprawy porusza problem heterogeniczności struktury danych, czyli obecności różnej liczby wektorów cech. Przedstawiono oryginalne strategie wykorzystania takich danych w klasyfikacji (dla obrazowania proteomicznego) oraz modelowaniu przeżycia (dla wielu gromadzeń w obrazowaniu PET/CT) oparte na agregacji wektorów cech lub agregacji wyników modelowania. Otrzymane wyniki pokazują, że wykorzystanie informacji pochodzącej ze wszystkich dostępnych wektorów cech poprzez zastosowanie agregacji pozwala na poprawę zdolności predykcyjnej modeli względem wykorzystania pojedynczego wektora cech.

Słowa kluczowe: heterogeniczne dane, uczenie maszynowe, klasyfikacja, indywidualizowany model, analiza przeżycia, agregacja

Summary

Cultures of cells desynchronised in cell cycle phase, tumour tissues composed of cells of different types and molecular profiles, cohorts of patients differing in disease status, genetic and environmental factors, diverse populations of regions or countries. Clinical, genomic, transcriptomic, proteomic and imaging data, from different sources, with different types, structure and dimensionality. Heterogeneity is an inherent aspect of biomedical research, the cause of many biological phenomena and the basis of personalised therapy. However, the analysis of heterogeneous data presents a number of challenges — from poor numerical conditioning and difficult estimation of model parameters, to the presence of different numbers of feature vectors for individual objects.

The key research problem addressed in this dissertation is the use of data representing heterogeneous objects in machine learning. The dissertation consists of a series of seven research articles, illustrating the author's own experience.

The first part addresses heterogeneity occurring at different levels. In particular, the dilemma between building a common model for heterogeneous groups or independent analysis of subgroups or individual objects is discussed. As a compromise combining these strategies, an original individualised approach to parameter estimation (using an epidemiological model as an example) is proposed, which involves estimating some parameters as common to all objects and some as independent. The use of individualised modelling helped overcome the problem of poor numerical conditioning, while allowing elements of individual characteristics to be retained, as reflected in the distribution of prediction errors.

The second part of the dissertation describes the problem of heterogeneity of the data structure, i.e. the presence of different numbers of feature vectors. Original strategies for the use of such data in classification (for proteomic imaging) and survival modelling (for

multiple uptakes in PET/CT imaging), based on aggregation of feature vectors or aggregation of modelling results are presented. The obtained results show that the use of information from all available feature vectors through aggregation improves the predictive ability of models relative to the use of a single feature vector.

Keywords: heterogeneous data, machine learning, classification, individualized model, survival analysis, aggregation

Rozdział 1

Wprowadzenie

Ludzkość. Setki społeczeństw, zamieszkujących różne środowiska, zorganizowanych w różny sposób. Osiem miliardów ludzi, w różnym wieku, różnym stanie zdrowia, prowadzących różny tryb życia. W każdym człowieku dziesiątki tkanek, dziesiątki trylionów komórek — własnych i bakteryjnych, w każdej z nich tysiące różnych metabolitów, białek, transkryptów. I w każdej komórce DNA, złożone z miliardów par zasad, które może podlegać mutacjom i modyfikacjom. Pojawiająca się na każdym możliwym poziomie heterogeniczność jest nieodłącznym elementem życia i przedmiotem licznych badań w dziedzinie biomedycyny.

Heterogeniczność ma między innymi istotne znaczenie kliniczne, zwłaszcza w chorobach nowotworowych [8, 9, 10, 11]. Ponieważ nie ma dwóch takich samych pacjentów, również każdy nowotwór jest inny. Co więcej, obecność heterogenicznych i ciągle ewoluujących populacji komórek, zarówno nowotworowych jak i tworzących mikrośrodowisko guza, warunkuje szybkość rozwoju choroby, skłonność do przerzutowania, a nawet podatność lub oporność na terapię. Nowe możliwości oznaczeń molekularnych i obrazowych, w połączeniu z prowadzonymi na całym świecie badaniami translacyjnymi, prowadzą do wyodrębniania coraz bardziej zawężonych podtypów nowotworów, co można zaobserwować choćby na przykładzie raka piersi [12, 13, 14]. Równocześnie zmienia się tendencja podejścia terapeutycznego, dążącego w kierunku medycyny personalizowanej, w której charakterystyka genetyczna, epigenetyczna, transkryptomiczna czy proteomiczna ma pozwalać na dobór leczenia o najwyższej skuteczności przy możliwie najmniejszych skutkach niepożądanych [15, 16, 17, 18].

Zróżnicowanie samych komórek, tkanek czy pacjentów nie jest jedynym źródłem heterogeniczności danych biomedycznych. W badaniach wykorzystywane są dane z wielu modalności, takich jak kliniczne, molekularne czy obrazowe, charakteryzujące się zróżnicowaną strukturą, wymiarowością czy typem [19, 20]. Analiza wielomodalnych danych wymaga zastosowania metod integracji oraz fuzji, przy czym dodatkowym wyzwaniem jest fakt, że pomiar dla poszczególnych modalności często wykonywany jest na różnych próbkach. Do heterogeniczności danych medycznych wykorzystywanych w uczeniu maszynowym przyczynia się też ich pochodzenie z różnych źródeł, na przykład z kilku szpitali [21, 22, 23]. Jeżeli dane dla kohorty pacjentów zbierane są na przestrzeni wielu lat, zmienność może również wynikać z ewolucji narzędzi diagnostycznych czy wytycznych terapeutycznych.

Zarówno zróżnicowanie badanych obiektów jak i danych, które je opisują, rodzi potrzebę stosowania dedykowanych metod w ich modelowaniu. Niniejszą rozprawę doktorską stanowi cykl siedmiu artykułów naukowych, obrazujących własne doświadczenia autorki z analizą heterogenicznych danych biomedycznych. W pierwszej części opisane są przykłady heterogeniczności na różnych poziomach, od poziomu komórkowego aż do populacji krajów, oraz wyzwania, jakie pojawiają się w analizie heterogenicznych danych różnych modalności. Druga część rozprawy poświęcona jest modelowaniu danych o heterogenicznej strukturze, szczególnie w sytuacji, w której modelowane obiekty opisane są różną liczbą wektorów cech.

Rozdział 2

Cele i tezy pracy

Celem niniejszej pracy jest:

1. Opis heterogeniczności występującej na różnych poziomach w danych biomedycznych.
2. Opracowanie algorytmów pozwalających na analizę i modelowanie heterogenicznych populacji oraz kohort.
3. Przedstawienie problematyki strukturalnej heterogeniczności danych.
4. Zaproponowanie metod agregacji umożliwiających wykorzystanie informacji z różnej liczby wektorów cech odpowiadających poszczególnym modelowanym obiektom.

W rozprawie przedstawiono następujące tezy:

1. Indywidualizacja modeli dla kohorty obiektów pozwala zmniejszyć ryzyko złego uwarunkowania numerycznego zadania estymacji parametrów.
2. Zastosowanie agregacji w przypadku różnej liczby wektorów cech dla poszczególnych obiektów skutkuje poprawą jakości predykcji modelu względem wykorzystania tylko jednego wektora na obiekt.

Rozdział 3

Heterogeniczność na różnych poziomach

Niezależnie od tego, czy zajmujemy się badaniami epidemiologicznymi, kohortowymi, analizą tkanek czy nawet komórek, w biomedycynie niemal zawsze będziemy mieć do czynienia z heterogenicznymi populacjami. Mogą to być populacje ludzi różniących się stanem zdrowia, tłem socjoekonomicznym lub czynnikami środowiskowymi; mogą to być populacje różnych gatunków bakterii tworzących mikrobiotę; mogą to być populacje różnych typów komórek w tkankach. Nawet jeżeli analizujemy komórki jednej linii komórkowej, w obrębie hodowli zaobserwujemy komórki o różnych profilach molekularnych wynikających chociażby z desynchronizacji w fazie cyklu komórkowego [24, 25, 26]. Mimo, że taki stan odzwierciedla heterogeniczność występującą w ludzkim organizmie, może podczas analizy przysłaniać zmiany związane z badanym efektem biologicznym.

3.1 Poziom komórkowy

[1] Zarczynska I, Gorska-Arcisz M, Cortez AJ, Kujawa KA, Wilk AM, Skladanowski AC, Stanczak A, Skupinska M, Wieczorek M, Lisowska KM, Sadej R, Kitowska K. p38 Mediates Resistance to FGFR Inhibition in Non-Small Cell Lung Cancer. Cells. Nov 30;10(12):3363. (2021)

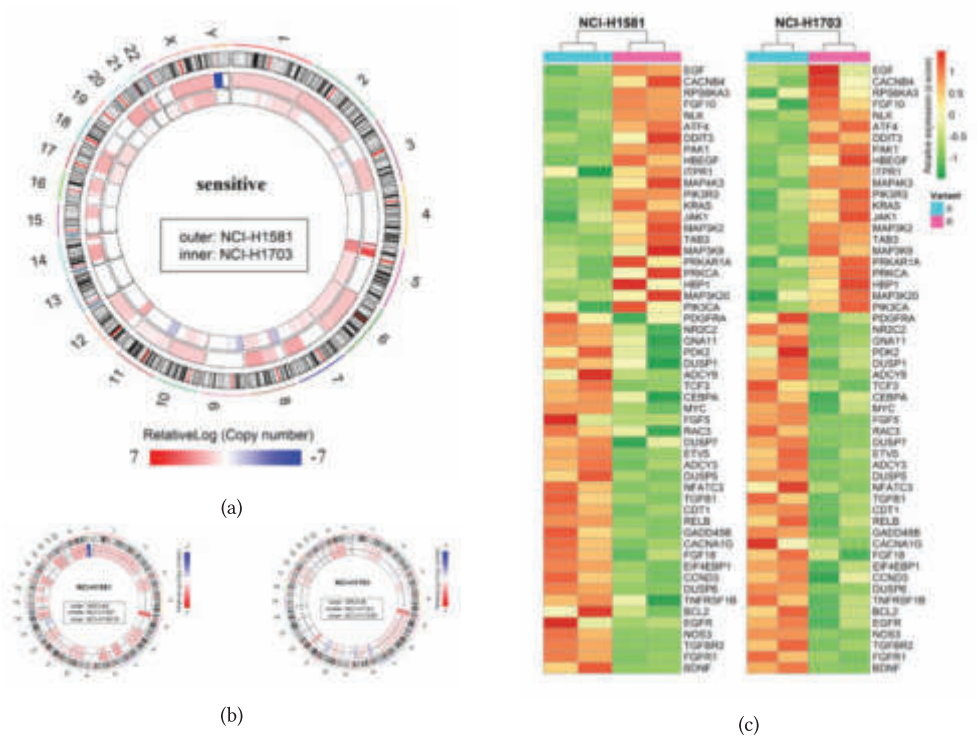
Cel. *Identyfikacja molekularnej charakterystyki nabytej oporności na inhibitor FGFR w liniach komórkowych niedrobnokomórkowego raka płuc.*

Inhibicja receptora czynnika wzrostu fibroblastów (FGFR) jest obiecującym kierunkiem terapeutycznym w raku płuc [27, 28]. Nowe leki, zanim otrzymają akceptację do wykorzystania w klinice, poddawane są szeregowi testów — badań klinicznych i przedklinicznych, na przykład na liniach komórkowych. W przypadku terapii celowanych niezwykle istotne jest między innymi określenie, jaka grupa pacjentów ma największą szansę odpowiedzi na leczenie — sam status amplifikacji FGFR nie jest jeszcze niezawodnym kryterium. W pracy [1] linie komórkowe wykorzystano do badania nabytej oporności na inhibitor CPL304110 w niedrobnokomórkowym raku płuc (NSCLC).

Spośród panelu linii nowotworowych NSCLC wykazujących amplifikację genu FGFR1 wybrano dwie, które okazały się najbardziej wrażliwe na inhibitor: NCI-H1581 i NCI-H1703. Stosując coraz wyższe stężenia CPL304110, wyprowadzono warianty oporne linii, a następnie zarówno wariant wrażliwy i oporny poddano badaniom molekularnym: aCGH (*Array Comparative Genomic Hybridization*, oparta na technolgii mikromacierzy technika służąca do oznaczania liczby kopii genów) oraz RNAseq (sekwencjonowanie RNA, wysokoprzepustowa technika umożliwiająca oznaczenie poziomu transkryptów).

Podobnie jak w wielu badaniach na liniach komórkowych, istotny problem stanowi wielkość próby [29, 30], w szczególności rozróżnienie pomiędzy powtórzeniem biologicznym i technicznym. Ponieważ linie komórkowe, nawet pochodzące z kilku różnych banków komórkowych, wyprowadzone zostały oryginalnie od jednego pacjenta, zaleca się, aby weryfikować uzyskane wyniki dla więcej niż jednej odpowiednio dobranej linii. Jednakże, w szczególności w przypadku linii nowotworowych, w których przeważnie obciążenie mutacyjne jest wysokie, nawet dla linii reprezentujących ten sam podtyp nowotworu obserwować można znaczne różnice (Fig. 3.1a). Również dla analizy transkryptomu, analiza głównych składowych (PCA) pokazała, że nawet wobec wariantu związanego z opornością, ponad 70% zmienności odpowiada kierunkowi separującemu dwie różne linie komórkowe.

Przy tak niewielkiej próbie, łączna analiza różnicowa dla obu linii komórkowych pod kątem różnic między wariantem wrażliwym i opornym prowadziła do wyników o niskim znaczeniu biologicznym — dla większości genów znaczna różnica dla jednej z linii była wystarczająca, aby uzyskać istotność statystyczną mimo braku różnic w drugiej.



(a)

(b)

(c)

Fig. 3.1: Heterogeniczność obserwowana w liniach komórkowych niedrobnokomórkowego raka płuc. (a) Zmienność liczby kopii genów między wrażliwymi („dzikimi") wariantami linii NCI-H1581 i NCI-H1703. (b) Analiza różnic pomiędzy wariantem wrażliwym i opornym osobno dla każdej linii. (c) Wynik analizy RNAseq - geny różnicujące dla obu linii, o takim samym kierunku zmiany.

Ostatecznie zdecydowano się na analizę niezależnie dla każdej linii komórkowej (Fig. 3.1b). Jednakże, po wyznaczeniu genów różnicujących w obu porównaniach okazało się, że wyniki w dalszym ciągu były trudne do interpretacji, ponieważ nawet mimo odpowiedniej siły zmiany, jej kierunek mógł być różny. Zastosowanie warunku wymagającego jednakowego kierunku zmiany dla genów różnicujących (Fig. 3.1c) w połączeniu z analizą funkcjonalną i zestawieniem danych genomicznych i transkryptomicznych pozwoliło na wyznaczenie szlaku p38 jako potencjalnego mediatora oporności na inhibicję FGFR.

Wniosek. *Ponieważ dla małych prób zmienność związana z różnicą pomiędzy typami komórek może przysłonić efekt biologiczny, czasem korzystna jest niezależna analiza dla poszczególnych podgrup i poszukiwanie wspólnych elementów przy pomocy odpowiednich warunków.*

Wkład autorki w cytowaną pracę. *Analiza bioinformatyczna zmienności liczby kopii – ekstrakcja cech genomicznych z surowych danych dla macierzy aCGH, preprocessing (normalizacja, korekcja GC i cy3/cy5, centrowanie), estymacja liczby kopii genów, analiza różnicowa; analiza bioinformatyczna danych RNAseq – preprocessing (analiza jakości, usuwanie adapterów, mapowanie i annotacja, wyznaczenie macierzy zliczeń), normalizacja, analiza różnicowa; analiza nienadzorowana i wizualizacja danych z eksperymentów wysokoprzepustowych; analiza szlaków sygnałowych – GSEA.*

3.2 Poziom tkankowy i różnice między pacjentami

Jednym z powodów, dla których nowotwory często są wykrywane w późnym stadium, jest brak wiarygodnych, uniwersalnych markerów diagnostycznych. Jako że komórki nowotworowe charakteryzują się specyficzną aktywnością metaboliczną, metabolomika jest rozważana jako jedno z potencjalnych źródeł biomarkerów [31].

[2] Mrowiec K, Debik J, Jelonek K, Kurczyk A, Ponge L, Wilk A, Krzempek M, Giskeødegård GF, Bathen TF i Widłak P. Profiling of serum metabolome of breast cancer: multi-cancer features discriminate between healthy women and patients with breast cancer. Front. Oncol. 14:1377373. (2024)

Cel. *Metabolomiczna charakterystyka raka piersi, identyfikacja sygnatury wspólnej dla różnych nowotworów, konstrukcja klasyfikatora metabolomicznego.*

W pracy przeanalizowano profile metabolomiczne z osocza zdrowych dawców, a także pacjentów chorych na nowotwory piersi, płuc, jelita grubego oraz głowy i szyi. Na podstawie analizy statystycznej wyselekcjonowano metabolity odróżniające zdrowe kontrole od każdego typu nowotworu (ze względu na zmiany profilu metabolomicznego zachodzące wraz z wiekiem, dla kontroli oraz pacjentek z rakiem piersi wyselekcjonowano podgrupy dopasowane strukturą wieku do pozostałych nowotworów). Wspólna „wielo-nowotworowa sygnatura" zawierała 6 aminokwasów (Ala, Asp, Glu, His, Phe, Leu+Ile), 2 diglicerydy, 2 triglicerydy, oraz 13 lizofosfatydylocholin (a także całkowity poziom lizofosfatydylocholin).

Następnie korzystając z wyznaczonej sygnatury skonstruowano klasyfikator pozwa-

lający rozróżnić zdrowe kontrole od pacjentek z rakiem piersi. Po przetestowaniu różnych modeli oraz wartości hiperparametrów, zdecydowano się na maszynę wektorów wspierających (SVM) z radialną funkcją jądra. Jakość klasyfikacji oszacowano w procedurze 500-krotnej kroswalidacji Monte Carlo przeprowadzonej na zbiorze próbek, które nie były wykorzystywane do selekcji wielo-nowotworowej sygnatury. Aby sprawdzić, czy klasyfikator sprawdzi się również dla innych europejskich populacji, model nauczony na zbiorze polskich pacjentek testowano na zbiorze pacjentek norweskich. Otrzymany klasyfikator cechował się bardzo wysoką zdolnością predykcyjną, osiągając w walidacji mediany czułości=0.97, swoistości=0.92 i AUC=0.98.

Wniosek. *Pomimo różnic między typami nowotworów, da się wyselekcjonować wspólną sygnaturę, dla której możliwa jest klasyfikacja z wysoką dokładnością polskich i norweskich pacjentek z rakiem piersi.*

Wkład autorki w cytowaną pracę. *Uczenie i walidacja modeli uczenia maszynowego, badanie wpływu hiperparametrów, opracowanie schematu testowania modeli w celu zbadania generalizowalności dla różnych kohort, udział w przygotowaniu manuskryptu.*

3.3 Poziom populacji — indywidualizowane modele

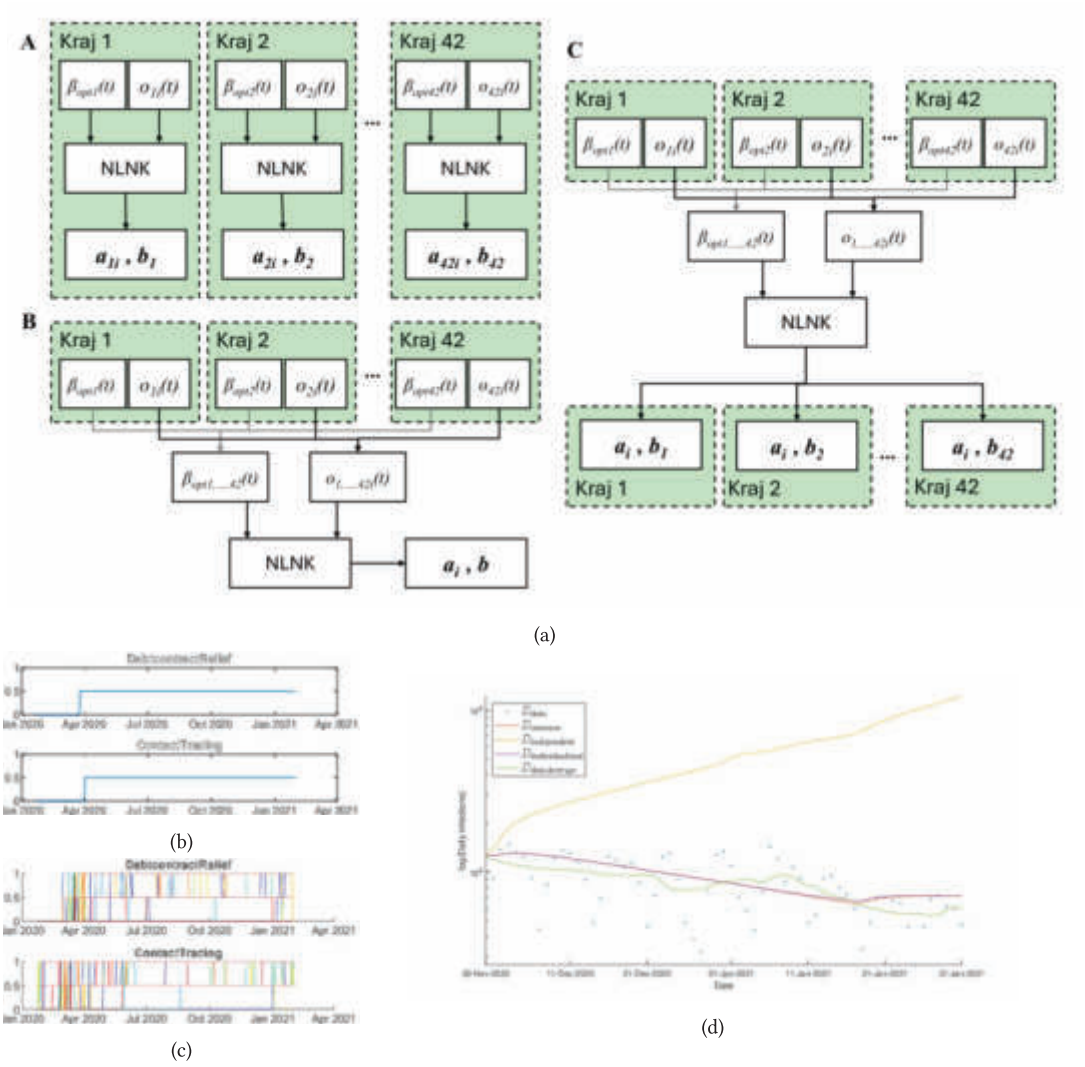
Mimo, iż niezależna analiza podgrup, jak linii komórkowych w pracy [1] czy typów raka w [2], może się wydać kusząca, ponieważ pozwala na zachowanie indywidualnego charakteru w heterogenicznych kohortach, nie zawsze jest ona wykonalna. Przyczyną może być na przykład nieestymowalność parametrów, co wymaga zastosowania alternatywnego podejścia.

[3] Wilk AM, Łakomiec K, Psiuk-Maksymowicz K i Fujarewicz K. Impact of government policies on the COVID-19 pandemic unraveled by mathematical modelling. Scientific Reports 12, 16987 (2022)

Cel. *Ocena wpływu niefarmakologicznych interwencji rządowych na rozprzestrzenianie się pandemii COVID-19.*

W pierwszej fazie pandemii COVID-19, przed wprowadzeniem szczepień, jedyną

metodą walki z rozprzestrzenianiem się wirusa SARS-CoV-2 były interwencje niefarmakologiczne, w tym obostrzenia takie jak zamknięcie szkół czy odwołanie wydarzeń publicznych, środki ekonomiczne oraz środki związane z systemem zdrowia [32]. Wiele badań poświęconych było oszacowaniu skuteczności poszczególnych interwencji, podstawową przeszkodą był jednak fakt, że zwykle wprowadzane one były w pakietach, przez co część była silnie skorelowana [33].



(a)

(b)

(c)

(d)

Fig. 3.2: Przewidywanie rozprzestrzeniania się pandemii COVID-19 w krajach europejskich przy pomocy indywidualizowanych modeli. (a) Trzy wykorzystywane podejścia estymacji parametrów: A — niezależne, B — wspólne, C — indywidualizowane (b) Nierozróżnialne funkcje przebiegu obostrzeń dla Polski (c) Te same obostrzenia z uwzględnieniem wszystkich analizowanych krajów. (d) Wyniki predykcji dla testowego okresu czasu dla Polski. Niebieskie punkty to dzienne zachorowania, zielona linia to 7-dniowa średnia ruchoma z liczby zachorowań, żółta, czerwona i fioletowa linia to odpowiednio predykcje dla modelu niezależnego, wspólnego i zindywidualizowanego.

Rozważmy model SEIR, jeden z najpopularniejszych kompartmentalnych modeli epidemiologicznych, w którym populacja podzielona jest na cztery kompartmenty:

- Susceptible, czyli osoby podatne na zakażenie,
- Exposed, czyli osoby, które miały kontakt z patogenem i zostały zakażone, ale same

jeszcze nie zarażają,

- Infected, czyli osoby zakażone, które mogą zarażać innych,
- Removed, czyli osoby, które nie są podatne na zakażenie, czy to przez nabycie odporności czy przez śmierć.

Model dla kraju k można zapisać w formie równań [34]:

$$\begin{cases} \dot{S}_k(t) = \frac{-\beta_k(t) S_k(t) I_k(t)}{N_k} \\ \dot{E}_k(t) = \frac{\beta_k(t) S_k(t) I_k(t)}{N_k} - k_{EI} E_k(t) \\ \dot{I}_k(t) = k_{EI} E_k(t) - k_{IR} I_k(t) \\ \dot{R}_k(t) = k_{IR} I_k(t) \end{cases} ; \quad k = 1, ..., K \tag{3.1}$$

gdzie warunki początkowe to $S_k(0) = N_k - I_0$, $E_k(0) = 0$, $I_k(0) = I_0$, $R_k(0) = 0$. Wielkość populacji N_k przyjęto jako stałą.

Parametr $\beta(t)$ określa szybkość rozprzestrzeniania się pandemii. Możemy przyjąć, że jego wartość zależna jest od pewnej wartości bazowej b, poziomów obostrzeń o_i oraz skuteczności tych obostrzeń a_i:

$$\beta(t) = b(1 - a_1 o_1(t) - a_2 o_2(t) - \cdots - a_r o_r(t)), \tag{3.2}$$

Aby zachować specyfikę poszczególnych krajów, naturalna byłaby estymacja parametrów dla każdego kraju niezależnie. W przypadku pokrywających się funkcji obostrzeń, przykładowo $o_i = o_j = o_{ij}$ (patrz Fig. 3.2b),

$$\beta(t) = b(1 - a_1 o_1(t) - \cdots - a_i o_i(t) - a_j o_j(t) - \cdots - a_r o_r(t)),$$

$$\beta(t) = b(1 - a_1 o_1(t) - \cdots - a_i o_{ij}(t) - a_j o_{ij}(t) - \cdots - a_r o_r(t)),$$

$$\beta(t) = b(1 - a_1 o_1(t) - \cdots - o_{ij}(t)(a_i + a_j) - \cdots - a_r o_r(t)),$$

Parametry a_i oraz a_j są wówczas nieestymowalne, ponieważ zmianę wartości jednego z nich można dowolnie kompensować wartością drugiego. Możliwym rozwiązaniem jest estymacja parametrów wspólnie dla całej kohorty krajów (Fig. 3.2c). Wprawdzie niweluje to problem nieestymowalności, ale zatraca się również indywidualny charakter poszczególnych obiektów.

Zaproponowano rozwiązanie kompromisowe, modelu indywidualizowanego, polegające na estymacji części parametrów razem dla wszystkich krajów (w tym przykładzie przyjęto, że wspólnymi parametrami są skuteczności obostrzeń a_i), a części niezależnie (tu: współczynnik bazowy b). Podejście takie pozwala na zmniejszenie liczby estymowanych parametrów, zachowanie pewnego odzwierciedlenia heterogeniczności obiektów kohorty i zmniejszenie ryzyka niewłaściwego uwarunkowania numerycznego problemu estymacji. Trzy wspomniane podejścia do estymacji parametrów przedstawiono schematycznie na Fig. 3.2a.

Najpierw, dla każdego kraju wyznaczono β_{opt} na podstawie wartości zachorowań raportowanych przez rządowe agencje. Następnie, parametry a_i oraz b estymowane są przy pomocy nieliniowej metody najmniejszych kwadratów (NLNK). Estymacji parametrów dokonano na okresie uczącym obejmującym czas od początku pandemii do końca listopada 2020, natomiast dwa kolejne miesiące (od początku grudnia 2020 do końca stycznia 2021, kiedy wprowadzono szczepienia) potraktowano jako okres testowy. Jako miarę jakości dopasowania przyjęto znormalizowany błąd średniokwadratowy (NRMSE). Dla Polski (Fig. 3.2d) modele wspólny i zindywidualizowany dały podobne rezultaty, bliskie obserwowanego przebiegu. Model niezależny natomiast obarczony jest znacznym błędem.

Tab. 3.1: Skuteczność poszczególnych strategii estymacji parametrów modelu SEIR (dla $k_{EI} = 0.2605$ i $k_{IR} = 0.1020$). Przedstawiono średnią i odchylenie standardowe (SD) dla NRMSE oraz liczbę krajów, w których dany model był najlepszy (w przypadku takiej samej wartości NRMSE dla dwóch modeli, liczone są oba).

Podsumowanie jakości predykcji			
	Wspólny model	**Niezależne modele**	**Indywidualizowane modele**
Średnie NRMSE	2.374	4.920	1.682
SD NRMSE	4.414	5.906	2.222
Najlepszy model	17	10	17

Dla całej kohorty krajów modele wspólny oraz indywidualizowane wykazują wysoką jakość predykcji, z przewagą modeli indywidualizowanych pod względem średniego błędu oraz rozrzutu błędów (Tabela 3.1). Mimo, iż zastosowanie wspólnego modelu również pozwoliło przezwyciężyć problemy numeryczne, jego ograniczona zdolność dopasowania do danych dla heterogenicznej kohorty skutkowała bardzo wysokimi wartościami błędów dla niektórych krajów (co odzwierciedla odchylenie standardowe NRMSE). Ponadto, w przypadku modeli niezależnych, wartości estymowanych parametrów często osiągały nierealistyczne poziomy. Przykładowo, interwencja „Zamknięcie szkół" miała współczynnik 0.225 dla modelu wspólnego, 0.149 dla modeli indywidualizowanych, oraz między -0.443 a 0.718 dla modeli niezależnych. Biorąc pod uwagę przyjętą funkcję rozprzestrzeniania się wirusa w zależności od obostrzeń (równanie 3.2) oznaczałoby to, że w niektórych krajach zamknięcie szkół przyspieszyło rozwój pandemii. Jakość dopasowania modeli niezależnych dało się poprawić poprzez zastosowanie ograniczeń wartości parametrów, ale jednym z założeń, którymi kierowano się w pracy, był brak konieczności wiedzy o dopuszczalnych wartościach parametrów *a priori*.

Wniosek. *Zastosowanie indywidualizowanego podejścia do estymacji parametrów modeli dla kohorty obiektów pozwala na ograniczenie ryzyka niewłaściwego uwarunkownia numerycznego bez całkowitej utraty indywidualnego charakteru obiektów.*

Wkład autorki w cytowaną pracę. *Przygotowanie zbioru danych – pozyskanie z zewnętrznych baz danych (zachorowania, obostrzenia, populacje), odfiltrowanie braków, zapis w formacie odpowiednim do dalszej analizy; estymacja wpływu obostrzeń – implementacja modelu wspólnego, niezależnych i indywidualizowanych; walidacja modeli na testowym przedziale czasowym – estymacja liczby zachorowań z modelu SEIR, ocena błędu dopasowania; wizualizacja wyników; przygotowanie manuskryptu.*

We wszystkich omówionych dotąd przypadkach, problem sprowadzał się do analizy i modelowania heterogenicznych obiektów oraz wykrycia ogólnych prawidłowości bez całkowitej utraty ich indywidualnego charakteru. Mimo występujących na wielu poziomach różnic pomiędzy obiektami, za każdym razem były dla nich dostępne analogiczne dane — zmiana liczby kopii genów, poziom ekspresji, profil metabolomiczny czy wektor poziomu obostrzeń. Zagadnienie heterogeniczności może jednak dotyczyć nie tylko samych obiektów, ale także danych, które je opisują.

Rozdział 4

Heterogeniczność strukturalna

Standardowo, dane wejściowe dla modeli predykcyjnych mają jednakową strukturę dla każdego analizowanego obiektu. W uproszczeniu, na podstawie wektora cech wyznaczana jest funkcja celu, a jej wartość pozwala na dokonanie predykcji, którą może być na przykład przypisanie kategorii w przypadku klasyfikacji czy wartości ciągłej dla regresji. Może się jednak zdarzyć, że dla pewnych obiektów możliwe jest pozyskanie więcej niż jednego wektora cech, przy czym ich liczba nie jest stała (patrz Fig 4.1). W rezultacie otrzymujemy dane o heterogenicznej strukturze, których wykorzystanie w modelu jest nietrywialne.

4.1 „Single-pixel approach" — klasyfikacja w oparciu o przestrzenną proteomikę

Jedną z dziedzin, w których coraz częściej pojawia się opisany rodzaj heterogeniczności strukturalnej, jest biologia molekularna. Rozwój nowoczesnych technologii pozwala na uzyskanie danych proteomicznych, metabolomicznych czy transkryptomicznych dla pojedynczych komórek [35, 36, 25, 37], lub dla siatki punktów w przypadku technologii przestrzennych [38, 39]. W zależności od technologii, dla każdej próbki otrzymujemy więc od kilkudziesięciu do nawet kilku tysięcy wektorów cech. Praca

[4] Kurczyk A., Gawin M., Chekan M., Wilk A., Łakomiec K., Mrukwa G., Frątczak K., Polanska J., Fujarewicz K., Pietrowska M. i Widlak, P. Classification of

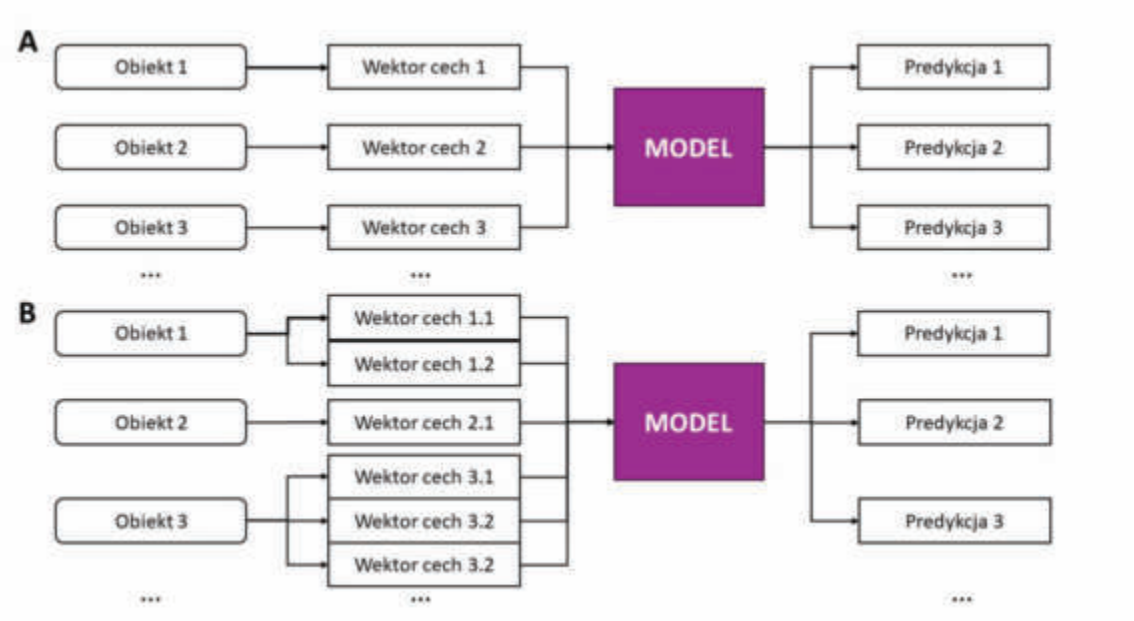


Fig. 4.1: Przykład heterogeniczności strukturalnej danych. A. Każdemu obiektowi odpowiada jeden wektor cech, można więc zastosować standardowe modele. B. Dla poszczególnych obiektów dostępne są analogiczne wektory cech, ale liczba wektorów jest różna. Modelowanie wymaga zatem dodatkowych kroków.

Thyroid Tumors Based on Mass Spectrometry Imaging of Tissue Microarrays; a Single-Pixel Approach. International journal of molecular sciences, 21(17), 6289, (2022)

opisuje sposób wykorzystania tego typu danych do klasyfikacji tkanek tarczycy.

Cel. *Klasyfikacja podtypów raka tarczycy w oparciu o heterogeniczne strukturalnie dane z obrazowania spektrometrią mas.*

Rak tarczycy jest najczęściej występującym nowotworem endokrynnym — w 2022 roku zdiagnozowano ponad 800 tysięcy przypadków na świecie [40], w Polsce natomiast rocznie występuje około 4700 przypadków (dane na 2023 rok) [41]. Rokowanie jest bardzo dobre dla większości pacjentów (w szczególności reprezentujących zróżnicowany podtyp), gorsze natomiast dla niskozróżnicowanych podtypów [42, 43]. Diagnostyka podtypu zwykle przeprowadzana jest na podstawie analizy preparatów histopatologicznych, może więc być subiektywna i w znacznym stopniu zależy od doświadczenia i aktualnej dyspozycji patologa, a także od jakości przygotowania preparatu i jego cha-

rakteru. Proces diagnostyczny mogłaby poprawić identyfikacja molekularnych biomarkerów charakterystycznych dla poszczególnych podtypów.

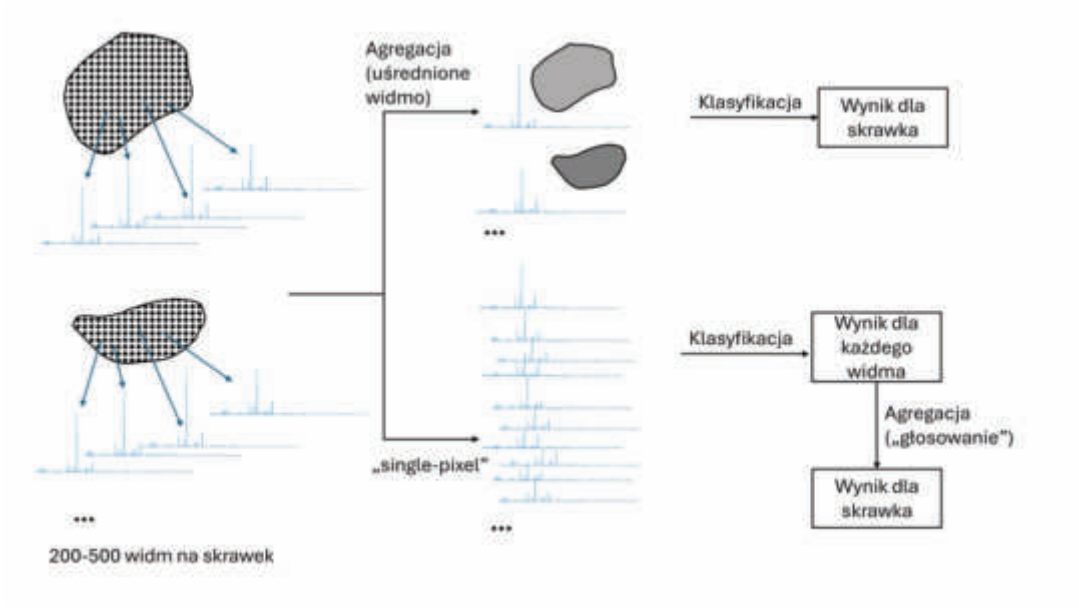
W pracy przeanalizowano macierze tkankowe zawierające łącznie 375 skrawków tkanek pochodzących od 134 pacjentów diagnozowanych w Narodowym Instytucie Onkologii im. Marii Skłodowskiej-Curie — Państwowym Instytucie Badawczym Oddziale w Gliwicach (NIO). Reprezentowały one siedem typów tkanki:

- NT (*normal thyroid*) — prawidłowa tkanka tarczycy,
- FA (*follicular adenoma*) — gruczolak pęcherzykowy, niezłośliwy guzek tarczycy,
- MTC (ang. *medullary thyroid carcinoma*) — rak rdzeniasty tarczycy, rzadko występujący, niskozróżnicowany złośliwy nowotwór tarczycy, o średnim rokowaniu,
- ATC (*anaplastic thyroid carcinoma*) — rak anaplastyczny tarczycy, rzadko występujący, niskozróżnicowany złośliwy nowotwór tarczycy, o słabym rokowaniu,
- FTC (*follicular thyroid carcinoma*) — rak pęcherzykowy tarczycy, często występujący, zróżnicowany złośliwy nowotwór tarczycy, o dobrym rokowaniu,
- PTC-CV (*papillary thyroid carcinoma classic variant*) — podstawowy, „klasyczny" wariant raka brodawkowatego tarczycy — często występującego, zróżnicowanego złośliwego nowotworu tarczycy o dobrym rokowaniu,
- PTC-FV (*papillary thyroid carcinoma follicular variant*) — wariant pęcherzykowy raka brodawkowatego tarczycy.

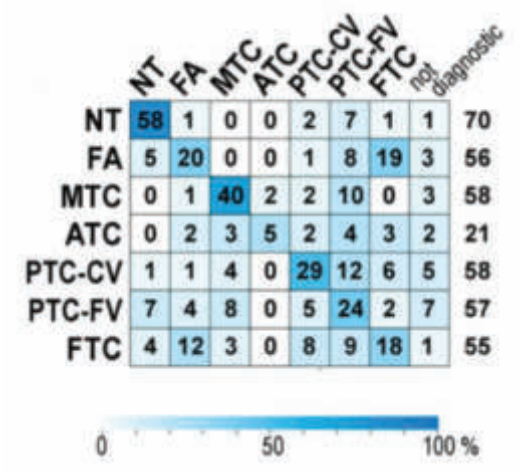
Preparaty zostały poddane obrazowaniu spektrometrią mas (*MS, mass spectrometry*) przy pomocy techniki MALDI-MSI, którą uzyskano średnio 360 widm MS na skrawek.

Na podstawie walidacji różnych typów modeli, do rozróżnienia widm wybrano maszynę wektorów wspierających (*SVM, support vector machine*) z liniową funkcją jądra. W celu uzyskania klasyfikacji wieloklasowej, problem został przekształcony do zadania 1vs1. Dla każdej pary klas został wytrenowany binarny model (łącznie 21 klasyfikatorów). Jeżeli decyzja była „jednogłośna", czyli pewna klasa otrzymała łącznie sześć głosów (we wszystkich modelach porównujących ją z pozostałymi klasami), była ona przypisywana do widma; w przeciwnym razie, widmo klasyfikowane było jako "niediagnostyczne". Rozwiązanie takie przyjęte zostało po konsultacji z patologiem, dla którego

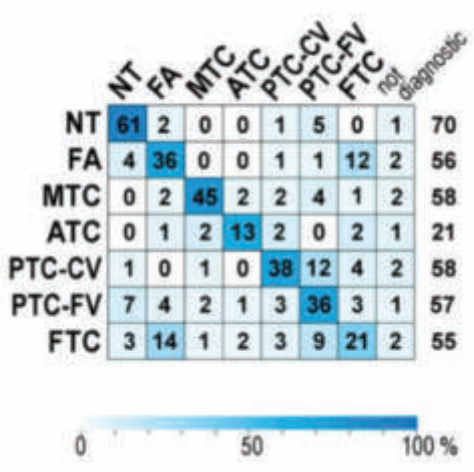
informacja, że próbka jest trudna do sklasyfikowania jest preferowana względem nieprawidłowej lub niepewnej diagnozy.



(a)



(b)



(c)

Fig. 4.2: Zastosowanie agregacji dla klasyfikacji danych pochodzących z obrazowania proteomicznego. (a) Uproszczony schemat działania. (b) Tablica pomyłek dla podejścia uśrednionego widma — w wierszach klasy eksperckie, w kolumnach przewidywane. (c) Tablica pomyłek dla podejścia „single-pixel" — w wierszach klasy eksperckie, w kolumnach przewidywane. Podejście „single-pixel" wykorzystujące pojedyncze widma pozwala na osiągnięcie wyższej dokładności klasyfikacji w porównaniu z uśrednionym widmem.

Ponieważ standardowe modele uczenia maszynowego wykorzystują jeden wektor

cech przypadający na obserwację, klasyfikacja całego skrawka wymaga zastosowania agregacji. Przetestowano dwie strategie (Fig. 4.2a):

1. Uśrednione widmo — dla skrawka generowane jest reprezentatywne widmo jako średnia arytmetyczna wszystkich pomiarów. Jest to rozwiązanie ułatwiające implementację klasyfikatora, w jego wyniku następuje jednak utrata większości pomiarów oraz informacji o heterogeniczności tkanki.

2. „Single-pixel" — strategia „pojedynczego piksela", w której model jest budowany na wszystkich dostępnych w zbiorze uczącym widmach. Podczas klasyfikacji nowego skrawka, każdy z kilkuset wchodzących w jego skład punktów jest przyporządkowany do klasy, a o ostatecznym wyniku decyduje „głosowanie", w którym przypisywana jest klasa większościowa (nawet, jeżeli jest to klasa „niediagnostyczne").

Zdolność predykcyjna modeli różni się w zależności od klasy (Fig. 4.2b i 4.2c). Stosunkowo dobrze rozróżniana jest prawidłowa tkanka tarczycy od tkanek nowotworowych (zarówno łagodnych jak i złośliwych), znaczne błędy obserwowane są natomiast dla podtypów raka brodawkowatego oraz między gruczolakiem i rakiem pęcherzykowym. Niezależnie od klasy można jednak zaobserwować, że podejście „single-pixel" osiąga wyższą jakość niż metoda oparta o uśrednione widma – dokładność klasyfikacji dla tych strategii to odpowiednio 0.67 i 0.52.

Wniosek. *Predykcja na podstawie zagregowanych wyników klasyfikacji dla wszystkich dostępnych pomiarów pozwala na osiągnięcie wyższej dokładności niż predykcja na podstawie wyniku klasyfikacji reprezentatywnego uśrednionego pomiaru.*

Wkład autorki w cytowaną pracę. *Analiza nienadzorowana zbioru danych (PCA), opracowanie i implementacja strategii uśrednionego widma oraz „single-pixel", implementacja systemu głosowania dla wieloklasowego modelu, zaprojektowanie schematu walidacji uwzględniającego powiązania między próbkami, wizualizacja.*

4.2 „Multi-lesion radiomics" — agregacja w radiomicznych modelach przeżycia

Inną (obok właściwości preparatu wykorzystanego do pomiaru) możliwą przyczyną dostępności różnej liczby wektorów cech opisujących poszczególne modelowane obiekty, jest zróżnicowany stan samych obiektów, na przykład pacjentów nowotworowych. Trzy następne prace wchodzące w skład niniejszego cyklu opisują kolejne kroki opracowania metodologii umożliwiającej wykorzystanie heterogenicznych strukturalnie danych w modelowaniu przeżycia.

[5] Wilk AM, Kozłowska E, Borys D, D'Amico A, Fujarewicz K, Gorczewska I, Debosz-Suwinska I, Suwinski R, Smieja J, Swierniak A. Radiomic signature accurately predicts the risk of metastatic dissemination in late-stage non-small cell lung cancer. Translational Lung Cancer Research 12(7):1372-1383. (2023)

Cel. *Przewidywanie ryzyka przerzutów odległych dla pacjentów z niedrobnokomórkowym rakiem płuc na podstawie danych klinicznych i radiomicznych.*

Rak płuc jest drugim co do zapadalności (po raku piersi) nowotworem na świecie, zajmuje jednak niekwestionowane pierwsze miejsce pod względem śmiertelności i odpowiada za ok. 20% zgonów spowodowanych nowotworami [40]. Pięcioletnie przeżycie w raku płuca kształtuje się na poziomie poniżej 20%, co można w znacznej mierze przypisać typowo późnej diagnozie spowodowanej niespecyficznymi objawami we wczesnej fazie choroby, a także znaczną inwazyjnością [44]. Granicznym momentem dla możliwości terapeutycznych jest pojawienie się przerzutów odległych, co odzwierciedlone jest między innymi w rekomendacjach Europejskiego Towarzystwa Onkologii Medycznej (*ESMO, European Society for Medical Oncology*), które podzielone są na raka niemetastatycznego i przerzutowego [45, 46].

W aktualnej praktyce, przerzuty wykrywane są na drodze regularnej kontroli przy pomocy badań obrazowych. Podejście takie ma pewne wady, związane między innymi z ograniczoną częstością wykonywania badań, wynikającą z przepustowości systemu zdrowia oraz bezpieczeństwa (badania obrazowe, mimo iż generalnie uznawane za nieinwazyjne, również nie mogą być wykonywane zbyt często), a także z możliwościami samego obrazowania. Biorąc pod uwagę rozdzielczość technologii takich jak pozytonowa

tomografia emisyjna (*PET - positron emission tomography)*, niewielkie, nawet centymetrowe ogniska mogą nie być wykrywalne [47]. Znalezienie czynników pozwalających na stratyfikację pacjentów pod względem ryzyka przerzutów odległych miałoby więc istotne znaczenie kliniczne.

W pracy [5] zbadana została możliwość wykorzystania cech klinicznych oraz cech pozyskanych z obrazowania PET/CT wykonanego pod kątem planowania radioterapii do przewidywania przerzutów odległych. Biorąc pod uwagę, że kohorty onkologiczne często są zbyt mało liczne, aby możliwa była konstrukcja wiarygodnego modelu opartego na głębokim uczeniu [48], zdecydowano się na podejście radiomiczne. Radiomika to gałąź nauki zajmująca się ekstrakcją z określonego obszaru obrazu, zwanego regionem zainteresowania (*ROI, region of interest*), liczbowych cech opisujących między innymi jego kształt i teksturę [49].

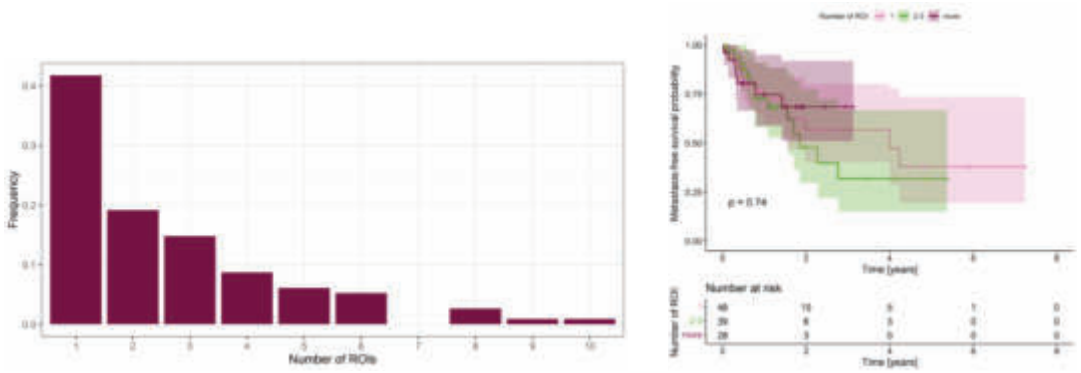
Przeanalizowano kohortę 115 pacjentów leczonych w NIO z powodu niedrobnokomórkowego raka płuca (*NSCLC, non-small cell lung cancer*). Artykuł zawiera podstawowe statystyki opisowe, analizę eksploracyjną oraz jednoczynnikową analizę statystyczną dostępnych cech. Analiza wieloczynnikowa została przeprowadzona na dwa sposoby — wykorzystując podejście oparte na klasyfikacji oraz na modelach przeżycia. Oba podejścia pokazały, że dostępne w badaniu cechy kliniczne, obejmujące wiek, płeć, podtyp, lokalizację guza, jego wielkość i stopień regionalnego rozsiania nowotworu według klasyfikacji TNM (*tumour, node, metastasis*), nie stanowią dobrych predyktorów ryzyka przerzutów. Korzystając z cech radiomicznych jesteśmy natomiast w stanie podzielić pacjentów na grupy istotnie statystycznie różniące się prawdopodobieństwem przeżycia wolnego od przerzutów odległych (*MFS, metastasis-free survival*).

Wniosek. *Cechy radiomiczne wyekstrahowane z obszaru zainteresowania obejmującego guz pierwotny wykazują potencjał predykcyjny dla przewidywania ryzyka przerzutów odległych w niedrobnokomórkowym raku płuc. Standardowo zbierane cechy kliniczne takie jak wielkość guza czy rozsiew do węzłów chłonnych nie wykazują jednak istotności dla przewidywania przerzutów.*

Wkład autorki w cytowaną pracę. *Analiza nienadzorowana zbioru danych —- PCA, analiza korelacji; statystyka opisowa, analiza jednoczynnikowa, przewdywanie przeżycia wolnego od przerzutów i wolnego od zdarzeń z wykorzystaniem metod klasyfikacji, kon-*

strukcja finalnego regresyjnego modelu przeżycia, przygotowanie pierwszej wersji manuskryptu.

Jednym z wniosków z analizy danych klinicznych jest stwierdzenie, że regionalne rozsianie nowotworu, wyrażone częściowo przez człon „N" w klasyfikacji TNM, nie wpływa na ryzyko przerzutów odległych. Obserwacja ta, jakkolwiek z pozoru nieintuicyjna, pokrywa się z aktualną teorią na temat mechanizmu metastazy w raku płuc, według której za powstawanie przerzutów odległych odpowiedzialny jest przede wszystkim rozsiew z guza pierwotnego, nie z węzłów chłonnych [50]. Oprócz guza pierwotnego i ewentualnie zajętych węzłów chłonnych, mogą jednak występować również dodatkowe ogniska raka (mówimy wtedy o nowotworze wieloogniskowym), które nie są odzwierciedlone w klasyfikacji TNM. Wszystkie te gromadzenia mogą stanowić regiony zainteresowania (Fig. 4.3). Chociaż sama ich liczba nie jest czynnikiem prognostycznym dla przerzutów odległych (Fig. 4.3b), postawiono hipotezę, że uwzględnienie wszystkich regionów w modelowaniu przeżycia może poprawić zdolność predykcyjną modelu.



(a) (b)

Fig. 4.3: U większości pacjentów w badanej kohorcie w obrazie PET płuc dało się zaobserwować przynajmniej dwa gromadzenia. (a) Częstość występowania poszczególnych liczb gromadzeń — regionów zainteresowania. (b) Wartość predykcyjna liczby ROI dla przewidywania przerzutów.

Podobnie jak w opisanym wcześniej przypadku klasyfikacji danych z obrazowania molekularnego [4], z perspektywy modelowania wyzwanie sprowadza się do heterogenicznej struktury danych, czyli dostępności różnej liczby (tu od jednego do dziesięciu) wektorów cech radiomicznych przypadających na pacjenta. Dla rozważanego problemu przewidywania ryzyka przerzutów, należy jednak wziąć pod uwagę pewne dodatkowe aspekty.

1. Inny typ zadania. W odróżnieniu od klasyfikacji, gdzie wynikiem jest przypisanie obiektu do z góry znanych grup (klas), modele przeżycia wpisują się w problem regresji. Wynikiem predykcji jest ciągła liczbowa wartość, interpretowana jako ryzyko wystąpienia analizowanego zdarzenia.

2. Nierównoważność wektorów cech. W przestrzennych technikach molekularnych, każdy pomiar dotyczy pewnego punktu siatki, którą „pokryta" jest tkanka. Można zatem przyjąć, że wszystkie wektory cech są równoważne i nie jest możliwe uporządkowanie ich według żadnego logicznego kryterium. Tymczasem dla obrazowania medycznego, jedną z oczywistych cech odróżniających wektory cech jest wielkość (objętość) regionu zainteresowania, z jakiego je wyznaczono.

3. Interpretowalność uśrednionego wektora cech. Dla danych transkryptomicznych czy proteomicznych, wartości cech można interpretować jako stężenie RNA lub białka. Uśrednione wartości z wielu punktów zachowują swoją interpretację, jako stężenie dla większej liczby komórek. W przypadku radiomiki, wyznaczane cechy są ściśle powiązane z analizowanym regionem zainteresowania — opisują między innymi jego kształt czy teksturę. Uśrednienie wartości z kilku ROI sprawia, że wiele cech radiomicznych traci interpretowalność.

Kolejna praca stanowi pewnego rodzaju dowód słuszności koncepcji, że mimo powyższych względów, agregacja wektorów cech albo wyników predykcji może być rozwiązaniem problemu heterogeniczności struktury danych dla radiomicznych modeli przeżycia.

[6] Wilk AM, Kozłowska E, Borys D, D'Amico A, Gorczewska I, Debosz-Suwińska I, Gałecki S, Fujarewicz K, Suwiński R i Świerniak A. Improving the Predictive Ability of Radiomics-Based Regression Survival Models Through Incorporating Multiple Regions of Interest. W: Strumiłło, P., Klepaczko, A., Strzelecki, M., Bociąga, D. (eds) The Latest Developments and Challenges in Biomedical Engineering. PCBEE 2023. Lecture Notes in Networks and Systems, vol 746. Springer, Cham. (2024)

Cel. *Wykorzystanie wszystkich dostępnych regionów zainteresowania z obrazowania PET/CT w modelu przeżycia przewidującym ryzyko przerzutów.*

W badaniu wykorzystano kohortę 115 pacjentów z niedrobnokomórkowym rakiem płuca, która została przeanalizowana w pracy [5]. Tym razem, oprócz guza pierwotnego, na obrazach PET/CT zostały przez radiologa okonturowane wszystkie gromadzenia (Fig. 4.4a), które następnie posłużyły do wyznaczenia cech radiomicznych.

Analogicznie jak w przypadku obrazowania MALDI [4], zaproponowane metody uwzględnienia w modelu wszystkich wektorów cech opierają się na dwóch możliwych strategiach (schematycznie przedstawionych na fig. 4.4a).

1. **Agregacja ROI**. Polega na wyborze lub wygenerowaniu jednego wektora cech, który będzie reprezentować danego pacjenta, czyli sprowadzeniu danych do homogenicznej struktury zwykle wykorzystywanej w analizie. W wyniku modelowania, dla każdego pacjenta uzyskujemy jedną wartość ryzyka. Zastosowane w pracy metody agregacji ROI to:

    - *largestROI* — w modelu wykorzystane jest największe ognisko, odpowiadające guzowi pierwotnemu. Jest to podejście wykorzystywane w większości badań radiomicznych poświęconych nowotworom.
    - *randomROI* — w modelu wykorzystane jest losowe ognisko.
    - *arithmeticMeanROI* — generowany jest reprezentatywny wektor cech, jako średnia arytmetyczna wszystkich przypadających na danego pacjenta.
    - *weightedMeanROI* — generowany jest reprezentatywny wektor cech, jako średnia ważona wszystkich przypadających na danego pacjenta, gdzie wagą jest objętość odpowiadających ROI.

2. **Agregacja ryzyka**. Model stosowany jest niezależnie dla wszystkich ROI, w wyniku czego dla pojedynczego pacjenta może występować kilka wartości ryzyka. Następnie wyniki te są agregowane tak, aby uzyskać pojedynczą wartość. Zastosowane metody to:

    - *allROIMin* — pacjentowi przypisywana jest najniższa z uzyskanych dla jego ROI wartości ryzyka.
    - *allROIMax* — pacjentowi przypisywana jest najwyższa z uzyskanych dla jego ROI wartości ryzyka.

- *allROIMean* — pacjentowi przypisywana jest średnia arytmetyczna z uzyskanych dla jego ROI wartości ryzyka.

Do analizy przeżycia wykorzystano regularyzowaną regresję Coxa (Coxnet), zawierającą w sobie również selekcję cech. Aby uzyskać bardziej miarodajne porównanie pomiędzy metodami, zastosowano kroswalidację typu Monte Carlo (*MCCV, Monte Carlo Cross-validation*), z 1000 losowymi podziałami zbioru pacjentów na zbiór uczący i testowy w stosunku 2:1 (w przypadku metod agregacji ryzyka ostatecznie wykorzystywane zbiory zawierają wszystkie ROI od odpowiednich pacjentów, mają więc różną liczność w poszczególnych iteracjach). Jako wskaźnik jakości predykcji przyjęto c-indeks Harrella, jedną z najczęściej stosowanych miar dopasowania modeli przeżycia. Jest to współczynnik zgodności opisujący stosunek liczby „zgodnych" par obserwacji (czyli takich, w których obserwacji o wyższym ryzyku odpowiada krótszy czas do zdarzenia) do wszystkich możliwych par. Określany jest jako przeżyciowy odpowiednik pola pod krzywą ROC — przyjmuje wartości od 0 do 1, przy czym 0.5 oznacza model losowy, a 1 oznacza nieomylny model.

Tab. 4.1: Porównanie metod przetwarzania wielu ROI w modelu przeżycia.

Metoda	Typ metody	Mediana c-indeksów	Min c-indeks	Max c-indeks
largestROI	agregacja ROI	0.581	0.206	0.862
randomROI	agregacja ROI	0.534	0.217	0.828
arithmeticMeanROI	agregacja ROI	0.557	0.290	**0.892**
weightedMeanROI	agregacja ROI	0.592	0.206	0.835
allROIMin	agregacja ryzyka	0.566	0.193	0.817
allROIMax	agregacja ryzyka	**0.617**	0.349	0.880
allROIMean	agregacja ryzyka	0.616	**0.369**	0.827

Tabela 4.1 przedstawia podsumowanie osiągniętych przez poszczególne metody jakości predykcji. Dla standardowego podejścia, opartego na uwzględnieniu jedynie guza pierwotnego, mediana c-indeksów z 1000 iteracji MCCV wyniosła 0.581. Gorsze wyniki uzyskano dla metod *randomROI*, *arithmeticMeanROI* oraz *allROIMin*. Nie jest to zaskakujące, ponieważ w przypadku losowego ROI ryzyko dla pacjenta mogło zostać oszacowane na przykład na podstawie bardzo małego gromadzenia, dla którego cechy teksturowe są mniej informatywne [51, 52]. Dla ROI wygenerowanego przy pomocy

średniej arytmetycznej, ROI o różnych rozmiarach traktowane są tak samo, co prowadzi do obniżenia jakości. Lepiej, przewyższając medianą c-indeksów model oparty tylko na guzie pierwotnym, zadziałała metoda *weightedMeanROI*, w której wektory cech skalowane były względem wielkości ROI. Najwyższą jakość predykcji uzyskano dla metod *allROIMax* oraz *allROIMean*. Co więcej, metody te okazały się być bardziej odporne na próbkowanie, osiągając wyższe minimalne c-indeksy.

Wniosek. *Różną liczbę wektorów cech radiomicznych wyznaczonych z gromadzeń wykrytych w obrazowaniu PET/CT można uwzględnić w modelu przeżyciowym poprzez zastosowanie agregacji ROI lub agregacji ryzyka.*

Wkład autorki w cytowaną pracę. *Konceptualizacja, opracowanie i implementacja metod agregacji; analiza statystyczna, testowanie modeli przeżycia, wizualizacja, przygotowanie manuskryptu.*

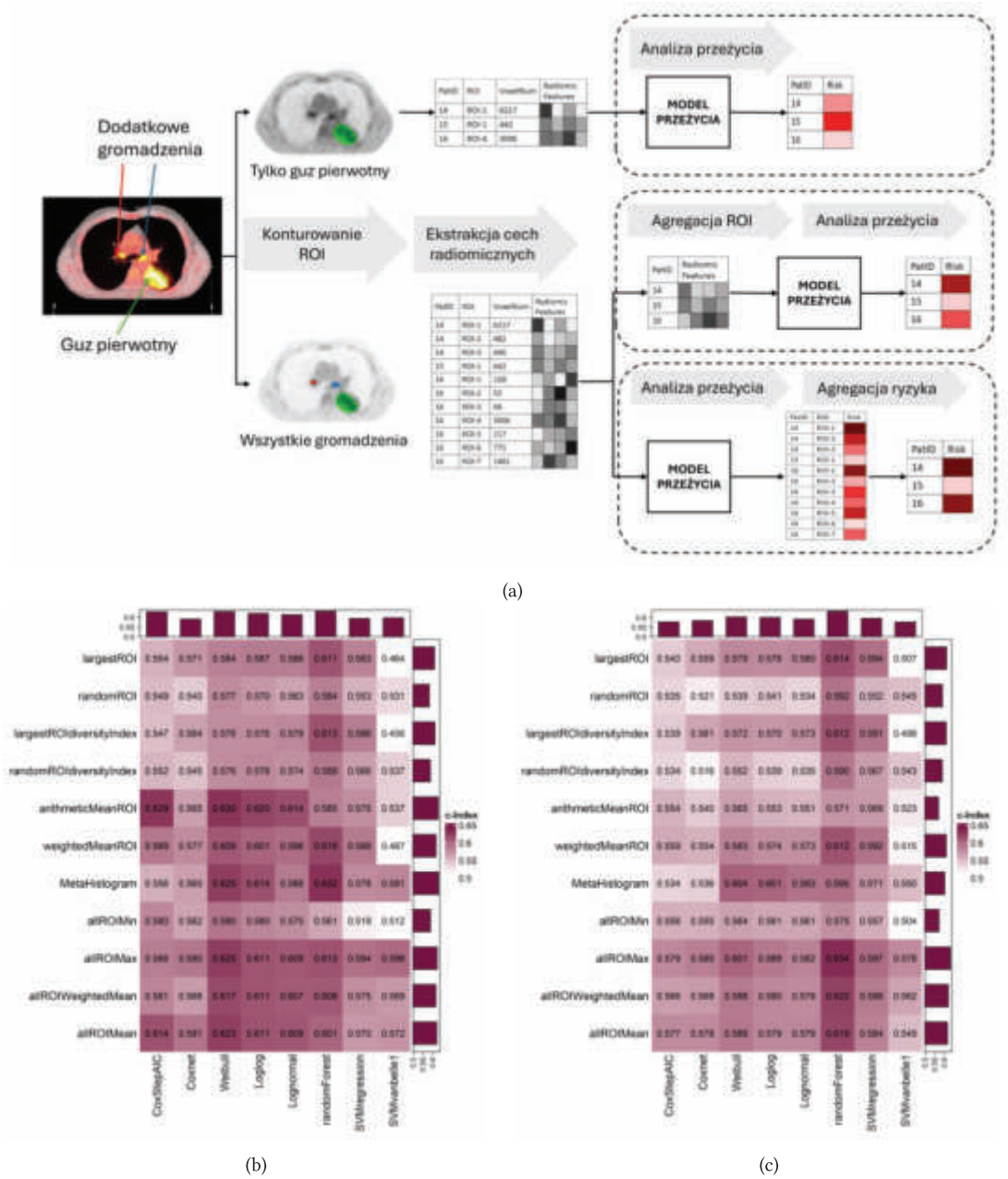


Fig. 4.4: „Multi-lesion radiomics" — uwzględnienie różnej liczby regionów zainteresowania w modelach przeżycia przez zastosowanie odpowiednich metod agregacji. (a) Główna idea. (b) Wyniki 1000-krotnej kroswalidacji dla zbioru PET — mediany c-indeksów dla poszczególnych schematów (c) Wyniki 1000-krotnej kroswalidacji dla zbioru PET_CT — mediany c-indeksów dla poszczególnych schematów. Niezależnie od zbioru i modelu, jakość osiąganą dla jednego ROI odpowiadającego guzowi pierwotnemu da się poprawić wykorzystując wszystkie gromadzenia.

Uzyskane wyniki były na tyle obiecujące, że postanowiono rozszerzyć wykonane

badanie, sprawdzając, czy podobna tendencja utrzyma się dla innych modeli przeżycia. Zastanawiająca była także najwyższa jakość predykcji osiągnięta przez metodą *allROIMax* oraz niska jakość dla metody *allROIMin*. W przypadku pacjentów o większej niż jeden liczbie gromadzeń naturalne jest, że wynikiem predykcji będą zarówno wyższe jak i niższe wartości ryzyka. Przypisanie takim pacjentom najwyższej z uzyskiwanych wartości ryzyka okazało się najskuteczniejszą strategią, mimo, że sama liczba ROI nie jest predyktorem ryzyka przerzutów (Fig. 4.3b). Postawiono zatem hipotezę, że nie liczba, a zróżnicowanie pomiędzy ROI, może być związane z wyższym ryzykiem przerzutów odległych.

[7] Wilk AM, Swierniak A, d'Amico A, Suwiński R, Fujarewicz K i Borys D. Towards the use of multiple ROIs for radiomics-based survival modelling: finding a strategy of aggregating lesions. Preprint arXiv: 2405.17668 [stat.AP]. (2024) Praca w recenzji w czasopiśmie Computers in Biology and Medicine

Cel. *Ocena heterogeniczności pomiędzy ROI, rozwój metod uwzględnienia wielu ognisk nowotworu w modelach przeżyciowych, sprawdzenie, czy agregacja poprawia jakość predykcji dla różnych modeli i parametrów ekstrakcji cech radiomicznych oraz porównanie z metodami opisanymi w literaturze.*

Badania ponownie przeprowadzono na opisanym wyżej zbiorze danych dla pacjentów NSCLC. Przetestowano natomiast dwa zbiory cech radiomicznych (dla obu przeprowadzono standaryzację obrazów PET względem znormalizowanej wartości absorpcji uwzględniającej masę ciała SUVbw) — w oryginalnej rozdzielczości PET (zbiór PET) oraz w rozdzielczości interpolowanej do CT przy pomocy metody najbliższych sąsiadów (zbiór PET_CT).

Ocena heterogeniczności pomiędzy ROI wymagała zdefiniowania liczbowych indeksów zróżnicowania. Jako że cechy radiomiczne mogą przyjmować ujemne wartości, niemożliwe było zastosowanie indeksów opartych na entropii, skupiono się zatem na różnych miarach odległości:

1. odległość Canberra,
2. odległość euklidesowa,
3. odległość Minkowskiego,

4. odległość Kendalla (oparta na współczynniku korelacji Kendalla),
5. odległość Spearmana (oparta na współczynniku korelacji Spearmana; ze względu na znaczne różnice rzędów wielkości między cechami współczynnik korelacji Pearsona okazał się niepraktyczny).

Dla każdego pacjenta przyjęto, że indeks heterogeniczności ma wartość zero jeżeli dostępny jest tylko jeden wektor cech (czyli jedno ROI), oraz wartość odpowiadającą średniej odległości dla każdej unikalnej pary ROI w przeciwnym przypadku. Dla analogii względem testu dla liczby ROI (Fig. 4.3b) ponownie podzielono zbiór pacjentów na trzy podgrupy (według tercyli wartości indeksów) i porównano przy pomocy testu log-rank. Dla zbioru PET grupy wydzielone na przykład na podstawie odległości euklidesowej różniły się istotnie MFS (p=0.026).

Jednym z celów pracy [7] było porównanie zaproponowanych metod z istniejącymi. Szczegółowe przeszukanie literatury pokazało jednak, że jakkolwiek idea „multi-region radiomics" jest znana i zyskuje coraz większą popularność, jej rozumienie jest przeważnie inne niż w opisywanym badaniu. Polega bowiem na wykorzystaniu cech radiomicznych wyekstrahowanych z wielu regionów, które są jednak analogiczne dla każdego pacjenta i stanowią raczej modyfikację definicji ROI (przykładowo guz i obszar okołoguzowy, lub wydzielone podobszary guza) niż odrębne ogniska. Rozumiane w ten sposób dodatkowe obszary nie generują zatem nowych wektorów cech, a jedynie zwiększają liczbę cech dostępnych dla każdego pacjenta. W jedynej znalezionej pracy dotyczącej integracji radiomiki dla wielu niezależnych obszarów, Zhao i współautorzy [53] wprowadzili metodę „meta histogramu". Szeregując wartości cechy radiomicznej w kolejności od największego do najmniejszego ROI, tworzony jest „meta histogram", dla którego następnie wyznaczane są średnia, wariancja, skośność, kurtoza, energia, entropia i suma, które wprowadzane są jako cechy do modelu. Metoda ta została przez autorów wykorzystana w problemie klasyfikacji dla pacjentów z gruczolakorakiem płuc, u których występowały co najmniej dwa ogniska przerzutowe.

Ostatecznie, oprócz metod agregacji opisanych w pracy [6], uwzględniono także metody:

- *largestROIdiversityIndex* — agregacja ROI, w której do wektora cech wyznaczonych dla guza pierwotnego konkatenowane są opisane wyżej indeksy heterogeniczności,

- *randomROIdiversityIndex* — analogicznie jak w poprzedniej metodzie, indeksy heterogeniczności dodawane są jednak do cech wyekstrahowanych z losowego ROI,
- *MetaHistogram* — metoda opisana w pracy [53], którą również można zaklasyfikować jako agregację ROI,
- *allROIWeightedMean* — agregacja ryzyka, w której pacjentowi przypisywana jest średnia ważona ryzyka dla wszystkich jego ROI, gdzie wagą jest objętość ROI.

Podobnie jak wcześniej, zastosowano kroswalidację Monte Carlo z 1000 iteracji, porównując łącznie osiem modeli przeżycia:

- *CoxStepAIC* — regresja Coxa z selekcją krokową na podstawie kryterium informacyjnego Akaike (*AIC, Akaike Information Criterion*), czyli metoda najczęściej stosowana w literaturze medycznej,
- *Coxnet* — regularyzowana regresja Coxa,
- *Weibull* — boosting oparty na modelu Weibulla,
- *Loglog* — boosting oparty na modelu log-log,
- *Lognormal* — boosting oparty na modelu log-normal,
- *randomForest* — losowy las przeżyciowy,
- *SVMregression* — przeżyciowa wersja maszyny wektorów wspierających z modelem regresji i addytywną funkcją jądra,
- *SVMvanbelle1* — przeżyciowa wersja maszyny wektorów wspierających z modelem van Belle i addytywną funkcją jądra.

Wyniki kroswalidacji (mediany c-indeksów) przedstawione są na Fig. 4.4b dla zbioru PET oraz Fig. 4.4c dla zbioru PET_CT. Jakość predykcji różni się w zależności od zbioru, czyli od metody ekstrakcji cech radiomicznych, oraz od modelu. Mimo to, stosując metody agregacji pozwalające na uwzględnienie informacji o wszystkich obecnych w obrazowaniu gromadzeniach, za każdym razem możliwe jest uzyskanie jakości wyższej niż dla metody *largestROI* w obrębie tego samego zbioru i modelu.

Wniosek. *Zastosowanie agregacji ROI lub agregacji ryzyka pozwala na osiągnięcie wyższej jakości predykcji niż w modelu opartym jedynie na guzie pierwotnym.*

Wkład autorki w cytowaną pracę. *Konceptualizacja — określenie problemu badawczego, przegląd istniejących rozwiązań, zaprojektowanie badania; Metodologia — opracowanie i implementacja metod agregacji; analiza statystyczna, testowanie modeli przeżycia, wizualizacja, przygotowanie manuskryptu.*

Rozdział 5

Podsumowanie

Heterogeniczność jest nieodłącznym aspektem danych biomedycznych. Wymusza ona nieustanne poszukiwanie kompromisu pomiędzy uogólnieniem zapewniającym wyższą moc wnioskowania statystycznego, a jak najdokładniejszym uchwyceniem niuansów i różnic, tak istotnych między innymi dla personalizacji terapii. Stanowi z jednej strony wyzwanie dla możliwości analizy, z drugiej motor napędowy rozwoju technologii i klucz do pełnego zrozumienia zjawisk i procesów biologicznych.

W niniejszej rozprawie, stanowiącej cykl siedmiu artykułów, przedstawiono badania, stanowiące oryginalny wkład autorki w dziedzinę inżynierii biomedycznej, które pozwoliły na realizację następujących celów:

1. Opis heterogeniczności występującej na poziomie komórkowym [1], tkankowym [2] oraz populacyjnym [3], dla obiektów scharakteryzowanych przez różne typy danych, w tym genomiczne, transkryptomiczne, metabolomiczne czy obrazowe [5].

2. Opracowanie algorytmu indywidualizowanej estymacji parametrów modelu [3], pozwalającego na uzyskanie niższych błędów predykcji niż dla niezależnej lub wspólnej estymacji.

3. Przedstawienie problematyki strukturalnej heterogeniczności danych, w szczególności sytuacji, w której liczba dostępnych wektorów cech różni się dla poszczególnych obiektów, na przykład w przestrzennych badaniach molekularnych [4] oraz w obrazowaniu medycznym regionalnie zaawansowanego nowotworu [6, 7].

4. Zaproponowanie metod agregacji wektorów cech lub wyników modelu, znajdujących zastosowanie zarówno dla modeli klasyfikacyjnych [4] jak i przeżyciowych [6, 7].

Opracowane algorytmy uczenia maszynowego mogą być stosowane dla różnych typów danych. Nie ograniczają się także do konkretnego problemu biologicznego czy modelu, dzięki czemu posiadają szeroki potencjał zastosowania w badaniach, w których ważne jest uwzględnienie heterogeniczności, między innymi:

- w badaniach epidemiologicznych,
- dla kohort pacjentów onkologicznych (lub w innych jednostkach chorobowych)
- w technikach molekularnych, generujących informacje dla pojedynczych komórek lub punktów w przestrzeni.

Zaprezentowane wyniki badań potwierdzają słuszność postawionych w rozprawie tez.

Indywidualizacja modeli dla kohorty obiektów pozwala zmniejszyć ryzyko złego uwarunkowania numerycznego zadania estymacji parametrów. W modelu pandemii COVID-19 uwzględniającym wpływ interwencji niefarmakologicznych na szybkość rozprzestrzeniania się wirusa, silnie skorelowane poziomy obostrzeń prowadziły do trudności w estymacji odpowiadających im parametrów dla pojedynczego kraju. Mimo, że problem nieestymowalności da się w tym przypadku rozwiązać konstruując wspólny model dla kohorty europejskich krajów, uogólnienie prowadzi do wysokich wartości błędów dla części krajów. Zastosowanie indywidualizowanych modeli pozwoliło na poprawę uwarunkowania numerycznego, jednocześnie umożliwiając lepsze dopasowanie do danych niż wspólny model.

Zastosowanie agregacji w przypadku różnej liczby wektorów cech dla poszczególnych obiektów skutkuje poprawą jakości predykcji modelu względem wykorzystania tylko jednego wektora na obiekt. W klasyfikatorze podtypu tkanki tarczycy wykorzystującym dane z obrazowania proteomicznego, agregacja predykcji dla pojedynczych widm pozwoliła na uzyskanie jakości klasyfikacji znacznie wyższej, niż wykorzystanie w predykcji jedynie uśrednionego widma. Również dla modeli przeżyciowych przewidujących ryzyko przerzutu odległego w niedrobnokomórkowym raku

płuc, wykorzystując w modelu wszystkie dostępne gromadzenia uzyskano istotną poprawę jakości predykcji względem takich samych modeli zbudowanych tylko na podstawie guza pierwotnego.

Stosując zaproponowane strategie, czy to indywidualizacji estymacji parametrów, czy agregacji wyników modelowania, uzyskano znacznie lepsze wyniki w porównaniu z podejściem opartym na maksymalnej generalizacji i uśrednieniu. Można więc uznać, że heterogeniczność nie jest wrogiem, a raczej sprzymierzeńcem w analizie danych biomedycznych.

Rozdział 6

Introduction

Humanity. Hundreds of societies, living in different environments, organized in different ways. Eight billion people, of different ages, different states of health, different lifestyles. In each person, dozens of tissues, tens of trillions of cells — own and bacterial; in each of them thousands of different metabolites, proteins, transcripts. And in each cell, DNA, composed of billions of base pairs, which can undergo mutations and modifications. Emerging at every possible level, heterogeneity is inherent in life and the subject of much research in the biomedical field.

Heterogeneity has important clinical implications, especially in cancer [8, 9, 10, 11]. Since no two patients are the same, also every cancer is different. Moreover, the presence of heterogeneous and constantly evolving cell populations (both cancer cells and cells forming the tumour microenvironment) determines the rate of disease progression, propensity to metastasize, and even sensitivity or resistance to therapy. New molecular and imaging capabilities, combined with translational research being conducted worldwide, are leading to the identification of increasingly narrowed cancer subtypes, as can be seen, for example, in breast cancer [12, 13, 14]. Simultaneously, the trend in the therapeutic approach is changing, moving towards personalized medicine, in which genetic, epigenetic, transcriptomic or proteomic characteristics allow for the selection of treatment with the highest effectiveness and the least possible adverse effects [15, 16, 17, 18].

The diversity of cells, tissues or patients themselves is not the only source of heterogeneity in biomedical data. Studies use data from multiple modalities, such as clinical, molecular or imaging, characterized by varying structure, dimensionality or type. [19,

20]. Analysing multi-modal data requires the use of integration and fusion methods, with the added challenge that assays for each modality are often performed on different samples. Also contributing to the heterogeneity of medical data used in machine learning is the fact that they come from different sources, such as several hospitals [21, 22, 23]. If data for a cohort of patients are collected over many years, variability may also stem from evolving diagnostic tools or therapeutic guidelines.

Both the heterogeneity of the studied objects and the data that describe them give rise to the need for dedicated methods in their modelling. This dissertation comprises a series of seven scientific articles, illustrating the author's own experience with the analysis of heterogeneous biomedical data. The first part outlines examples of heterogeneity at different levels, from the cellular level to the populations of countries, and the challenges that arise in analysing heterogeneous data of different modalities. The second part of the dissertation is dedicated to modelling heterogeneous data, especially when the analysed entities are described by a different number of feature vectors.

Rozdział 7

Aims and theses of the work

The aims of this work are to:

1. Describe the heterogeneity occurring at different levels in biomedical data.
2. Develop algorithms to analyze and model heterogeneous populations and cohorts.
3. Present the problem of structural data heterogeneity.
4. Propose aggregation methods that enable the use of information from a different number of feature vectors corresponding to the different objects being modelled.

In the dissertation the following theses are presented:

1. Individualizing models for a cohort of entities reduces the risk of poor numerical conditioning of the parameter estimation task.
2. The use of aggregation in the case of a different number of feature vectors for each entity results in improved model prediction quality relative to the use of only one vector per object.

Rozdział 8

Heterogeneity on different levels

Whether we are performing epidemiological studies, cohort studies, tissue or even cell analysis, in biomedicine we will almost always be dealing with heterogeneous populations. These may be populations of people who differ in health status, socioeconomic background or environmental factors; they may be populations of different species of bacteria that make up the microbiota; they may be populations of different cell types in tissues. Even if we analyze cells of a single cell line, within a culture we will observe cells with different molecular profiles resulting, for example, from desynchronization in the cell cycle phase [24, 25, 26]. Although such a state reflects the heterogeneity found in the human body, it can obscure changes associated with the biological effect of interest during analysis.

8.1 Cellular level

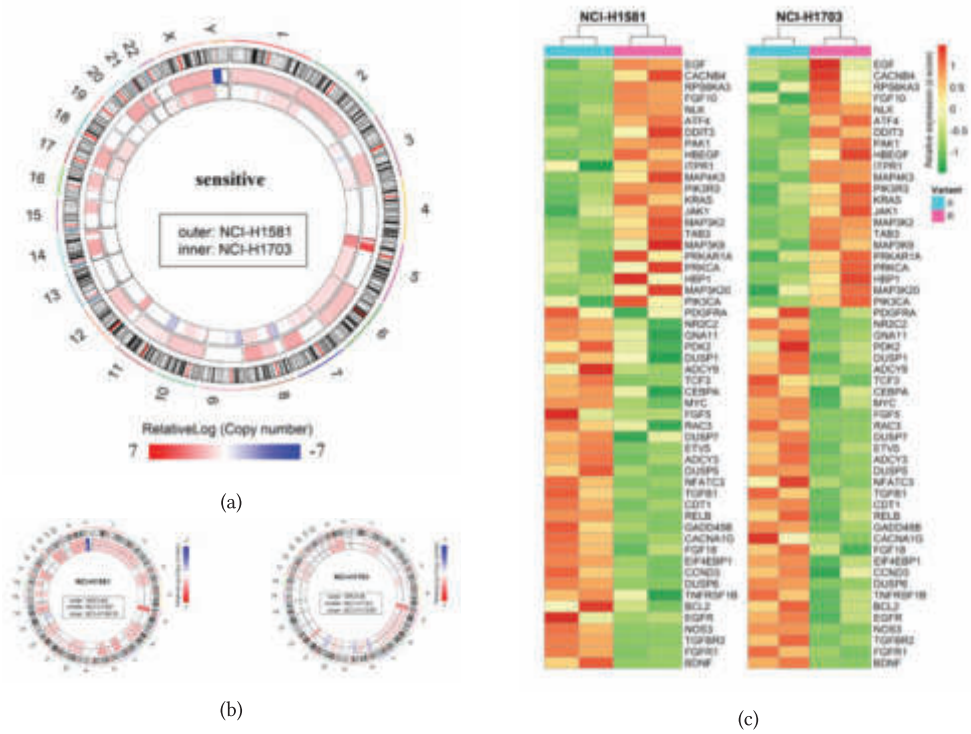
[1] Zarczynska I, Gorska-Arcisz M, Cortez AJ, Kujawa KA, Wilk AM, Skladanowski AC, Stanczak A, Skupinska M, Wieczorek M, Lisowska KM, Sadej R, Kitowska K. p38 Mediates Resistance to FGFR Inhibition in Non-Small Cell Lung Cancer. Cells. Nov 30;10(12):3363. (2021)

Aim. *Identification of molecular characteristics of acquired FGFR inhibitor resistance in non-small cell lung cancer cell lines.*

Fibroblast growth factor receptor (FGFR) inhibition is a promising therapeutic direc-

tion in lung cancer [27, 28]. Before they are approved for use in the clinic, new drugs undergo a series of clinical trials and preclinical studies, for example on cell lines. In the case of targeted therapies, it is extremely important to determine which group of patients has the best chance of responding to treatment — FGFR amplification status alone is not a perfect criterion. In the paper, [1] cell lines were used to study acquired resistance to the CPL304110 inhibitor in non-small cell lung cancer (NSCLC).

From a panel of NSCLC cell lines exhibiting amplification of the FGFR1 gene, two were selected that proved most sensitive to the inhibitor: NCI-H1581 and NCI-H1703. Using increasing concentrations of CPL304110, resistant variants of the lines were derived, and both the sensitive and resistant variants were then subjected to molecular testing: aCGH (*Array Comparative Genomic Hybridization*, a microarray-based technique for determining gene copy number variation) and RNAseq (RNA sequencing, a high-throughput technique that allows for transcript level quantification).



(a) (b) (c)

Fig. 8.1: Heterogeneity observed in non-small cell lung cancer cell lines. (a) Gene copy number variation for sensitive („wild-type") variants of the NCI-H1581 and NCI-H1703 lines. (b) Analysis of differences between sensitive and resistant variants separately for each line. (c) Result of RNAseq analysis - differentially expressed genes for both lines with the same direction of change.

As in many cell line studies, sample size is a significant issue [29, 30], in particular, the distinction between biological and technical replicates. Since cells from a particular line, even from several different cell banks, were originally derived from a single patient, it is recommended that the results obtained be verified for more than one appropriately selected line. However, particularly in the case of cancer lines, where the mutational burden is usually high, even for lines representing the same tumour subtype, significant differences can be observed (Fig. 8.1a). Also for the transcriptome analysis, principal component analysis (PCA) showed that even despite the resistance-related variability, more than 70% of the variance corresponds to the direction separating the two different cell lines.

For such a small sample, joint differential analysis between sensitive and resistant variants for both cell lines led to results of low biological significance — for most genes, a significant difference for one of the lines was sufficient to achieve statistical significance despite the absence of differences in the other. Finally, it was decided to analyze each cell line independently (Fig. 8.1b). However, after determining the differentiation genes in both comparisons, we found that the results were still difficult to interpret, because even though the strength of the change was sufficient, the direction of the change could vary. Using a filter requiring equal direction of change for differentially expressed genes (Fig. 8.1c) combined with functional analysis and the compilation of genomic and transcriptomic data, allowed the identification of the p38 pathway as a potential mediator of resistance to FGFR inhibition.

Conclusion. *Because for small samples, the variability associated with differences between cell types can obscure the biological effect, it is sometimes advantageous to analyze individual subgroups independently and look for common elements using appropriate filters.*

Author's contribution to the cited work. *Bioinformatics analysis of copy number variation - extraction of genomic features from raw data for aCGH array, preprocessing (normalization, GC and cy3/cy5 correction, centering), gene copy number estimation, differential analysis; Bioinformatics analysis of RNAseq data - preprocessing (quality analysis, adapter removal, mapping and annotation, construction of count matrix), normalization, differential expression analysis; unsupervised analysis and visualization of data from high-*

throughput experiments; analysis of signaling pathways - GSEA.

8.2 Tissue level and differences between patients

One of the reasons why cancers are often detected at a late stage is the lack of reliable, universal diagnostic markers. As cancer cells are characterized by specific metabolic activity, metabolomics is being considered as one potential source of biomarkers [31].

[2] Mrowiec K, Debik J, Jelonek K, Kurczyk A, Ponge L, Wilk A, Krzempek M, Giskeødegård GF, Bathen TF, Widłak P. Profiling of serum metabolome of breast cancer: multi-cancer features discriminate between healthy women and patients with breast cancer. Front. Oncol. 14:1377373. (2024)

Aim. *Metabolomic characterization of breast cancer, identification of a signature common to different cancers, construction of a metabolomic classifier.*

In this study, metabolomic profiles were analysed for the plasma of healthy donors, as well as patients with breast, lung, colorectal and head and neck cancers. On the basis of statistical analysis, metabolites distinguishing healthy controls from each type of cancer were selected (due to changes in the metabolomic profile with age, subgroups matching the age structure of other cancers were selected for controls and breast cancer patients). The common „multi-cancer signature" included 6 amino acids (Ala, Asp, Glu, His, Phe, Leu+Ile), 2 diglycerides, 2 triglycerides, and 13 lysophosphatidylcholines (as well as the total level of lysophosphatidylcholines).

Then, using the determined signature, a classifier was constructed to distinguish between healthy controls and breast cancer patients. After testing various models and hyperparameter values, a support vector machine (SVM) with a radial kernel function was decided upon. The quality of the classifier was estimated by a 500-fold Monte Carlo cross-validation procedure performed on a set of samples that were not used for multi-cancer signature selection. To test whether the classifier would also work for other European populations, the model trained on a set of Polish patients was tested on a set of Norwegian patients. The resulting classifier had very high predictive ability, reaching median sensitivity=0.97, specificity=0.92 and AUC=0.98 in validation.

Conclusion. *Despite the differences between cancer types, it is possible to select a common signature for which it is possible to classify with high accuracy Polish and Norwegian breast cancer patients.*

Author's contribution to the cited work. *Training and validating machine learning models, investigating the impact of hyperparameters, developing a model testing scheme to test generalizability for different cohorts, contributing to the preparation of the manuscript.*

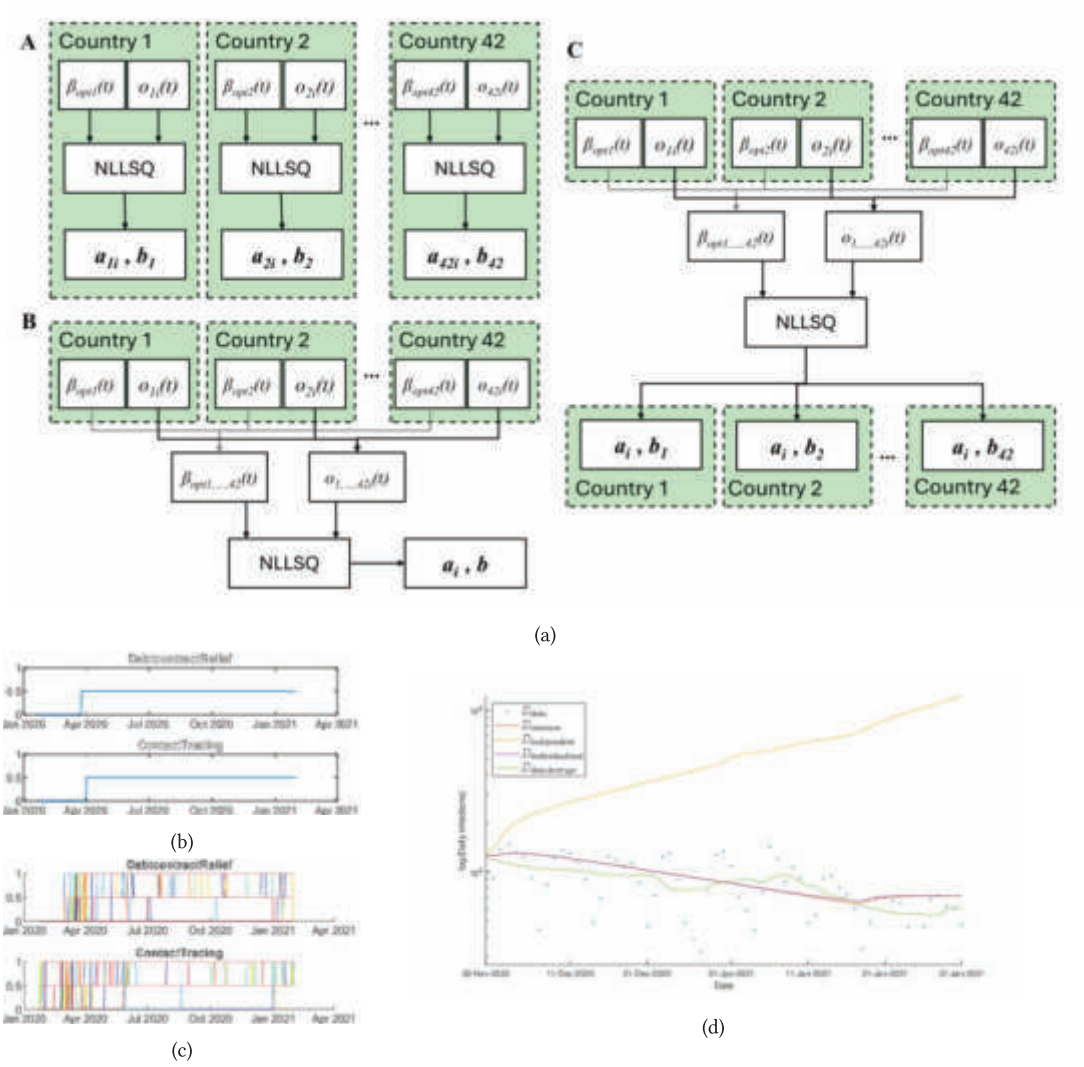
8.3 Population level — individualized models

Although an independent analysis of subgroups, such as cell lines in the work of [1] or cancer types in [**Metabolomics**], may seem tempting because it allows for individuality in heterogeneous cohorts, it is not always feasible. The reason may be, for example, the non-estimability of the parameters, which requires an alternative approach.

[3] Wilk AM, Łakomiec K, Psiuk-Maksymowicz K, Fujarewicz K. Impact of government policies on the COVID-19 pandemic unraveled by mathematical modelling. Scientific Reports 12, 16987 (2022)

Aim. *Assessing the impact of non-pharmaceutical government interventions on the spread of the COVID-19 pandemic.*

In the first phase of the COVID-19 pandemic, prior to the introduction of vaccination, the only method of combating the spread of the SARS-CoV-2 virus was through non-pharmaceutical interventions, including restrictions such as school closures or cancellation of public events, economic measures, and measures related to the health system [32]. Many studies have been devoted to estimating the effectiveness of individual interventions, but a primary obstacle was that they were usually introduced in packages, making some highly correlated [33].



(a)

(b)

(c)

(d)

Fig. 8.2: Predicting the spread of the COVID-19 pandemic in European countries using individualized models. (a) Three parameter estimation approaches used: A — independent, B — common, C — individualized (b) Indistinguishable policy functions in Poland (c) The same restrictions with all countries analyzed. (d) Prediction results for the test time period for Poland. Blue points are daily infections, green line is the 7-day moving average of the number of infections, yellow, red and purple lines are the predictions for the independent, common and individualized model, respectively.

Consider the SEIR model, one of the most popular compartmental epidemiological models, in which the population is divided into four compartments:

- Susceptible,
- Exposed,

- Infected,
- Removed.

The model for country k can be described with equations [34]:

$$\begin{cases} \dot{S}_k(t) = \frac{-\beta_k(t) S_k(t) I_k(t)}{N_k} \\ \dot{E}_k(t) = \frac{\beta_k(t) S_k(t) I_k(t)}{N_k} - k_{EI} E_k(t) \\ \dot{I}_k(t) = k_{EI} E_k(t) - k_{IR} I_k(t) \\ \dot{R}_k(t) = k_{IR} I_k(t) \end{cases} ; \quad k = 1, ..., K \tag{8.1}$$

with initial conditions $S_k(0) = N_k - I_0$, $E_k(0) = 0$, $I_k(0) = I_0$, $R_k(0) = 0$. The population size N_k was assumed constant.

Parameter $\beta(t)$ determines the rate of spread of a pandemic. We can assume that its value depends on a certain baseline value b, stringency of the policies o_i and the effectiveness of these policies a_i:

$$\beta(t) = b(1 - a_1 o_1(t) - a_2 o_2(t) - \cdots - a_r o_r(t)), \tag{8.2}$$

In order to preserve country characteristics, it would be natural to estimate parameters for each country independently. In the case of overlapping restriction functions, for example $o_i = o_j = o_{ij}$ (see Fig. 8.2b),

$$\beta(t) = b(1 - a_1 o_1(t) - \cdots - a_i o_i(t) - a_j o_j(t) - \cdots - a_r o_r(t)),$$
$$\beta(t) = b(1 - a_1 o_1(t) - \cdots - a_i o_{ij}(t) - a_j o_{ij}(t) - \cdots - a_r o_r(t)),$$
$$\beta(t) = b(1 - a_1 o_1(t) - \cdots - o_{ij}(t)(a_i + a_j) - \cdots - a_r o_r(t)),$$

Parameters a_i and a_j are then non-estimable, since a change in the value of one of them can be freely compensated by the value of the other. A possible solution is to estimate the parameters jointly for the entire cohort of countries (Fig. 8.2c). While this

mitigates the problem of non-estimability, the individual character of each object is also lost.

A compromise solution, the individualized model, was proposed, consisting in estimating part of the parameters together for all countries (in this example, it was assumed that the common parameters are the effectiveness of the restrictions a_i), and part independently (here: the base coefficient b). Such an approach makes it possible to reduce the number of parameters to be estimated, preserve some reflection of the heterogeneity of the cohort objects and reduce the risk of inappropriate numerical conditioning of the estimation problem. The three approaches to parameter estimation are shown schematically in Fig. 8.2a.

First, $beta_{opt}$ was determined for each country based on the infection values reported by government agencies. Then, the parameters a_i and b were estimated using the nonlinear least squares (NLLSQ) method. Parameter estimation was performed on a training period covering the time from the beginning of the pandemic to the end of November 2020, while the following two months (from the beginning of December 2020 to the end of January 2021, when vaccination was introduced) were treated as a test period. The normalized root mean square error (NRMSE) was used as a measure of the goodness of fit. For Poland (Fig. 8.2d), the common and individualized models produced similar results, close to the observed pattern. The independent model, on the other hand, had a significant error.

Tab. 8.1: Effectiveness of strategies for estimating SEIR model parameters (for $k_{EI} = 0.2605$ and $k_{IR} = 0.1020$). The mean and standard deviation (SD) for the NRMSE and the number of countries where the model was the best are shown (in the case of the same NRMSE value for two models, both are counted).

Summary of prediction quality			
	Common model	**Independent models**	**Individualized models**
Mean NRMSE	2.374	4.920	1.682
SD NRMSE	4.414	5.906	2.222
Best model	17	10	17

For the entire cohort of countries, the common and individualized models show high prediction quality, with the individualized model having an advantage in terms of mean error and error distribution (Table 8.1). Although the use of a common model also over-

came numerical problems, its limited ability to fit to the data for a heterogeneous cohort resulted in very high error values for some countries (as reflected by the standard deviation of the NRMSE). In addition, for independent models, the values of the estimated parameters often reached unrealistic levels. For example, the intervention „School closing" had a coefficient of 0.225 for the common model, 0.149 for the individualized models, and between -0.443 and 0.718 for the independent models. Taking into account the assumed function of the transmission of the virus depending on the restrictions (equation 8.2), this would mean that in some countries, school closures accelerated the spread of the pandemic. The goodness of fit for the independent models could have been improved by using constraints on the values of the parameters, but one of the assumptions in the work was to not require any knowledge *a priori* with regard to the parameters.

Conclusion. *Using an individualized approach to estimating model parameters for a cohort of objects makes it possible to reduce the risk of poor numerical conditioning without completely losing the individual character of the objects.*

Author's contribution to the cited work. *Preparation of the dataset - acquisition from external databases (incidence, restrictions, populations), filtering out countries with missing data, converting to a format suitable for further analysis; estimation of the impact of policies - implementation of the common, independent and individualized model; validation of the models on the test period - estimation of the number of infections from the SEIR model, evaluation of the error; visualization of the results; preparation of the manuscript.*

In all the cases discussed so far, the problem amounted to analyzing and modelling heterogeneous objects and detecting general patterns without completely losing their individual character. Despite the differences between the objects occurring at many levels, analogous data were available for them each time — copy number variation, expression levels, metabolomic profile or vector of policy stringencies. The issue of heterogeneity, however, can apply not only to the objects themselves, but also to the data that describe them.

Rozdział 9

Structural heterogeneity

By default, the input data for predictive models have the same structure for each object under analysis. In simple terms, an objective function is determined based on the feature vector, and its value determines the prediction, which can be, for example, a category assignment in case of classification or a continuous value for regression. However, it may be possible to obtain more than one feature vector for certain objects, the number of which is not fixed (see Fig 9.1). The result is heterogeneous data, the use of which in the model is non-trivial.

9.1 „Single-pixel approach" — classification based on spatial proteomics

One of the fields in which the described type of structural heterogeneity is becoming increasingly common is molecular biology. Modern technology developments make it possible to obtain proteomic, metabolomic or transcriptomic data for single cells [35, 36, 25, 37], or a grid of points in spatial assays [38, 39]. Thus, depending on the technology, for each sample, we obtain from tens to even thousands of feature vectors. The article

[4] Kurczyk A., Gawin M., Chekan M., Wilk A., Łakomiec K., Mrukwa G., Frątczak K., Polanska J., Fujarewicz K., Pietrowska M. and Widlak, P. Classification of Thyroid Tumors Based on Mass Spectrometry Imaging of Tissue Microarrays; a Single-Pixel Approach. International journal of molecular sciences, 21(17),

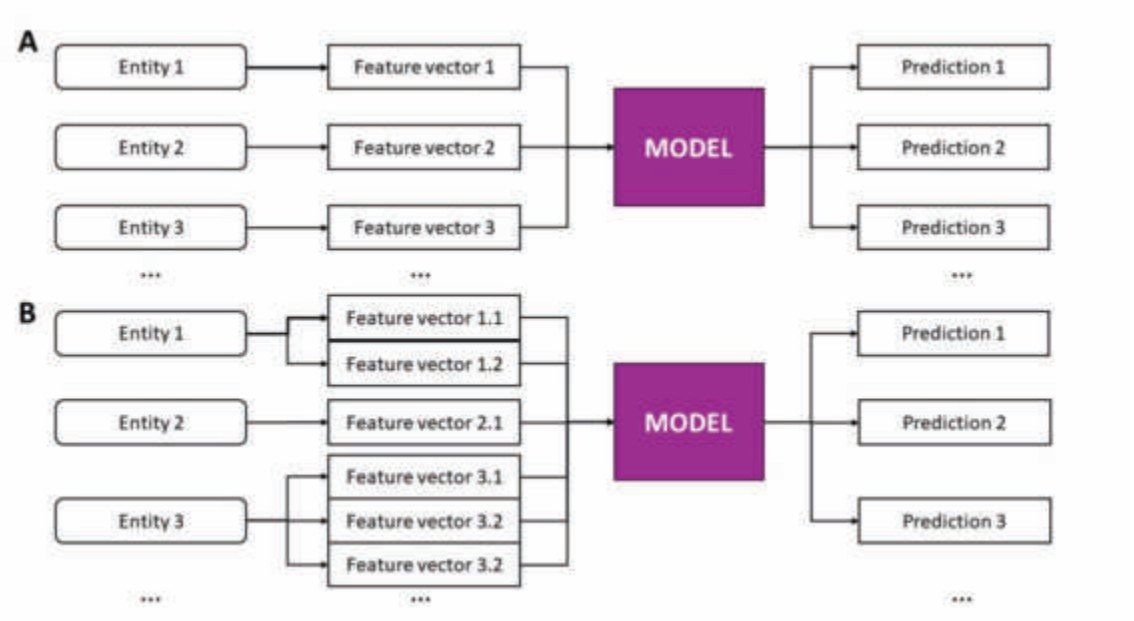


Fig. 9.1: An example of structural heterogeneity in the data. A. Each object corresponds to one feature vector, so standard models can be used. B. Analogous feature vectors are available for each object, but the number of vectors varies. Modelling requires additional steps.

6289, (2022)

describes how to use this type of data to classify thyroid tissue specimen.

Aim. *Classification of thyroid cancer subtypes based on structurally heterogeneous mass spectrometry imaging data.*

Thyroid cancer is the most common endocrine neoplasm — more than 800,000 cases were diagnosed worldwide in 2022 [40], while in Poland there are about 4,700 cases per year (data as of 2023) [41]. The prognosis is very favorable for most patients (especially those representing the differentiated subtype), but worse for poorly differentiated subtypes[42, 43]. Subtype diagnosis is usually based on the analysis of histopathologic specimens, so it can be subjective and largely depends on the experience and current disposition of the pathologist, as well as the quality of preparation of the sample and its nature. The diagnostic process could be improved by identifying molecular biomarkers specific to each subtype.

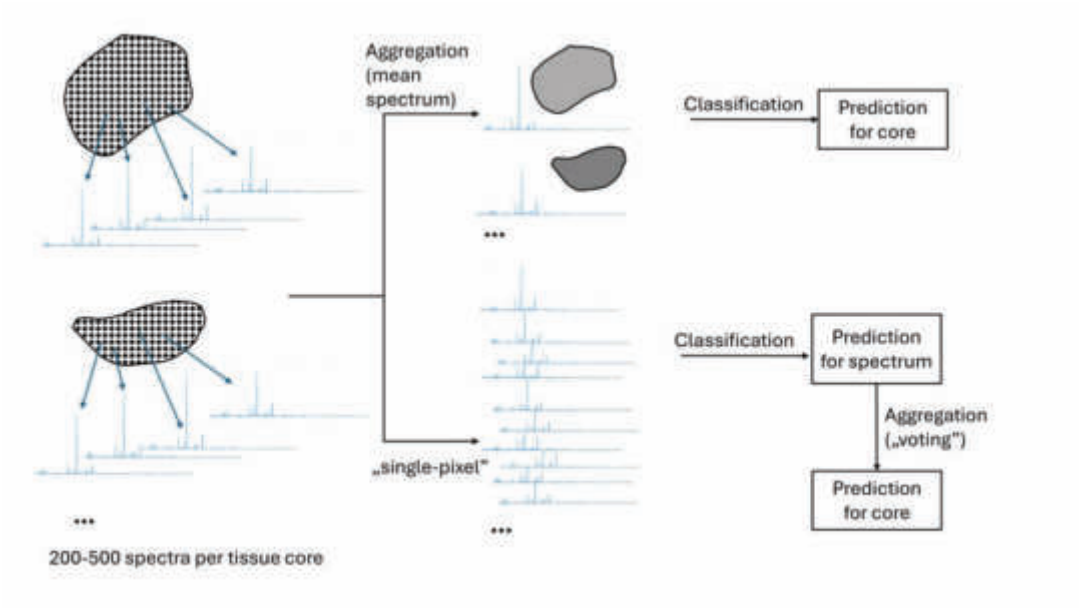
The study analyzed tissue matrices containing a total of 375 tissue cores from 134

patients diagnosed in the Maria Skłodowska-Curie National Research Institute of Oncology, Gliwice Branch (NIO). They represented seven types of tissue:

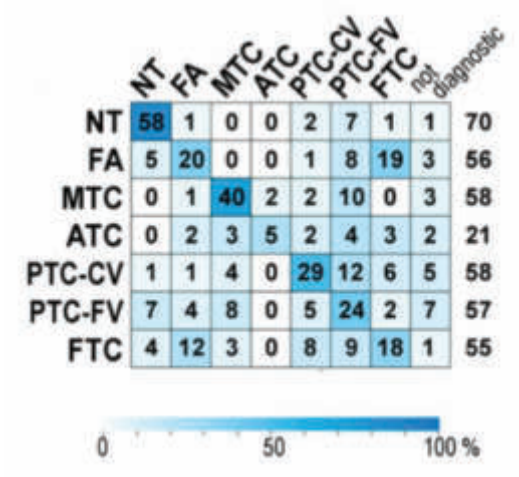
- NT (*normal thyroid*) — healthy thyroid tissue
- FA (*follicular adenoma*) — non-malignant thyroid nodule,
- MTC (ang. *medullary thyroid cancer*) — rare, poorly differentiated thyroid cancer with average prognosis,
- ATC (*anaplastic thyroid carcinoma*) — rare, poorly differentiated thyroid cancer with poor prognosis,
- FTC (*follicular thyroid carcinoma*) — common, differentiated thyroid cancer with excellent prognosis,
- PTC-CV (*papillary thyroid carcinoma classic variant*) — a subtype of papillary thyroid cancer, common, differentiated thyroid cancer with excellent prognosis,
- PTC-FV (*papillary thyroid carcinoma follicular variant*).

The specimens were subjected to mass spectrometry (MS) imaging using the MALDI-MSI technique, which yielded an average of 360 MS spectra per slice.

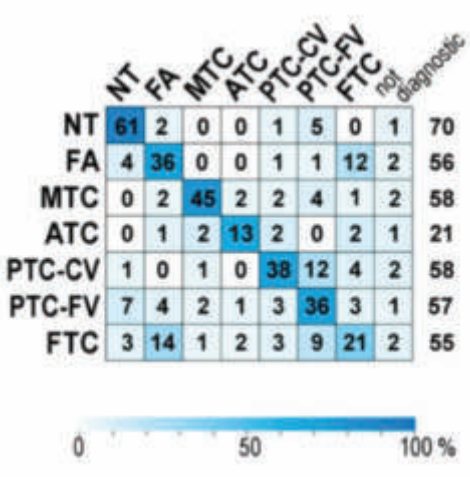
Based on the validation of different types of models, a support vector machine (SVM) with a linear kernel function was chosen to discriminate between spectra. To achieve multi-class classification, the problem was converted to a 1vs1 task. A binary model (21 classifiers in total) was trained for each pair of classes. If the decision was „unanimous" that is, a certain class received a total of six votes (in all models comparing it with the other classes), it was assigned to the spectrum; otherwise, the spectrum was classified as „non-diagnostic". This solution was adopted after consultation with a pathologist, for whom the information that a sample is difficult to classify is preferable to an incorrect or uncertain diagnosis.



(a)



(b)



(c)

Fig. 9.2: Application of aggregation for classification of proteomic imaging data. (a) Simplified flowchart. (b) Confusion matrix for the mean spectrum approach — expert classes in rows, predicted classes in columns. (c) Confusion matrix for the „single-pixel" approach — expert classes in rows, predicted classes in columns. The „single-pixel" approach using individual spectra achieves higher classification accuracy compared to averaged spectra.

Since standard machine learning models use a single feature vector per observation, classification of the entire tissue core requires the use of aggregation. Two strategies were tested (Fig. 9.2a):

1. Mean spectrum — a representative spectrum is generated for a slice as the arithmetic average of all spectra. This is a solution that simplifies the implementation of the classifier; however, it results in the loss of most of the measurements and information about the heterogeneity of the tissue.

2. „Single-pixel" — a strategy in which the model is built using all the available spectra in the training set. When classifying a new tissue core, each of the several hundred of its spectra is assigned to a class, and the final result is decided by a „vote" in which the majority class is assigned (even if it is a „non-diagnostic" class).

The predictive ability of the models varies by class (Fig. 9.2b and 9.2c). Relatively good performance is achieved between normal thyroid tissue and neoplastic tissue (both benign and malignant), while significant errors are observed for subtypes of papillary carcinoma and between adenoma and follicular carcinoma. Regardless of class, however, it can be observed that the „single-pixel" approach achieves higher quality than the mean spectrum approach – the classification accuracy for these strategies is 0.67 and 0.52, respectively.

Conclusion. *Prediction based on aggregated classification results for all available measurements achieves higher accuracy than prediction based on the classification result of a representative averaged measurement.*

Author's contribution to the cited work. *Unsupervised analysis of the dataset (PCA), development and implementation of mean spectrum and „single-pixel" strategies, implementation of a voting system for a multi-class model, design of a validation scheme that takes into account the relationships between samples, visualization.*

9.2 „Multi-lesion radiomics" — aggregation in radiomics-based survival models

Another (besides the properties of the specimen used for measurement) possible reason for the availability of a different number of feature vectors describing the various objects being modeled is the varying state of the objects themselves, such as cancer pa-

tients. The next three papers in this series describe subsequent steps in developing a methodology to use structurally heterogeneous data in survival modelling.

[5] Wilk AM, Kozłowska E, Borys D, D'Amico A, Fujarewicz K, Gorczewska I, Debosz-Suwinska I, Suwinski R, Smieja J, Swierniak A. Radiomic signature accurately predicts the risk of metastatic dissemination in late-stage non-small cell lung cancer. Translational Lung Cancer Research 12(7):1372-1383. (2023)

Aim. *Predicting the risk of distant metastasis for patients with non-small cell lung cancer based on clinical and radiomic data.*

Lung cancer is the second most common cancer (after breast cancer) in the world, but ranks unquestionably first in terms of mortality and accounts for about 20% of cancer-related deaths worldwide [40]. The five-year survival rate for lung cancer is less than 20%, which can be largely attributed to the typically late diagnosis due to nonspecific symptoms in the early stages of the disease, as well as the significant invasiveness [44]. The pivotal moment for therapeutic options is the occurrence of distant metastases, as reflected, among other things, in the recommendations of the European Society for Medical Oncology (*ESMO*), which are divided into non-metastatic and metastatic cancer [45, 46].

In current practice, metastases are detected by regular follow-up using imaging. This approach has some drawbacks, related to, among other things, the limited frequency of testing due to the capacity of the health system and safety concerns (imaging studies, although generally considered non-invasive, also cannot be performed too frequently), as well as the capabilities of imaging itself. Given the resolution of technologies such as positron emission tomography (PET), small, even centimeter-sized foci may not be detectable [47]. Finding factors to help stratify patients in terms of their risk of distant metastasis would therefore be clinically meaningful.

In the work [5] the possibility was investigated of using clinical features and features extracted from PET/CT imaging performed for radiotherapy planning to predict distant metastasis. Given that oncology cohorts are often too small for the construction of a reliable model based on deep learning [48], a radiomics approach was decided upon. Radiomics is a branch of science concerned with extracting from a specific image area, called the region of interest (*ROI*), numerical features describing, among others, its shape

and texture [49].

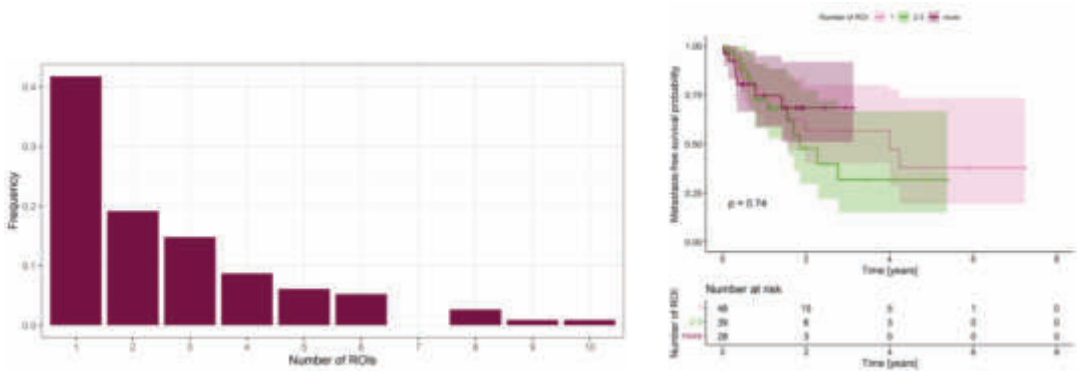
A cohort of 115 patients treated at NIO for non-small cell lung cancer was analyzed. The article includes basic descriptive statistics, exploratory analysis and univariate statistical analysis of available characteristics. Multivariate analysis was performed in two ways — using a classification-based approach and survival models. Both approaches showed that the clinical characteristics available in the study, including age, sex, subtype, tumour location, tumour size and degree of regional cancer spread according to TNM classification (*tumour, node, metastasis*), are not good predictors of metastasis risk. Using radiomic features however, we are able to divide patients into groups that are statistically significantly different in their probability of metastasis-free survival (*MFS*).

Conclusion. *Radiomic features extracted from the region of interest covering the primary tumour show potential for predicting the risk of distant metastasis in non-small cell lung cancer. However, routinely collected clinical features such as tumour size and lymph node metastasis do not show significance for predicting distant metastasis.*

Author's contribution to the cited work. *Unsupervised analysis of the dataset —- PCA, correlation analysis; descriptive statistics, univariate analysis, prediction of metastasis-free and event-free survival using classification methods, construction of the final regression survival model, preparation of the first draft of the manuscript.*

One of the conclusions from the analysis of the clinical data is that the regional spread of the cancer, expressed in part by the „N" part of the TNM classification, does not affect the risk of distant metastasis. This observation, however seemingly counter-intuitive, is in line with current theory on the mechanism of metastasis in lung cancer, according to which dissemination from the primary tumour, not from the lymph nodes, is primarily responsible for the formation of distant metastases [50]. However, in addition to the primary tumour and possibly involved lymph nodes, there may also be additional foci of cancer (we are then talking about multifocal cancer), which are not reflected in the TNM classification. All of these uptakes may be regions of interest (Fig. 9.3). Although their number alone is not a predictor of distant metastasis (Fig. 9.3b), it was hypothesized that including all regions in survival modelling could improve the predictive ability of the model.

As in the earlier described case of classification of molecular imaging data [4], from



(a) (b)

Fig. 9.3: In the majority of patients in the study cohort, at least two uptakes could be observed in the PET images of the lungs. (a) Prevalence of each number of uptakes — regions of interest. (b) Predictive value of ROI number for predicting metastasis.

a modelling perspective, the challenge comes down to the heterogeneous structure of the data, that is, the availability of different numbers (here from one to ten) of radiomic feature vectors per patient. For the metastasis risk prediction problem under consideration, however, some additional aspects must be taken into account.

1. A different type of task. Unlike classification, where the result is the assignment of an object to predetermined groups (classes), survival models fit into a regression problem. The result of prediction is a continuous numerical value, interpreted as the risk of occurrence of the analyzed event.

2. Non-equivalence of feature vectors. In spatial molecular techniques, each measurement refers to a certain point of the grid, which „covers” the tissue. Therefore, it can be assumed that all feature vectors are equivalent, and it is not possible to order them according to any logical criterion. Meanwhile, for medical imaging, one obvious feature that distinguishes feature vectors is the size (volume) of the region of interest from which they are extracted.

3. Interpretability of the averaged feature vector. For transcriptomic or proteomic data, feature values can be interpreted as RNA or protein concentrations. Averaged values from multiple points retain their interpretation as concentrations, albeit for a larger number of cells. In the case of radiomics, the determined features are closely related to the analyzed region of interest, describing its shape or

texture. By averaging the values from several ROIs, many radiomic features lose their interpretability.

The next paper provides a proof of concept of a kind that, despite the above considerations, aggregation of either feature vectors or prediction scores can be a solution to the problem of heterogeneity of data structure for radiomic survival models.

[6] Wilk AM, Kozłowska E, Borys D, D'Amico A, Gorczewska I, Debosz-Suwińska I, Gałecki S, Fujarewicz K, Suwiński R, Świerniak A. Improving the Predictive Ability of Radiomics-Based Regression Survival Models Through Incorporating Multiple Regions of Interest. W: Strumiłło, P., Klepaczko, A., Strzelecki, M., Bociąga, D. (eds) The Latest Developments and Challenges in Biomedical Engineering. PCBEE 2023. Lecture Notes in Networks and Systems, vol 746. Springer, Cham. (2024)

Aim. *To use all available regions of interest from PET/CT imaging in a survival model predicting metastasis risk.*

The study used the cohort of 115 patients with non-small cell lung cancer, which was analyzed in the paper [5]. This time, in addition to the primary tumor, all accumulations were contoured on the PET/CT images by the radiologist (Fig. 9.4a), and then were used to extract radiomic features.

Analogous to MALDI imaging [4], the proposed methods for including all feature vectors in the model are based on two possible strategies (schematically depicted on the fig. 9.4a).

1. **ROI aggregation**. It involves selecting or generating a single feature vector to represent a given patient, that is, reducing the data to the homogeneous structure usually used in analysis. The modelilng results in a single risk value assigned to each patient. The ROI aggregation methods used in the study are:

    - *largestROI* — the largest uptake, corresponding to the primary tumour, is used in the model. This is the approach used in most radiomic studies dedicated to tumorus.
    - *randomROI* — a random uptake is used in the model.

- *arithmeticMeanROI* — a representative vector of features is generated, as the arithmetic average of all vectors for the patient.
- *weightedMeanROI* — a representative vector of features is generated, as the weighted average of all vectors for the patient, with weights corresponding to ROI volumes.

2. **Risk aggregation**. The model is applied independently to all ROIs, resulting in several risk values for a single patient. These results are then aggregated to produce a single value. The methods used are:

   - *allROIMin* — the patient is assigned the lowest risk value obtained for his ROIs.
   - *allROIMax* — the patient is assigned the highest risk value obtained for his ROIs.
   - *allROIMean* — The patient is assigned the average risk value obtained for his ROIs.

Regularized Cox regression (Coxnet), which also includes feature selection, was used for survival analysis. In order to obtain a more meaningful comparison between methods, a Monte Carlo cross-validation (*MCCV*) was used, with 1000 random splits of the patient set into a training set and a test set in a 2:1 ratio (in the case of risk aggregation methods, the sets ultimately used contain all ROIs from the corresponding patients, so they have different cardinalities in each iteration). Harrell's c-index, one of the most popular quality measures for survival models, was used as an indicator of prediction quality. It is a concordance index describing the ratio of the number of „concordant" pairs of observations (i.e., those in which a higher-risk observation corresponds to a shorter time-to-event) to all possible pairs. It is referred to as the survival equivalent of the area under the ROC curve — it takes values from 0 to 1, with 0.5 indicating a random model and 1 indicating a perfect model.

Table 9.1 presents a summary of the prediction quality achieved by each method. For the standard approach, based on considering only the primary tumour, the median c-index from 1000 iterations of MCCV was 0.581. Worse results were obtained for the methods of *randomROI*, *arithmeticMeanROI* and *allROIMin*. This is not surprising, since in

Tab. 9.1: Comparison of methods for handling multiple ROIs in a survival model.

Method	Method type	Median c-index	Min c-index	Max c-index
largestROI	ROI aggregation	0.581	0.206	0.862
randomROI	ROI aggregation	0.534	0.217	0.828
arithmeticMeanROI	ROI aggregation	0.557	0.290	**0.892**
weightedMeanROI	ROI aggregation	0.592	0.206	0.835
allROIMin	risk aggregation	0.566	0.193	0.817
allROIMax	risk aggregation	**0.617**	0.349	0.880
allROIMean	risk aggregation	0.616	**0.369**	0.827

the case of a random ROI, the risk to the patient could have been estimated, for example, on the basis of a very small collection, for which textural features are less informative [51, 52]. For ROIs generated using the arithmetic mean, ROIs of different sizes are treated the same, leading to lower quality. Better, outperforming the model based only on the primary tumour, was the method *weightedMeanROI*, in which feature vectors were scaled relative to the size of the ROI. The highest prediction quality was obtained for the *allROIMax* and *allROIMean* methods. Moreover, these methods proved to be more robust to sampling, achieving higher minimum c-indexes.

Conclusion. *The different number of radiomic feature vectors determined from the uptakes detected in PET/CT imaging can be included in the survival model by using ROI aggregation or risk aggregation.*

Author's contribution to the cited work. *Conceptualization, development and implementation of aggregation methods; statistical analysis, testing of survival models, visualization, preparation of the manuscript.*

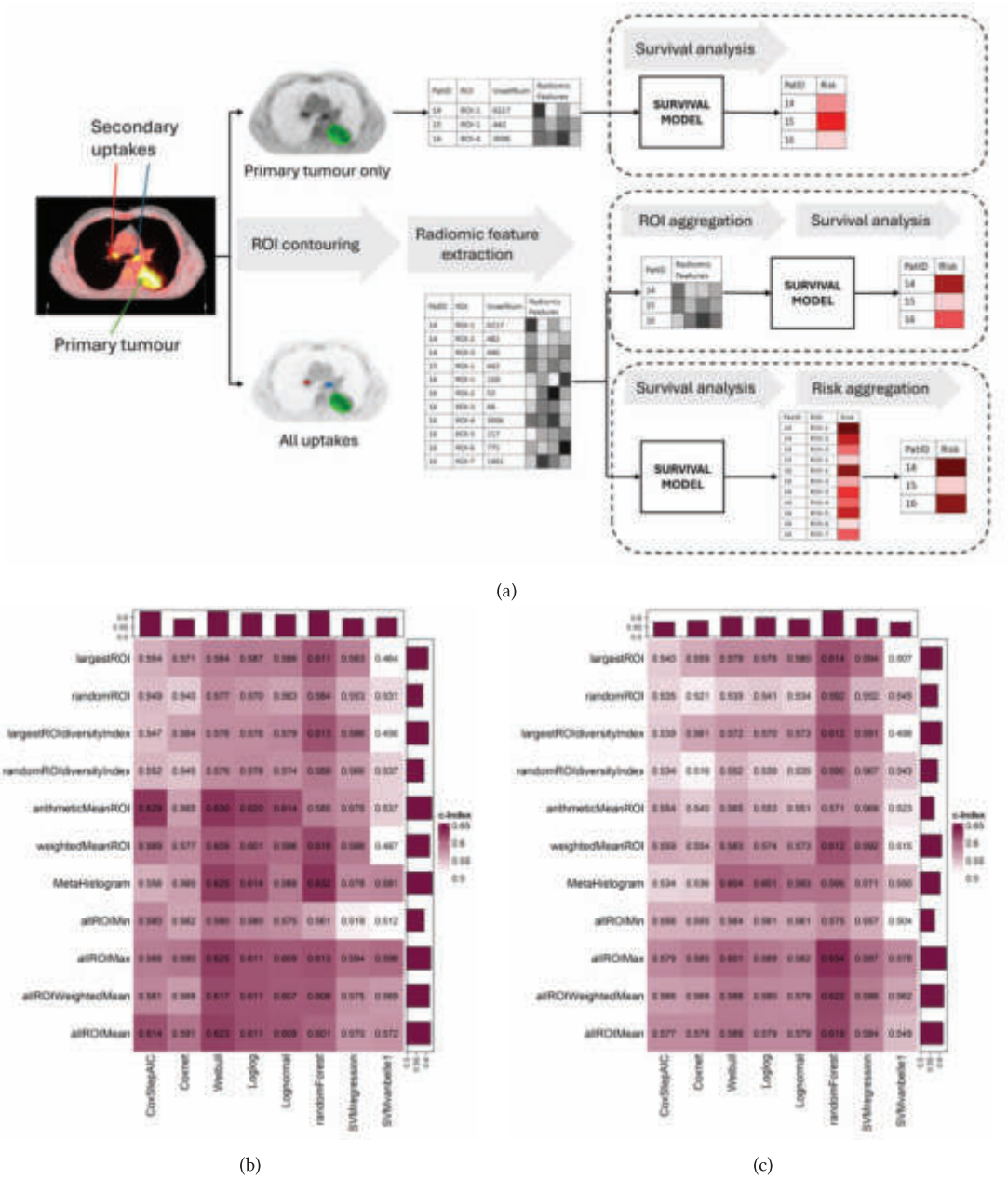


Fig. 9.4: „Multi-lesion radiomics" — including different numbers of regions of interest in survival models by using appropriate aggregation methods. (a) Main idea. (b) Results of 1000-fold cross-validation for the PET set — median c-indexes for each scheme (c) Results of 1000-fold cross-validation for the PET_CT set — median c-indexes for each scheme. Regardless of the dataset and model, the quality achieved for a single ROI corresponding to the primary tumour can be improved using all uptakes.

The obtained results were so promising that it was decided to extend the study, chec-

king whether a similar trend would hold for other survival models. The highest prediction quality achieved by the *allROIMax* method and the low quality for the *allROIMin* method were also noteworthy. For patients with more than one uptake, it is natural that the prediction result will be both higher and lower risk values. Assigning such patients the highest of the resulting risk values proved to be the most effective strategy, despite the fact that the number of ROIs alone is not a predictor of metastasis risk (Fig. 9.3b). Thus, it was hypothesised that not the number, but the variation between ROIs, may be associated with a higher risk of distant metastasis.

[7] Wilk AM, Swierniak A, d'Amico A, Suwiński R, Fujarewicz K i Borys D. Towards the use of multiple ROIs for radiomics-based survival modelling: finding a strategy of aggregating lesions. Preprint arXiv: 2405.17668 [stat.AP]. (2024) Work currently under review in journal Computers in Biology and Medicine

Aim. *Evaluation of heterogeneity between ROIs, development of methods to include multiple tumour foci in survival models, testing whether aggregation improves prediction quality for different models and radiomic feature extraction parameters, and comparison with methods reported in the literature.*

The study was again performed on the NSCLC dataset described above. However, two sets of radiomic features were tested (for both, PET images were standardized against a standardized uptake value taking into account body weight, SUVbw) — at the original PET resolution (PET set) and at a resolution interpolated to CT using the nearest neighbor method (PET_CT set).

Assessing heterogeneity between ROIs required defining numerical diversity indices. Since radiomic features can take negative values, it was not possible to use entropy-based indexes, so the focus was on various distance measures:

1. Canberra distance,
2. Euclidean distance,
3. Minkowski distance,
4. Kendall distance (based on the Kendall correlation coefficient),

5. Spearman distance (based on Spearman's correlation coefficient; due to significant order-of-magnitude differences between traits, Pearson's correlation coefficient has proven impractical).

For each patient, the heterogeneity index was assumed to have a value of zero if only one feature vector (i.e., one ROI) was available, and a value corresponding to the average for each unique pair of ROIs otherwise. As an analogy to the test for the number of ROIs (Fig. 9.3b), the set of patients was again divided into three subgroups (according to the terciles of index values) and compared using the log-rank test. For the PET set, groups separated by Euclidean distance, for example, differed significantly in MFS (p=0.026).

One of the aims of the work [7] was to compare the proposed methods with existing ones. However, a detailed search of the literature showed that, although the idea of „multi-region radiomics" is well-known and gaining popularity, its understanding is mostly different from that of the described study. This is because it relies on the use of radiomic features extracted from multiple regions, which are nevertheless analogous for each patient and represent a modification of the ROI definition (for example, tumour and peritumoral area, or tumour subregions) rather than separate foci. Understood in this way, additional areas therefore do not generate new feature vectors, but only increase the number of features available for each patient. In the only paper found on radiomics integration for multiple independent areas, Zhao and co-authors [53] introduced the method of „meta histogram". By ordering the radiomic feature values from the largest to the smallest ROI, a „meta histogram" is created, for which the mean, variance, skewness, kurtosis, energy, entropy and sum are then calculated and entered as features into the model. This method was used by the authors in a classification problem for patients with lung adenocarcinoma who had at least two metastatic foci.

Ultimately, in addition to the aggregation methods described in the paper [6], the following methods were also included:

- *largestROIdiversityIndex* — ROI aggregation, in which the heterogeneity indices described above are concatenated to the feature vector determined for the primary tumour,

- *randomROIdiversityIndex* — Similar to the previous method, however, heterogeneity indices are added to the features extracted from a random ROI,

- *MetaHistogram* — method described in [53], which can also be classified as ROI aggregation,
- *allROIWeightedMean* — r isk aggregation, in which a patient is assigned a weighted average of risks for all his ROIs, where the weight is the volume of the ROI.

As before, Monte Carlo cross-validation with 1000 iterations was used, comparing a total of eight survival models:

- *CoxStepAIC* — Cox proportional hazard regression with sequential forward selection based on Akaike Information Criterion (*AIC*), which is a method commonly used in medical literature,
- *Coxnet* — regularized Cox regression,
- *Weibull* — boosting based on Weibull model,
- *Loglog* — boosting based on log-log model,
- *Lognormal* — boosting oparty na modelu log-normal,
- *randomForest* — random survival forest,
- *SVMregression* — survival version of support vector machine with a regression model and additive kernel function,
- *SVMvanbelle1* — Survival version of support vector machine with a van Belle model and additive kernel function.

The cross-validation results (median c-indices) are depicted in Fig. 4.4b for the PET dataset and Fig. 4.4c for the PET_CT dataset. The quality of prediction varies depending on the set, that is, on the method of radiomic feature extraction, and on the model. Nonetheless, using aggregation methods that allow for the inclusion of information about all the uptakes present in the imaging, it is possible to obtain a higher quality than for the *largestROI* method within the same set and model every time.

Conclusion. *The use of ROI aggregation or risk aggregation achieves a higher quality of prediction than in a model based only on the primary tumour.*

Author's contribution to the cited work. *Conceptualization — definition of the research problem, review of existing approaches, design of the study; Methodology — development and implementation of aggregation methods; statistical analysis, testing of survival models, visualization, preparation of the manuscript.*

Rozdział 10

Summary

Heterogeneity is an inherent aspect of biomedical data. It necessitates a constant search for a compromise between generalisation to ensure higher statistical power, and the most accurate capturing of nuances and differences, so important for personalised therapy. It represents, on the one hand, a challenge to the possibilities of analysis, on the other hand, a driving force for the development of technology and a key to the full understanding of biological phenomena and processes.

This dissertation, comprising a series of seven articles, presents research that represents the author's original contribution to the field of biomedical engineering, which has achieved the following goals:

1. Description of heterogeneity occurring on cellular level [1], tissue level [2] and population level [3], for objects characterised by different types of data, including genomic, transcriptomic, metabolomic or imaging data [5].

2. Development of an algorithm for individualized estimation of model parameters [3], allowing for lower prediction errors than for independent or joint estimation.

3. Presentation of the problem of structural heterogeneity of data, in particular the situation when the number of available feature vectors differs for individual objects, for example, in spatial molecular research [4] or in medical imaging of regionally advanced cancer [6, 7].

4. Proposing methods for aggregating feature vectors or model outputs, applicable to both classification [4] and survival models [6, 7].

The developed machine learning algorithms can be applied to a variety of data types. They are also not limited to a specific biological problem or model, so they have a broad potential for application in research where it is important to take heterogeneity into account, among others:

- in epidemiological studies,
- for oncologic patient cohorts (or other diseases)
- in molecular techniques that generate information for single cells or points in space.

The presented research results confirm the validity of the theses presented in the dissertation.

Individualizing models for a cohort of objects reduces the risk of poor numerical conditioning of the parameter estimation task. In model of the COVID-19 pandemic accounting for the effect of non-pharmaceutical interventions on the rate of virus transmission, highly correlated stringency levels led to difficulties in estimating the corresponding parameters for a single country. Although the problem of non-estimability can be solved in this case by constructing a common model for a cohort of European countries, generalization leads to high error values for some countries. The use of individualized models has improved numerical conditioning while allowing a better fit to the data than a common model.

The use of aggregation in the case of a different number of feature vectors for each entity results in improved model prediction quality relative to the use of only one vector per object. In a thyroid tissue subtype classifier using proteomic imaging data, aggregating predictions for individual spectra yielded classification quality significantly higher than using only averaged spectra in the prediction. Also, for survival models predicting the risk of distant metastasis in non-small cell lung cancer, using all available uptakes in the model yielded a significant improvement in prediction quality compared to the same models built on the basis of the primary tumor alone.

Using the proposed strategies, whether individualization of parameter estimation or aggregation of modeling results, significantly better results were obtained compared to the approach based on maximum generalization and averaging. Thus, it can be concluded that heterogeneity is not an enemy, but rather an ally in biomedical data analysis.

Prace wchodzące w skład cyklu

[1] Izabela Zarczynska, Monika Gorska-Arcisz, Alexander Jorge Cortez, Katarzyna Aleksandra Kujawa, Agata Małgorzata Wilk, Andrzej Cezary Skladanowski, Aleksandra Stanczak, Monika Skupinska, Maciej Wieczorek, Katarzyna Marta Lisowska, Rafal Sadej i Kamila Kitowska. „p38 Mediates Resistance to FGFR Inhibition in Non-Small Cell Lung Cancer". W: *Cells* 10.12 (2021). ISSN: 2073-4409. DOI: 10.3390/cells10123363.

[2] Katarzyna Mrowiec, Julia Debik, Karol Jelonek, Agata Kurczyk, Lucyna Ponge, Agata Wilk, Marcela Krzempek, Guro F. Giskeødegård, Tone F. Bathen i Piotr Widłak. „Profiling of serum metabolome of breast cancer: multi-cancer features discriminate between healthy women and patients with breast cancer". W: *Frontiers in Oncology* 14 (2024). ISSN: 2234-943X. DOI: 10.3389/fonc.2024.1377373.

[3] Agata Małgorzata Wilk, Krzysztof Łakomiec, Krzysztof Psiuk-Maksymowicz i Krzysztof Fujarewicz. „Impact of government policies on the COVID-19 pandemic unraveled by mathematical modelling". W: *Scientific Reports* 12.1 (2022). ISSN: 2045-2322. DOI: 10.1038/s41598-022-21126-2.

[4] Agata Kurczyk, Marta Gawin, Mykola Chekan, Agata Wilk, Krzysztof Łakomiec, Grzegorz Mrukwa, Katarzyna Frątczak, Joanna Polanska, Krzysztof Fujarewicz, Monika Pietrowska i Piotr Widlak. „Classification of Thyroid Tumors Based on Mass Spectrometry Imaging of Tissue Microarrays; a Single-Pixel Approach". W: *International Journal of Molecular Sciences* 21.17 (2020). ISSN: 1422-0067. DOI: 10.3390/ijms21176289.

[5] Agata Małgorzata Wilk, Emilia Kozłowska, Damian Borys, Andrea D'Amico, Krzysztof Fujarewicz, Izabela Gorczewska, Iwona Debosz-Suwinska, Rafał Suwinski, Ja-

rosław Smieja i Andrzej Swierniak. „Radiomic signature accurately predicts the risk of metastatic dissemination in late-stage non-small cell lung cancer". W: *Translational Lung Cancer Research* 12.7 (2023), s. 1372–1383. DOI: 10.21037/tlcr-23-60.

[6] Agata Małgorzata Wilk, Emilia Kozłowska, Damian Borys, Andrea D'Amico, Izabela Gorczewska, Iwona Debosz-Suwińska, Seweryn Gałecki, Krzysztof Fujarewicz, Rafał Suwiński i Andrzej Świerniak. „Improving the Predictive Ability of Radiomics-Based Regression Survival Models Through Incorporating Multiple Regions of Interest". W: *The Latest Developments and Challenges in Biomedical Engineering*. Red. Paweł Strumiłło, Artur Klepaczko, Michał Strzelecki i Dorota Bociąga. Cham: Springer Nature Switzerland, 2024, s. 163–173. ISBN: 978-3-031-38430-1.

[7] Agata Małgorzata Wilk, Andrzej Swierniak, Andrea d'Amico, Rafał Suwiński, Krzysztof Fujarewicz i Damian Borys. *Towards the use of multiple ROIs for radiomics-based survival modelling: finding a strategy of aggregating lesions*. 2024. DOI: 10.48550/arXiv.2405.17668. arXiv: 2405.17668 [stat.AP].

Bibliografia

[8] Andriy Marusyk i Kornelia Polyak. „Tumor heterogeneity: Causes and consequences”. W: *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer* 1805.1 (2010), s. 105–117. ISSN: 0304-419X. DOI: https://doi.org/10.1016/j.bbcan.2009.11.002.

[9] Albert Rübben i Arturo Araujo. „Cancer heterogeneity: converting a limitation into a source of biologic information”. W: *Journal of Translational Medicine* 15.1 (2017). ISSN: 1479-5876. DOI: 10.1186/s12967-017-1290-9.

[10] Santiago Ramón y Cajal, Marta Sesé, Claudia Capdevila, Trond Aasen, Leticia De Mattos-Arruda, Salvador J. Diaz-Cano, Javier Hernández-Losa i Josep Castellví. „Clinical implications of intratumor heterogeneity: challenges and opportunities”. W: *Journal of Molecular Medicine* 98.2 (2020), s. 161–177. ISSN: 1432-1440. DOI: 10.1007/s00109-020-01874-2. URL: http://dx.doi.org/10.1007/s00109-020-01874-2.

[11] Fabiana Lüönd, Stefanie Tiede i Gerhard Christofori. „Breast cancer as an example of tumour heterogeneity and tumour cell plasticity during malignant progression”. W: *British Journal of Cancer* 125.2 (2021), s. 164–175. ISSN: 1532-1827. DOI: 10.1038/s41416-021-01328-7.

[12] Roman Rouzier, Charles M. Perou, W. Fraser Symmans i in. „Breast Cancer Molecular Subtypes Respond Differently to Preoperative Chemotherapy”. W: *Clinical Cancer Research* 11.16 (sierp. 2005), s. 5678–5685. ISSN: 1078-0432. DOI: 10.1158/1078-0432.CCR-04-2421.

[13] Vandana G. Abramson, Brian D. Lehmann, Tarah J. Ballinger i Jennifer A. Pietenpol. „Subtyping of triple-negative breast cancer: Implications for therapy”. W:

Cancer 121.1 (2015), s. 8–16. DOI: https://doi.org/10.1002/cncr.28914.

[14] Julia Y Tsang i Gary M Tse. „Update on triple-negative breast cancers—-highlighting subtyping update and treatment implication". W: *Histopathology* 82.1 (2023), s. 17–35. DOI: https://doi.org/10.1111/his.14784.

[15] Laura H. Goetz i Nicholas J. Schork. „Personalized medicine: motivation, challenges, and progress". W: *Fertility and Sterility* 109.6 (2018), s. 952–963. ISSN: 0015-0282. DOI: https://doi.org/10.1016/j.fertnstert.2018.05.006.

[16] Isaac S. Chan i Geoffrey S. Ginsburg. „Personalized Medicine: Progress and Promise". W: *Annual Review of Genomics and Human Genetics* 12.Volume 12, 2011 (2011), s. 217–244. ISSN: 1545-293X. DOI: https://doi.org/10.1146/annurev-genom-082410-101446.

[17] Margaret A. Hamburg i Francis S. Collins. „The Path to Personalized Medicine". W: *New England Journal of Medicine* 363.4 (2010), s. 301–304. ISSN: 1533-4406. DOI: 10.1056/nejmp1006304. URL: http://dx.doi.org/10.1056/NEJMp1006304.

[18] Valentina Gambardella, Noelia Tarazona, Juan Miguel Cejalvo, Pasquale Lombardi, Marisol Huerta, Susana Roselló, Tania Fleitas, Desamparados Roda i Andres Cervantes. „Personalized Medicine: Recent Progress in Cancer Therapy". W: *Cancers* 12.4 (2020). ISSN: 2072-6694. DOI: 10.3390/cancers12041009.

[19] Sören Richard Stahlschmidt, Benjamin Ulfenborg i Jane Synnergren. „Multimodal deep learning for biomedical data fusion: a review". W: *Briefings in Bioinformatics* 23.2 (sty. 2022), bbab569. ISSN: 1477-4054. DOI: 10.1093/bib/bbab569.

[20] Maria A. Wörheide, Jan Krumsiek, Gabi Kastenmüller i Matthias Arnold. „Multi-omics integration in biomedical research – A metabolomics-centric review". W: *Analytica Chimica Acta* 1141 (2021), s. 144–162. ISSN: 0003-2670. DOI: https://doi.org/10.1016/j.aca.2020.10.038.

[21] Rui Yan, Liangqiong Qu, Qingyue Wei, Shih-Cheng Huang, Liyue Shen, Daniel L. Rubin, Lei Xing i Yuyin Zhou. „Label-Efficient Self-Supervised Federated Learning for Tackling Data Heterogeneity in Medical Imaging". W: *IEEE Transactions on Medical Imaging* 42.7 (2023), s. 1932–1943. DOI: 10.1109/TMI.2022.3233574.

[22] Lin Yue, Dongyuan Tian, Weitong Chen, Xuming Han i Minghao Yin. „Deep learning for heterogeneous medical data analysis". W: *World Wide Web* 23.5 (2020), s. 2715–2737. ISSN: 1573-1413. DOI: 10.1007/s11280-019-00764-z.

[23] Suraj Rajendran, Weishen Pan, Mert R. Sabuncu, Yong Chen, Jiayu Zhou i Fei Wang. „Learning across diverse biomedical data modalities and cohorts: Challenges and opportunities for innovation". W: *Patterns* 5.2 (2024), s. 100913. ISSN: 2666-3899. DOI: https://doi.org/10.1016/j.patter.2023.100913.

[24] G. Chiorino, J.A.J. Metz, D. Tomasoni i P. Ubezio. „Desynchronization Rate in Cell Populations: Mathematical Modeling and Experimental Data". W: *Journal of Theoretical Biology* 208.2 (2001), s. 185–199. ISSN: 0022-5193. DOI: https://doi.org/10.1006/jtbi.2000.2213.

[25] Florian Buettner, Kedar N Natarajan, F Paolo Casale, Valentina Proserpio, Antonio Scialdone, Fabian J Theis, Sarah A Teichmann, John C Marioni i Oliver Stegle. „Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells". W: *Nature Biotechnology* 33.2 (2015), s. 155–160. ISSN: 1546-1696. DOI: 10.1038/nbt.3102.

[26] Chance M. Nowak, Tyler Quarton i Leonidas Bleris. „Impact of variability in cell cycle periodicity on cell population dynamics". W: *PLOS Computational Biology* 19.6 (czer. 2023), s. 1–16. DOI: 10.1371/journal.pcbi.1011080.

[27] Marcello Tiseo, Francesco Gelsomino, Roberta Alfieri, Andrea Cavazzoni, Cecilia Bozzetti, Anna Maria De Giorgi, Pier Giorgio Petronini i Andrea Ardizzoni. „FGFR as potential target in the treatment of squamous non small cell lung cancer". W: *Cancer Treatment Reviews* 41.6 (2015), s. 527–539. ISSN: 0305-7372. DOI: https://doi.org/10.1016/j.ctrv.2015.04.011.

[28] Arpita Desai i Alex A. Adjei. „FGFR Signaling as a Target for Lung Cancer Therapy". W: *Journal of Thoracic Oncology* 11.1 (2016), s. 9–20. ISSN: 1556-0864. DOI: https://doi.org/10.1016/j.jtho.2015.08.003.

[29] Stanley E. Lazic, Charlie J. Clarke-Williams i Marcus R. Munafò. „What exactly is 'N' in cell culture and animal experiments?" W: *PLOS Biology* 16.4 (kw. 2018), s. 1–14. DOI: 10.1371/journal.pbio.2005282.

[30] Samuel J. Lord, Katrina B. Velle, R. Dyche Mullins i Lillian K. Fritz-Laylin. „SuperPlots: Communicating reproducibility and variability in cell biology". W: *Journal of Cell Biology* 219.6 (kw. 2020), e202001064. ISSN: 0021-9525. DOI: 10.1083/jcb.202001064. eprint: https://rupress.org/jcb/article-pdf/219/6/e202001064/1833825/jcb_202001064.pdf.

[31] Akram Tayanloo-Beik, Masoumeh Sarvari, Moloud Payab, Kambiz Gilany, Sepideh Alavi-Moghadam, Mahdi Gholami, Parisa Goodarzi, Bagher Larijani i Babak Arjmand. „OMICS insights into cancer histology; Metabolomics and proteomics approach". W: *Clinical Biochemistry* 84 (2020), s. 13–20. ISSN: 0009-9120. DOI: https://doi.org/10.1016/j.clinbiochem.2020.06.008.

[32] Thomas Hale, Noam Angrist, Rafael Goldszmidt, Beatriz Kira, Anna Petherick, Toby Phillips, Samuel Webster, Emily Cameron-Blake, Laura Hallas, Saptarshi Majumdar i Helen Tatlow. „A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker)". W: *Nature Human Behaviour* 5.4 (kw. 2021), s. 529–538. ISSN: 2397-3374. DOI: 10.1038/s41562-021-01079-8.

[33] Gonzalo Castex, Evgenia Dechter i Miguel Lorca. „COVID-19: The impact of social distancing policies, cross-country analysis". W: *Economics of Disasters and Climate Change* 5.1 (kw. 2021), s. 135–159. ISSN: 2511-1299. DOI: 10.1007/s41885-020-00076-x.

[34] Herbert W. Hethcote. „The Mathematics of Infectious Diseases". W: *SIAM Review* 42.4 (2000), s. 599–653. DOI: 10.1137/S0036144500371907.

[35] Luke F. Vistain i Savaş Tay. „Single-Cell Proteomics". W: *Trends in Biochemical Sciences* 46.8 (2021), s. 661–672. ISSN: 0968-0004. DOI: 10.1016/j.tibs.2021.01.013. URL: http://dx.doi.org/10.1016/j.tibs.2021.01.013.

[36] Anoop P. Patel, Itay Tirosh, John J. Trombetta, Alex K. Shalek, Shawn M. Gillespie, Hiroaki Wakimoto, Daniel P. Cahill, Brian V. Nahed, William T. Curry, Robert L. Martuza, David N. Louis, Orit Rozenblatt-Rosen, Mario L. Suvà, Aviv Regev i Bradley E. Bernstein. „Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma". W: *Science* 344.6190 (2014), s. 1396–1401. DOI: 10.1126/science.1254257.

[37] Michael J. Taylor, Jessica K. Lukowski i Christopher R. Anderton. „Spatially Resolved Mass Spectrometry at the Single Cell: Recent Innovations in Proteomics and Metabolomics". W: *Journal of the American Society for Mass Spectrometry* 32.4 (2021), s. 872–894. DOI: 10.1021/jasms.0c00439.

[38] Anjali Rao, Dalia Barkley, Gustavo S. França i Itai Yanai. „Exploring tissue architecture using spatial transcriptomics". W: *Nature* 596.7871 (2021), s. 211–220. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03634-9.

[39] Dale S Cornett, Michelle L Reyzer, Pierre Chaurand i Richard M Caprioli. „MALDI imaging mass spectrometry: molecular snapshots of biochemical systems". W: *Nature Methods* 4.10 (2007), s. 828–833. ISSN: 1548-7105. DOI: 10.1038/nmeth1094.

[40] Freddie Bray, Mathieu Laversanne, Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Isabelle Soerjomataram i Ahmedin Jemal. „Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries". W: *CA: A Cancer Journal for Clinicians* 74.3 (2024), s. 229–263. DOI: https://doi.org/10.3322/caac.21834.

[41] Joanna Didkowska, Klaudia Barańska, Marta J. Miklewska i Urszula Wojciechowska. „Cancer incidence and mortality in Poland in 2023". W: *Biuletyn Polskiego Towarzystwa Onkologicznego Nowotwory* 9.2 (2024), s. 87–105. ISSN: 2543–8077. URL: https://journals.viamedica.pl/biuletyn_pto/article/view/100824.

[42] Barbara Jarząb, Marek Dedecjus, Andrzej Lewiński i in. „Diagnosis and treatment of thyroid cancer in adult patients — Recommendations of Polish Scientific Societies and the National Oncological Strategy. 2022 Update [Diagnostyka i leczenie raka tarczycy u chorych dorosłych — Rekomendacje Polskich Towarzystw Na-

ukowych oraz Narodowej Strategii Onkologicznej. Aktualizacja na rok 2022]". W: *Endokrynologia Polska* 73.2 (2022), s. 173–300. ISSN: 2299-8306.

[43] Cari M. Kitahara i Arthur B. Schneider. „Epidemiology of Thyroid Cancer". W: *Cancer Epidemiology, Biomarkers & Prevention* 31.7 (lip. 2022), s. 1284–1297. ISSN: 1055-9965. DOI: 10.1158/1055-9965.EPI-21-1440. eprint: https://aacrjournals.org/cebp/article-pdf/31/7/1284/3175348/1284.pdf. URL: https://doi.org/10.1158/1055-9965.EPI-21-1440.

[44] Tao Lu, Xiaodong Yang, Yiwei Huang, Mengnan Zhao, Ming Li, Ke Ma, Jiacheng Yin, Cheng Zhan i Qun Wang. „Trends in the incidence, treatment, and survival of patients with lung cancer in the last four decades". W: *Cancer Management and Research* Volume 11 (2019), s. 943–953. ISSN: 1179-1322. DOI: 10.2147/cmar.s187317.

[45] P.E. Postmus, K.M. Kerr, M. Oudkerk, S. Senan, D.A. Waller, J. Vansteenkiste, C. Escriu i S. Peters. „Early and locally advanced non-small-cell lung cancer (NSCLC): ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up". W: *Annals of Oncology* 28 (2017), s. iv1–iv21. ISSN: 0923-7534. DOI: 10.1093/annonc/mdx222.

[46] D. Planchard, S. Popat, K. Kerr, S. Novello, E.F. Smit, C. Faivre-Finn, T.S. Mok, M. Reck, P.E. Van Schil, M.D. Hellmann i S. Peters. „Metastatic non-small cell lung cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up". W: *Annals of Oncology* 29 (2018), s. iv192–iv237. ISSN: 0923-7534. DOI: 10.1093/annonc/mdy275.

[47] Charlotte S. van der Vos, Daniëlle Koopman, Sjoerd Rijnsdorp, Albert J. Arends, Ronald Boellaard, Jorn A. van Dalen, Mark Lubberink, Antoon T. M. Willemsen i Eric P. Visser. „Quantification, improvement, and harmonization of small lesion detection with state-of-the-art PET". W: *European Journal of Nuclear Medicine and Molecular Imaging* 44.S1 (2017), s. 4–16. ISSN: 1619-7089. DOI: 10.1007/s00259-017-3727-z.

[48] Xuxin Chen, Ximin Wang, Ke Zhang, Kar-Ming Fung, Theresa C. Thai, Kathleen Moore, Robert S. Mannel, Hong Liu, Bin Zheng i Yuchen Qiu. „Recent advances

and clinical applications of deep learning in medical image analysis". W: *Medical Image Analysis* 79 (2022), s. 102444. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2022.102444.

[49] Hugo J. W. L. Aerts, Emmanuel Rios Velazquez, Ralph T. H. Leijenaar i in. „Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach". W: *Nature Communications* 5.1 (2014). DOI: 10.1038/ncomms5006.

[50] Wen-Fang Tang, Min Wu, Hua Bao i in. „Timing and Origins of Local and Distant Metastases in Lung Cancer". W: *Journal of Thoracic Oncology* 16.7 (2021), s. 1136–1148. ISSN: 1556-0864. DOI: 10.1016/j.jtho.2021.02.023.

[51] Vishwa Parekh i Michael A. Jacobs. „Radiomics: a new application from established techniques". W: *Expert Review of Precision Medicine and Drug Development* 1.2 (2016), s. 207–226. ISSN: 2380-8993. DOI: 10.1080/23808993.2016.1164013.

[52] Laura J. Jensen, Damon Kim, Thomas Elgeti, Ingo G. Steffen, Bernd Hamm i Sebastian N. Nagel. „Stability of Radiomic Features across Different Region of Interest Sizes—A CT and MR Phantom Study". W: *Tomography* 7.2 (2021), s. 238–252. ISSN: 2379-139X. DOI: 10.3390/tomography7020022.

[53] Meixin Zhao, Kilian Kluge, Laszlo Papp, Marko Grahovac, Shaomin Yang, Chunting Jiang, Denis Krajnc, Clemens P. Spielvogel, Alexander Haug Boglarka Ecsedi, Shiwei Wang, Marcus Hacker, Weifang Zhang i Xiang Li. „Multi-lesion radiomics of PET/CT for non-invasive survival stratification and histologic tumor risk profiling in patients with lung adenocarcinoma". W: *European Radiology* 32.10 (2022), s. 7056–7067. ISSN: 1432-1084. DOI: 10.1007/s00330-022-08999-7. URL: http://dx.doi.org/10.1007/s00330-022-08999-7.

Spis wybranych skrótów i symboli

- DNA (ang. *deoxyribonucleic acid*) — kwas deoksyrybonukleinowy.
- FGFR (ang. *fibroblast growth factor receptor*) — receptor czynnika wzrostu fibroblastów.
- NSCLC (ang. *non-small cell lung cancer*) — niedrobnokomórkowy rak płuc, najczęściej występujący podtyp tego nowotworu,
- aCGH (ang. *array comparative genomic hybridization*) — macierzowa porównawcza hybrydyzacja genomowa, technika biologii molekularnej oparta na technologii mikromacierzowej, pozwalająca na badanie zmian liczby kopii genów.
- RNA (ang. *ribonucleic acid*) — kwas rybonukleinowy.
- RNAseq (ang. *RNA sequencing*) — sekwencjonowanie RNA, wysokoprzepustowa technika molekularna, pozwalająca na badanie poziomu ekspresji genów.
- PCA (ang. *Principal Component Analysis*) — analiza głównych składowych, technika redukcji wymiarowości.
- GSEA (ang. *Gene Set Enrichment Analysis*) — analiza nadreprezentacji zestawu genów, metoda analizy szlaków sygnałowych.
- SVM (ang. *support vector machine*) — maszyna wektorów wspierających, model uczenia maszynowego.
- ROC (ang. *Receiver Operating Characteristic*) — charakterystyka operacyjna odbiornika.

- AUC (ang. *area under the ROC curve*) — pole pod krzywą ROC, miara jakości klasyfikacji.
- SEIR (ang. *Susceptible-Exposed-Infected-Removed*) — popularny kompartmentalny model epidemiologiczny, dzielący populację na cztery kompartmenty: podatnych, narażonych, zainfekowanych i usuniętych.
- NLNK (ang. *NLLSQ — non-linear least squares*) — nieliniowa metoda najmniejszych kwadratów, metoda dopasowania modelu do danych.
- NRMSE (ang. *normalized root mean square error*) — znormalizowany błąd średniokwadratowy, miara jakości dopasowania modelu.
- SD (ang. *standard deviation* — odchylenie standardowe.
- NIO (ang. *Maria Skłodowska-Curie National Research Institute of Oncology, Gliwice Branch*) — Narodowy Instytut Onkologii im. Marii Skłodowskiej-Curie — Państwowy Instytut Badawczy, Oddział w Gliwicach.
- NT (ang. *normal thyroid*) — prawidłowa (łagodna) tkanka tarczycy.
- FA (ang. *follicular adenoma*) — gruczolak pęcherzykowy, niezłośliwy guzek tarczycy.
- MTC (ang. *medullary thyroid carcinoma*) — rak rdzeniasty tarczycy, rzadko występujący, niskozróżnicowany złośliwy nowotwór tarczycy, o średnim rokowaniu.
- ATC (ang. *anaplastic thyroid carcinoma*) — rak anaplastyczny tarczycy, rzadko występujący, niskozróżnicowany złośliwy nowotwór tarczycy, o słabym rokowaniu.
- FTC (ang. *follicular thyroid carcinoma*) — rak pęcherzykowy tarczycy, często występujący, zróżnicowany złośliwy nowotwór tarczycy, o dobrym rokowaniu.
- PTC (ang. *papillary thyroid carcinoma*) — rak brodawkowaty tarczycy, często występujący, zróżnicowany złośliwy nowotwór tarczycy, o dobrym rokowaniu.
- PTC-CV (ang. *papillary thyroid carcinoma classic variant*) — podstawowy, „klasyczny” wariant raka brodawkowatego tarczycy.

- PTC-FV (ang. *papillary thyroid carcinoma follicular variant*) — wariant pęcherzykowy raka brodawkowatego tarczycy.
- MS (ang. *mass spectrometry*) — spektrometria mas, technika analityczna wyznaczająca iloraz masy do ładunku cząstek znajdujących się w próbce, jedno z podstawowych wysokoprzepustowych narzędzi wykorzystywanych w proteomice.
- MALDI-MSI (ang. *matrix-assisted laser desorption/ionization - mass spectrometry imaging*) — desorpcja/jonizacja laserowa wspomagana matrycą - obrazowanie spektrometrią mas, technologia spektrometrii mas umożliwiająca pozyskanie widm MS dla siatki punktów pokrywających próbkę, na przykład skrawek tkanki.
- ESMO (ang. *European Society for Medical Oncology*) - Europejskie Towarzystwo Onkologii Medycznej, stowarzyszenie zajmujące się między innymi przygotowaniem rekomendacji terapeutycznych dla nowotworów.
- PET (ang. *positron emission tomography*) — pozytonowa tomografia emisyjna, rodzaj obrazowania medycznego wykorzystujący radioizotopy do określenia aktywności metabolicznej.
- CT (ang. *computed tomography*) — tomografia komputerowa, rodzaj obrazowania medycznego wykorzystujący promieniowanie rentgenowskie.
- ROI (ang. *region of interest* — region zainteresowania, obszar obrazu poddawany analizie.
- TNM (ang. *Tumour, Node, Metastasis*) — klasyfikacja stopnia zaawansowania nowotworów, opisująca wielkość guza pierwotnego (T), przerzuty w węzłach chłonnych (N) i przerzuty odległe (M).
- MFS (ang. *metastasis-free survival*) — przeżycie wolne od przerzutów odległych.
- MCCV (ang. *Monte Carlo cross-validation*) — kroswalidacja typu Monte Carlo, rodzaj walidacji krzyżowej polegającym na wielokrotnych, losowych podziałach zbioru danych na uczący i testowy/walidacyjny.

Spis rysunków

Spis tabel

Suplement

Teksty publikacji wchodzących w skład cyklu

cells

MDPI

Article

p38 Mediates Resistance to FGFR Inhibition in Non-Small Cell Lung Cancer

Izabela Zarczynska [1], Monika Gorska-Arcisz [1], Alexander Jorge Cortez [2], Katarzyna Aleksandra Kujawa [3], Agata Małgorzata Wilk [2,4], Andrzej Cezary Skladanowski [1], Aleksandra Stanczak [5], Monika Skupinska [6], Maciej Wieczorek [5], Katarzyna Marta Lisowska [3], Rafal Sadej [1,*] and Kamila Kitowska [1,*]

1 Department of Molecular Enzymology and Oncology, Intercollegiate Faculty of Biotechnology, University of Gdansk and Medical University of Gdansk, Debinki 1, 80-211 Gdansk, Poland; izabela.zarczynska@gumed.edu.pl (I.Z.); monika.gorska@gumed.edu.pl (M.G.-A.); acskla@gumed.edu.pl (A.C.S.)
2 Department of Biostatistics and Bioinformatics, Maria Sklodowska-Curie National Research Institute of Oncology, Gliwice Branch, Wybrzeze Armii Krajowej 15, 44-102 Gliwice, Poland; alexander.cortez@io.gliwice.pl (A.J.C.); Agata.Wilk@io.gliwice.pl (A.M.W.)
3 Center for Translational Research and Molecular Biology of Cancer, Maria Sklodowska-Curie National Research Institute of Oncology, Gliwice Branch, Wybrzeze Armii Krajowej 15, 44-102 Gliwice, Poland; Katarzyna.Kujawa@io.gliwice.pl (K.A.K.); Katarzyna.Lisowska@io.gliwice.pl (K.M.L.)
4 Department of Systems Biology and Engineering, Silesian University of Technology, 44-100 Gliwice, Poland
5 Clinical Development Department, Celon Pharma S.A., Marymoncka 15, 05-152 Kazuń Nowy, Poland; aleksandra.stanczak@celonpharma.com (A.S.); maciej.wieczorek@celonpharma.com (M.W.)
6 Preclinical Development Departament, Celon Pharma S.A., Marymoncka 15, 05-152 Kazuń Nowy, Poland; monika.skupinska@celonpharma.com
* Correspondence: rafal.sadej@gumed.edu.pl (R.S.); kamila.kitowska@gumed.edu.pl (K.K.)

Abstract: FGFR signalling is one of the most prominent pathways involved in cell growth and development as well as cancer progression. FGFR1 amplification occurs in approximately 20% of all squamous cell lung carcinomas (SCC), a predominant subtype of non-small cell lung carcinoma (NSCLC), indicating FGFR as a potential target for the new anti-cancer treatment. However, acquired resistance to this type of therapies remains a serious clinical challenge. Here, we investigated the NSCLC cell lines response and potential mechanism of acquired resistance to novel selective FGFR inhibitor CPL304110. We found that despite significant genomic differences between CPL304110-sensitive cell lines, their resistant variants were characterised by upregulated p38 expression/phosphorylation, as well as enhanced expression of genes involved in MAPK signalling. We revealed that p38 inhibition restored sensitivity to CPL304110 in these cells. Moreover, the overexpression of this kinase in parental cells led to impaired response to FGFR inhibition, thus confirming that p38 MAPK is a driver of resistance to a novel FGFR inhibitor. Taken together, our results provide an insight into the potential direction for NSCLC targeted therapy.

Keywords: FGFR; p38; acquired resistance; lung cancer

Citation: Zarczynska, I.; Gorska-Arcisz, M.; Cortez, A.J.; Kujawa, K.A.; Wilk, A.M.; Skladanowski, A.C.; Stanczak, A.; Skupinska, M.; Wieczorek, M.; Lisowska, K.M.; et al. p38 Mediates Resistance to FGFR Inhibition in Non-Small Cell Lung Cancer. *Cells* **2021**, *10*, 3363. https://doi.org/10.3390/cells10123363

Academic Editor: T.K.S. Kumar

Received: 28 October 2021
Accepted: 26 November 2021
Published: 30 November 2021

1. Introduction

Lung cancer is the most common cause of cancer-related deaths in men and women, with 2,206,771 new cases and 1,796,144 deaths in 2020 worldwide [1]. Non-small cell lung carcinoma (NSCLC), which comprises approximately 84% of all lung cancers, is categorised into adenocarcinoma (ADC, approximately 40–50% of all cases), squamous cell carcinoma (SCC, approximately 20–30% of all cases), and large cell carcinoma (LCC, 10% of all cases). Gene alterations and rearrangements in *EGFR*, *ALK*, *ROS1*, *BRAF*, or *NTRK* typical for ADC were discovered in the last two decades, which led to the development of targeted therapies using tyrosine kinase inhibitors (TKI) [2]. Such drugs, applied as an alternative to conventional chemotherapy and radiotherapy in first-line treatment of advanced ADC,

have significantly improved clinical outcomes. Contrary to ADC, SCCs harbour distinct types of gene alterations: amplifications (*MET*, *HER2*, and *FGFR*) [3] and gene mutations (*CDKN2A*, *PTEN*, *KEAP1*, *MLL2*, *HLA-A*, *NFE2L2*, *NOTCH1*, and *RB1*) [4]. Therefore, several clinical trials are underway, developing the targeted therapies specifically for lung squamous cell carcinomas (SCC).

The FGFR family consists of four transmembrane receptor tyrosine kinases (FGFR1-4) activating multiple signalling pathways, including RAS/RAF/MAPK, PI3K/AKT, and STAT involved in the regulation of proliferation, cell survival, migration, and invasion [5]. FGFR-related genomic alterations, i.e., gene amplification, chromosomal translocation, gain-of-function mutations, and gene fusions, lead to constitutive receptor activation or enhanced signalling [6,7]. FGFR1 amplification is one of the most common genomic alterations in SCC, occurring in 10–20% of cases. Several studies demonstrated that SCCs also harbour FGFR2 and FGFR3 fusions [7,8]. Since deregulated FGFR signalling has been implicated in oncogenesis and cancer progression, FGFR emerged as a promising target for anti-cancer therapies [9]. A few FGFR tyrosine kinase inhibitors (TKIs) have been approved for clinical use, and several are currently undergoing preclinical and clinical investigation in various FGFR-associated tumours (NCT02965378, NCT03762122, NCT02154490, NCT03827850). Although small-molecule inhibitors of FGFR activity represents a promising anti-cancer strategy, emerging resistance to applied drug remains a growing challenge.

So far, several mechanisms of acquired resistance to therapies targeting FGFR have been presented. In general, acquired resistance to TKIs can develop either as a result of secondary mutations within the ATP-binding domain, preventing the receptor from inhibitor binding, or through activation of alternative signalling pathways that circumvent the FGFR signalling cascade [10–13]. In vitro studies performed in lung cancer cell lines revealed that MET upregulation followed by reactivation of the ERK/MAPK pathway is involved in the development of resistance to BGJ398 [14]. Moreover, the AKT pathway was shown to mediate resistance to BGJ398 in lung and urothelial cancer cell lines [15]. These results are in line with studies indicating that incomplete suppression of the key survival pathways: PI3K/AKT and MAPK by BGJ398 or PD173074 in lung and colorectal cancer cell lines may be associated with acquired resistance to FGFR inhibitors [16].

In this study, we analysed the sensitivity of a panel of lung cancer cell lines to CPL304110 (Celon Pharma, Poland), a novel pan-FGFR inhibitor that is currently in phase I of a clinical trial in adults with advanced solid malignancies (NCT04149691) [17]. We identified NCI-H1581 and NCI-H1703 non-small cell lung cancer cell lines as highly sensitive to inhibition of FGFR. We found that despite genomic and transcriptomic divergence between parental cells, CPL304110-resistant variants of these cell lines displayed increased expression of the genes encoding members of the p38 signalling pathway. Moreover, the overexpression of p38 kinase resulted in resistance to CPL304110 in parental cells, while inhibition of p38 MAPK resensitised resistant cells to CPL304110 treatment. These confirmed the importance of p38 MAPK in the process of acquisition of resistance to inhibition of FGFR signalling in NSCLC cell lines.

2. Materials and Methods

2.1. Cell Lines and Cell Culture Reagents

DMS114, NCI-H1581, NCI-H1703, NCI-H2170, and NCI-H520 cell lines were obtained from ATCC. NCI-H1581 cells were routinely maintained in DMEM/F12; DMS114 in Waymouth's MB752/1; whereas NCI-H1703, NCI-H2170, and NCI-H520 were maintained in RPMI 1640 medium. All culture media contained 10% of FBS and penicillin/streptomycin (100 U/mL/100 µg/mL). Cells were grown at 37 °C in a humidified atmosphere of 5% CO_2. All culture media and corresponding supplements were purchased from Merck KGaA (Darmstadt, Germany) or Biowest (Riverside, MO, USA). CPL304110 (WO/2014/141015) inhibitor was provided by Celon Pharma S.A., Poland [17]. SB202190 and SB203580 inhibitors were purchased from Selleckchem (Houston, TX, USA).

2.2. Cell Proliferation Assay

Cell viability was estimated using the 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide (MTT) colorimetric assay Merck KgaA (Darmstadt, Germany). Cells were seeded in 96-well plates in triplicates and on the following day treated with DMSO or indicated inhibitor for 48 and 96 h. MTT stock solution was added to each well so that the final concentration of MTT in the medium was 0.5 mg/mL. After 2 h incubation at 37 °C, the medium was discarded, and MTT formazan was dissolved in DMSO. The absorbance was measured at 590 nm using a microplate reader.

2.3. Clonogenic Assay

Cells were seeded in 6-well plates. On the following day, the medium was replaced with a regular medium or medium containing indicated inhibitors. Media were changed every three days. Cells were cultured for 10–14 days, depending on the cell line, fixed with 4% paraformaldehyde and stained with 0.4% crystal violet.

2.4. Culturing Cells in Three-Dimensional BD Matrigel®

Cell growth in three-dimensional (3D) BD Matrigel® (BD Matrigel Matrix Growth Factor Reduced, BD Bioscience, Corning, NY, USA) was carried out as previously described [18]. Cells were cultured in regular medium or medium containing indicated inhibitors. Media were replaced every three days. After 14 days of culture, cell growth was evaluated by measuring colonies size (at least 100) using ZEISS PrimoVert microscope and ImageJ software.

2.5. Generation of CPL304110-Resistant Cell Lines

To develop resistance to the FGFR inhibitor, CPL304110, NCI-H1581, and NCI-H1703 cell lines were exposed to increasing concentrations of CPL304110 (starting from 50 nM). Cells were maintained in a medium containing the inhibitor, which was replaced every three days. When the growth kinetics of treated cells were similar to wild-type cells, the concentration of CPL304110 was increased until a final concentration of 2.5 μM for NCI-H1581 and 5 μM for NCI-H1703 was achieved. After 4-6 months of such culture, resistant cells were established and termed NCI-H1581R and NCI-H1703R.

2.6. Overexpression of p38

To generate cells overexpressing p38 MAPK NCI-H1581 and NCI-H1703 cells were seeded onto 6 cm plates and, after 24 h, transfected with pMT3-p38-HA plasmid (Addgene, #12658) using Lipofectamine 3000 (Invitrogen, Thermo Fisher Scientific, MA, USA). The overexpression was confirmed with Western blotting.

2.7. Western Blotting Analysis

For Western blotting analysis, cells were harvested at 60–70% confluency and lysed with 2x concentrated Laemmli buffer supplemented with 2 mM PMSF, 10 μg/mL aprotinin, 10 μg/mL leupeptin, 5 mM EGTA, 1 mM EDTA, 2 mM $Na_4P_2O_7$, 5 mM NaF, and 5 mM Na_3VO_4. Samples with equal amounts of protein were loaded per well, resolved in SDS-PAGE, and then transferred onto a nitrocellulose membrane. The membranes were blocked for 1 h in 5% skimmed milk and probed with specific primary antibodies at 4 °C. Antibodies for anti-FGFR1 (sc-57132), anti-FGFR3 (sc-13121), anti-FGFR4 (sc-136988), and anti-FRS2-α (sc-17841) were obtained from Santa Cruz Biotechnology (Dallas, TX, USA). The antibody against β-actin (A5316) was obtained from Merck KGaA (Darmstadt, Germany). All the remaining antibodies were from Cell Signaling Technology (Danvers, MA, USA): anti-Akt-Ser473 (#9271) anti-Akt (#92720), anti-CDK4 (#12790), anti-CDK6 (#3136), anti-Erk1/2-Thr202/Tyr204 (#4377), anti-Erk1/2 (#9102), anti-FGFR-Tyr653/654 (#3471), anti-FGFR2 (#23328), anti-FRS2-α-Tyr196 (#3864), anti-FRS2-α-Tyr436 (#3861), anti-PLC-γ-1-Tyr783 (#2821), anti-PLC-γ-1 (#2822), anti-p27-Kip1 (#3686), anti-p38-MAPK-Thr180/Tyr182 (#9211), anti-p38-MAPK (#9212), anti-Rb-Ser780 (#9307). Appropriate

secondary Alexa Fluor®-conjugated antibodies (680 or 790 nm) (Jackson ImmunoResearch, #111-625-144, #715-655-150) and Odyssey® CLx imaging system (LI-COR® Biosciences, NE, USA) were used to detect protein bands.

2.8. Cell Cycle Analysis

The cell cycle was analysed by quantification of DNA content using flow cytometry. Cells were fixed in 70% ethanol for 24 h at −20 °C, RNase A (1 mg/mL) was added, EURX Ltd. (Gdansk, Poland), and cells were stained with propidium iodide (2,5 μg/mL) (PI; #P4170, Sigma-Aldrich; Merck KGaA). The cell cycle was analysed with FACSCalibur™; Becton Dickinson and Company (San Jose, CA, USA). The results were analysed using the CellQuest™ Pro Software version 6.0 Becton Dickinson and Company (San Jose, CA, USA).

2.9. Array-Based Comparative Genomic Hybridisation (aCGH)

The aCGH was used to reveal copy number changes (amplifications and deletions) in the genome of sensitive versus resistant cells (NCI-H1581 vs. NCI-H1581R and NCI-H1703 vs. NCI-H1703R). DNA from the cells was isolated using GeneJET Genomic DNA Purification Kit, Thermo Fisher Scientific (Waltham, MA, USA). Genomic DNA was analysed by hybridisation to the 60K SurePrint G3 Unrestricted CGH arrays, Agilent Technologies, Inc. (Santa Clara, CA, USA) service provided by Genomed S.A., Warsaw, Poland. Normal human Caucasian male genome GRCh38 (hg38) was used as the reference.

2.10. aCGH Data Analysis

Raw image files were preprocessed using Agilent Feature Extraction software (version 11.0.1.1). Data were checked for quality, and features were extracted for CGH_1100_Jul11 protocol. Further analysis was conducted with Bioconductor package rCGH version 1.20.0 released on 28 October 2020 according to standard protocol dedicated for Agilent dual-colour hybridisation chips [19]. First, the signals were adjusted for GC content and cy3/cy5 bias, after which the log2 relative ratios (LRR) could be computed. Next, the genome profiles were segmented using the Circular Binary Segmentation algorithm, and finally, the LRRs were centred using an expectation–maximisation algorithm. Results were visualised using Rcircos version 1.2.1 released on 12 March 2019 and pheatmap version 1.0.12 released on 4 January 2019 (https://cran.r-project.org/web/packages/pheatmap/index.html, accessed on 28 October 2021) R packages [20]. Genes were considered differential between variants if the absolute value of LRR difference was greater than 0.5. All analyses were performed using R environment for statistical computing version 4.0.3 "Bunny-Wunnies Freak Out" released on 10 October 2020 (R Foundation for Statistical Computing, Vienna, Austria, http://www.r-project.org, accessed on 28 October 2021).

2.11. Transcriptome Sequencing (RNA-Seq)

For RNA-seq experiment, NCI-H1581 and NCI-H1703 cells (sensitive and resistant cell line variants) were seeded onto 6 cm-diameter dishes. The next day, the medium was replaced with a fresh medium, and after 24 h, cells were harvested. This procedure was repeated to obtain a second replicate. Total RNA was extracted from the cells using RNeasy Plus Mini Kit, Qiagen (Hilden, Germany) with simultaneous DNase I digestion, according to the manufacturer's instructions. RNA purity and concentration were estimated with a Nanodrop ND-2000 spectrophotometer, Thermo Fisher Scientific (Waltham, MA, USA). RNA quality was assessed using the 2100 Bioanalyzer with the RNA 6000 Nano Kit, Agilent Technologies (Santa Clara, CA, USA). All the samples had an RNA integrity number (RIN) above 7.0. cDNA library preparation and transcriptome sequencing were completed by Genomed, Warsaw, Poland. RNA-seq was performed using the Illumina HiSeq4000 Platform with the standard paired-end protocol (58 mln paired reads, 100 bp read length).

2.12. RNA-seq Data Analysis

After standard quality control, the raw sequencing data were quantified using the Salmon tool against the reference genome GRCh38 (hg38) [21]. Quantified transcripts were imported into the R environment with tximport [22]. Low-abundance genes were pre-filtered, keeping only rows with at least 10 reads total. Gene counts were normalised using the median-of-ratios method [23]. For unsupervised analysis, regularised logarithm (rlog) transformed data were used. Principal Component Analysis (PCA) was applied to assess the main sources of variability in data. Hierarchical clustering of 500 genes exhibiting the largest overall variance was performed to explore relationships between samples–heatmaps were generated with pheatmap R package (https://cran.r-project.org/web/packages/pheatmap/index.html, accessed on 28 October 2021). Both the unsupervised methods revealed significant differences between cell lines, effectively overshadowing differences between variants, for which reason further analysis was performed separately for the two cell lines. Differentially expressed genes were identified using DESeq2 package [23] version 1.30.1 released on 20 February 2021, with FDR adjusted (Benjamini–Hochberg correction) *p*-value cut-off 0.1. Overrepresented signalling pathways were identified with Gene Set Enrichment Analysis (GSEA) method [24] implemented in Bioconductor package clusterProfiler version 3.14.3 released on 8 January 2020 [25]. Gene sets between 10 and 700 bp were considered, and the permutation number was set to 10,000. Pathways were downloaded from MsigDB version 7.3 [26]. For global analysis, the C2:CP (Curated gene sets: Canonical pathways) collection was used. All analyses were performed using R environment for statistical computing version 4.0.3 "Bunny-Wunnies Freak Out" released on 10 October 2020 (R Foundation for Statistical Computing, Vienna, Austria, http://www.r-project.org, accessed on 28 October 2021).

2.13. Statistical Analysis

All data are expressed as mean ± SD from at least three independent experiments. Comparative data were analysed with the unpaired Student's *t*-test using the Statistica™ software, v.10; TIBCO Software Inc. (Palo Alto, CA, USA). Two-sided $p \leq 0.05$ was considered a statistically significant difference.

3. Results

3.1. Differential Response of NSCLC Cell Lines to CPL304110 Inhibitor

Approximately 20% of all non-small cell lung cancer cases are characterised by FGFR1 amplification, indicating FGFR as a promising target for anti-cancer therapies [16]. Therefore, we analysed the response of five lung cancer cell lines to the new FGFR inhibitor, CPL304110. Evaluation of cell proliferation revealed that NCI-H1581 and NCI-H1703 cells strongly responded to CPL304110, showing a significant growth reduction with IC50 < 1 μM, while DMS114, NCI-H2170, and NCI-H520 cells were less sensitive to FGFR inhibition (IC50 ≥ 1 μM) (Figure 1A). In accordance with these results, significant CPL304110-mediated 3D growth inhibition, even in 0.1 μM, relatively low drug concentration, was found for both cell lines (Figure 1B). Since NCI-H1703 cells exhibit highly invasive growth and do not form classical spheroids in 3D culture, the quantitative growth analysis was possible only for the NCI-H1581 cell line (Figure S1). Further investigation revealed that response to CPL304110 correlated with FGFR protein expression, as Western blotting analysis showed the highest FGFR1-4 expression levels in NCI-H1581 and NCI-H1703 among analysed cell lines (Figure 1C).

Although NCI-H1581 and NCI-H1703 cells were both found sensitive to CPL304110, array based-comparative genome hybridisation (aCGH) analysis revealed significant genetic differences between them. Both cell lines showed significant copy number variation (CNV) compared to the normal human genome, which was distinct for each cell line, as illustrated by the circos plot (Figure 1D). Although the NCI-H1581 genome was characterised by the prevalence of gain of DNA copy number events, loss of copy number was

observed only within chromosome Y. On the other hand, the NCI-H1703 genome was rich in both gains and losses of DNA copy number.

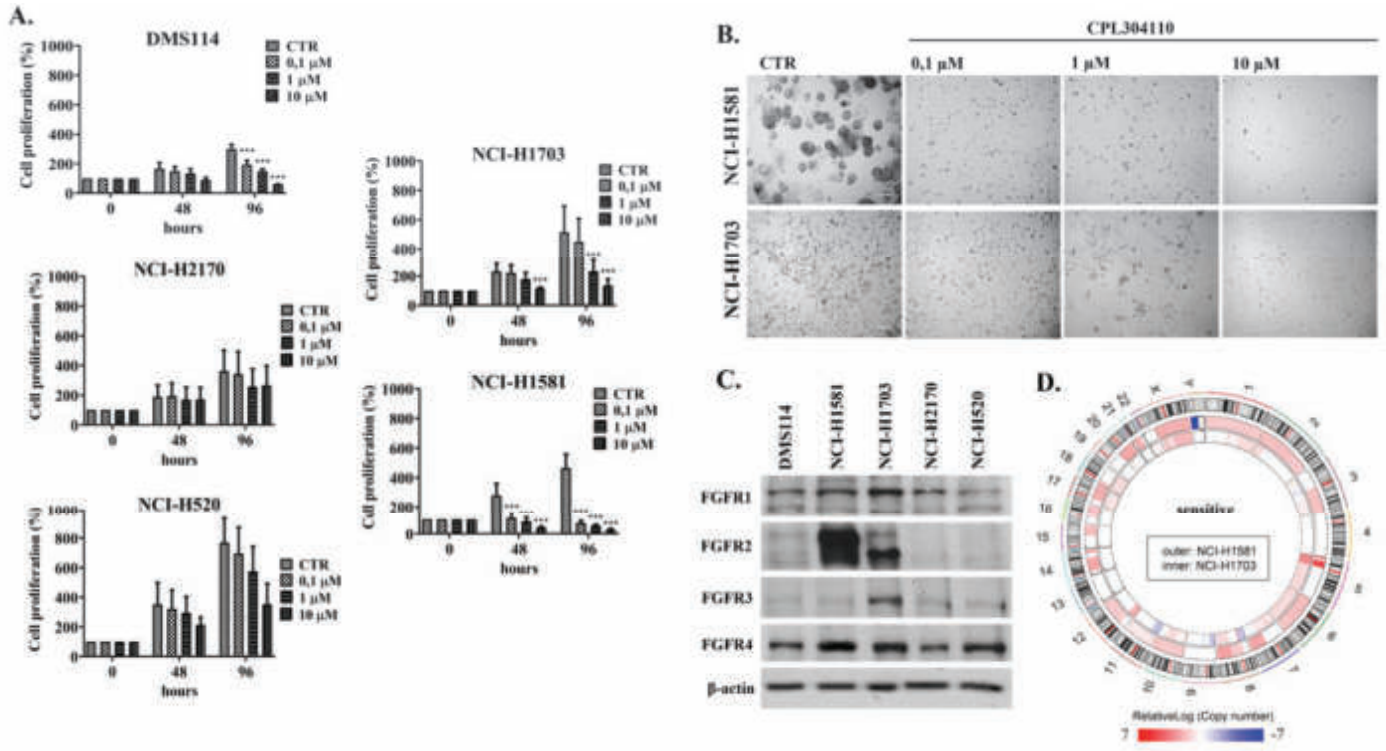


Figure 1. Lung cancer cell line response to a novel FGFR inhibitor. (**A**) DMS114, NCI-H1581, NCI-H1703, NCI-H2170, and NCI-H520 cell lines were exposed to the indicated CPL304110 concentrations for 48 and 96 h. The anti-proliferative effect of the inhibitor was assessed using the MTT cell viability test. Data are expressed as mean ± SD, *** $p \leq 0.001$ compared to non-treated cells, n = 3. (**B**) NCI-H1581 and NCI-H1703 cells were grown in 3D BD Matrigel® for 14 days in the presence of CPL304110 in indicated concentrations. Representative pictures were taken. Scale bar represents 100 μm, n = 3. (**C**) Western blot analysis of FGFR1-4 protein expression was performed for lysates of all five lung cancer cell lines. Experiments were conducted in triplicates. Representative blots are shown. (**D**) The circos plot depicts aCGH-derived DNA copy number profiles (relativeLog) of the analysed lung cancer cell lines. The outer circle represents chromosome cytobands (centromeres are shown as red bars) of the reference human genome GRCh38; middle and inner circle show a copy number changes in the NCI-H1581 and NCI-H1703 cells (respectively), in comparison to the reference genome. Numbers and letters on the outside indicate chromosomes.

3.2. CPL304110 Induces Cell Cycle Arrest in Sensitive Cells

CPL304110-resistant cell line variants (NCI-H1581R and NCI-H1703R) were developed to investigate the mechanism of acquired resistance to FGFR inhibition. Resistance to CPL304110 has been confirmed for both cell lines with analysis of 3D cell growth in BD Matrigel® (Figure 2A and Figure S2A) and proliferation assay (Figure 2B). Analysis of cell cycle revealed that it was affected by inhibition of FGFR only in sensitive cells.

CPL304110-induced cell cycle arrest in G0/G1 phase was observed for NCI-H1581 and NCI-H1703 but not in resistant cell variants (Figure 2C). This has been supported by analysis of CPL304110 impact on expression/activity of proteins involved directly in regulation of cell cycle. Upon CPL304110 treatment, sensitive cells displayed a downregulated expression of CDK4 and CDK6 together with a decrease in Rb phosphorylation and upregulation of p27, which is a well-known cell cycle inhibitor. For both resistant variants, treatment with CPL304110 did not affect the level of any of the cell cycle-related proteins (Figure S2B).

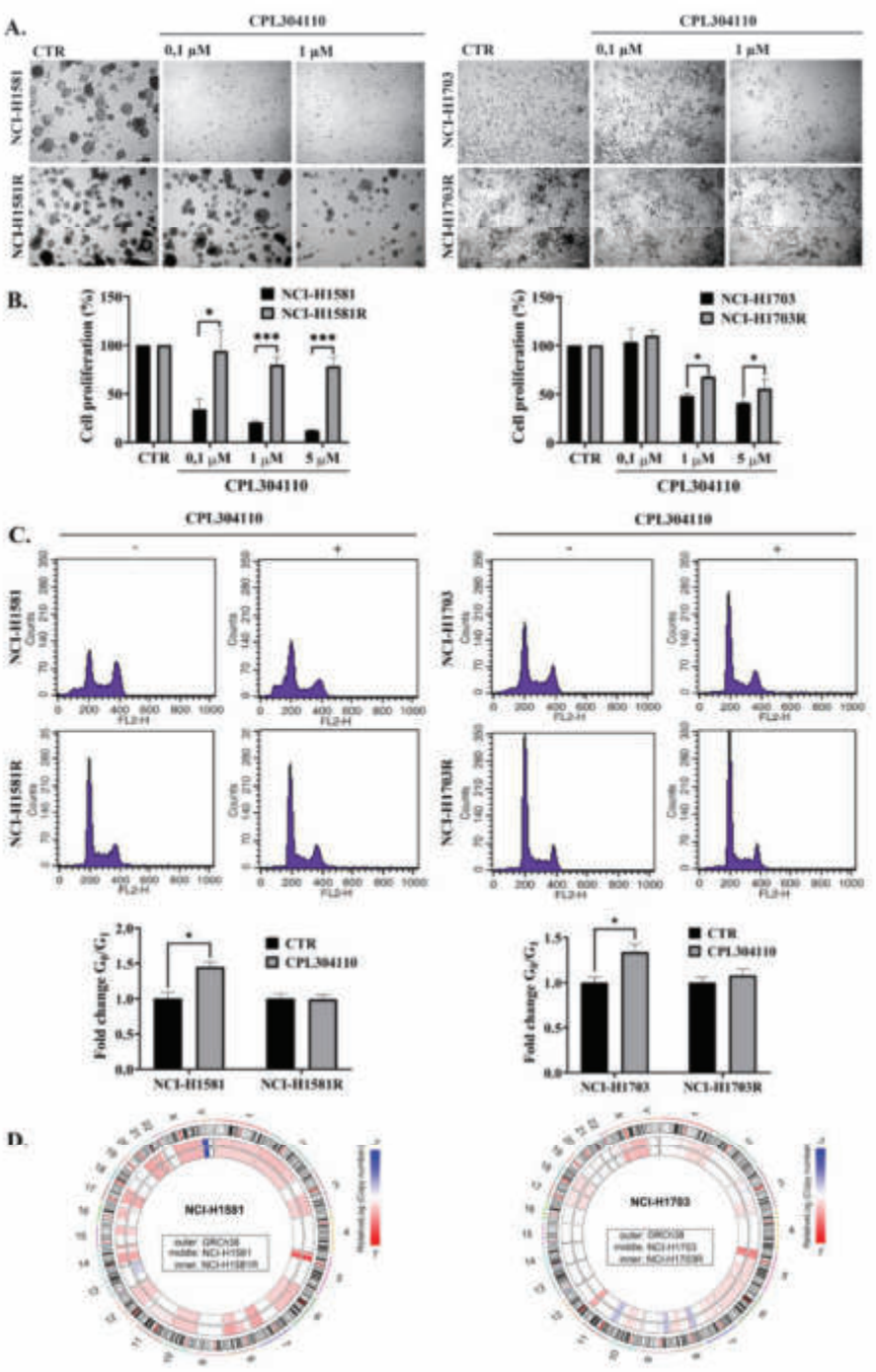


Figure 2. CPL304110-mediated cell cycle arrest. Resistance to CPL304110 was induced in NCI-H1581 and NCI-H1703 by chronic exposure to CPL304110. (**A**) Response of sensitive and resistant cells to the FGFR inhibitor was analysed in 3D BD Matrigel®. After 14 days of culture, representative pictures were taken. Scale bar represents 100 μm, n = 3. (**B**) Proliferation of sensitive and resistant cells in the presence of CPL304110 was evaluated with MTT test. Data are expressed as mean ± SD, * $p \leq 0.01$, *** $p \leq 0.001$, n = 3. (**C**) Cell cycle analysis of CPL304110-treated cells. Sensitive and resistant cell variants were serum-starved and subsequently treated with CPL304110 (0.1 μM for NCI-H1581 and NCI-H1581R; 1 μM for NCI-H1703 and NCI-H1703R) for 48 h. Bar graphs represent the fold change of cells arrested in the G0/G1 phase. Data are expressed as mean ± SD, * $p \leq 0.01$, n = 3. (**D**) Circos plots representing a copy number variation (relativeLog) in NCI-H1581 vs. NCI-H1581R (left circos) and NCI-H1703 vs. NCI-H1703R (right circos). The outer circle represents chromosome cytobands (the centromeres are shown as a red bar) of the reference human genomeGRCh38 middle and the inner circle represents a copy number variation in the sensitive and resistant variant (respectively), in comparison to reference genome. Numbers and letters on the outside indicate chromosomes.

In Principal Component Analysis, the main source of variance in gene expression profile was related to the type of cell line. First Principal Component (PC1) was responsible for 74% of the variance in gene expression and was related to the NCI-H1581 vs. NCI-H1703 difference. PC2 was identified as related to NCI-H1581R vs. NCI-H1581 difference (13% of variance), while PC3 was related to NCI-H1703R vs. NCI-H1703 difference (only 5% of variance) (Figure S2C). Additionally, hierarchical clustering showed that samples group preferentially by cell line type but not by their sensitivity to FGFR inhibition (resistant vs. sensitive) (Figure S2D). The comparison of CNV patterns between sensitive and resistant variants revealed several differences, which might be related to the acquisition of resistance to CPL304110 (Figure 2D). Interestingly, different changes were observed in each cell line. The majority of genetic events observed in NCI-H1581R were connected with a reduction of DNA copy number compared to NCI-H1581 but still resulting in more DNA copies than in normal reference genome. This trend was observed in large parts of chromosomes 2, 3, 6, 8, 12, and 15 and small portions of chromosomes 10 and 18. On the contrary, NCI-H1703R showed an increase in DNA copy number within the portion of chromosomes 3, 5, 8, 22, and X, normalisation of DNA copy number in the fragment of chromosomes 3, 4, and 5, and decrease of DNA copy number in part of chromosome 11. Such a diverse pattern of copy number alterations accompanying acquired resistance to CPL304110 may suggest different mechanisms of drug resistance in analysed cell lines.

3.3. Identification of Potential Biomarker of Acquired Resistance to CPL304110 Inhibitor

In further experiments, the molecular mechanism of acquired cell resistance to CPL304110 was investigated. Multiple studies demonstrated that the mechanisms of acquired resistance to TKIs are associated with reactivation of downstream signalling, which initially is switched off by applied inhibitors [27,28]. When CPL304110-sensitive cells were compared to their resistant variants, it was found that the phosphorylation of FGFR and its direct downstream effector proteins Fibroblast Growth Factor Receptor Substrate 2-α (FRS2-α) and phospholipase C-γ-1 (PLC-γ-1) were decreased in NCI-H1581R and NCI-H1703R (Figure 3A). Additionally, we investigated the expression and phosphorylation levels of AKT, ERK, and p38 MAPK, which have been previously implicated in mediating the acquired resistance to RTK inhibitors in lung cancer cells [14,15,29,30]. For both CPL304110-resistant cell lines, no changes in expression level, as well as phosphorylation, were observed for AKT and ERK compared to corresponding sensitive variants.

Interestingly, NCI-H1581R and NCI-H1703R demonstrated an increased expression and phosphorylation of p38 kinase, indicating its upregulation as a possible common mechanism of acquired resistance to FGFR inhibition. Additionally, Western blot analysis of p38 expression level in all five originally investigated lung cancer cell lines (Figure S3) confirmed that it is related to acquired resistance and does not correlate with the initial cells response to CPL304110.

To investigate the genomic changes related to CPL304110 resistance, a copy number analysis of p38-related genes was performed (Figure 3B,C). In both NCI-H1581R and NCI-H1703R, the analysis revealed a decrease in copy number of the following genes: *MYC*, *PTK2*, *ADCY8*, *MAFA*, *MAPK8IP1*, *EIF4EBP1*, and *FGFR1*. Our RNA-sequencing analyses of p38 signalling (Figure S4A–F) revealed a panel of 54 genes common for both analysed cell lines with the same pattern of changes in expression following the acquisition of resistance to CPL304110. Among these genes, members of MAPK, TGFβ, PI3K, EGF, and FGF pathways were found, as well as genes involved in the regulation of cell cycle and apoptosis (Figure 3D). Interestingly, three genes (*MYC*, *EIF4EBP1*, and *FGFR1*) showed a decrease in both copy number and expression in resistant cell variants.

3.4. p38 Mediates Resistance to FGFR Inhibition in Lung Cancer Cells

Previous results led to an assumption that p38 kinase may be involved in the mechanism of resistance to CPL304110. To further verify whether p38 mediates resistance to FGFR inhibitor, we analysed how SB202190, a chemical inhibitor of p38, affects cell growth

in 3D BD Matrigel®, proliferation, and response to CPL304110. The inhibitory effect of SB202190 on p38 activity was analysed with Western blot (Figure S5). Interestingly, we observed that the inhibition of p38 resulted in reduced cell growth in 3D and impaired the colony formation of CPL304110-resistant cells. Furthermore, the inhibition of p38 with SB202190 restored the sensitivity of NCI-H1581R and NCI-H1703R cells to CPL304110 as dual inhibition of FGFR and p38 led to reduced cell growth in 3D and proliferation (Figure 4A,B and Figure S6), as well as impaired colony formation (Figure S7). In concordance with these results, we observed reduced cell growth in 3D upon treatment with another commercially available p38 inhibitor, SB203580 (Figure S8). These results further indicated p38 involvement in CPL304110 resistance.

As inhibition of p38 resensitised CPL304110-resistant cells to FGFR inhibition, we investigated whether the overexpression of p38 was sufficient to confer resistance to CPL304110 in sensitive cells. We established two cell lines (NCI-H1581/p38↑ and NCI-H1703/p38↑) with a stable overexpression of p38 (and concomitant increase in p38 phosphorylation) (Figure 5A) to mimic p38 status in CPL304110-resistant cells. Interestingly, the ectopic overexpression of p38 led to significantly impaired sensitivity to CPL304110, as the proliferation and 3D growth of NCI-H1581/p38↑ and NCI-H1703/p38↑ were not or barely affected by FGFR inhibition (Figure 5B,C and Figure S9). The obtained results indicate that acquired resistance to FGFR inhibition in lung cancer cells is mediated by p38, suggesting that dual inhibition of FGFR and p38 might serve as a feasible direction in lung cancer targeted therapy.

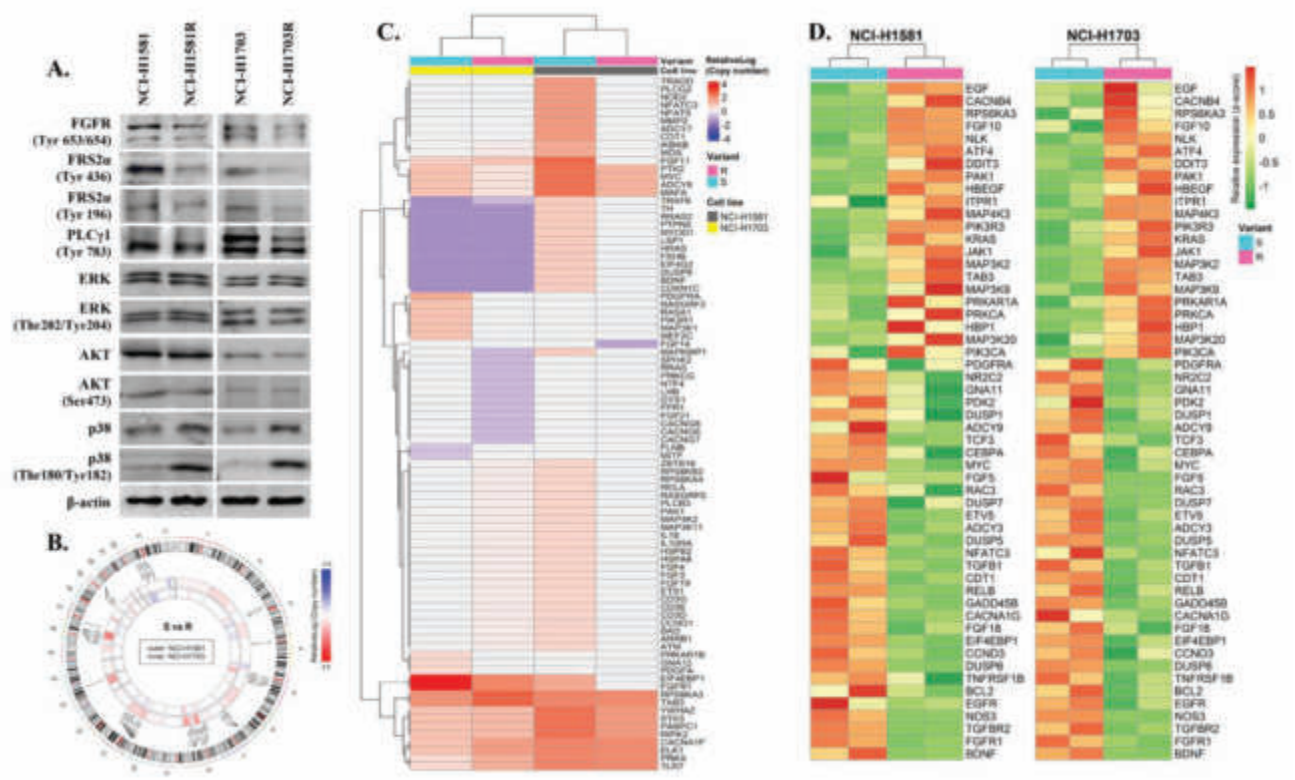


Figure 3. Involvement of p38 kinase in resistance to FGFR inhibition. (**A**) Protein expression/phosphorylation levels of FGFR and its direct downstream effectors were analysed with Western blot. Experiments were conducted in triplicates. Representative blots are shown. (**B**) A circos plot showing a copy number variation (relativeLog) in resistant variant of NCI-H1581 (outer circle) and NCI-H1703 (inner circle) cell lines in comparison to respective sensitive variant. Enlarged circos plot with a legible gene names is shown in Supplementary Materials. (**C**) A heat map showing genes with copy number variation (CNV) between sensitive versus resistant variant (data subjected to rlog transformation). (**D**) A heat map showing differentially expressed genes between sensitive versus resistant cell line variants (only genes with the mutual direction of change in both analysed cell lines are shown; data subjected to rlog transformation).

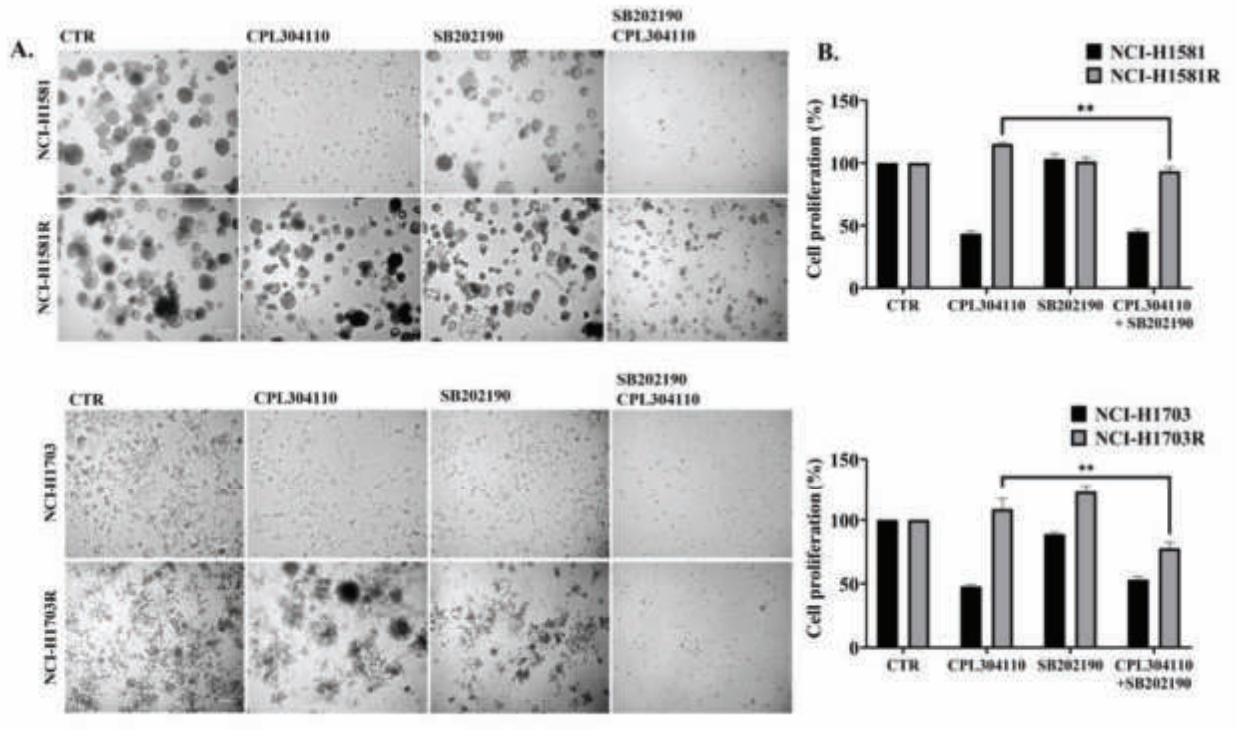


Figure 4. p38 activity mediates in CPL304110-induced cell growth inhibition. (**A**) Sensitive and resistant variants of NCI-H1581 and NCI-H1703 cells were grown with CPL304110 (0.1 μM for NCI-H1581 and NCI-H1581R; 1 μM for NCI-H1703 and NCI-H1703R) and/or SB202190 (2 μM) in 3D BD Matrigel®. Cell growth was measured with ImageJ software after 14 days of culture. Representative pictures were taken. Scale bar represents 100 μm, n = 3. (**B**) Proliferation analysis was evaluated by MTT in sensitive and resistant cells exposed to CPL304110 (0.1 μM for NCI-H1581 and NCI-H1581R; 1 μM for NCI-H1703 and NCI-H1703R) and/or SB202190 (2 μM) for 96 h. Data are expressed as mean ± SD, ** $p \leq 0.005$, n = 3.

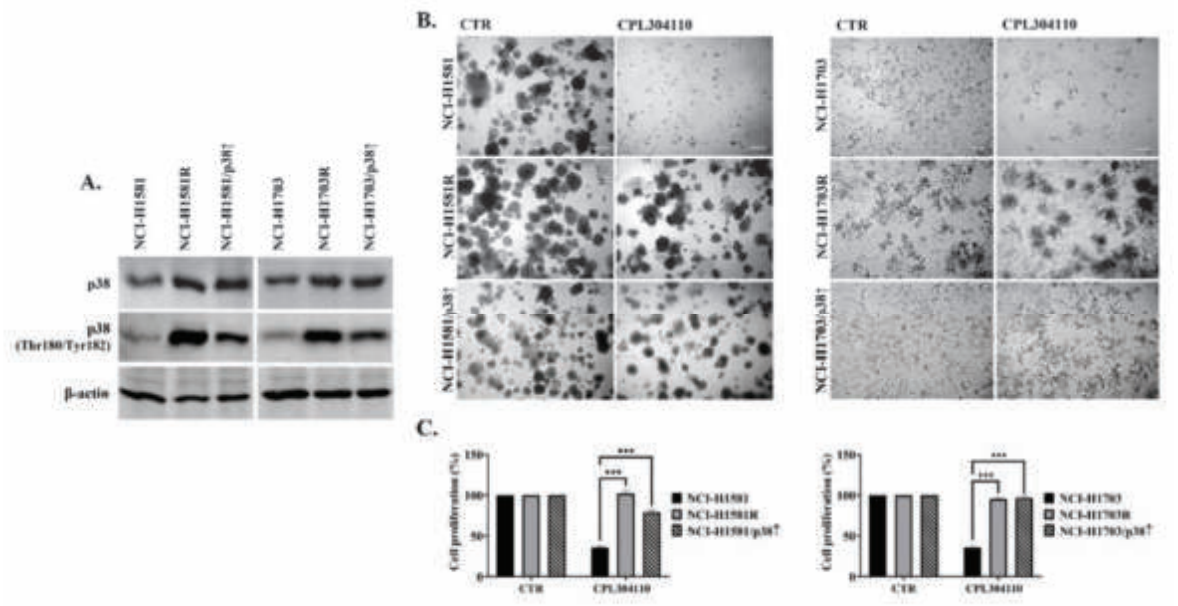


Figure 5. p38 MAPK overexpression induces resistance to FGFR inhibition. (**A**) p38 kinase overexpression was established in NCI-H1581 and NCI-H1703 cells and confirmed with Western blot. Experiments were conducted in triplicates. Representative blots are shown. (**B**) Cell growth in 3D BD Matrigel® and (**C**) cell proliferation in the presence of CPL304110 (0.1 μM for NCI-H1581, NCI-H1581R, NCI-H1581/p38↑; 1 μM for NCI-H1703, NCI-H1703R, NCI-H1703/p38↑) was assessed. Cells were cultured in 3D BD Matrigel® for 14 days. Representative pictures were taken. Scale bar represents 100 μm, n = 3. Cell proliferation was assessed using MTT viability assay after 96 h. Data are expressed as mean ± SD, *** $p \leq 0.001$, n = 3.

4. Discussion

Deregulation of FGFR signalling has a significant impact on cancer development and progression. Therefore, extensive preclinical and clinical studies were undertaken to introduce the new generation of selective FGFR inhibitors to anti-cancer therapy. FGFR1 amplification is one of the most common genomic alterations in SCC. Therefore, FGFR1 has been considered as one of the promising therapeutic targets [31–34]. Herein, we present preclinical studies for a novel FGFR inhibitor, CPL304110, which is currently in phase I of clinical trials in adults with advanced solid malignancies. After the analysis of response to CPL304110 in the panel of five lung cancer cell lines, the two most sensitive, NCI-H1581 and NCI-H1703, were used for further studies. CNV and gene expression analysis revealed a significant variability between NCI-H1581 and NCI-H1703, which could suggest differences in the molecular mechanism underlying the emergence of acquired resistance to FGFR inhibitor. Subsequently, the resistant variants of these cells were developed in order to investigate a possible molecular mechanism of acquired resistance to FGFR inhibitor. We found that both CPL304110-resistant cell lines (NCI-H1581R and NCI-H1703R) were characterised by "switched off" FGFR signalling. Although several studies indicated the reactivation of PI3K/AKT and MAPK pathways as a possible bypass mechanism of resistance to FGFR inhibition [15,27,28,35], we did not observe any significant changes in the expression/phosphorylation level of AKT or ERK1/2 in resistant cell lines. Interestingly, both NCI-H1581R and NCI-H1703R demonstrated an increase in expression and phosphorylation of p38 kinase. Moreover, the ectopic overexpression of p38 kinase in CPL304110-sensitive cells conferred the resistance to the inhibitor, whereas the inhibition of p38 activity with SB202190 and SB203580 led to CPL304110-mediated growth suppression of both resistant variants. Overall, p38 kinase is activated in response to a variety of extracellular stimuli and is mainly described as a stress-activated kinase. In accordance with our observations, a previous study showed that dual inhibition of EGFR and p38 had a synergistic inhibitory effect on the proliferation of bladder cancer cell lines [36]. Moreover, another research unveiled that p38 MAPK confers intrinsic resistance to EGFR TKIs, lapatinib and gefitinib, in K-Ras mutant colon cancer cell lines, by concurrent stimulation of EGFR gene transcription and protein dephosphorylation [37]. In NSCLC cell lines and a mouse PDX model, Yeung et al. showed that acquired resistance to gefitinib is mediated by YAP-MKK3/6-p38 MAPK-STAT3 signalling, and both inhibition and knockdown of p38 result in cell resensitisation and overcoming resistance [38]. Meanwhile, Malchers et al. reported that PD173074, a pan-FGFR inhibitor, induced apoptosis in lung cancer cells overexpressing both FGFR1 and MYC but not FGFR1 alone, suggesting that MYC is required for cell response to FGFR inhibition [39]. These results are in concordance with our data demonstrating a decrease in CNV and gene expression of both *FGFR1* and *MYC* in resistant variants of NCI-H1581 and NCI-H1703. It was also reported that the mechanism of resistance to FGFR inhibitor was accompanied by *NRAS* amplification and *DUSP6* deletion (member of a subfamily of protein tyrosine phosphatases known as dual-specificity phosphatases, negative regulators of MAPK signalling [40]), which led to the MAPK pathway reactivation [14]. Moreover, the silencing of *DUSP6* led to an increase in p38 phosphorylation [41]. Interestingly DUSP6-mediated negative regulation of p38 has been demonstrated in hepatocellular carcinoma, where enhanced polyubiquitination and the degradation of DUSP6 contributed to the increased phosphorylation of p38. This led to cell proliferation and cell cycle progression of cancer cells via activation of p38 pathway [42]. In line with these studies, we observed that both resistant cell variants showed a significant decrease in *DUSP6* gene expression with a subsequent increase in p38 protein expression. These results confirmed that the bypass mechanism of resistance to CPL304110 in these cell lines is driven by p38 kinase. We can speculate that resistance to FGFR inhibition is mediated by MAPK pathway activation, possibly in concert with the loss of DUSP6. Taking into account that our RNA-sequencing analyses of p38 signalling revealed a panel of 54 genes common for both analysed cell lines with the same pattern of changes in expression following the acquisition of resistance to CPL304110, there are several other gene candidates possibly

involved in the mechanism of FGFR inhibition resistance. Among these genes, members of MAPK, TGFβ, PI3K, EGF, and FGF pathways, as well as genes involved in the regulation of cell cycle and apoptosis, were found. Further evaluation of these pathways would be required to identify the complexity of the resistance mechanism.

Herein, for the first time, we demonstrate that despite initial genomic and transcriptomic differences between CPL304110-sensitive cell lines revealed by CNV analysis by aCGH and RNA sequencing, in both NCI-H1581 and NCI-H1703 cells, activation of p38 kinase is involved in the development of acquired resistance to FGFR inhibition. In conclusion, targeting p38 activity may be a potential therapeutic approach to circumvent FGFR inhibitor resistance.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/cells10123363/s1, Figure S1: Quantitative analysis of NCI-H1581 cells grown in 3D BD Matrigel®in the presence of CPL304110; Figure S2: Quantitative analysis of NCI-H1581 and cells grown in 3D BD Matrigel®; Western Blot analysis of proteins involved in cell cycle regulation; Analysis of variance between sensitive and CPL304110-resistant cells; Figure S3: Western blot analysis of p38 protein expression in all tested cell lines; Figure S4: Hierarchical clustering of gene sets/pathways related to p38, p38 MAPK pathway, MAPK pathway, and p38 substrates; Figure S5: Western blot analysis was performed to assess phosphorylation of p38 in sensitive and resistant cell lines upon SB202190 treatment; Figure S6: Quantitative analysis of NCI-H1581 and NCI-H1581R cells lines grown in 3D BD Matrigel®in the presence of CPL304110 and/or SB202190; Figure S7: Colony formation assay was performed for sensitive and resistant cells treated with CPL304110 and/or SB202190; Figure S8: Sensitive and resistant variants of NCI-H1581 and NCI-H1703 cells were grown with CPL304110 and/or SB203580 in 3D BD Matrigel®; Figure S9: Quantitative analysis of NCI-H1581, NCI-H1581R, and NCI-H1581/p38↑ cells grown in 3D BD Matrigel®in the presence of CPL304110.

Author Contributions: Conceptualisation, K.M.L., R.S., K.K.; methodology, I.Z., M.G.-A., K.A.K., A.J.C., A.M.W., R.S., K.K.; validation, I.Z., M.G.-A., K.A.K., A.J.C., A.M.W., M.S., K.K.; formal analysis, I.Z., M.G.-A., K.A.K., A.J.C., A.M.W., R.S., K.K.; bioinformatical analyses, A.J.C., A.M.W.; investigation, I.Z., M.G.-A., K.A.K., A.J.C., A.M.W., K.K.; resources, A.C.S., A.S., M.W., K.M.L., R.S.; data curation, I.Z., M.G.-A., K.A.K., A.J.C., A.M.W., R.S., K.K.; writing—original draft preparation, I.Z., A.J.C., K.M.L., K.K.; writing—review and editing, I.Z., K.K., R.S.; visualisation, I.Z, K.K., A.J.C., A.M.W., R.S.; supervision, K.M.L., R.S., K.K.; project administration, A.C.S., A.S., M.W., K.M.L., R.S.; funding acquisition, A.C.S., A.S., M.W., K.M.L., R.S. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported within the "CELONKO" project (STRATEGMED2/266776/17/NCBR/2015), co-financed by the Polish National Center of Research and Development and the pharmaceutical company Celon Pharma S.A. A.J.C. was co-financed by the European Union through the European Social Fund (grant no. POWR.03.02.00-00-I029).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available in the article and Supplementary Materials. Further inquiries can be directed to the corresponding authors. Sequencing data have been deposited in GEO under accession codes: GSE189270 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE189270, accessed on 23/11/2021); GSE189278 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE189268, accessed on 23/11/2021); GSE189269 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE189269, accessed on 23/11/2021).

Acknowledgments: The authors would like to express their gratitude to Anna Piotrowska from the Department of Histology of Medical University of Gdansk for the help and directions regarding flow cytometry experiments. pMT3 p38 was a gift from John Kyriakis (Addgene plasmid # 12658).

Conflicts of Interest: A.S., M.S., and M.W. are employees of the Innovative Drugs R&D Department, Celon Pharma (the company responsible for CPL304110 development and synthesis). The remaining authors declare no competing interests.

References

1. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J. Clin.* **2021**, *71*, 209–249. [CrossRef]
2. Cohen, P.; Cross, D.; Jänne, P.A. Kinase drug discovery 20 years after imatinib: Progress and future directions. *Nat. Rev. Drug Discov.* **2021**, *20*, 551–569. [CrossRef] [PubMed]
3. Sankar, K.; Gadgeel, S.M.; Qin, A. Molecular therapeutic targets in non-small cell lung cancer. *Expert Rev. Anticancer Ther.* **2020**, *20*, 647–661. [CrossRef]
4. Network, C.G.A.R. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **2012**, *489*, 519–525. [CrossRef]
5. Turner, N.; Grose, R. Fibroblast growth factor signalling: From development to cancer. *Nat. Rev. Cancer* **2010**, *10*, 116–129. [CrossRef]
6. Dienstmann, R.; Rodon, J.; Prat, A.; Perez-Garcia, J.; Adamo, B.; Felip, E.; Cortes, J.; Iafrate, A.J.; Nuciforo, P.; Tabernero, J. Genomic aberrations in the FGFR pathway: Opportunities for targeted therapies in solid tumors. *Ann. Oncol.* **2014**, *25*, 552–563. [CrossRef]
7. Helsten, T.; Elkin, S.; Arthur, E.; Tomson, B.N.; Carter, J.; Kurzrock, R. The FGFR Landscape in Cancer: Analysis of 4,853 Tumors by Next-Generation Sequencing. *Clin. Cancer Res.* **2016**, *22*, 259–267. [CrossRef] [PubMed]
8. Wu, Y.M.; Su, F.; Kalyana-Sundaram, S.; Khazanov, N.; Ateeq, B.; Cao, X.; Lonigro, R.J.; Vats, P.; Wang, R.; Lin, S.F.; et al. Identification of targetable FGFR gene fusions in diverse cancers. *Cancer Discov.* **2013**, *3*, 636–647. [CrossRef]
9. Touat, M.; Ileana, E.; Postel-Vinay, S.; André, F.; Soria, J.C. Targeting FGFR Signaling in Cancer. *Clin. Cancer Res.* **2015**, *21*, 2684–2694. [CrossRef]
10. Zhou, Y.; Wu, C.; Lu, G.; Hu, Z.; Chen, Q.; Du, X. FGF/FGFR signaling pathway involved resistance in various cancer types. *J. Cancer* **2020**, *11*, 2000–2007. [CrossRef]
11. Yue, S.; Li, Y.; Chen, X.; Wang, J.; Li, M.; Chen, Y.; Wu, D. FGFR-TKI resistance in cancer: Current status and perspectives. *J. Hematol. Oncol.* **2021**, *14*, 23. [CrossRef] [PubMed]
12. Dai, S.; Zhou, Z.; Chen, Z.; Xu, G.; Chen, Y. Fibroblast Growth Factor Receptors (FGFRs): Structures and Small Molecule Inhibitors. *Cells* **2019**, *8*, 614. [CrossRef] [PubMed]
13. Kitowska, K.; Gorska-Arcisz, M.; Antoun, D.; Zarczynska, I.; Czaplinska, D.; Szczepaniak, A.; Skladanowski, A.C.; Wieczorek, M.; Stanczak, A.; Skupinska, M.; et al. MET-Pyk2 Axis Mediates Acquired Resistance to FGFR Inhibition in Cancer Cells. *Front. Oncol.* **2021**, *11*, 633410. [CrossRef] [PubMed]
14. Malchers, F.; Ercanoglu, M.; Schütte, D.; Castiglione, R.; Tischler, V.; Michels, S.; Dahmen, I.; Brägelmann, J.; Menon, R.; Heuckmann, J.M.; et al. Mechanisms of Primary Drug Resistance in *FGFR1*-Amplified Lung Cancer. *Clin. Cancer Res.* **2017**, *23*, 5527–5536. [CrossRef] [PubMed]
15. Datta, J.; Damodaran, S.; Parks, H.; Ocrainiciuc, C.; Miya, J.; Yu, L.; Gardner, E.P.; Samorodnitsky, E.; Wing, M.R.; Bhatt, D.; et al. Akt Activation Mediates Acquired Resistance to Fibroblast Growth Factor Receptor Inhibitor BGJ398. *Mol. Cancer Ther.* **2017**, *16*, 614–624. [CrossRef]
16. Weiss, J.; Sos, M.L.; Seidel, D.; Peifer, M.; Zander, T.; Heuckmann, J.M.; Ullrich, R.T.; Menon, R.; Maier, S.; Soltermann, A.; et al. Frequent and Focal FGFR1 Amplification Associates With Therapeutically Tractable FGFR1 Dependency in Squamous-cell Lung Cancer. *Sci. Transl. Med.* **2010**, *2*, 62ra93. [CrossRef]
17. Yamani, A.; Zdżalik-Bielecka, D.; Lipner, J.; Stańczak, A.; Piórkowska, N.; Stańczak, P.S.; Olejkowska, P.; Hucz-Kalitowska, J.; Magdycz, M.; Dzwonek, K.; et al. Discovery and optimization of novel pyrazole-benzimidazole CPL304110, as a potent and selective inhibitor of fibroblast growth factor receptors FGFR (1–3). *Eur. J. Med. Chem.* **2021**, *210*, 112990. [CrossRef]
18. Mieszkowska, M.; Piasecka, D.; Potemski, P.; Debska-Szmich, S.; Rychlowski, M.; Kordek, R.; Sadej, R.; Romanska, H.M. Tetraspanin CD151 impairs heterodimerization of ErbB2/ErbB3 in breast cancer cells. *Transl. Res.* **2019**, *207*, 44–55. [CrossRef]
19. Commo, F.; Guinney, J.; Ferté, C.; Bot, B.; Lefebvre, C.; Soria, J.C.; André, F. rCGH: A comprehensive array-based genomic profile platform for precision medicine. *Bioinformatics* **2016**, *32*, 1402–1404. [CrossRef]
20. Zhang, H.; Meltzer, P.; Davis, S. RCircos: An R package for Circos 2D track plots. *BMC Bioinform.* **2013**, *14*, 244. [CrossRef]
21. Patro, R.; Duggal, G.; Love, M.I.; Irizarry, R.A.; Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **2017**, *14*, 417–419. [CrossRef] [PubMed]
22. Soneson, C.; Love, M.I.; Robinson, M.D. Differential analyses for RNA-seq: Transcript-level estimates improve gene-level inferences. *F1000Research* **2015**, *4*, 1521. [CrossRef] [PubMed]
23. Love, M.I.; Huber, W.; Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **2014**, *15*, 550. [CrossRef]
24. Subramanian, A.; Tamayo, P.; Mootha, V.K.; Mukherjee, S.; Ebert, B.L.; Gillette, M.A.; Paulovich, A.; Pomeroy, S.L.; Golub, T.R.; Lander, E.S.; et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 15545–15550. [CrossRef] [PubMed]
25. Yu, G.; Wang, L.G.; Han, Y.; He, Q.Y. clusterProfiler: An R package for comparing biological themes among gene clusters. *Omics J. Integr. Biol.* **2012**, *16*, 284–287. [CrossRef]
26. Liberzon, A.; Subramanian, A.; Pinchback, R.; Thorvaldsdóttir, H.; Tamayo, P.; Mesirov, J.P. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **2011**, *27*, 1739–1740. [CrossRef]

27. Fearon, A.E.; Carter, E.P.; Clayton, N.S.; Wilkes, E.H.; Baker, A.M.; Kapitonova, E.; Bakhouche, B.A.; Tanner, Y.; Wang, J.; Gadaleta, E.; et al. PHLDA1 Mediates Drug Resistance in Receptor Tyrosine Kinase-Driven Cancer. *Cell Rep.* **2018**, *22*, 2469–2481. [CrossRef]
28. Bockorny, B.; Rusan, M.; Chen, W.; Liao, R.G.; Li, Y.; Piccioni, F.; Wang, J.; Tan, L.; Thorner, A.R.; Li, T.; et al. RAS-MAPK Reactivation Facilitates Acquired Resistance in. *Mol Cancer Ther.* **2018**, *17*, 1526–1539. [CrossRef]
29. Niederst, M.J.; Engelman, J.A. Bypass mechanisms of resistance to receptor tyrosine kinase inhibition in lung cancer. *Sci. Signal.* **2013**, *6*, re6. [CrossRef]
30. Adachi, Y.; Watanabe, K.; Kita, K.; Kitai, H.; Kotani, H.; Sato, Y.; Inase, N.; Yano, S.; Ebi, H. Resistance mediated by alternative receptor tyrosine kinases in FGFR1-amplified lung cancer. *Carcinogenesis* **2017**, *38*, 1063–1072. [CrossRef]
31. Yang, Y.; Lu, T.; Li, Z.; Lu, S. FGFR1 regulates proliferation and metastasis by targeting CCND1 in FGFR1 amplified lung cancer. *Cell Adhes. Migr.* **2020**, *14*, 82–95. [CrossRef]
32. Wang, K.; Ji, W.; Yu, Y.; Li, Z.; Niu, X.; Xia, W.; Lu, S. FGFR1-ERK1/2-SOX2 axis promotes cell proliferation, epithelial-mesenchymal transition, and metastasis in FGFR1-amplified lung cancer. *Oncogene* **2018**, *37*, 5340–5354. [CrossRef] [PubMed]
33. Wang, Y.; Gao, W.; Xu, J.; Chen, X.; Yang, Y.; Zhu, Y.; Yin, Y.; Guo, R.; Liu, P.; Shu, Y.; et al. The Role of FGFR1 Gene Amplification as a Poor Prognostic Factor in Squamous Cell Lung Cancer: A Meta-Analysis of Published Data. *Biomed. Res. Int.* **2015**, *2015*, 763080. [CrossRef] [PubMed]
34. Schildhaus, H.U.; Nogova, L.; Wolf, J.; Buettner, R. FGFR1 amplifications in squamous cell carcinomas of the lung: Diagnostic and therapeutic implications. *Transl. Lung Cancer Res.* **2013**, *2*, 92–100. [CrossRef]
35. Kas, S.M.; de Ruiter, J.R.; Schipper, K.; Schut, E.; Bombardelli, L.; Wientjens, E.; Drenth, A.P.; de Korte-Grimmerink, R.; Mahakena, S.; Phillips, C.; et al. Transcriptomics and Transposon Mutagenesis Identify Multiple Mechanisms of Resistance to the FGFR Inhibitor AZD4547. *Cancer Res.* **2018**, *78*, 5668–5679. [CrossRef]
36. Mora Vidal, R.; Regufe da Mota, S.; Hayden, A.; Markham, H.; Douglas, J.; Packham, G.; Crabb, S.J. Epidermal Growth Factor Receptor Family Inhibition Identifies P38 Mitogen-activated Protein Kinase as a Potential Therapeutic Target in Bladder Cancer. *Urology* **2018**, *112*, 225.e221–225.e227. [CrossRef]
37. Yin, N.; Lepp, A.; Ji, Y.; Mortensen, M.; Hou, S.; Qi, X.M.; Myers, C.R.; Chen, G. The K-Ras effector p38γ MAPK confers intrinsic resistance to tyrosine kinase inhibitors by stimulating. *J. Biol. Chem.* **2017**, *292*, 15070–15079. [CrossRef] [PubMed]
38. Yeung, Y.T.; Yin, S.; Lu, B.; Fan, S.; Yang, R.; Bai, R.; Zhang, C.; Bode, A.M.; Liu, K.; Dong, Z. Losmapimod Overcomes Gefitinib Resistance in Non-small Cell Lung Cancer by Preventing Tetraploidization. *EBioMedicine* **2018**, *28*, 51–61. [CrossRef] [PubMed]
39. Malchers, F.; Dietlein, F.; Schöttle, J.; Lu, X.; Nogova, L.; Albus, K.; Fernandez-Cuesta, L.; Heuckmann, J.M.; Gautschi, O.; Diebold, J.; et al. Cell-autonomous and non-cell-autonomous mechanisms of transformation by amplified FGFR1 in lung cancer. *Cancer Discov.* **2014**, *4*, 246–257. [CrossRef]
40. Nunes-Xavier, C.; Romá-Mateo, C.; Ríos, P.; Tárrega, C.; Cejudo-Marín, R.; Tabernero, L.; Pulido, R. Dual-specificity MAP kinase phosphatases as targets of cancer treatment. *Anticancer Agents Med. Chem.* **2011**, *11*, 109–132. [CrossRef]
41. Bagnyukova, T.V.; Restifo, D.; Beeharry, N.; Gabitova, L.; Li, T.; Serebriiskii, I.G.; Golemis, E.A.; Astsaturov, I. DUSP6 regulates drug sensitivity by modulating DNA damage response. *Br. J. Cancer* **2013**, *109*, 1063–1071. [CrossRef] [PubMed]
42. Hu, X.; Tang, Z.; Ma, S.; Yu, Y.; Chen, X.; Zang, G. Tripartite motif-containing protein 7 regulates hepatocellular carcinoma cell proliferation via the DUSP6/p38 pathway. *Biochem. Biophys. Res. Commun.* **2019**, *511*, 889–895. [CrossRef] [PubMed]

frontiers | Frontiers in Oncology

TYPE Original Research
PUBLISHED 04 April 2024
DOI 10.3389/fonc.2024.1377373

OPEN ACCESS

EDITED BY
Johannes Fahrmann,
University of Texas MD Anderson Cancer Center, United States

REVIEWED BY
Zheng Wang,
Shanghai Jiao Tong University, China
Mohammed Razeeth Shait Mohammed,
University of California, San Diego, United States

*CORRESPONDENCE
Karol Jelonek
karol.jelonek@gliwice.nio.gov.pl
Piotr Widłak
piotr.widlak@gumed.edu.pl

†These authors have contributed equally to this work and share first authorship

RECEIVED 27 January 2024
ACCEPTED 25 March 2024
PUBLISHED 04 April 2024

CITATION
Mrowiec K, Debik J, Jelonek K, Kurczyk A, Ponge L, Wilk A, Krzempek M, Giskeødegård GF, Bathen TF and Widłak P (2024) Profiling of serum metabolome of breast cancer: multi-cancer features discriminate between healthy women and patients with breast cancer. *Front. Oncol.* 14:1377373. doi: 10.3389/fonc.2024.1377373

Profiling of serum metabolome of breast cancer: multi-cancer features discriminate between healthy women and patients with breast cancer

Katarzyna Mrowiec[1†], Julia Debik[2,3†], Karol Jelonek[1*], Agata Kurczyk[4], Lucyna Ponge[1], Agata Wilk[4,5], Marcela Krzempek[4], Guro F. Giskeødegård[2], Tone F. Bathen[2,6] and Piotr Widłak[7*]

[1]Center for Translational Research and Molecular Biology of Cancer, Maria Sklodowska-Curie National Research Institute of Oncology, Gliwice, Poland, [2]Department of Circulation and Medical Imaging, The Norwegian University of Science and Technology, Trondheim, Norway, [3]Department of Public Health and Nursing, The Norwegian University of Science and Technology, Trondheim, Norway, [4]Department of Biostatistics and Bioinformatics, Maria Sklodowska-Curie National Research Institute of Oncology, Gliwice, Poland, [5]Department of Systems Biology and Engineering, Silesian University of Technology, Gliwice, Poland, [6]Clinic of Radiology and Nuclear Medicine, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway, [7]2nd Department of Radiology, Medical University of Gdansk, Gdansk, Poland

Introduction: The progression of solid cancers is manifested at the systemic level as molecular changes in the metabolome of body fluids, an emerging source of cancer biomarkers.

Methods: We analyzed quantitatively the serum metabolite profile using high-resolution mass spectrometry. Metabolic profiles were compared between breast cancer patients (n=112) and two groups of healthy women (from Poland and Norway; n=95 and n=112, respectively) with similar age distributions.

Results: Despite differences between both cohorts of controls, a set of 43 metabolites and lipids uniformly discriminated against breast cancer patients and healthy women. Moreover, smaller groups of female patients with other types of solid cancers (colorectal, head and neck, and lung cancers) were analyzed, which revealed a set of 42 metabolites and lipids that uniformly differentiated all three cancer types from both cohorts of healthy women. A common part of both sets, which could be called a multi-cancer signature, contained 23 compounds, which included reduced levels of a few amino acids (alanine, aspartate, glutamine, histidine, phenylalanine, and leucine/isoleucine), lysophosphatidylcholines (exemplified by LPC(18:0)), and diglycerides. Interestingly, a reduced concentration of the most abundant cholesteryl ester (CE(18:2)) typical for other cancers was the least significant in the serum of breast cancer patients. Components present in a multi-cancer signature enabled the establishment of a well-performing breast cancer classifier, which predicted cancer with a very high precision in independent groups of women (AUC>0.95).

Discussion: In conclusion, metabolites critical for discriminating breast cancer patients from controls included components of hypothetical multi-cancer signature, which indicated wider potential applicability of a general serum metabolome cancer biomarker.

KEYWORDS

biomarker, breast cancer, high-resolution mass spectrometry, metabolomics, multicancer signature, serum metabolome, The HUNT study

1 Introduction

In women's population breast cancer (BC) represents 25% of newly diagnosed cancer cases and about 15% of cancer-related deaths worldwide. In many developed countries, this cancer ranks first on the list of morbidity and mortality among all malignancies. Moreover, according to epidemiologic forecasts, both values will increase over the next decades (1). Therefore, intensive research is needed on the mechanisms of development of this very heterogenous malignancy (2) to clarify many aspects of its molecular biology further and identify biomarkers for risk assessment and early detection of this cancer (3).

Since 1920, when Otto Warburg noticed the accumulation of lactate in tumor tissue due to increased glucose consumption by aerobic glycolysis, cancer metabolomics has made great progress (4, 5). The metabolome combines information resulting from both endogenous processes and exogenous interactions, thus providing insight into cellular mechanisms and their modifications caused by a wide range of stimuli. In addition, metabolic changes are visible earlier than phenotypic ones, and their examination is possible in a quick and minimally invasive way (e.g., by determination in body fluids). Studying the profiles of metabolites enables the creation of so-called "metabolic fingerprints", i.e. changes in the metabolome characteristic of a specific state of the body. Numerous studies have been conducted to characterize cancer-related changes, using different cohorts and on various types of material (tissue, blood, etc.) (6, 7). Metabolic features of breast cancer were addressed in several reports, including information on cancer-related changes in the metabolism of amino acids, fatty acids, or glycerolipids (8–11). Unfortunately, reported results are ambiguous, hence the metabolic fingerprint for breast cancer and its specificity regarding other cancer types have yet to be fully characterized (12).

In the current study, we performed a quantitative analysis of metabolites present in serum samples of breast cancer patients and two cohorts of healthy women, which allowed us to identify differences in the metabolic profiles of healthy women and patients diagnosed with breast cancer. Moreover, women with three other types of solid cancers (colorectal cancer, head and neck cancer, and lung cancer) were included in the study, which revealed "multi-cancer" characteristics of certain metabolic features observed in patients with breast cancer.

2 Materials and methods

2.1 Characteristics of analyzed groups

The clinical material was collected at the Maria Sklodowska-Curie National Research Institute of Oncology, Gliwice Branch between 2010 and 2020. Blood samples were collected from women patients with breast (BC), colorectal (CC), head and neck (HC), and lung (LC) cancers before the start of cancer therapy. Two groups of healthy donors were included in the study: healthy volunteers living in the Silesia region, Poland (Ctr_P), recruited in the same period as cancer patients, and a subset of healthy women selected from participants of the HUNT2 study performed between 1995 and 1997 in the Trøndelag region, Norway (Ctr_N). The Trøndelag Health Study (HUNT) is a collaboration between HUNT Research Centre (Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology NTNU), Trøndelag County Council, Central Norway Regional Health Authority, and the Norwegian Institute of Public Health (13). The latter set included women selected from a group of 450 healthy participants analyzed in a previously published study (14) to match the age of BC patients (see diagram in Supplementary Figure S1). Consequently, two independent cohorts of healthy controls were included: Ctr_P and Ctr_N. The characteristics of the study cohort are presented in Table 1. Peripheral blood was collected into a 5 mL BD Vacutainer Tube, incubated for 30 min at room temperature to allow clotting, and then centrifuged at 1000× g for 10 min to remove the clot. The serum was aliquoted and stored at −80°C before further processing. The study was conducted following the Declaration of Helsinki, and approved by the Ethics Committee of Maria Sklodowska-Curie National Research Institute of Oncology, Gliwice Branch (KB/493-53/10 and KB/430-84/20) and the Regional Committee for Medical and Health Research Ethics (REK#1995/8395 and REK#2017/2231). All participants provided informed consent indicating their voluntary participation.

TABLE 1 Characteristics of the study cohorts.

	Control (Poland)		Control (Norway)		Breast Cancer		Colorectal Cancer	Head & Neck Cancer	Lung Cancer
Abbreviation	Ctr_P		Ctr_N		BC		CC	HC	LC
N	95	35*	112	35*	112	35*	30	32	35
Age (years) mean [S.D.]	48.3 [6.5]	53.7 [3.6]*	49.3 [11.0]	63.0 [9.0]*	49.3 [11.0]	60.5 [7.4]*	64.8 [10.5]	59.4 [11.9]	65.2 [8.8]
Clinical stage									
I	–	–	–	–	0	0*	6	0	9
II	–	–	–	–	56	19*	11	2	8
III	–	–	–	–	49	14*	12	9	13
IV	–	–	–	–	7	2*	1	21	5

*Sub-cohort of controls and BC cases selected for comparison with other solid cancers to enable similar age distribution in all groups.

2.2 Quantitative high-resolution mass spectrometry

Quantitative analysis of metabolites for all serum samples was performed using the Absolute IDQ p400 HR kit (Biocrates Life Sciences AG, Innsbruck, Austria) following the procedure recommended by the producer (Supplementary Data: Protocol for metabolite detection and quantification by the Absolute IDQ p400 HR kit). This is a commercial assay with an automated workflow, whose quality, stability, and repeatability were validated in the international ring trial (15). Orbitrap Q Exactive Plus high-resolution mass spectrometer (Thermo Fisher Scientific, Waltham, MA, USA) and 1290 Infinity UHPLC (Agilent, Santa Clara, CA, USA) system was used to measure concentrations of selected metabolites (including amino acids, biogenic amines, hexoses, acylcarnitines, diglycerides, triglycerides, (lyso) phosphatidylcholines, sphingolipids, and cholesteryl esters) in 10 µl human serum. Samples were measured in batches designed to secure the same proportion of different groups with a randomized order of samples within each group. The obtained chromatograms and spectra were processed using Xcalibur 4.1. and MetIDQ DB110-2976 software (Biocrates Life Sciences AG) resulting in a matrix of concentrations of metabolites in µM. To control the quality of quantitative analyses the coefficient of variation (CV) of all Quality Control (QC) measurements for all metabolites was calculated (16).

2.3 HRMS data processing

The MS dataset contained measurements of the levels of 389 metabolites present in 416 samples. Firstly, detection and imputation of missing values were performed. According to the recommendations of Chen and coworkers (17) a threshold of 50% was adopted for values missing not at random (i.e., values below the limit of detection). In the case of data missing completely at random (i.e., generated as a result of the internal standard error), a threshold of 10% was adopted. In the first case, missing values were imputed by random numbers generated from normal distribution truncated to a segment between 0 and the median value of the limits of quantitation for all test plates. In the second case, missing values were imputed using the k-nearest neighbor approach (the nearest observed data were identified using a correlation distance metric, and the mean value of the three nearest neighbors was used based on measurements collected for the same group). Metabolites that were non-compliant with these criteria were excluded from further analyses. Finally, 284 metabolites were qualified for quantitative analysis, and the remaining 105 compounds were left for binary analysis, which statistically tests whether the absence/presence status of a metabolite is a group-related feature. In the next step, the data were transformed using the log base 2 function, and then the batch effect was corrected using an empirical Bayes method, assuming that samples measured using a single 96-well sample preparation plate represent one batch (18).

2.4 Statistical and bioinformatics analyses

The quantitative analysis of metabolites that differentiated BC cases (n=112) and either control group (n=95 or n=112 for Ctr_P and Ctr_N, respectively) was performed using the Mann–Whitney U test, and then the Benjamini–Hochberg procedure was performed to reduce the number of false positive results. To analyze metabolites that differentiated either control group (n=35 for both Ctr_P and Ctr_N) from BC (n=35), CC (n=30), HC (n=32), and LC (n=35) cases, the Kruskal–Wallis test, followed by the post-hoc Conover test for pairwise comparisons was implemented. All statistical hypotheses were tested at the 5% significance level. In addition, the "r" effect size was calculated according to the formula: r=z/square root of N (where z is the value of the test statistic and N is the total number of observations in two compared groups) with interpretation according to the Cohen's criterion (small effect – $|r|>0.1$, medium effect – $|r|>0.3$, large effect – $|r|>0.5$) (19). Fisher's exact test was applied for metabolites that did not qualify for quantitative analysis to determine if there was a nonrandom association between the absence/presence of metabolites and

analyzed groups. Hierarchical clustering was performed to assess similarities between the analyzed groups. The median value of metabolite abundances for samples within a particular group was calculated (raw abundances were previously transformed to z-scores). Each analyzed group was characterized by a vector consisting of the calculated median value of metabolite abundances, and then similarities between groups were analyzed using agglomerative hierarchical cluster analysis with the Minkowski distance between pairs of observations and the average linkage clustering method. To assess the predictive quality of the multi-cancer signature for distinguishing breast cancer samples and controls, a classifier was constructed on a dataset containing samples not used for the signature selection (60 Ctr_P, 60 Ctr_N, 77 BC cases). A support vector machine (SVM) model with a radial kernel function was trained on half of the BC cases (n=39) and an equal number of Ctr_N controls, and tested on the remaining half of the BC cases (n=38) and an equal number of Ctr_P controls (see diagram in Supplementary Figure S1); this design enabled the validation of the universality of classification model using different populations of healthy women. The prediction quality on the test set was evaluated in terms of accuracy, sensitivity, specificity, positive and negative predictive value (PPV and NPV, respectively), and area under the receiver operating characteristic curve (AUC). To obtain a reliable estimation of classification quality, the procedures (sampling, training, and testing/validation) were repeated 500 times. All analyses were performed using the R Statistical Software (version 4.1.2, R Foundation for Statistical Computing, Vienna, Austria). The metabolic pathway enrichment analysis was performed using the MetaboAnalyst 5.0 platform for all quantitative data (https://www.metaboanalyst.ca/MetaboAnalyst/ModuleView.xhtml; last access October 6, 2023).

3 Results

3.1 HRMS-based analysis of serum metabolites revealed compounds discriminating between controls and breast cancer cases

Serum metabolite and lipid profiles were analyzed quantitatively by HRMS in a set of samples collected from breast cancer patients and two cohorts of healthy women living in Poland or Norway; three other types of solid cancer were also used for the comparison (baseline characteristics of the study groups are presented in Table 1; a similar age distribution was ensured between the compared groups). This approach enabled the detection of 389 metabolites, among which 284 compounds quantified in the majority of samples were used in quantitative analyses. The median value of concentrations of quantified compounds and the strength of differences between controls and BC cases are presented in Figure 1A. Good separation of cancer and control samples in the unsupervised Principal Component Analysis (PCA) was noted; moreover, a separation was also observed between both groups of control samples (Figure 1B, also see Supplementary Figure S1A for the results of the OPLS-DA analysis). There were 60 serum metabolites whose levels were significantly different between BC cases and Polish controls (medium or large effect size), including 49 metabolites downregulated and 11 upregulated in BC cases. On the other hand, there were 114 metabolites with levels significantly different between BC cases and Norwegian controls, including 88 downregulated and 26 upregulated in cancer samples (see Supplementary Table S1 for details). Moreover, when the remaining set of 105 metabolites not qualified for quantitative analysis was tested for the absence/presence status, 5 metabolites (AC(15:0), AC(18:1), DG(38:0), PC(44:3), Cer(44:0)) were significantly under-represented in BC cases compared to Polish controls. On the other hand, 4 metabolites (AC(7:0), LPC-O(17:1), PC(35:0), PC(44:10)) were under-represented while 4 metabolites (AC(4:0-OH), spermidine, PC-O(44:5), PC-O(32:0)) were over-represented in BC cases compared to Norwegian controls (Supplementary Table S2).

Differences between (Polish) breast cancer cases and Norwegian controls in concentrations of serum metabolites were stronger than differences between cases and Polish controls (Figure 1C); the medians of the effect sizes equal 0.228 and 0.153, respectively. Partially different sets of metabolites that discriminated against BC cases and either group of controls were related to significant differences between both control groups. We found 85 compounds with different concentrations between Ctr_P and Ctr_N groups (Supplementary Table S1). Nevertheless, a large set of metabolites commonly differentiated BC cases from both control groups: there were 33 overlapped metabolites significantly downregulated in BC cases and 10 metabolites significantly upregulated in BC cases in comparison to both control cohorts (medium or large effect size; Supplementary Table S1). Importantly, the identification of the same features in two different cohorts of healthy women indicated a universal significance of the signature, which could be called the "breast cancer signature". This signature included 13 amino acids, 12 lysophosphatidylcholines, and 6 diglycerides downregulated in BC cases as well as 9 acylcarnitines upregulated in BC cases. Noteworthy, the increased level of hexoses (incl. glucose) detected in cancer samples was significantly higher compared to Polish controls than compared to Norwegian controls (large and small effect size, respectively). The volcano plot in Figure 1D and Supplementary Figure S3 illustrates the metabolites that showed the most robust differences between controls and BC cases. These included glutamic acid (Glu) and aspartic acid (Asp) whose concentration was markedly reduced in the serum of BC patients compared to both groups of healthy controls (Figure 1E). Moreover, assuming the potential functional redundancy of lipids from the same class, aggregated amounts of major classes of detected lipids were also compared (Figure 1F). We found that total levels of lysophosphatidylcholines (LPC) and diglycerides (DG) were markedly reduced in sera of breast cancer patients compared to both groups of controls (effect size r<-0.3). Similar total levels of lipid classes were observed in sera of healthy individuals from both control cohorts (except for lysophosphatidylcholines slightly upregulated in Norwegian controls).

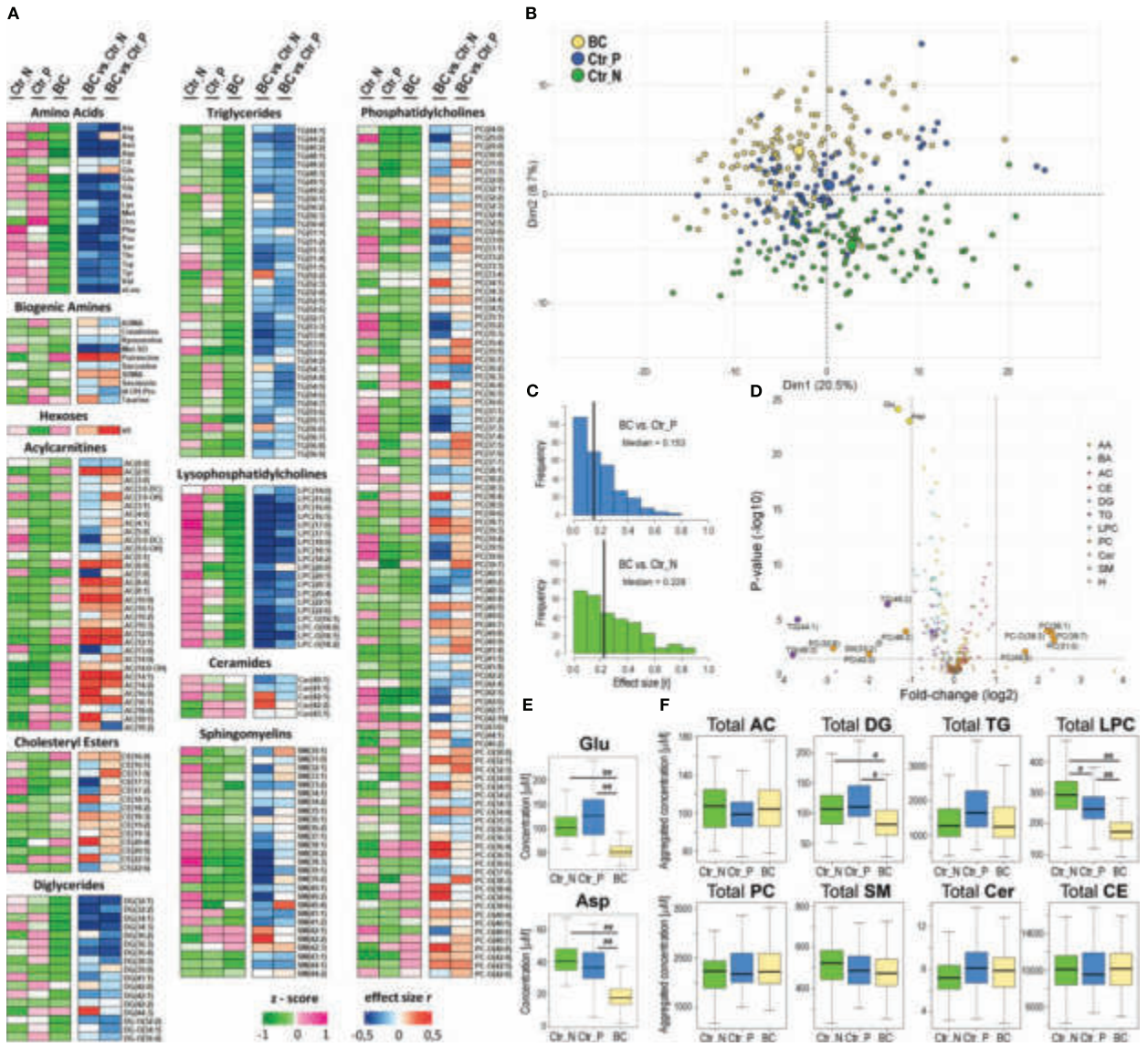


FIGURE 1
Characterization of the serum metabolome profile analyzed by mass spectrometry in breast cancer patients and healthy women. Panel **(A)** – Levels of metabolites in serum samples from 112 Norwegian and 95 Polish controls (Ctr_N and Ctr_P, respectively) and 112 cancer cases (BC); heatmap visualizes median levels of analyzed metabolites in each group (raw abundances were converted into z-scores) and magnitudes of differences between groups (quantified as "r" effect size). Panel **(B)** – Plot of the data in two dimensions of the first two principal components of PCA analysis to visually identify clusters; cases and controls are marked separately (large circles represent the average of each group). Panel **(C)** – The histograms for metabolites that showed the increased significance of differences between BC cases and either group of controls (vertical lines represent the median value of "r" effect sizes). Panel **(D)** – The volcano plot representing metabolites with significantly different concentrations between BC and Ctr_P; shown is the fold-change and corresponding p-value. Panel **(E)** – Concentrations of glutamic acid (Glu) and aspartic acid (Asp) in samples of BC cases and controls. Panel **(F)** – Differences in aggregated concentration of different classes of lipids between BC cases and controls. Boxplots represent minimum, lower quartile, median, upper quartile, and maximum; medium (|r|>0.3) or large (|r|>0.5) effect size is marked with one hash (#) or two hash (##) symbols, respectively; AC – acylcarnitines, DG – diglycerides, TG – triglycerides, LPC – lysophosphatidylcholines, PC – phosphatidylcholines, SM – sphingomyelins, Cer – ceramides, CE – cholesteryl esters.

3.2 HRMS-based analysis of serum lipids and metabolites revealed a common set of compounds that differentiated controls from breast cancer cases and three other types of solid cancers

Knowing serum metabolites that differentiated breast cancer patients from healthy controls, we aimed to check its potential specificity for this particular cancer compared to features of serum metabolome observed in other types of solid cancers. Therefore, additional groups of women patients diagnosed with either colorectal cancer (CC), head and neck cancer (HC), or lung cancer (LC) were included in the analysis. Assuming that the age of participants is a potential confounding factor in this type of study, smaller sub-cohorts of healthy controls and BC cases were selected to enable a similar age distribution in all groups (Table 1). The median value of

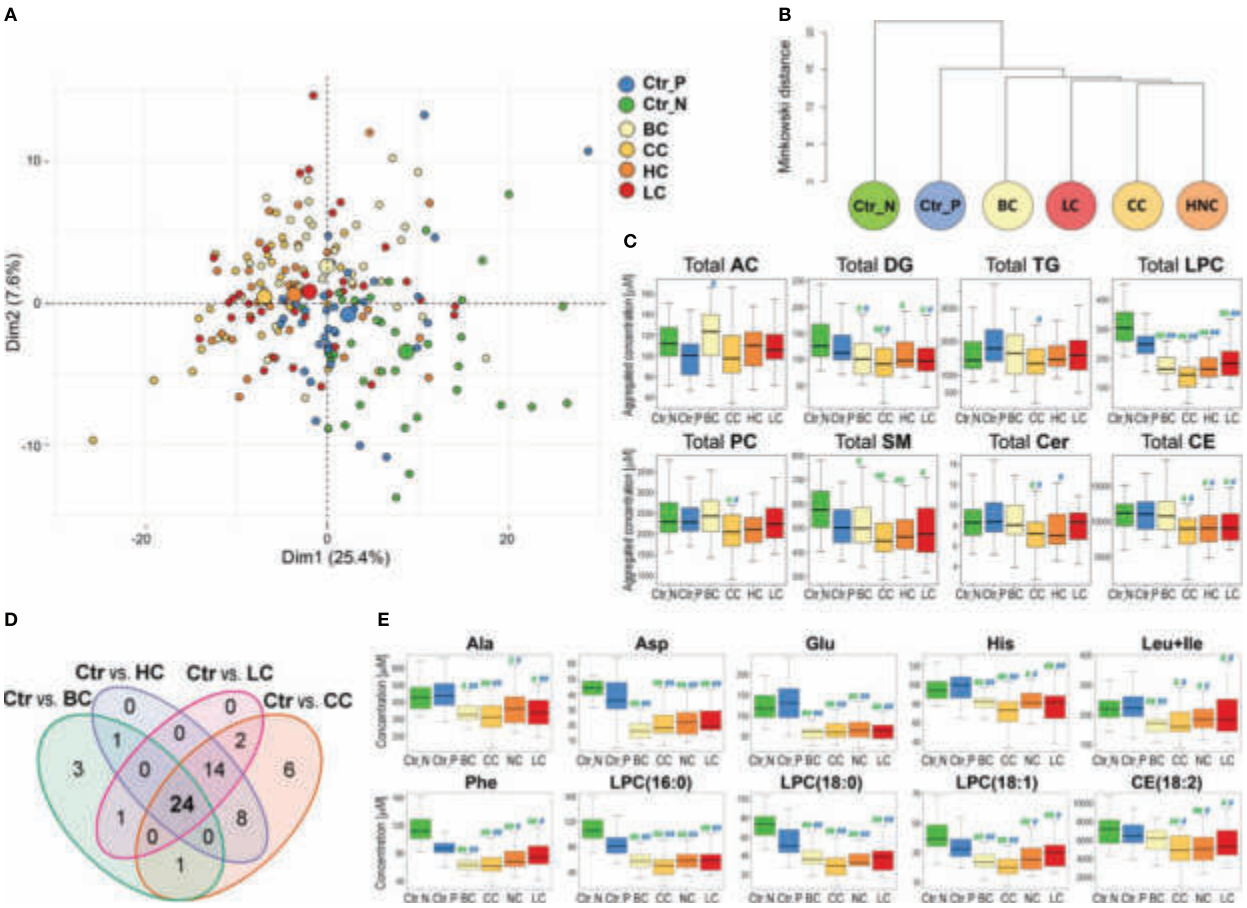


FIGURE 2
The serum metabolome features differentiate healthy women and patients with four types of solid cancers. Panel **(A)** – Plot of the data in two dimensions of the first two principal components of PCA analysis of samples of women diagnosed with breast (BC), colorectal (CC), head and neck (HC), and lung (LC) cancers, and healthy controls (large circles represent the average of each group). Panel **(B)** – Dendrogram of the hierarchical binary cluster tree of six groups based on the median value of metabolite concentrations; the height of each U represents the Minkowski distance between the two analyzed groups being connected within a cluster. Panel **(C)** – Differences in aggregated concentrations of the major classes of lipids between controls and four types of cancers; medium (|r|>0.3) or large (|r|>0.5) effect size of differences between control and particular cancer type is marked with one hash (#) or two hash (##) symbols, respectively (green marks on left and blue marks on right represent the significance of differences between cancer cases and Ctr_N or Ctr_P, respectively). Panel **(D)** – The Venn diagram showing the overlap of metabolites that discriminated four types of solid cancers from both groups of controls consistently (large and medium effect size). Panel **(E)** – Concentrations of selected metabolites that differentiated controls and four types of cancers (the effect size is marked as in Panel **C**).

concentrations of quantified compounds in all six groups and the strength of differences between controls and each cancer type are presented in Supplementary Figure S4. When clusters of samples were observed in the PCA analysis, both control groups were distinct from any cancer cases; Norwegian controls were the most dissimilar (Figure 2A; the analysis was based on all quantitated metabolites; also see Supplementary Figure S1B for the results of the OPLS-DA analysis). Similarly, unsupervised hierarchical clustering of "averaged group representatives" revealed the largest distance of controls from all cancer types (Figure 2B); CC cases and HC cases appeared the most similar. When the average strength of differences between controls and cancer cases were compared (based on the histogram of metabolites differentiating controls and cases with increasing effect size), the biggest effect was noted for CC and HC cases, while the effect was lower for BC cases (median value of the effect sizes equal to 0.157 and 0.259 for Ctr_P and Ctr_N, respectively) (Supplementary Figures S5A, B). Aggregated amounts of the major classes of lipids were also analyzed which revealed reduced levels of total serum lysophosphatidylcholines and diglycerides as a general cancer-related feature. On the other hand, certain lipid features characteristic of CC, HC, and LC were not observed in BC cases. These included downregulation of ceramides, sphingomyelins, and cholesteryl esters; the latter feature (i.e., downregulation of cholesteryl esters compared to both control groups) was statistically significant in all solid cancers except BC cases (Figure 2C).

Furthermore, we searched for specific metabolites whose serum levels differentiated both control cohorts from all cancer types (i.e., components of a hypothetical "multi-cancer" signature); cancer cases were analyzed against each control cohort separately (Supplementary Table S3). We found that 29 features discriminated between Polish controls and all four types of solid cancers (large and medium effect size); all but one (AC(14:1)) showed reduced concentration in cancer samples (Supplementary Figure S5C). On the other hand, 98 features

discriminated between Norwegian controls and all four types of solid cancers, which included reduced total concentrations of lysophosphatidylcholines, diglycerides, and sphingomyelins in cancer samples (Supplementary Figure S5D). Importantly, when these two sets of metabolites common for all cancer types were combined, 24 overlapped features were revealed (Figure 2D). This common "multi-cancer signature" included 6 amino acids (Ala, Asp, Glu, His, Phe, Leu +Ile), 2 DGs, 2 TGs, and 13 LPCs (and total LPC level) (Supplementary Table S4).

This is noteworthy, that only a small fraction of compounds (5 metabolites) differentiating BC cases from both cohorts of controls did not belong to this multi-cancer signature. Hence, the major fraction of metabolites that differentiated controls and BC cases showed similar differences between controls and other types of solid cancers. On the other hand, a subset of features that discriminated both cohorts of controls from cancers LC, HC, and CC but not from BC cases was relatively large (14 features). These cancer-specific features that were missed in the case of breast cancer included reduced concentration of CE(18:2), which is the most abundant cholesteryl ester detected in the serum (in general, similar concentrations of cholesteryl esters were noted in BC cases and controls). Examples of metabolites that discriminated control and cancer samples, including components of a hypothetical multi-cancer signature, are presented in Figure 2E.

3.3 Breast cancer classifier based on serum metabolome components present in the multi-cancer signature

Metabolites present in a hypothetical multi-cancer signature were used to test multicomponent binary classifiers discriminating breast cancer cases from healthy controls. Considering a high correlation of serum concentrations of 13 LPCs present in this signature, only an aggregated LPC level was included in the tested model. Therefore tested signature was composed of 11 features: Ala, Asp, Glu, His, Phe, Leu+Ile, DG(34:3), DG(36:4), TG(44:2), TG (46:2), and total LPC. Samples used for the identification of the multi-cancer signature were not used for training and testing of the classifier to avoid information leaks (scheme in Supplementary Figure S1). The classification model was trained using Norwegian controls and tested using Polish controls; this design strengthened the validation of the universality of the classification model. Five hundred repeats of the train/test procedure were implemented to assess the prediction power of the classifier. The indices of the classification model obtained during its vali-dation are presented in Figure 3A. In general, we found a very high prediction power of the resulting breast cancer classifier: sensitivity=0.97, specificity=0.92, and AUC=0.98. We concluded that a set of metabolites selected to differentiate healthy women from patients with four types of solid cancers (multi-cancer signature) classified an independent group of BC cases with very high precision.

3.4 Similar metabolic pathways were associated with breast cancer and three other types of solid cancers

The analysis of metabolic pathways associated with a set of 43 compounds whose levels differed between both groups of controls and BC cases (i.e., breast cancer signature) revealed several terms primarily related to amino acid metabolism. Pathway analysis was also performed with a set of 42 compounds whose levels differed between both groups of controls and three other types of cancer (CC, HC, and LC) not taking into account BC. Practically the same pathways were associated with sets of compounds differentiating controls from BC and compounds commonly differentiating controls from other types of cancers (Figure 3B); however, lipids that are not properly annotated in the used bioinformatics tool are not illustrated in either case. Nevertheless, the top-3 pathways were the same in both sets: "aminoacyl-tRNA bio-synthesis", "alanine, aspartate, and glutamate metabolism", and "histidine metabolism", confirming the functional similarity of serum metabolites characteristic of BC and other types of solid cancers.

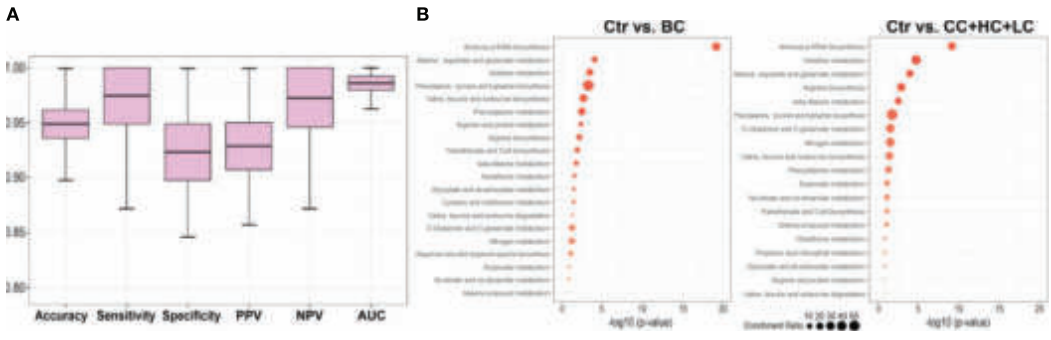


FIGURE 3
Comparison of the serum metabolite features characteristic of breast cancer patients and patients with other types of solid cancers. Panel **(A)** – Performance of breast cancer classifier built of features present in the multi-cancer signature. The classifier was trained and tested using Norwegian (Ctr_N) and Polish (Ctr_P) controls, respectively. Panel **(B)** – Metabolic pathways associated with compounds whose levels were different between controls and BC cases (left) or controls and other types of cancers (CC, HC, LC); shown are pathways with the most significant enrichment (size of a dot corresponds to the enrichment ratio).

3.5 Possible confounding factors affecting the study

The age of participants is a strong factor affecting the serum concentrations of several metabolites, which is manifested by a positive correlation between age and concentration of lipids, particularly DGs (only concentrations of LPCs were age-independent) (14). However, the compared groups have a similar age distribution in the present study, which excluded age as a confounding factor during the comparison between controls and cancer cases. Nevertheless, we have performed additional analysis where differences between the BC cases and controls were analyzed separately in sub-cohorts of "younger" (<50 year-old) and "older" (≥50 year-old) women, which putatively mirrored their pre- and post-menopause statuses (Supplementary Table S5). This analysis revealed the same patterns of difference between the BC cases and both groups of controls (Ctr_P and Ctr_N) in either age-defined sub-cohorts, which indicated that features of cancer signature were age-independent). Another biological factor that putatively affects serum concentrations of metabolites is a fasting period before the blood collection, which was not controlled in the current study for cancer patients and Polish controls. However, based on data obtained from healthy participants of the HUNT2 study (14), we found that levels of lysophosphatidylcholines differentiating controls and cancer cases (e.g., LPC(16:0), LPC(18:0), and total LPC level) were not affected by fasting. Similarly, levels of amino acids essential for cancer classification (Asp, Glu, Phe), which were generally decreased in cancer samples, were barely affected by fasting (a slight increase with time of fasting could be noted) (Supplementary Figure S6), which further reduced the putative confounding significance of this factor for hypothetical cancer signature. Moreover, since the sample's storage periods extended two years some metabolites might have been affected by long-term storage (20). Importantly, however, any changes induced by long-term storage were randomly distributed over the cases (BC, CC, HC, LC) and Ctr_P samples, which reduced the impact of this confounding factor on the observed differences between cancer patients and healthy controls. HUNT samples (Ctr_N) were stored for a longer time compared to Polish samples. However, proposed cancer signature included only compounds that jointly differentiated cancer samples from both groups of controls, which reduced the potential impact of differences in the storage period. Furthermore, measuring groups of samples as separate batches could result in differences strengthened by analytical factors such as instrumental drift. However, in the current study, all types of cancer and control samples were similarly distributed among sets/batches of measurements, and then potential batch effects among these sets were corrected during the data processing procedure.

4 Discussion

Metabolomics, which addresses the most dynamically changing system in the human body – the metabolome, represents an emerging opportunity for the understanding of human disease (6). The implementation of analytical approaches based on NMR and mass spectrometry for the detection and quantification of metabolites present in blood and other body fluids enabled the identification of multicomponent signatures that could be considered a goldmine of biomarkers of different cancers, including breast cancer (21, 22). However, though molecular features of breast cancer have been widely reported in the literature, the specificity of metabolic serum fingerprint for breast cancer has not been characterized comprehensively yet (12).

Here we applied an HRMS-based quantitative approach to compare metabolic profiles of serum from breast cancer patients and two cohorts of healthy women, which revealed a set of metabolites whose serum levels were significantly different between cancer cases and controls. Cancer-related features were clearly distinguished despite significant differences between Polish and Norwegian cohorts of healthy women used as a control. Differences between the cohorts of healthy women are putatively related to differences in lifestyle-related factors, including diet and physical activity, since the potential influence of ethnic/genetic background was rather limited (however, due to lack of demographic details analysis of hypothetical lifestyle-related factors could not be performed) or changes in the metabolite concentrations during the extend sample storage (Norwegian samples collected in the frame of HUNT2 study were stored for a longer period than Polish cases and controls). We found that concentrations of most of the amino acids (Glu and Asn in particular), diglycerides, triglycerides, and lysophosphatidylcholines were generally decreased while concentrations of hexoses (incl. glucose), and certain acylcarnitines were increased in sera of breast cancer patients. A study that applied the earlier version of the quantitative MS-based platform than the one used in our study (the Biocrates p180 assay), revealed significantly reduced concentrations of several amino acids (Ala, Asn, Glu, His, Leu, Lys, Met, Orn, Phe, Thr, Trp, Val) and biogenic amines (kynurenine and Met-SO) in plasma of breast cancer patients (23), which was fully coherent with results of the current study. However, metabolic patterns of breast cancer reported in several other studies only partially overlapped with the present study. For example, multiplatform (NMR, LC-MS, and GC-MS) analysis of plasma metabolome performed in a group of Hispanic women with breast cancer revealed an increased concentration of several acylcarnitines (which was coherent with our study) but also triglycerides, and lysophosphatidylcholines (which was contrary to our study) (24). Another study based on the combination of LC-MS and GC-MS showed increased levels of Gln and acylcarnitines while decreased levels of lysophosphatidylcholines and amino acids in plasma of breast cancer patients, which was consistent with our study, yet levels of glucose were decreased in cancer (25). Such inconsistencies among reports could be due to different analytical platforms and different demography/pathology characteristics of studied cohorts (for example, the age of donors is a major confounding factor affecting the profile of metabolites in serum samples (14)). Nevertheless, the majority of studies showed reduced levels of amino acids in the plasma or serum of breast cancer patients (8, 23, 26). Moreover, similar to our report, increased levels of glucose (27, 28) were documented in other studies. Hence, though several

differences among published reports exist, impaired metabolism of amino acids (manifested by their reduced serum/plasma concentrations) and glycolysis (manifested by increased concentration of glucose) appeared as general metabolic features observed in the blood of breast cancer patients.

Though major cancer-related changes in cellular metabolism are common for malignant cells (29, 30), specific differences could be observed when metabolic profiles of serum/plasma are compared among patients with different types of cancer. For example, differences in serum levels of certain amino acids and lipids were reported in patients with different leukemias (31). Here we observed several differences in serum metabolome patterns among four types of solid cancers. This could be exemplified by different serum lipid profiles: reduced levels of cholesteryl esters and phosphatidylcholines characteristic for patients with head and neck, lung, or colorectal cancer were not observed in patients with breast cancer. Nevertheless, a set of metabolites that significantly differentiated healthy controls from patients with all types of investigated solid cancers was identified. This set of metabolites comprised several amino acids (Ala, Asp, Glu, His, Leu, Ile, Phe) and lysophosphatidylcholines (including the most abundant LPC(16:0) and LPC(18:0)) with reduced serum concentrations in cancer patients. Hence, we concluded that metabolic features of serum that most markedly differentiated healthy women and breast cancer patients represent a set of metabolites common for women with different solid cancers, which could be considered as a "multi-cancer signature". This conclusion was further confirmed because metabolites present in this signature could be used to build a specific breast cancer classifier that showed very high prediction power when validated using independent groups of women.

The characteristic feature of the hypothetical multi-cancer signature was the reduced serum level of amino acids and lipids, a phenomenon widely described in the literature. In general, reduced levels of metabolites in the serum of cancer patients could reflect their transfer from blood to the tumor site caused by their higher consumption by cancer cells, where they work as biosynthesis substrates, fatty acid carriers, energy sources, and/or signaling molecules. Increased uptake from blood and enhanced metabolism of amino acids is a hallmark of many cancers, including cancers addressed in the current study (32). Under different types of stress conditions, amino acids facilitate the survival and proliferation of cancer cells due to their essential role in nucleotide and protein synthesis or DNA methylation. Moreover, some amino acids function as precursors of polyamines or nitric oxide, as well as act as signaling molecules (33). Decreased level of circulating amino acids is linked to increased uptake by cancer cells due to overexpression of amino acid transporters, including *SLC1A5* (Gln uptake), *SLC1A4* (Ser uptake), *SLC7A5* (Leu, Ile, and Val uptake), or *SLC7A1* (Arg uptake) (4). Another tumor-related feature is generally reduced levels of serum lipids (so-called hypolipidemia) resulting from increased utilization of lipids by cancer cells (34). Few studies showed decreased serum levels of glycerides (35) and cholesterols (36) in breast cancer patients. However, the most characteristic feature observed in cancer patients is a reduced level of circulating lysophosphatidylcholines with a chain of palmitic, stearic, or oleic acids (LPC(16:0), LPC (18:0), and LPC(18:1), respectively). Reduced serum levels of LPCs putatively reflected their transfer to tumor tissue and higher consumption by cancer cells. This effect could also result from intensified conversion of LPC by autotaxin (ATX) to lysophosphatidic acid (LPA), since increased ATX expression was observed in different cancers, including breast cancer where it was linked to the promotion of metastasis (37). Moreover, lysophosphatidylcholine acyltransferase 1 (LPCAT1), which catalyzes the conversion of LPC to PC, is overexpressed in different tumors including breast cancer (38–40). Nevertheless, the metabolism of phosphatidylcholines is significantly disturbed in cancer cells and their increased incorporation into plasma membranes enhances proliferation and motility. Therefore, the changed serum levels of their precursors (e.g., choline) and/or derivatives (e.g., lysophosphatidylcholines) are considered promising cancer markers (41, 42). This is noteworthy that reduced levels of LPC(18:0) were associated with an increased risk of different tumors including breast, prostate, colorectal, and lung cancers (14, 43, 44). The reduced level of serum cholesterol was another cancer-related feature observed in our study, though this effect was milder in patients with breast cancer compared to other cancers. Nevertheless, reduced levels of cholesterol may result from the fact that cholesterol is a key precursor of estrogen (45). Moreover, cholesterol-reach LDLs impact the proliferation of breast cancer cells due to the overexpression of Akt and ERK pathway intermediates (46), and high expression of LDL receptors was detected in breast cancer cells (47). Furthermore, increased serum concentrations of acylcarnitines, compounds involved in lipid and energy metabolism (48), were also characteristic of cancer patients. Noteworthy, similar metabolic pathways were associated with sets of compounds characteristic of breast cancer and characteristic of other types of solid cancers. Therefore, one should assume that metabolites whose serum levels differentiated healthy women from patients with breast cancer and other solid tumors were undoubtedly associated with metabolic pathways generally impaired in cancer cells.

Concentrations of serum metabolites are markedly affected by several biological and preanalytical confounding factors. These confounders include but are not limited to age, fasting status, and extended sample storage which were considered in the present study. However, other pre-analytical factors related to sample processing (49, 50) not controlled in the current study may be present as the cohorts were collected in the frame of different studies. Hence, the significance of the multi-cancer signature proposed in our pilot study should be further validated in the independent prospective study involving cohorts matched regarding age and medical conditions other than cancer status, where samples are collected and processed in fully standardized and controlled conditions. Nevertheless, a major strength of our results is the discovery of cancer signatures obtained in comparison with two different cohorts of control samples, and revealing the overlap between serum metabolome signatures of breast cancer and other types of solid cancers.

5 Conclusions

The high-throughput metabolomics approach implemented in the current study revealed a set of serum metabolites that discriminated between healthy women and breast cancer patients. The identified breast cancer signature included metabolites associated with known cancer-related pathways. Despite some differences in serum metabolome profiles among women with different solid cancers, a common set of metabolic features that discriminated cancer patients from healthy controls was established. Noteworthy, metabolites critical for discriminating breast cancer patients from controls included components of a hypothetical multi-cancer signature, which indicates wider potential applicability of a general metabolic cancer biomarker after its comprehensive validation.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material. Further inquiries can be directed to the corresponding authors.

Ethics statement

The studies involving humans were approved by Ethics Committee of Maria Sklodowska-Curie National Research Institute of Oncology, Gliwice Branch (KB/493-53/10 and KB/430-84/20) and the Regional Committee for Medical and Health Research Ethics (REK#1995/8395 and REK#2017/2231). The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

KM: Formal analysis, Investigation, Methodology, Writing – original draft. JD: Investigation, Writing – review & editing. KJ: Investigation, Methodology, Project administration, Writing – review & editing. AK: Data curation, Formal analysis, Methodology, Writing – review & editing. LP: Resources, Writing – review & editing. AW: Formal analysis, Writing – review & editing. MK: Formal analysis, Writing – review & editing. GG: Conceptualization, Writing – review & editing. TB: Conceptualization, Funding acquisition, Resources, Writing – review & editing. PW: Conceptualization, Funding acquisition, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research and the APC were funded by the Norwegian Financial Mechanism 2014-2021, National Science Centre, Poland, grant number 2019/34/H/NZ7/00503.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fonc.2024.1377373/full#supplementary-material

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin.* (2021) 71:209–49. doi: 10.3322/caac.21660

2. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature.* (2012) 490:61–70. doi: 10.1038/nature11412

3. Song JL, Chen C, Yuan JP, Sun SR. Progress in the clinical detection of heterogeneity in breast cancer. *Cancer Med.* (2016) 5:3475–88. doi: 10.1002/cam4.943

4. Cha YJ, Kim ES, Koo JS. Amino acid transporters and glutamine metabolism in breast cancer. *Int J Mol Sci.* (2018) 19:907–24. doi: 10.3390/ijms19030907

5. Tayanloo-Beik A, Sarvari M, Payab M, Gilany K, Alavi-Moghadam S, Gholami M, et al. OMICS insights into cancer histology; Metabolomics and proteomics approach. *Clin Biochem.* (2020) 84:13–20. doi: 10.1016/j.clinbiochem.2020.06.008

6. Damiani C, Gaglio D, Sacco E, Alberghina L, Vanoni M. Systems metabolomics: from metabolomic snapshots to design principles. *Curr Opin Biotechnol.* (2020) 63:190–9. doi: 10.1016/j.copbio.2020.02.013

7. Kosmides AK, Kamisoglu K, Calvano SE, Corbett SA, Androulakis IP. Metabolomic fingerprinting: challenges and opportunities. *Crit Rev BioMed Eng.* (2013) 41:205–21. doi: 10.1615/CritRevBiomedEng.v41.i3

8. Eniu DT, Romanciuc F, Moraru C, Goidescu I, Eniu D, Staicu A, et al. The decrease of some serum free amino acids can predict breast cancer diagnosis and progression. *Scand J Clin Lab Invest.* (2019) 79:17–24. doi: 10.1080/00365513.2018.1542541

9. Wang X, Zhao X, Chou J, Yu J, Yang T, Liu L, et al. Taurine, glutamic acid and ethylmalonic acid as important metabolites for detecting human breast cancer based on

the targeted metabolomics. *Cancer biomark.* (2018) 23:255–68. doi: 10.3233/CBM-181500

10. Lv W, Yang T. Identification of possible biomarkers for breast cancer from free fatty acid profiles determined by GC-MS and multivariate statistical analysis. *Clin Biochem.* (2012) 45:127–33. doi: 10.1016/j.clinbiochem.2011.10.011

11. Qiu Y, Zhou B, Su M, Baxter S, Zheng X, Zhao X, et al. Mass spectrometry-based quantitative metabolomics revealed a distinct lipid profile in breast cancer patients. *Int J Mol Sci.* (2013) 14:8047–61. doi: 10.3390/ijms14048047

12. Yang L, Wang Y, Cai H, Wang S, Shen Y, Ke C. Application of metabolomics in the diagnosis of breast cancer: a systematic review. *J Cancer.* (2020) 11:2540–51. doi: 10.7150/jca.37604

13. Åsvold BO, Langhammer A, Rehn TA, Kjelvik G, Grøntvedt TV, Sørgjerd EP, et al. Cohort profile update: the HUNT study, Norway. *Int J Epidemiol.* (2023) 52:e80–91. doi: 10.1093/ije/dyac095

14. Mrowiec K, Kurczyk A, Jelonek K, Debik J, Giskeødegård GF, Bathen TF, et al. Association of serum metabolome profile with the risk of breast cancer in participants of the HUNT2 study. *Front Oncol.* (2023) 13:1116806. doi: 10.3389/fonc.2023.1116806

15. Thompson JW, Adams KJ, Adamski J, Asad Y, Borts D, Bowden JA, et al. International ring trial of a high resolution targeted metabolomics and lipidomics platform for serum and plasma analysis. *Anal Chem.* (2019) 91:14407–16. doi: 10.1021/acs.analchem.9b02908

16. Zhang X, Dong J, Raftery D. Five easy metrics of data quality for LC–MS-based global metabolomics. *Anal Chem.* (2020) 92:12925–33. doi: 10.1021/acs.analchem.0c01493

17. Chen H, Quandt SA, Grzywacz JG, Arcury TA. A distribution-based multiple imputation method for handling bivariate pesticide data with values below the limit of detection. *Environ Health Perspect.* (2011) 119:351–6. doi: 10.1289/ehp.1002124

18. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics.* (2007) 8:118–27. doi: 10.1093/biostatistics/kxj037

19. Cohen J. *Statistical power analysis for the behavioral sciences. 2nd ed.* Hillsdale, NJ: Lawrence Erlbaum Associates (1988).

20. Haid M, Muschet C, Wahl S, Römisch-Margl W, Prehn C, Möller G, et al. Long-Term Stability of Human Plasma Metabolites during Storage at −80 °C. *J Proteome Res.* (2018) 17:203–11. doi: 10.1021/acs.jproteome.7b00518

21. McCartney A, Vignoli A, Biganzoli L, Love R, Tenori L, LuChinat C, et al. Metabolomics in breast cancer: A decade in review. *Cancer Treat Rev.* (2018) 67:88–96. doi: 10.1016/j.ctrv.2018.04.012

22. Chen Z, Li Z, Li H, Jiang Y. Metabolomics: a promising diagnostic and therapeutic implement for breast cancer. *Onco Targets Ther.* (2019) 12:6797–811. doi: 10.2147/OTT

23. Yuan B, Schafferer S, Tang Q, Scheffler M, Nees J, Heil J, et al. A plasma metabolite panel as biomarkers for early primary breast cancer detection. *Int J Cancer.* (2019) 144:2833–42. doi: 10.1002/ijc.31996

24. Cala MP, Aldana J, Medina J, Sánchez J, Guio J, Wist J, et al. Multiplatform plasma metabolic and lipid fingerprinting of breast cancer: A pilot control-case study in Colombian Hispanic women. *PloS One.* (2018) 13:e0190958. doi: 10.1371/journal.pone.0190958

25. Fan Y, Zhou X, Xia TS, Chen Z, Li J, Liu Q, et al. Human plasma metabolomics for identifying differential metabolites and predicting molecular subtypes of breast cancer. *Oncotarget.* (2016) 7:9925–38. doi: 10.18632/oncotarget.v7i9

26. Miyagi Y, Higashiyama M, Gochi A, Akaike M, Ishikawa T, Miura T, et al. Plasma free amino acid profiling of five types of cancer patients and its application for early detection. *PloS One.* (2011) 6:e24143. doi: 10.1371/journal.pone.0024143

27. Hou Y, Zhou M, Xie J, Chao P, Feng Q, Wu J. High glucose levels promote the proliferation of breast cancer cells through GTPases. *Breast Cancer (Dove Med Press).* (2017) 9:429–36. doi: 10.2147/BCTT

28. Park J, Shin Y, Kim TH, Kim DH, Lee A. Plasma metabolites as possible biomarkers for diagnosis of breast cancer. *PloS One.* (2019) 14:e0225129. doi: 10.1371/journal.pone.0225129

29. Ortmayr K, Dubuis S, Zampieri M. Metabolic profiling of cancer cells reveals genome-wide crosstalk between transcriptional regulators and metabolism. *Nat Commun.* (2019) 10:1841. doi: 10.1038/s41467-019-09695-9

30. Han J, Li Q, Chen Y, Yang Y. Recent metabolomics analysis in tumor metabolism reprogramming. *Front Mol Biosci.* (2021) 8. doi: 10.3389/fmolb.2021.763902

31. Xiong H, Zhang HT, Xiao HW, Huang CL, Huang MZ. Serum metabolomics coupling with clinical laboratory indicators reveal taxonomic features of leukemia. *Front Pharmacol.* (2022) 13:794042. doi: 10.3389/fphar.2022.794042

32. Budhu A, Terunuma A, Zhang G, Hussain SP, Ambs S, Wang XW. Metabolic profiles are principally different between cancers of the liver, pancreas and breast. *Int J Biol Sci.* (2014) 10:966–72. doi: 10.7150/ijbs.9810

33. Wei Z, Liu X, Cheng C, Yu W, Yi P. Metabolism of amino acids in cancer. *Front Cell Dev Biol.* (2020) 8:603837. doi: 10.3389/fcell.2020.603837

34. Wang W, Bai L, Li W, Cui J. The lipid metabolic landscape of cancers and new therapeutic perspectives. *Front Oncol.* (2020) 10:605154. doi: 10.3389/fonc.2020.605154

35. Ni H, Liu H, Gao R. Serum lipids and breast cancer risk: A meta-analysis of prospective cohort studies. *PloS One.* (2015) 10:e0142669. doi: 10.1371/journal.pone.0142669

36. Li X, Liu ZL, Wu YT, Wu H, Dai W, Arshad B, et al. Status of lipid and lipoprotein in female breast cancer patients at initial diagnosis and during chemotherapy. *Lipids Health Dis.* (2018) 17:91. doi: 10.1186/s12944-018-0745-1

37. Drosouni A, Panagopoulou M, Aidinis V, Chatzaki E. Autotaxin in breast cancer: role, epigenetic regulation and clinical implications. *Cancers (Basel).* (2022) 14:5437–51. doi: 10.3390/cancers14215437

38. Abdelzaher E, Mostafa MF. Lysophosphatidylcholine acyltransferase 1 (LPCAT1) upregulation in breast carcinoma contributes to tumor progression and predicts early tumor recurrence. *Tumour Biol.* (2015) 36:5473–83. doi: 10.1007/s13277-015-3214-8

39. Mansilla F, da Costa KA, Wang S, Kruhøffer M, Lewin TM, Orntoft TF, et al. Lysophosphatidylcholine acyltransferase 1 (LPCAT1) overexpression in human colorectal cancer. *J Mol Med (Berlin Germany).* (2009) 87:85–97. doi: 10.1007/s00109-008-0409-0

40. Shen L, Gu P, Qiu C, Ding WT, Zhang L, Cao WY, et al. Lysophosphatidylcholine acyltransferase 1 promotes epithelial-mesenchymal transition of hepatocellular carcinoma *via* the Wnt/β-catenin signaling pathway. *Ann Hepatol.* (2022) 27:100680. doi: 10.1016/j.aohep.2022.100680

41. Ackerstaff E, Glunde K, Bhujwalla ZM. Choline phospholipid metabolism: a target in cancer cells? *J Cell Biochem.* (2003) 90:525–33. doi: 10.1002/jcb.10659

42. Zhao Z, Xiao Y, Elson P, Tan H, Plummer SJ, Berk M, et al. Plasma lysophosphatidylcholine levels: potential biomarkers for colorectal cancer. *J Clin Oncol.* (2007) 25:2696–701. doi: 10.1200/JCO.2006.08.5571

43. Kühn T, Floegel A, Sookthai D, Johnson T, Rolle-Kampczyk U, Otto W, et al. Higher plasma levels of lysophosphatidylcholine 18:0 are related to a lower risk of common cancers in a prospective metabolomics study. *BMC Med.* (2016) 14:13. doi: 10.1186/s12916-016-0552-3

44. Widłak P, Jelonek K, Kurczyk A, Żyła J, Sitkiewicz M, Bottoni E, et al. Serum metabolite profiles in participants of lung cancer screening study; comparison of two independent cohorts. *Cancers (Basel).* (2021) 13:2714–27. doi: 10.3390/cancers13112714

45. Ben Hassen C, Goupille C, Vigor C, Durand T, Guéraud F, Silvente-Poirot S, et al. Is cholesterol a risk factor for breast cancer incidence and outcome? *J Steroid Biochem Mol Biol.* (2023) 232:106346. doi: 10.1016/j.jsbmb.2023.106346

46. dos Santos CR, Domingues G, Matias I, Matos J, Fonseca I, de Almeida JM, et al. LDL-cholesterol signaling induces breast cancer proliferation and invasion. *Lipids Health Dis.* (2014) 13:16. doi: 10.1186/1476-511X-13-16

47. Liu J, Xu A, Lam KS, Wong NS, Chen J, Shepherd PR, et al. Cholesterol-induced mammary tumorigenesis is enhanced by adiponectin deficiency: role of LDL receptor upregulation. *Oncotarget.* (2013) 4:1804–18. doi: 10.18632/oncotarget.v4i10

48. Dambrova M, Makrecka-Kuka M, Kuka J, Vilskersts R, Nordberg D, Attwood MM, et al. Acylcarnitines: nomenclature, biomarkers, therapeutic potential, drug targets, and clinical trials. *Pharmacol Rev.* (2022) 74:506–51. doi: 10.1124/pharmrev.121.000408

49. Debik J, Isaksen SH, Strømmen M, Spraul M, Schäfer H, Bathen TF, et al. Effect of delayed centrifugation on the levels of NMR-measured lipoproteins and metabolites in plasma and serum samples. *Anal Chem.* (2022) 94:17003–10. doi: 10.1021/acs.analchem.2c02167

50. Wang F, Debik J, Andreassen T, Euceda LR, Haukaas TH, Cannet C, et al. Effect of repeated freeze-thaw cycles on NMR-measured lipoproteins and metabolites in biofluids. *J Proteome Res.* (2019) 18:3681–8. doi: 10.1021/acs.jproteome.9b00343

scientific reports

Check for updates

OPEN

Impact of government policies on the COVID-19 pandemic unraveled by mathematical modelling

Agata Małgorzata Wilk[1,2]✉, Krzysztof Łakomiec[1,3], Krzysztof Psiuk-Maksymowicz[1,3] & Krzysztof Fujarewicz[1,3]✉

Since the very beginning of the COVID-19 pandemic, control policies and restrictions have been the hope for containing the rapid spread of the virus. However, the psychological and economic toll they take on society entails the necessity to develop an optimal control strategy. Assessment of the effectiveness of these interventions aided with mathematical modelling remains a non-trivial issue in terms of numerical conditioning due to the high number of parameters to estimate from a highly noisy dataset and significant correlations between policy timings. We propose a solution to the problem of parameter non-estimability utilizing data from a set of European countries. Treating a subset of parameters as common for all countries and the rest as country-specific, we construct a set of individualized models incorporating 13 different pandemic control measures, and estimate their parameters without prior assumptions. We demonstrate high predictive abilities of these models on an independent validation set and rank the policies by their effectiveness in reducing transmission rates. We show that raising awareness through information campaigns, providing income support, closing schools and workplaces, cancelling public events, and maintaining an open testing policy have the highest potential to mitigate the pandemic.

Over two years, more than 320 million confirmed cases and 5.5 million deaths[1] into the COVID-19 pandemic, health systems are now better equipped with knowledge and means to suppress the SARS-CoV-2 virus transmission, most importantly by mass vaccination. However, optimal control strategy remains a major issue, and was even more challenging before the development of specialized measures. The highly contagious virus spread rapidly throughout the world, gaining the status of a global pandemic less than four months after the first reported case[2]. Governing bodies were faced with the task of containing the pandemic using means generally deemed effective against other contagious diseases.

For detected cases and known exposed individuals, isolation and quarantine were generally implemented. Given the route of infection and high number of asymptomatic cases, considerable efforts have been focused on minimizing non-essential human contact. This resulted in a variety of social distancing policies, including business and school closing, cancellation of public events and restrictions on gatherings, as well as limiting mobility through international and internal travel controls. In extreme cases, emergency states and complete lock-downs were imposed[3]. Due to the enormous socio-economic impact of these interventions and growing controversy surrounding, for example, mandatory facial coverings, it is crucial to determine their effectiveness in mitigating the spread of COVID-19. With the traditional, case–control study design being infeasible for a pandemic happening in real time, the solution must be found through mathematical modelling.

The ability of non-pharmaceutical interventions to reduce coronavirus spread has been the subject of many studies harnessing a wide range of methodologies. The most common approaches include compartmental models[4–11], agent-based models[12–15], mobility or social networks[4,7,8,16], as well as mechanistic models[17], particle physics[18] and regression[19,20]. Usually, the research is focused on a specific policy or a small set of policies, such as mask use[12,14,20], school closing[13,15,19,21], travel controls[8,14,16], and social distancing/lockdown[5,7,17,18,20]. The main factor standing in the way of including more restrictions in a single model is parameterization and numerical

[1]Department of Systems Biology and Engineering, Silesian University of Technology, 44-100 Gliwice, Poland. [2]Department of Biostatistics and Bioinformatics, Maria Sklodowska-Curie National Research Institute Gliwice Branch, 44-100 Gliwice, Poland. [3]Biotechnology Center, Silesian University of Technology, 44-100 Gliwice, Poland. ✉email: agata.wilk@polsl.pl; krzysztof.fujarewicz@polsl.pl

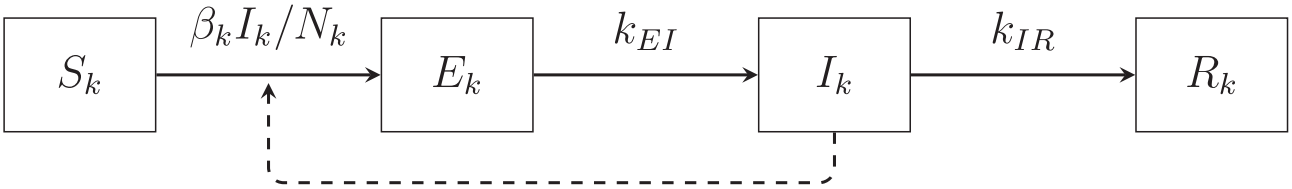


Figure 1. Structure of the SEIR model describing a single country. Susceptible individuals (S_k) are exposed (E_k) to the virus at a rate $\beta_k I_k/N_k$. They become infectious (I_k) at a rate k_{EI} and recover/are removed (R_k) at a rate k_{IR} .

conditioning; as noted by Castex et al.[10], timings of control policies are highly correlated. Jorge et al.[11] solve this issue by constructing a synthetic, time-dependent stringency index. Köhler et al.[9] balance the complexity of their model against the size of the dataset by taking advantage of prior knowledge and introducing certain constraints on parameter values. These solutions, although effective, are not applicable in cases of poorly known systems where it is impossible to formulate reasonable assumptions.

Here we propose a workflow for prediction of the efficiency of different policies in reducing transmission rates of SARS-CoV-2 infections using a SEIR model incorporating 13 different pandemic control interventions. We demonstrate a multi-step method of parameter estimation without any prior assumptions through individualized modelling of a cohort of European countries, based on adjoint sensitivity analysis, non-linear least squares and coordination. We confirm the satisfying predictive ability of our approach compared to classical modelling strategies over a separate validation time period. Using the developed algorithm we rank control policies by their efficiency in reducing virus transmission rates.

Methods

Epidemic model. We simulated the COVID-19 pandemic in the kth country using a version of the Susceptible-Exposed-Infectious-Removed (SEIR) model[22], described by the following system of ordinary differential equations:

$$\begin{cases} \dot{S}_k(t) = \frac{-\beta_k(t) S_k(t) I_k(t)}{N_k} \\ \dot{E}_k(t) = \frac{\beta_k(t) S_k(t) I_k(t)}{N_k} - k_{EI} E_k(t) \\ \dot{I}_k(t) = k_{EI} E_k(t) - k_{IR} I_k(t) \\ \dot{R}_k(t) = k_{IR} I_k(t) \end{cases} ; \quad k = 1, ..., K \tag{1}$$

with initial conditions $S_k(0) = N_k - I_0$, $E_k(0) = 0$, $I_k(0) = I_0$, $R_k(0) = 0$.

In the above equations the variables S_k, E_k, I_k, and R_k represent the numbers of individuals who are susceptible, exposed, infectious and removed from compartments for the kth country, respectively (Fig. 1). N_k is equal to the sum of all compartments of the SEIR model (1) for the kth country

$$N_k = S_k(t) + E_k(t) + I_k(t) + R_k(t) = \text{const.} \tag{2}$$

The coefficients k_{EI} and k_{IR} are parameters of the SEIR model (1) which stand for the inverse of times of viral latency (defined as the time to becoming contagious, not to symptom onset) and of recovery from infection, respectively. The function $\beta_k(t)$ represents the time-dependent virus transmission intensity for country k.

Data. The statistics on reported COVID-19 cases were taken from the JHU CSSE data repository[1]. Specifically, we considered two manners of reporting cases in the kth country: daily infections on the ith day, denoted as $D_k(t_i)$ ($D_{dk}(t_i)$ where it is necessary to indicate observed data, as opposed to $D_{mk}(t_i)$ estimated from modelling) or cumulative cases $C_k(t_i)$ ($C_{dk}(t_i)$ or $C_{mk}(t_i)$ where necessary). Of course, $D_k(t_i)$ and $C_k(t_i)$ satisfy the relation

$$C_k(t_i) = \sum_{j=1}^{i} D_k(t_j) \ .$$

Although the data contained obvious artifacts resulting from policy changes in reporting cases and retrospective updates, these were not considered exclusion criteria. We considered a time frame from the beginning of the pandemic in Europe to the end of January 2021 when vaccinations (which are not included in the model) started to take effect. To prevent information leakage (evaluating a model using the same data it was trained on), the time period between 1 February 2020 and 31 November 2020 was the basis for parameter estimation, and the final two months (between 1 December 2020 and 31 January 2021) were used to validate the models.

We used publicly available pandemic control information provided by the Oxford COVID-19 Government Response Tracker[3], focusing on European countries. We selected $r = 13$ policies which may potentially influence the spread of SARS-CoV-2, with the exception of restrictions on international travel, not accounted for in the model. We included income support and debt relief, since economic measures, while not directly limiting virus transmission, affect the observance of restrictions.

Containment and closure	Economic	Health system
School closing [0–3]	Income support [0–2]	Public information campaigns [0–2]
Workplace closing [0–3]	Debt/contract relief [0–2]	Testing policy [0–3]
Cancelling public events [0–2]		Contact tracing [0–2]
Restrictions on gatherings [0–4]		Facial coverings [0–4]
Close public transport [0–2]		
Stay at home requirements [0–3]		
Restrictions on internal movement [0–2]		

Table 1. 13 government policies to mitigate the pandemic used in the mathematical model. Numbers in brackets represent the original value ranges of each policy.

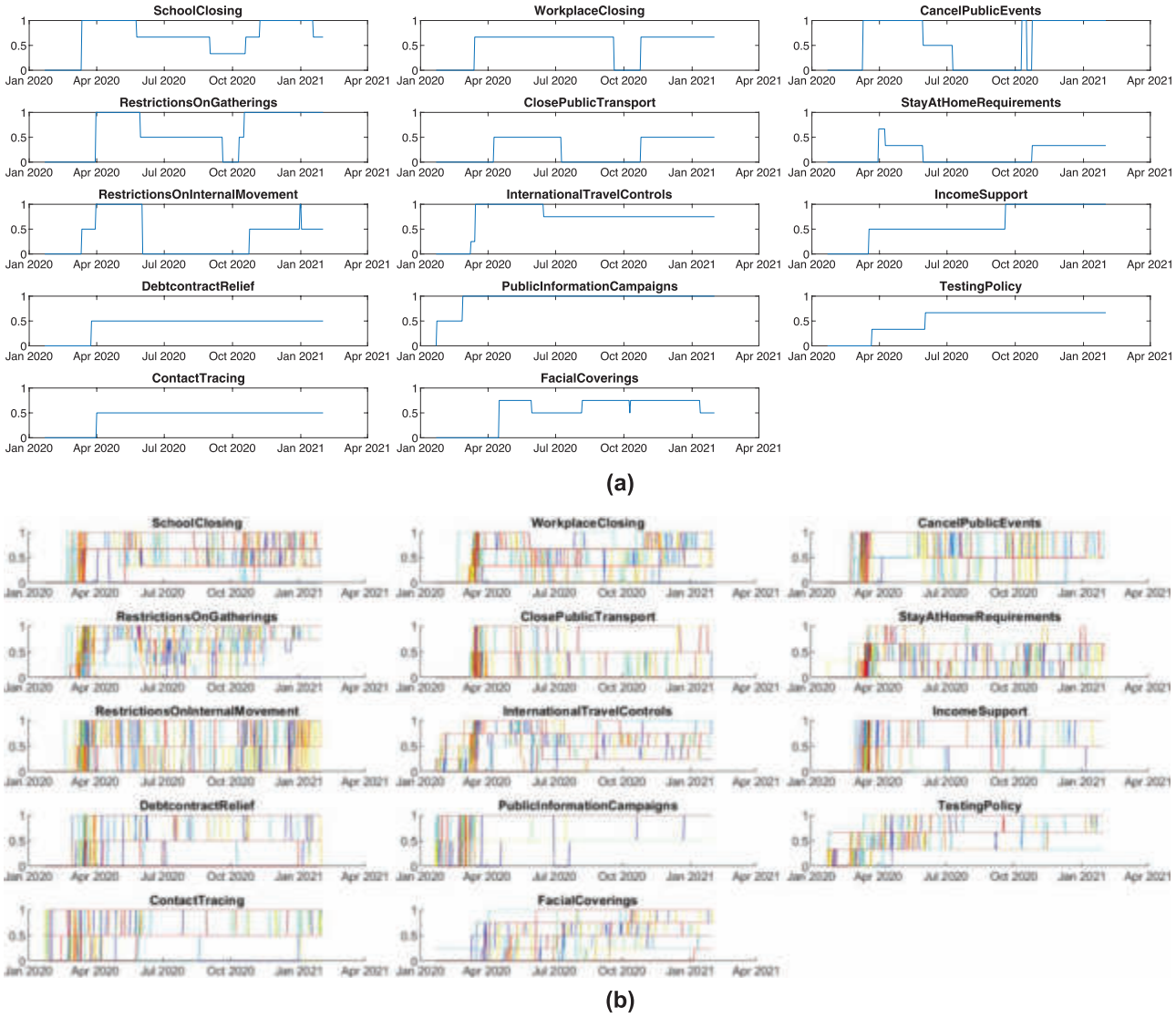


Figure 2. (**a**) Moments of enabling/disabling specific actions to prevent the spread of the COVID-19 pandemic in Poland. The enabling times are the same for debt contract relief and contact tracing, making them indistinguishable for the model. Several policies never reach their maximum level, which makes interpretation difficult. (**b**) Restriction functions for all European countries (each color represents a distinct country). Due to the large number of countries, the figure shown is only illustrative to demonstrate that the problem of indistinguishability of the impact of restrictions, observed for a single country, is now no longer relevant. Separate charts for individual countries can be seen in the Supplementary File.

For increased interpretability, we scaled all values to the range [0–1], with varying degrees of policies denoted by fractions. Countries with incomplete data were excluded from the analysis, resulting in a total of $K = 42$ countries to which the model was applied.

The extents of particular government policies in a given country can be treated as time-dependent functions, and for policy i and country k we denote them as $o_{ki}(t)$. Since the policy levels are discrete, $o_{ki}(t)$ takes the form of step functions, examples of which can be seen in Fig. 2a.

Impact of restrictions. The time-dependent virus transmission intensity $\beta(t)$ varies for each country depending on the implemented restrictions and individual factors including temperature, humidity, population density etc. We describe it as a function of restrictions incorporated during the pandemic with a generic form:

$$\beta(t) = b(1 - a_1 o_1(t) - a_2 o_2(t) - \cdots - a_r o_r(t)), \tag{3}$$

where time functions $o_i(t)$ for $i = 1, 2, \ldots, r$ are governments' policies (Table 1), coefficients a_i are weights reflecting their efficiency, and b is a constant value representing the native unaffected virus transmission intensity (possibly influenced by other factors). When all government policies are inactive then, according to the formula (3), the virus transmission rate is unchanged and constant $\beta(t) = b$. In addition, taking into account that functions $o_i(t)$ takes values from 0 to 1 (1 means the strongest level of the particular restriction), it is possible to interpret estimated parameters a_i as effectiveness of the policy $o_i(t)$. For example, if $a_i = 0.1$ then the strongest level of the corresponding policy $o_i(t)$ will reduce the virus transmission rate $\beta(t)$ by 10%.

Modelling approaches for a cohort of entities. The Eq. (3), substituted into the SEIR model (1), means that for each country it is necessary to estimate 14 parameters $(a_1, a_2, \ldots, a_{13}, b)$, in addition to the parameters k_{EI} and k_{IR}. However, looking at the moments of enabling or disabling selected actions for a single country (in the present case, Poland) presented in Fig. 2a, it is easy to notice that, for example, the actions "Debt / contract relief" and "Contact tracing" were activated practically at the same time. As a result, an appropriately parameterized model will not be able to distinguish the impact of these two actions, and the two parameters responsible for these actions will be a pair of non-estimable[23] parameters (an increase in one can be compensated by a decrease in the other).

We countered this problem by incorporating data from the entire cohort of countries (Fig. 2b) and developing an individualized approach to modeling.

Consider K real objects, systems, or processes. Each of them is described using the same model M (mathematical or computational), and the differences in their behavior result from: different initial conditions and different signals affecting the objects (control signals). Each of the models is additionally described by the m-element vector P_k. Parameters are estimated on the basis of data sets obtained independently for each object: $\mathscr{U}_1, \mathscr{U}_2, ..., \mathscr{U}_K$, which form a common data set $\mathscr{U} = \{\mathscr{U}_1 \cup \mathscr{U}_2 \cup ... \cup \mathscr{U}_K\}$.

The task of fitting K models to data $\mathscr{U}$, can be solved by creating:

(A) a common model,
(B) independent models,
(C) the proposed individualized models.

A. Common model. This consists of estimating a common vector of parameters, the same for each of the K models, in the form:

$$P_1 = P_2 = \cdots = P_K = [a_1, a_2, ..., a_m]^T \tag{4}$$

based on the common dataset of $\mathscr{U}$. The number of estimated unique parameters for the entire set of K models is relatively small and is equal to m.

This modeling method is typical, for example, in statistics (e.g. linear regression), where one common model is built based on a sample drawn from the population, then applied to each element of the population.

B. Independent models. This approach is based on the estimation of the parameter vector in the form

$$P_k = [b_{k1}, b_{k2}, ..., b_{km}]^T \tag{5}$$

on the basis of the $\mathscr{U}_k$ independently for each of the K models, $k = 1, 2, ..., K$. The number of estimated parameters in this case is very large and is equal to Km.

Such an approach is often used for technical objects and where experimental data is cheap and readily available.

C. Individualized models. In this case, we assume that among the m parameters characterizing a single model, r parameters $a_1, a_2, ..., a_r$ are common parameters and the remaining q parameters $b_1, b_2, ..., b_q$ are individual parameters. Of course, $r + q = m$. The entire parameter vector for the kth model has the form

$$P_k = [a_1, a_2, \ldots, a_r, b_{k1}, b_{k2}, \ldots, b_{kq}]^T \tag{6}$$

The number of estimated parameters in this case is a compromise and equals $r + qK$. Basic differences between approaches A, B and C are presented in Table 2.

Approach	Common model	Independent models	Individualized models
Main idea	Estimates model parameters from a combined dataset of all countries	Estimate parameters for each country separately	Estimate a subset of parameters separately for each country and the rest from a combined dataset
Number of parameters	Low	High	Compromise
Ability to fit model to data	Low	High	Compromise
Overfitting risk	Low	High	Compromise
Risk of numerically ill-conditioned parameter estimation problem	Low	High	Compromise

Table 2. Comparison of approaches to modeling a set of entities. Undesired characteristics are indicated in red.

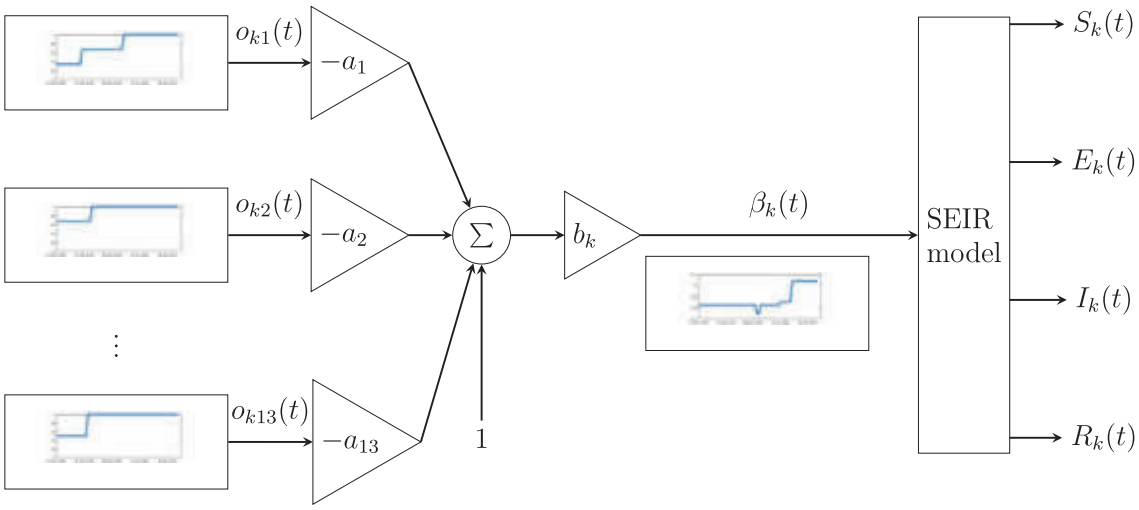


Figure 3. Block diagram of the entire COVID-19 mathematical model for the kth country. The non-stationary parameter $\beta_k(t)$ is calculated as a function of government policies $o_{ki}(t)$. The diagram presents *individualized* version of the model (7), where parameters a_i are common for all countries and the parameter b_k is an individual parameter for the kth country. The structure of the *common* and the *independent* models are the same—the only difference is related to parameters of the function calculating $\beta_k(t)$.

For individualized simulation of the pandemic, we assumed that parameters a_i representing the efficiency of individual policies were common. Biases b_k, which also serve as scaling factors for policy efficiencies, were estimated separately for each country. Therefore, based on Eq. (3), the formula for virus transmission intensity in kth country can be written as:

$$\beta_k(t) = b_k(1 - a_1 o_{k1}(t) - a_2 o_{k2}(t) - \cdots - a_r o_{kr}(t)) \ . \tag{7}$$

The full individualized model, containing the static part (7) calculating $\beta_k(t)$ followed by the SEIR model (1), is presented in Fig. 3.

In addition to k_{EI} and k_{IR} which can be viewed as virus properties, the numbers of estimated parameters were: 14 for the common model, 588 for independent models, and 55 for individualized models.

Parameter estimation. Our goal was to fit the daily infections predicted by the model $D_{mk}(t_i)$ to the observed data $D_{dk}(t_i)$ by minimization of the quadratic objective function:

$$MSE = \sum_{k=1}^{K} \sum_{i=1}^{M} (D_{mk}(t_i) - D_{dk}(t_i))^2. \tag{8}$$

It is worth explaining at this point how the SEIR model (1) is used to predict daily infections $D_{mk}(t_i)$. The cumulative cases can be expressed as a sum of $I(t)$ and $R(t)$ compartments of the model (1):

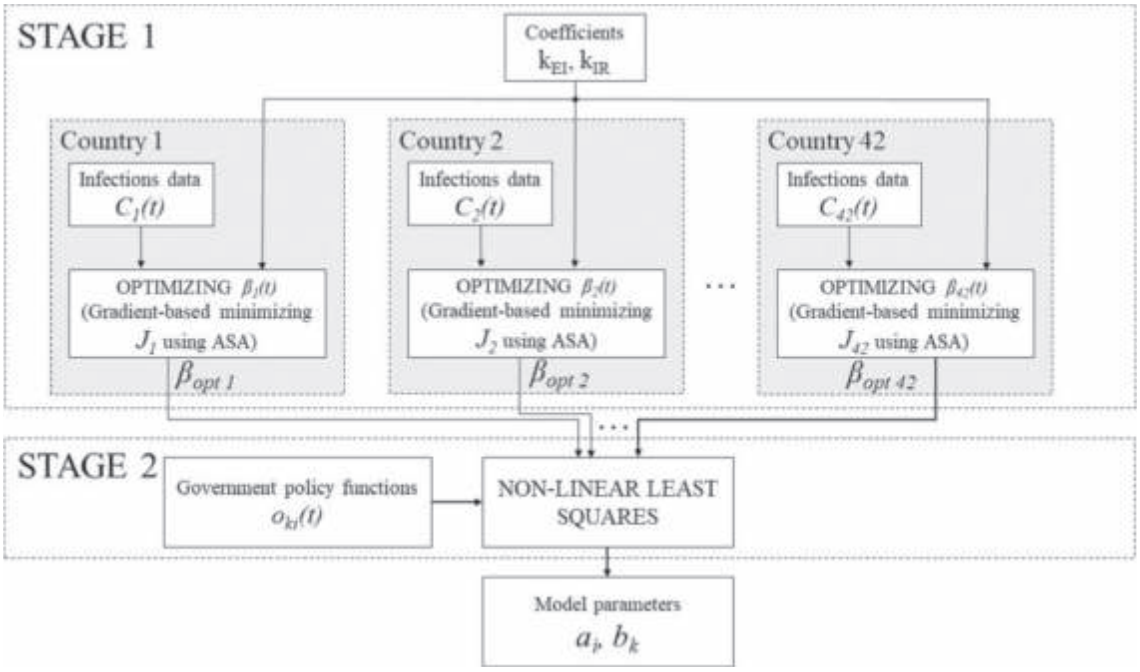


Figure 4. Two-stage parameter estimation workflow for a single set of k_{EI} and k_{IR} values. For each country, optimal $\beta(t)$ is estimated based on cumulative cases using adjoint sensitivity analysis. Based on all the $\beta_{opt}(t)$ functions and restriction functions $o_{ki}(t)$, model parameters are estimated with non-linear least squares.

$$C_{mk}(t_i) = I_k(t_i) + R_k(t_i)\ . \tag{9}$$

On the other hand, the daily cases $D_{mk}(t_i)$ may be estimated as a rate of changes of the sum from the above equation, which, taking into account that the time unit is equal to one day, gives

$$D_{mk}(t_i) \approx \frac{\mathrm{d}(I_k(t_i) + R_k(t_i))}{\mathrm{d}t} = k_{EI} E_k(t_i)\ . \tag{10}$$

To reduce computation time we divided the parameter estimation process into two stages (Fig. 4).

Two-stage procedure. First, in *STAGE 1*, we found the function $\beta_{opt}(t)$ separately for each country by minimizing the following objective function:

$$J_k = \sum_{i=1}^{M} \left(C_{mk}(t_i) - C_{dk}(t_i) \right)^2 , \tag{11}$$

where $C_{mk}(t_i)$ and $C_{dk}(t_i)$ are the numbers of cumulative infections for country k at time t_i predicted by the SEIR model and obtained from data, respectively. The reason why we used here the numbers of cumulative cases was the high noise in data of daily cases (Fig. 5).

To minimize the function J_k we used gradient descent method in which the gradient $\nabla_{\beta(t)} J$ was computed using adjoint sensitivity analysis (ASA)[24–30,31]. This stage was independent from government policies (and consequently, from the modelling approach) and produced a near-perfect fit (Fig. 5).

Next, in *STAGE 2*, we employed the non-linear least squares method to obtain the values of parameters P, based on the set of functions $\beta_{opt}(t)$ and restriction functions $o_{ki}(t)$.

We repeated this procedure for different values of parameters k_{EI} and k_{IR}. The values for which we obtained the best global model performance then served as a starting point for their additional fine-tuning together with optimizing $\beta_k(t)$ for all countries. This optimization has been done using two-level direct method of coordination[32] where k_{EI} and k_{IR} were treated as upper level (coordination) decision variables and $\beta_k(t)$ were the bottom level (local) optimized signals. Effectively, while this process can be considered an extension of *STAGE1* in which coefficients k_{EI} and k_{IR} are not entirely arbitrary, it remains consistent with the general notion of a two-stage parameter estimation. The first stage remains independent from government policies.

Analyses were performed using the Matlab environment, version 2021a.

Model evaluation. For a quantitative comparison of models, the root-mean-square error was calculated for the validation period ($M_v = 62$ days, from 1 December 2020 to 31 January 2021) against daily infections (estimated and observed, denoted as D_m and D_d, respectively).

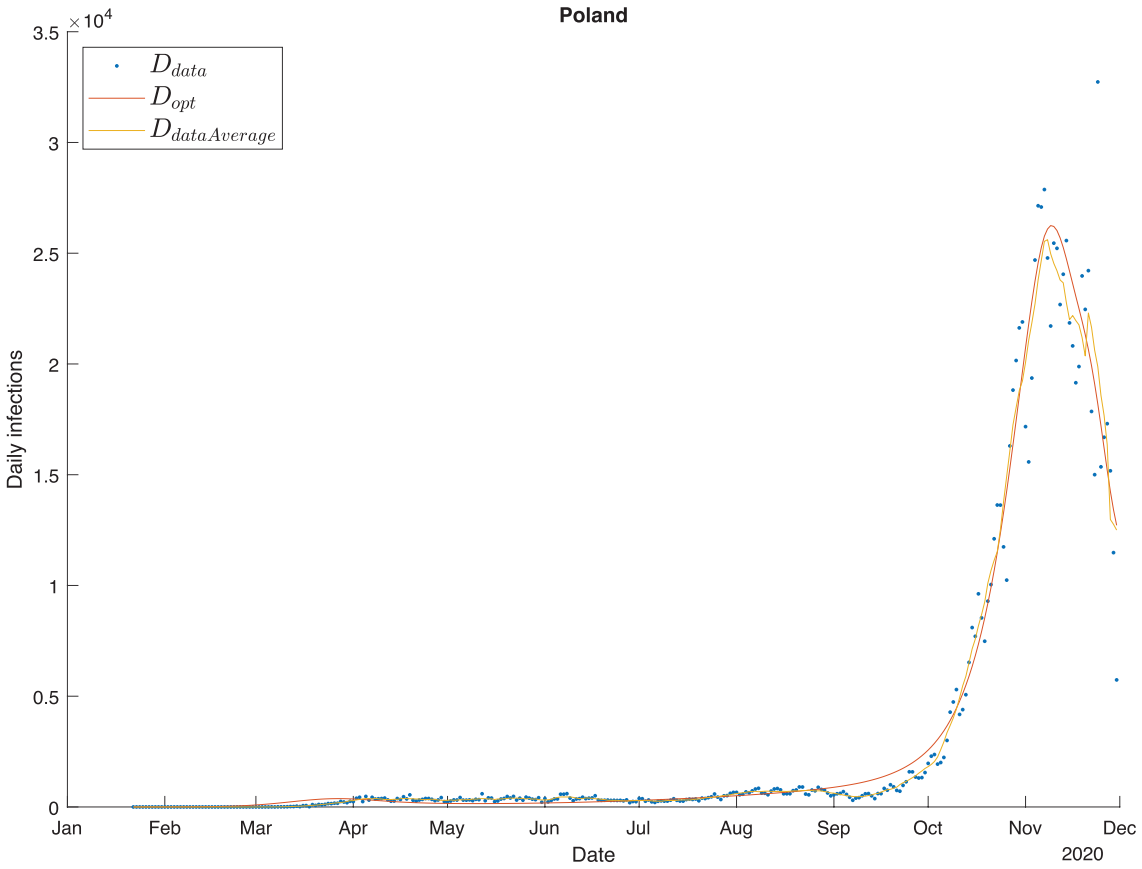


Figure 5. Example fit (Poland) based on β_{opt} over the training time period. Blue points represent daily infections, the yellow line the seven-day moving average of daily infections, and the red line is the fit obtained by substituting the $\beta_{opt}(t)$ into the SEIR model. The red and yellow lines practically overlap.

k_{EI}	k_{IR}	Common model	Independent models	Individualized models
0.05	0.1	2.579	**3.712**	2.762
0.15	0.1	2.376	4.599	1.876
0.20	0.1	2.400	4.232	1.722
0.25	0.1	**2.356**	5.091	**1.675**
0.15	0.01	4.983	5.669	5.225

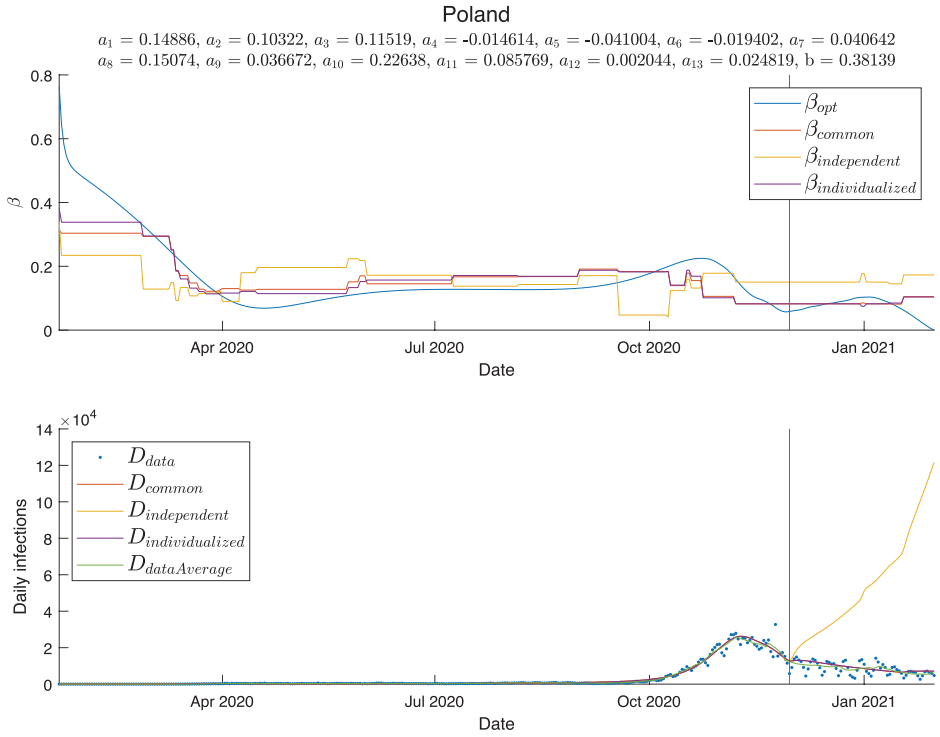
Table 3. Average NRMSE for different values of parameters k_{EI} and k_{IR} parameters. The lowest error for each modelling approach is shown in bold. The lowest error overall was achieved for the individualized approach for $k_{EI} = 0.25$ and $k_{IR} = 0.1$.

$$RMSE = \sqrt{\frac{\sum_{i=M+1}^{M+M_v}(D_m(t_i) - D_d(t_i))^2}{M_v}}. \tag{12}$$

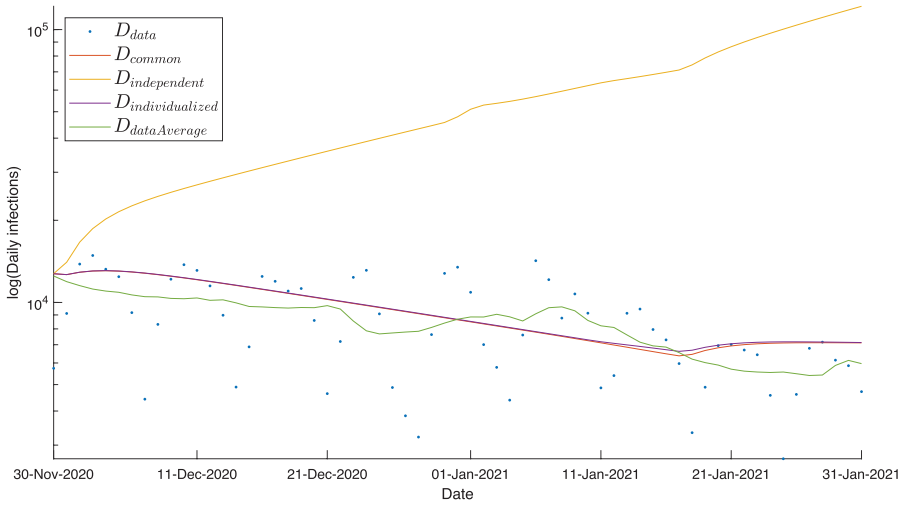
Due to varying number of infections in different countries, RMSE values were normalized by the average number of infections in the validation period.

$$NRMSE = \frac{RMSE}{\frac{1}{M_v}\sum_{i=M+1}^{M+M_v} D_d(t_i)}. \tag{13}$$

The mean of NRMSE over all countries was considered a general measure of model quality.



(a) Estimation results for Poland. Top panel: fitting a function of restrictions to the signal β_{opt}. Bottom panel: Ability of the model to predict daily infections. The black vertical line indicates the beginning of the validation period. As the number of daily infections is highly variable, a seven day moving average is also presented



(b) Estimation results for Poland: a close-up of the validation period. For better visibility, the y axis is presented as a log scale. Blue points represent daily infections, the green line is a 7-day moving average. The common and individualized approaches (red and purple, respectively), are able to predict infections quite accurately, and the independent model (yellow line) considerably deviates from the observed data.

Figure 6. Prediction results.

Results

Adjusting k_{EI} and k_{IR} parameters. The predictive ability of the model depends heavily on the values of parameters k_{EI} and k_{IR} (Table 3) with overall best performance achieved by the individualized model for values $k_{EI} = 0.25$ and $k_{IR} = 0.1$. It is worth noting that for common and individualized approaches, the lowest errors coincide with k_{EI} and k_{IR} corresponding to latency and recovery times within ranges estimated for SARS-CoV-2[33]. This is, however, not the case for independent models.

Country	Common	Independent	Individualized	Country	Common	Independent	Individualized
NRMSE by country							
Albania	4.09	**0.62**	0.93	Latvia	**0.70**	8.13	0.88
Andorra	**0.71**	8.94	0.90	Lithuania	0.70	15.90	**0.42**
Austria	**0.50**	0.80	0.55	Luxembourg	2.96	9.65	**1.75**
Belarus	24.39	2.96	**0.81**	Malta	**0.39**	10.39	0.69
Belgium	1.75	**0.76**	5.06	Moldova	4.97	19.14	**1.27**
Bosnia and Herzegovina	17.32	**2.34**	11.86	Monaco	**0.93**	1.06	1.03
Bulgaria	3.78	3.93	**1.62**	Netherlands	0.51	4.70	**0.43**
Croatia	**3.01**	22.97	3.67	Norway	1.45	1.97	**0.34**
Cyprus	**0.89**	0.97	1.01	Poland	**0.36**	6.59	**0.36**
Czech Rep.	**0.54**	0.58	0.66	Portugal	0.87	**0.55**	0.59
Denmark	0.70	0.87	**0.66**	Romania	**0.33**	0.53	0.40
Estonia	1.27	15.52	**0.46**	Russia	0.35	**0.24**	1.14
Finland	**0.37**	1.29	0.39	San Marino	**0.83**	0.87	0.93
France	**0.68**	0.97	1.40	Serbia	1.33	2.11	**0.47**
Germany	1.09	**0.66**	3.81	Slovenia	**0.67**	13.27	0.90
Greece	0.68	0.94	**0.37**	Spain	1.14	**1.10**	**1.10**
Hungary	3.15	**1.31**	2.56	Sweden	1.45	4.34	**1.44**
Iceland	1.27	15.73	**1.11**	Switzerland	1.15	**0.88**	1.05
Ireland	1.32	**1.27**	1.34	Turkey	**3.47**	3.84	3.54
Italy	**2.14**	4.28	8.42	Ukraine	**1.32**	2.09	2.67
Kosovo	3.53	10.58	**1.17**	UK	0.70	0.96	**0.48**
		Common model		**Independent models**		**Individualized models**	
Summary							
Average NRMSE		2.374		4.920		1.682	
Standard deviation		4.414		5.906		2.222	

Table 4. Performances of the final models (with optimized values of coefficients $k_{EI} = 0.2605$ and $k_{IR} = 0.1020$). For each country, lowest error values are in bold. Average NRMSE and standard deviation for each approach is also presented.

The final estimates, obtained with the coordination method, are $k_{EI} = 0.2605$ and $k_{IR} = 0.1020$, which corresponds to approximately 3.8 days of latency period and 9.8 days of infectious period. These parameter values were used in the subsequent analyses.

Predictive ability of the model. The estimation results for Poland, presented in Fig. 6, are a representative example of a general tendency observed in the behavior of our models. With few exceptions, while predictions obtained with common and individualized approaches are reasonably close to the observed daily infection numbers, the independent model displays typical signs of overfitting. The results for all countries are presented in the Supplementary Material.

Performances of all the models are outlined in Table 4. Common and individualized approaches are similar in terms of the number of countries where they are the best fit. Nevertheless, errors of the individualized approach where it proved less efficient tend to be smaller, which is reflected in the average NRMSE of 1.682, as compared to 2.374 for the common model, as well as the standard deviations. The independent modelling approach appears inferior in all respects.

Impact of control policies. The weights a_i representing the effectiveness of different interventions varied between approaches, although the general inference is similar for common and individualized approaches (Fig. 7). For the common model they were in the range between −0.04 and 0.22, for the independent models between −2.41 and 1.66, and for the individualized models between −0.04 and 0.23.

For the individualized models, Table 5 shows the ranking of policies from the most to the least effective.

Discussion

Correct model parametrization and numerical conditioning is a non-trivial task, particularly in complex systems with sparse data available for estimation. This issue became a major obstacle in determining the optimal strategy for non-pharmaceutical interventions dedicated to mitigating the spread of the COVID-19 pandemic, as the coincidence of several policies rendered the impact of individual ones indistinguishable in terms of mathematical modelling. Furthermore, highly noisy data may cause difficulties to obtain a converged solution.

The perhaps most intuitive approach, which is modelling each country independently, may prove inadequate for estimation of a large number of parameters, particularly when no constraints on their values are imposed.

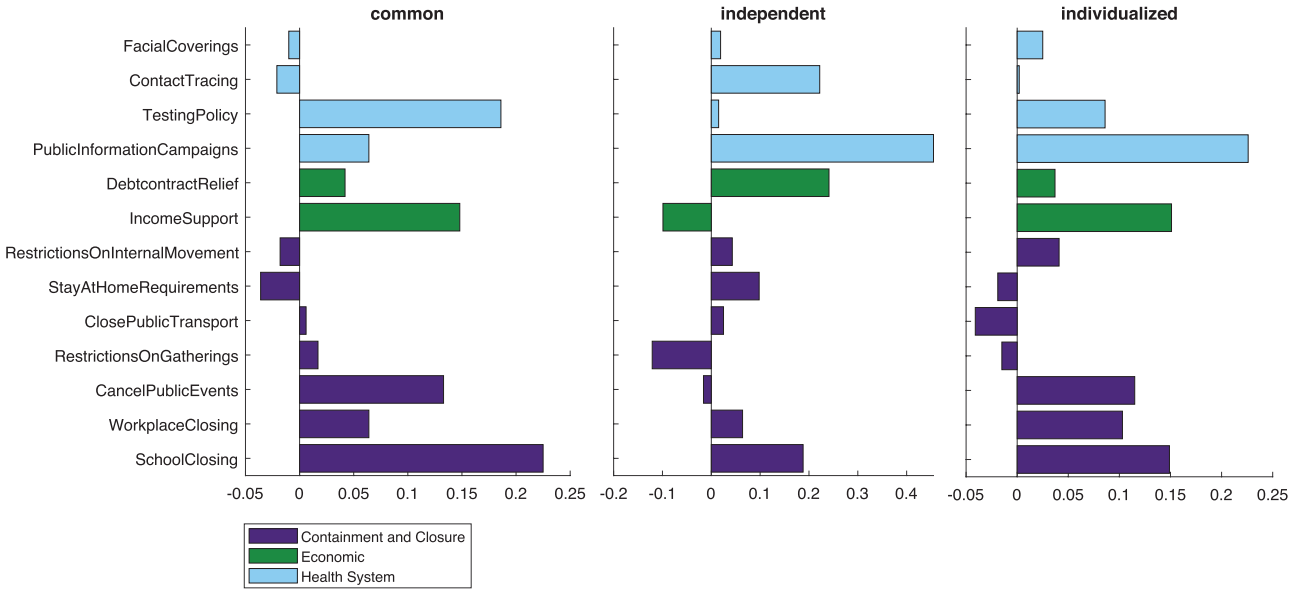


Figure 7. Estimated weights for control policies using different approaches. High positive values correspond to effective policies. Since in independent models each country has its own parameter set, for demonstrative purposes values for Poland are presented.

Rank	Policy	a_i
1	Public information campaigns	0.226
2	Income support	0.151
3	School closing	0.149
4	Cancel public events	0.115
5	Workplace closing	0.103
6	Testing policy	0.086
7	Restrictions on internal movement	0.041
8	Debt/contract relief	0.037
9	Facial coverings	0.025
10	Contact tracing	0.002
11	Restrictions on gatherings	− 0.015
12	Stay at home requirements	− 0.019
13	Close public transport	− 0.041
	Sum of all a_i	0.859

Table 5. Ranking of policies according to effectiveness.

Poor numerical conditioning may result in unrealistic estimations. This became apparent for the parameters k_{EI} and k_{IR}, for which good estimates can be found in literature. The mean latent period for COVID-19 has been estimated as 3.3 days[34], and the mean incubation period (which is reported to be up to 2 days longer than the latent period) as 5.5 days[35], 5.8 days[33] or 5.6 days[36]. The latent period (calculated as $\frac{1}{k_{EI}}$) for which the lowest errors were obtained, was approximately 4 days for the common and individualized models, and 20 days for the independent models. While the common and individualized approach produce the best results for latency times close to the actual values reported for COVID-19, the independent approach worked best for an unrealistic value. Moreover, NRMSE values for the independent approach vary considerably even for small changes of k_{EI} within a realistic range (0.15, 0.2 and 0.25), reinforcing the expectation of its numerical instability.

We developed a workflow for assessing the effect of pandemic control policies in Europe utilizing data from the entire cohort of countries. Our individualized approach yields satisfactory results with no assumptions regarding parameter values. Meanwhile, for independent models poor numerical conditioning was evident in that the estimated weights a_i often reached negative values (even as low as −2.41), which in practice would suggest that the interventions increased virus transmission rates. Indeed, while introduction of a non-negativity constraint significantly improved the performance of independent models, it had little effect on individualized models. A common model has similar numerical advantages, however it does not capture individual country characteristics.

Comparing the values a_i for individual policies, there is a pronounced inconsistency between modelling approaches, particularly for the independent models. For example, "School closing" had a weight of 0.225 for the

common model, 0.149 for the individualized model, and between −0.443 and 0.718 for the independent models. As discussed above, the data used for fitting the independent models is too scarce and noisy for the considered number of parameters to be estimable. Hence, the estimations obtained with this approach cannot reliably be used for inference. The differences between weights estimated by common and individualized models, although present, do not lead to drastically different conclusions—indeed, both models indicate that school closing has a large impact on virus transmission.

Application of the methodology proposed here enabled a ranking of government policies according to their impact. The most effective measures are public information campaigns followed closely by income support, school and workplace closures, cancellation of public events and open testing policy. The high rank of information campaigns emphasises the importance of knowledge and public access to verified information in mitigating a crisis. As hypothesised at the stage of pre-selecting policies to be included in the model, providing financial support to those affected by the pandemic proved effective, likely through reducing the necessity of bypassing restrictions to secure livelihood. An open testing policy, particularly not limiting testing to symptomatic individuals, ensures higher detection rates and consequently more effective case isolation.

Perhaps surprisingly, some of the intuitively powerful measures such as facial coverings, restrictions on gatherings and stay at home requirements ranked relatively low. One possible explanation is that the restrictions pertain to formal policies, not to the degree of their observance. Indeed, a certain pattern may be noticed. Assuming a threshold of 0.05, the ineffective policies comprise restrictions on internal movement, debt/contract relief, facial coverings, contact tracing, restrictions on gatherings, stay at home requirements and closing public transport. Almost every policy deemed ineffective is less tangible, difficult to enforce and monitor. Restrictions on internal movement, gatherings and stay at home requirements would require frequent and strict controls beyond the capabilities of any country's police force. Precise and exact contact tracing is practically impossible to achieve since it relies on either the entire population providing a perfect and constant account of all their encounters or using a geolocation device at any given time. Facial coverings, as possibly the most debated restriction, have met with considerable resistance. Moreover, especially at the beginning of the pandemic, limited sanitary resources were rerouted to the health system with the rest of the general population using homemade coverings. In contrast, the policies deemed effective are straightforward and well-defined. Public information campaigns, income support, school closing, cancellation of public events, workplace closing and testing policy are all top-down, independent of the individual and easily traceable. The low position in the ranking of using facial coverings may also be attributed to infections occurring primarily by prolonged contact, for example with family, during events, or at schools or workplaces (which ranked high), where masks are often neglected.

Notably, the sum of all a_i values is 0.859, which suggests implementing all restrictions simultaneously at the highest level would decrease the virus transmission rate to approximately 14% of its original value.

This study is not without limitations. The SEIR model used to model the pandemic is relatively simple as it does not consider repeated infections or population structure. Furthermore, our estimations were based on confirmed cases constituting only a fraction of the actual number of infections, many of which were asymptomatic. One solution could be correcting the number of cases for testing capacity. However, there are considerable inconsistencies and gaps in reporting numbers of tests: in many countries the number of tests was reported only several months into the pandemic, in others only cumulative number was reported weekly. In total, to incorporate testing data, even allowing for a certain percentage of missing values, 23 countries would have to be excluded from the analysis. The number of deaths could also be used as an alternative indicator of the course of the pandemic. Yet, the reporting of COVID-related deaths also varied by country, even evolving within a single country—the reported deaths were sometimes interpreted as ones directly caused by SARS-CoV-2, otherwise as any death coinciding with infection. To some extent, testing accessibility is incorporated into our model as one of the policies ("Testing policy"). Lower levels, usually observed at the beginning of the pandemic, indicate limited access to tests. Higher levels, denoting unlimited access to testing, are typically observed at the later time when the testing system was fully developed. The importance of testing capacity, allowing for more effective isolation of infected individuals, is reflected in the high coefficient for this policy.

This work may be extended in several directions. Additional compartments may be included in the model, representing for example vaccinated individuals, asymptomatic cases, quarantined individuals, movement of people between countries and so on. Alternative forms of the restriction impact function may also be considered, including a multiplicative form.

Nevertheless, our findings lay the foundation for a new approach to parameter estimation and provide a tool for planning pandemic control strategy.

Data availability

The data used in this study was taken from publicly available databases: Oxford COVID-19 Government Response Tracker (OxCGRT) https://github.com/OxCGRT/covid-policy-tracker. JHU CSSE COVID-19 data repository https://github.com/CSSEGISandData/COVID-19. Definite data and MATLAB code were made available in a GitHub repository https://github.com/AgataWilk/COVID19_ImpactOfPolicies.git.

Received: 17 February 2022; Accepted: 22 September 2022
Published online: 10 October 2022

References

1. Dong, E., Du, H. & Gardner, L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis.* **20**, 533–534. https://doi.org/10.1016/S1473-3099(20)30120-1 (2020).

2. World Health Organization. WHO Director-General's opening remarks at the media briefing on COVID-19. https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19-11-march-2020 (2020).
3. Hale, T. *et al.* A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). *Nat. Hum. Behav.* **5**, 529–538. https://doi.org/10.1038/s41562-021-01079-8 (2021).
4. Silva, C. J. *et al.* Optimal control of the COVID-19 pandemic: Controlled sanitary deconfinement in Portugal. *Sci. Rep.* **11**, 3451. https://doi.org/10.1038/s41598-021-83075-6 (2021).
5. Grimm, V., Mengel, F. & Schmidt, M. Extensions of the SEIR model for the analysis of tailored social distancing and tracing approaches to cope with COVID-19. *Sci. Rep.* **11**, 4214. https://doi.org/10.1038/s41598-021-83540-2 (2021).
6. Chen, S., Li, Q., Gao, S., Kang, Y. & Shi, X. State-specific projection of COVID-19 infection in the United States and evaluation of three major control measures. *Sci. Rep.* **10**, 22429. https://doi.org/10.1038/s41598-020-80044-3 (2020).
7. Kennedy, D. M., Zambrano, G. J., Wang, Y. & Neto, O. P. Modeling the effects of intervention strategies on COVID-19 transmission dynamics. *J. Clin. Virol.* **128**, 104440. https://doi.org/10.1016/j.jcv.2020.104440 (2020).
8. Linka, K., Peirlinck, M., Costabal, F. S. & Kuhl, E. Outbreak dynamics of COVID-19 in Europe and the effect of travel restrictions. *Comput. Methods Biomech. Biomed. Eng.* **23**, 710–717. https://doi.org/10.1080/10255842.2020.1759560 (2020)
9. Köhler, J. *et al.* Robust and optimal predictive control of the COVID-19 outbreak. *Annu. Rev. Control.* **51**, 525–539. https://doi.org/10.1016/j.arcontrol.2020.11.002 (2021).
10. Castex, G., Dechter, E. & Lorca, M. COVID-19: The impact of social distancing policies, cross-country analysis. *Econ. Disasters Clim. Change* **5**, 135–159. https://doi.org/10.1007/s41885-020-00076-x (2021).
11. Jorge, D. C. *et al.* Assessing the nationwide impact of COVID-19 mitigation policies on the transmission rate of SARS-CoV-2 in Brazil. *Epidemics* **35**, 100465. https://doi.org/10.1016/j.epidem.2021.100465 (2021).
12. Panovska-Griffiths, J. *et al.* Modelling the potential impact of mask use in schools and society on COVID-19 control in the UK. *Sci. Rep.* **11**, 8747. https://doi.org/10.1038/s41598-021-88075-0 (2021).
13. Mukherjee, U. K. *et al.* Evaluation of reopening strategies for educational institutions during COVID-19 through agent based simulation. *Sci. Rep.* **11**, 6264. https://doi.org/10.1038/s41598-021-84192-y (2021).
14. Bouchnita, A. & Jebrane, A. A multi-scale model quantifies the impact of limited movement of the population and mandatory wearing of face masks in containing the COVID-19 epidemic in Morocco. *Math. Model. Nat. Phenom.*https://doi.org/10.1051/mmnp/2020016 *(2020).*
15. Abdollahi, E., Haworth-Brockman, M., Keynan, Y., Langley, J. M. & Moghadas, S. M. Simulating the effect of school closure during COVID-19 outbreaks in Ontario, Canada. *BMC Med.* **18**, 230. https://doi.org/10.1186/s12916-020-01705-8 (2020).
16. Chinazzi, M. *et al.* The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science* **368**, 395–400. https://doi.org/10.1126/science.aba9757 (2020).
17. Wang, S. & Ramkrishna, D. A model to rate strategies for managing disease due to COVID-19 infection. *Sci. Rep.* **10**, 22435. https://doi.org/10.1038/s41598-020-79817-7 (2020).
18. De-Leon, H. & Pederiva, F. Particle modeling of the spreading of coronavirus disease (COVID-19). *Phys. Fluids* **32**, 087113. https://doi.org/10.1063/5.0020565 (2020).
19. Auger, K. A. *et al.* Association between statewide school closure and COVID-19 incidence and mortality in the US. *JAMA* **324**, 859–870. https://doi.org/10.1001/jama.2020.14348 (2020).
20. Babino, A. & Magnasco, M. O. Masks and distancing during COVID-19: A causal framework for imputing value to public-health interventions. *Sci. Rep.* **11**, 5183. https://doi.org/10.1038/s41598-021-84679-8 (2021).
21. Viner, R. M. *et al.* School closure and management practices during coronavirus outbreaks including COVID-19: A rapid systematic review. *Lancet Child Adolesc. Health* **4**, 397–404. https://doi.org/10.1016/S2352-4642(20)30095-X (2020).
22. Hethcote, H. W. The mathematics of infectious diseases. *SIAM Rev.* **42**, 599–653. https://doi.org/10.1137/S0036144500371907 (2000).
23. Jacquez, J. A. & Greif, P. Numerical parameter identifiability and estimability: Integrating identifiability, estimability, and optimal sampling design. *Math. Biosci.* **77**, 201–227. https://doi.org/10.1016/0025-5564(85)90098-7 (1985).
24. Fujarewicz, K. & Galuszka, A. Generalized backpropagation through time for continuous time neural networks and discrete time measurements. In *International Conference on Artificial Intelligence and Soft Computing* 190–196 https://doi.org/10.1007/978-3-540-24844-6_24 (Springer, 2004).
25. Fujarewicz, K., Kimmel, M. & Swierniak, A. On fitting of mathematical models of cell signaling pathways using adjoint systems. *Math. Biosci. Eng.* **2**, 527 https://doi.org/10.3934/mbe.2005.2.527(2005).
26. Fujarewicz, K., Kimmel, M., Lipniacki, T. & Swierniak, A. Adjoint systems for models of cell signaling pathways and their application to parameter fitting. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **4**, 322–335 https://doi.org/10.1109/tcbb.2007.1016 (2007).
27. Fujarewicz, K. *Application of Certain Methods of Neural Networks in Control and Bioinformatics* (Silesian University of Technology, 2010) (**In Polish**).
28. Łakomiec, K. & Fujarewicz, K. Parameter estimation of non-linear models using adjoint sensitivity analysis. In *Advanced Approaches to Intelligent Information and Database Systems* 59–68 https://doi.org/10.1007/978-3-319-05503-9_6 (Springer, 2014).
29. Fujarewicz, K. & Łakomiec, K. Adjoint sensitivity analysis of a tumor growth model and its application to spatiotemporal radiotherapy optimization. *Math. Biosci. Eng.* **13**, 1131–1142 https://doi.org/10.3934/mbe.2016034 (2016).
30. Fujarewicz, K. & Łakomiec, K. Spatiotemporal sensitivity of systems modeled by cellular automata. *Math. Methods Appl. Sci.* **41**, 8897–8905. https://doi.org/10.1002/mma.5358 (2018).
31. Łakomiec, K., Wilk, A., Psiuk-Maksymowicz & K., Fujarewicz, K. Finding the Time-Dependent Virus Transmission Intensity via Gradient Method and Adjoint Sensitivity Analysis. In: Pietka, E., Badura, P., Kawa, J., Wieclawek, W. (eds) Information Technology in Biomedicine. ITIB 2022. Advances in Intelligent Systems and Computing, vol 1429. Springer, Cham. https://doi.org/10.1007/978-3-031-09135-3_41 (2022)
32. Findeisen, W. *et al. Control and Coordination in Hierarchical Systems* (Wiley, 1980).
33. McAloon, C. *et al.* Incubation period of COVID-19: A rapid systematic review and meta-analysis of observational research. *BMJ Open*https://doi.org/10.1136/bmjopen-2020-039652 *(2020).*
34. Zhao, S. *et al.* Estimating the generation interval and inferring the latent period of COVID-19 from the contact tracing data. *Epidemics* **36**, 100482. https://doi.org/10.1016/j.epidem.2021.100482 (2021).
35. Lauer, S. A. *et al.* The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application. *Ann. Intern. Med.* **172**, 577–582. https://doi.org/10.7326/M20-0504 (2020).
36. Quesada, J. *et al.* Incubation period of COVID-19: A systematic review and meta-analysis. *Rev. Clín. Esp. (English Edition)* **221**, 109–117. https://doi.org/10.1016/j.rceng.2020.08.002 (2021).

Acknowledgements

This work was supported by Polish National Science Centre, Grant Number: UMO-2020/37/B/ST6/01959 and Silesian University of Technology statutory research funds. Calculations were performed on the Ziemowit computer cluster in the Laboratory of Bioinformatics and Computational Biology, created in the EU Innovative Economy Programme POIG.02.01.00-00-166/08 and expanded in the POIG.02.03.01-00-040/13 project.

Data analysis was partially carried out using the Biotest Platform developed within Project n. PBS3/B3/32/2015 financed by the Polish National Centre of Research and Development (NCBiR). This work was carried out in part by the Silesian University of Technology internal research funding (A.M.W., K.Ł., K.P.M., K.F.). We would also like to thank Ron Hancock who read the manuscript and helped with the linguistic proofreading.

Author contributions

A.M.W.: data acquisition, restriction modelling, interpretation, visualization and original draft preparation. K.Ł.: SEIR model implementation, adjoint sensitivity analysis, optimization, validation and manuscript preparation. K.P.M.: investigation, interpretation and manuscript preparation. K.F.: Conceptualization, methodology, interpretation and manuscript preparation.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-21126-2.

Correspondence and requests for materials should be addressed to A.M.W. or K.F.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

International Journal of
Molecular Sciences

MDPI

Article

Classification of Thyroid Tumors Based on Mass Spectrometry Imaging of Tissue Microarrays; a Single-Pixel Approach

Agata Kurczyk [1,†], Marta Gawin [1,†], Mykola Chekan [1], Agata Wilk [2], Krzysztof Łakomiec [2], Grzegorz Mrukwa [2], Katarzyna Frątczak [2], Joanna Polanska [2], Krzysztof Fujarewicz [2], Monika Pietrowska [1] and Piotr Widlak [1,*]

1 Maria Skłodowska-Curie National Research Institute of Oncology, Gliwice Branch, 44-102 Gliwice, Poland; agata.kurczyk@io.gliwice.pl (A.K.); marta.gawin@io.gliwice.pl (M.G.); mykola.chekan@io.gliwice.pl (M.C.); monika.pietrowska@io.gliwice.pl (M.P.)

2 Faculty of Automatic Control, Electronics and Computer Science, Silesian University of Technology, 44-100 Gliwice, Poland; agatamkwilk@gmail.com (A.W.); krzysztof.lakomiec@polsl.pl (K.Ł.); grzegorz.mrukwa@polsl.pl (G.M.); katarzyna.fratczak@polsl.pl (K.F.); joanna.polanska@polsl.pl (J.P.); krzysztof.fujarewicz@polsl.pl (K.F.)

* Correspondence: piotr.widlak@io.gliwice.pl

† These authors contributed equally to this work.

Received: 14 July 2020; Accepted: 28 August 2020; Published: 31 August 2020

Abstract: The primary diagnosis of thyroid tumors based on histopathological patterns can be ambiguous in some cases, so proper classification of thyroid diseases might be improved if molecular biomarkers support cytological and histological assessment. In this work, tissue microarrays representative for major types of thyroid malignancies—papillary thyroid cancer (classical and follicular variant), follicular thyroid cancer, anaplastic thyroid cancer, and medullary thyroid cancer—and benign thyroid follicular adenoma and normal thyroid were analyzed by mass spectrometry imaging (MSI), and then different computation approaches were implemented to test the suitability of the registered profiles of tryptic peptides for tumor classification. Molecular similarity among all seven types of thyroid specimens was estimated, and multicomponent classifiers were built for sample classification using individual MSI spectra that corresponded to small clusters of cells. Moreover, MSI components showing the most significant differences in abundance between the compared types of tissues detected and their putative identity were established by annotation with fragments of proteins identified by liquid chromatography-tandem mass spectrometry in corresponding tissue lysates. In general, high accuracy of sample classification was associated with low inter-tissue similarity index and a high number of components with significant differences in abundance between the tissues. Particularly, high molecular similarity was noted between three types of tumors with follicular morphology (adenoma, follicular cancer, and follicular variant of papillary cancer), whose differentiation represented the major classification problem in our dataset. However, low level of the intra-tissue heterogeneity increased the accuracy of classification despite high inter-tissue similarity (which was exemplified by normal thyroid and benign adenoma). We compared classifiers based on all detected MSI components (n = 1536) and the subset of the most abundant components (n = 147). Despite relatively higher contribution of components with significantly different abundance and lower overall inter-tissue similarity in the latter case, the precision of classification was generally higher using all MSI components. Moreover, the classification model based on individual spectra (a single-pixel approach) outperformed the model based on mean spectra of tissue cores. Our result confirmed the high feasibility of MSI-based approaches to multi-class detection of cancer types and proved the good performance of sample classification based on individual spectra (molecular image pixels) that overcame problems related to small amounts of heterogeneous material, which limit the applicability of classical proteomics.

Keywords: thyroid cancer; molecular classifiers; molecular imaging; bioinformatics; proteomics; mass spectrometry imaging; biomarkers

1. Introduction

Though thyroid nodules are very common in the overall population, malignant tumors occur in less than 1% of such nodules. Nevertheless, thyroid cancer is the most common endocrine cancer and contributes to 1–2% of all new malignancies diagnosed each year worldwide. The majority of thyroid carcinomas originate from follicular epithelial cells and include well-differentiated papillary thyroid carcinomas (PTC; up to 80% of all thyroid malignancies) and follicular thyroid carcinomas (FTC; about 15% of all thyroid malignancies). Moreover, undifferentiated anaplastic thyroid carcinoma (ATC; 1–2% of all thyroid cancers), which is the most aggressive thyroid malignancy, also develop from epithelial cells. Additionally, medullary thyroid carcinoma (MTC) is derived from the neuroendocrine parafollicular C-cell comprises 3–5% of all thyroid cancers [1,2]. Currently, most of the thyroid tumors are diagnosed by pathomorphological assessment alone, and such classification is the primary step in the assessment of prognosis and selection of the treatment [3]. The primary diagnosis in the majority of patients with thyroid cancers is based on fine-needle aspiration cytology (FNAC) of thyroid nodules; then, further diagnosis is performed based on histopathological intra- or post-operative examination of the resected thyroid tissue [4,5]. However, in some cases, cytological and histological patterns are ambiguous, and proper classification is problematic [6]. Therefore, the classification of thyroid tumors, particularly those exhibiting unusual morphological patterns, might be improved if additional molecular tests could be applied.

It has been generally accepted that new biomarkers identified with the use of high-throughput "omics" approaches could markedly support the classification of thyroid cancers based on histopathological patterns [7–9]. However, one should be aware that the initial diagnosis of thyroid tumors is based on FNAC; hence, hypothetical biomarkers have to be compatible with cytological material collected by fine needle biopsy. Therefore, the major obstruction associated with such a biospecimen is a low amount of material (usually up to a few hundred cells) and its high heterogeneity (usually a mixture of tumor cells and other types of cells). It has been proposed that challenges associated with the molecular classification of thyroid cancer using cytological material could be approached by mass spectrometry imaging (MSI). It is an emerging technology in biomedical research and has been recently evolving into a powerful tool in the study of various types of diseases. The major benefit of MSI is the possibility of combining molecular and morphological information. Molecular images generated by mass spectrometry are spatially resolved and correlated with the respective histological images; hence, molecular profiles could be allocated to specific cells or tissue regions. Moreover, different molecular species, e.g., proteins, peptides, lipids, drugs, and their metabolites, can be imaged, significantly broadening the amount of information derived from a tissue [10–13]. Among many applications of MSI, there was molecular characterization and classification of different types of solid tumors, including thyroid cancers [14–16]. It might be expected that the limitations of fine needle biopsy–derived material could be overcome by MSI due to its ability to provide information collected from a low number of specific cells present in cytological samples. The idea that MSI is capable of distinguishing between different papillary tumors using proteomic signatures of cytological FNA samples has been validated experimentally in a series of reports from Fabio Pagni and Fulvio Magni laboratories who used a Matrix-Assisted Laser Desorption/Ionization Time-of-Flight (MALDI-TOF) MSI approach [17–19]. Moreover, metabolome profiling by Desorption Electrospray Ionization (DESI) MSI was also recently tested for the diagnosis of thyroid tumors based on FNA samples [20].

Molecular classification of cancer that is based on MSI analysis of cytological and histological samples has several challenges related to the implemented methodology both at (pre)analytical [21,22] and data processing [23,24] steps. The latter step involves the necessity to solve several problems related

to the processing of heterogeneous samples and optimization of feature selection for cancer classifiers. Here, we implemented several biostatistical approaches to test the suitability of spectral data generated by MALDI-TOF MSI for multi-class comparison and classification of different thyroid tumors. Tissue microarrays were used as a surrogate of cytological smears assuming a comparable number of spectra registered by MSI in both types of biospecimens. The classification of thyroid tumors using tissue microarrays analyzed by MALDI-TOF MSI was previously reported by Galli et al. [25], who compared material from benign tumors (follicular adenoma and hyperplastic lesions) and papillary thyroid cancer (both classical and follicular variant). However, mean spectra for tissue cores were used in that case. In the present work, the assessment of similarities and differences between seven types of thyroid specimens was based on features of individual spectra (i.e., pixels of molecular images), which seems to be a more relevant approach to potential applications in molecular diagnostics of cancer.

2. Results

Tissue specimens representative for seven types of thyroid tissue (ROIs) were analyzed by MALDI-MSI. There were about 400 tissue cores analyzed derived from 134 patients with either follicular adenoma (FA), classical or follicular variant of papillary thyroid carcinoma (PTC-CV and PTC-FV, respectively), follicular thyroid carcinoma (FTC), anaplastic thyroid carcinoma (ATC), or medullary thyroid carcinoma (MTC); normal thyroid (NT) samples were collected from tissue region showing normal thyroid histology. About 135,000 spectra were registered (360 spectra per one tissue core, on average); the general characteristics of the analyzed biomaterial are presented in Supplementary Table S1. Mean spectra obtained for each type of ROI are presented in Supplementary Figure S1. Gaussian mixture modeling was used for the identification of spectral components that resulted in 1536 components that corresponded to tryptic peptides; the averaged spectrum used for the model construction is depicted in Figure 1. Importantly, this approach provided a substantial reduction of data dimensionality, which reduced the computation load for the processing of multiple spectra.

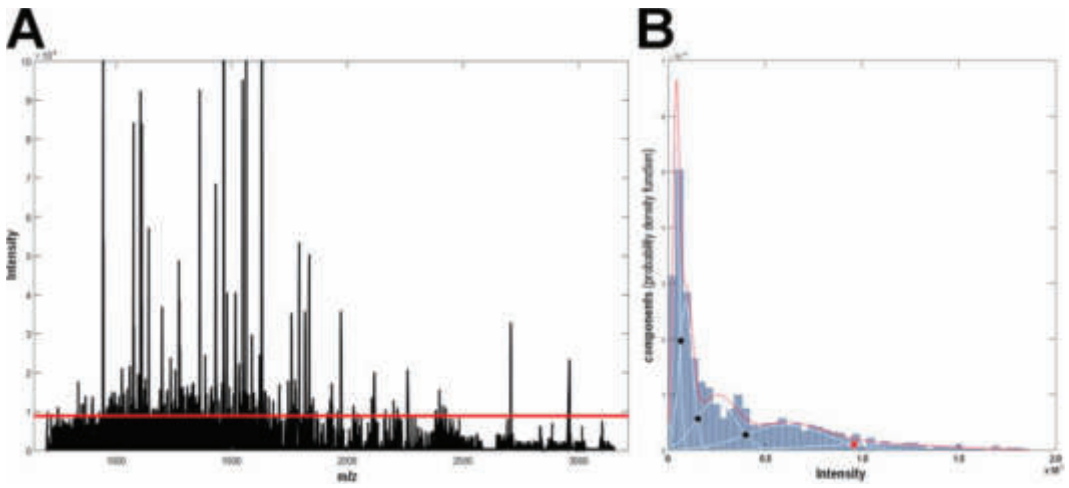


Figure 1. Mass spectrometry imaging (MSI) spectrum. (**A**) The average mass spectrum. (**B**) Distribution of m/z components with different average intensity; the red dotted curve represents an intensity distribution model with its constituents represented by blue lines (dots represent thresholds delineating intensity-related classes of components). The red line in Panel A represents the threshold separating 147 most abundant components (>9610 a.u.).

In the first step, general similarities between spectra derived from different types of ROIs were estimated. The principal component analysis revealed a clear separation of spectra representative for normal thyroid and spectra representative for ATC, MTC, and the majority of well-differentiated thyroid cancers (WDTC, including FTC and both variants of PTC); the most differentiating for normal and cancer spectra was PC2 responsible for 16.3% of the general variability (Figure 2) (PC1 was putatively

associated with inter-individual differences present within each patients' subset). However, spectra representative for FA overlapped with both normal thyroid and cancer ROIs. Spectra representative for WDTC, ATC, and MTC could not be separated reasonably along with any principal component (Figure 2).

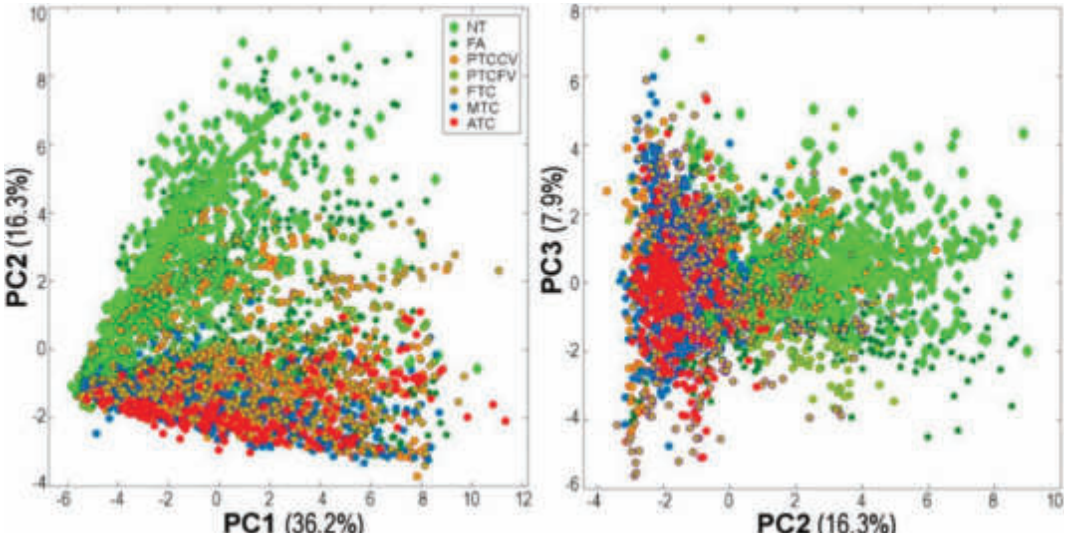


Figure 2. Principal component analysis (PCA) of spectra representative for all regions of interest (ROIs). Ten spectra were randomly selected from each tissue core for clarity; here, three principal components together describe 60% of the variability.

Next, the similarity index between pairwise compared spectra from seven types of ROIs was estimated based on all registered molecular components (i.e., 1536 m/z components) to address tissue heterogeneity; the resulting cumulative distribution functions (CDFs) of similarity are depicted in Figure 3A–D. To assess intra-tissue heterogeneity, similarities between spectra from the same type of ROI were compared (Figure 3A). We found the highest homogeneity for normal thyroid tissue (median similarity of 0.93) and the highest heterogeneity for ATC (median similarity of 0.77); even higher heterogeneity (i.e., lower intra-ROI similarity) was observed if spectra from all cancer ROIs (or all WDTC) were combined (Supplementary Table S2). To assess inter-tissue heterogeneity, similarities between spectra from different ROIs were compared (Supplementary Table S3). When normal thyroid was compared with different thyroid tumors, the highest similarity was observed between NT and benign FA (median similarity of 0.85) and the lowest similarity was observed between NT and the undifferentiated ATC (median similarity of 0.53), while similarities between NT and differentiated cancers were flanked by these two extremes (Figure 3B). When benign FA was compared with malignant tumors, the highest similarity was observed between FA and follicular variant of PTC and FTC (median similarity of 0.79 and 0.75, respectively), while the lowest similarity was observed between FA and ATC (median similarity of 0.66) (Figure 3C). Median similarities between different types of malignant tumors were in the range between 0.79 and 0.69. Similarities between three major types of thyroid cancer were comparable—the median similarities between WDTC vs. ATC, WDTC vs. MTC, and ATC vs. MTC were 0.74, 0.77, and 0.79, respectively (Figure 3D).

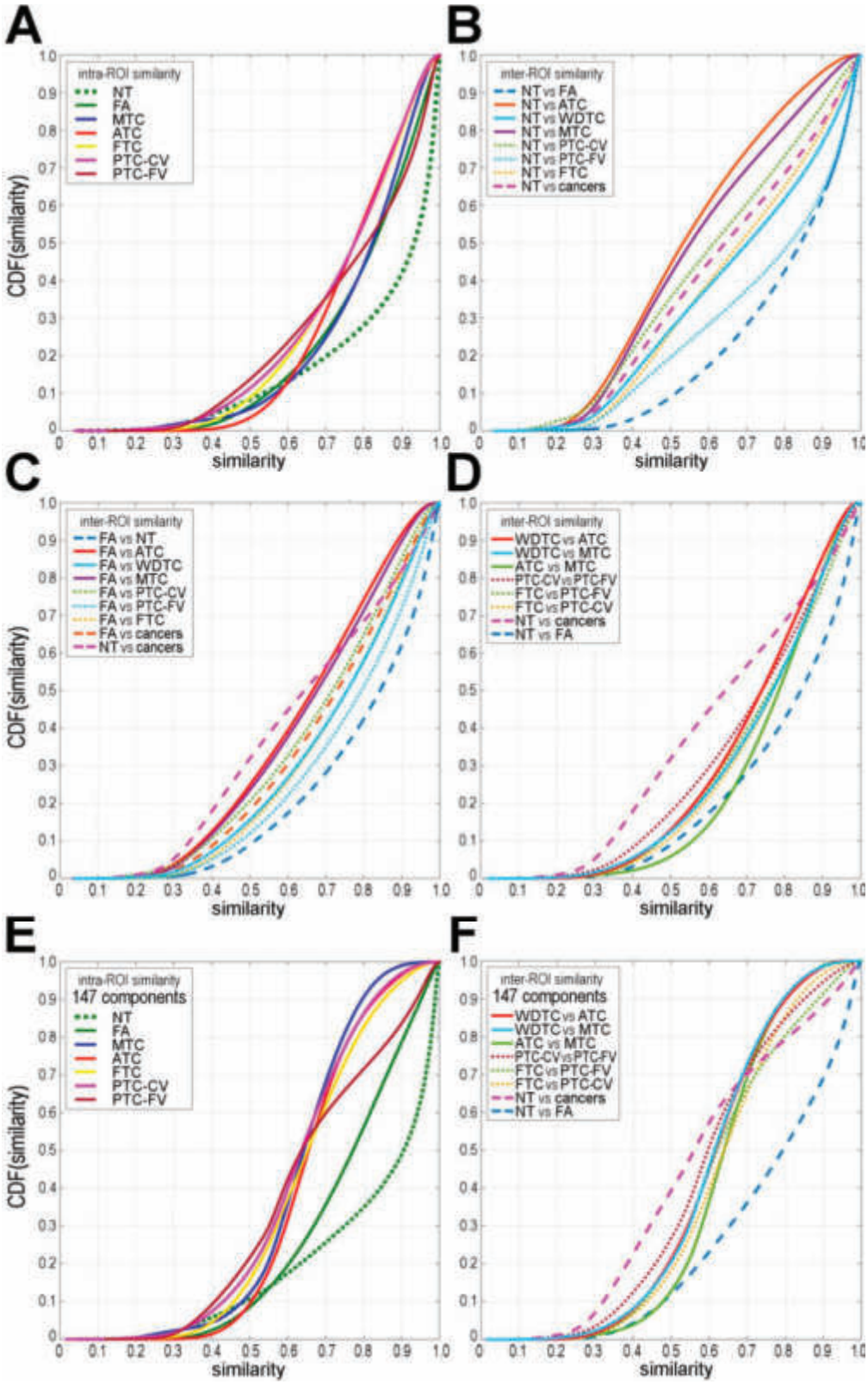


Figure 3. Molecular similarity among different types of thyroid tissue. Depicted are the cumulative distribution functions of similarity (cumulative distribution function (CDF(similarity)), either intra-ROI (panel **A**) or inter-ROI (panels **B**–**D**) when all 1536 components were considered. Intra-ROI (panel **E**) and inter-cancer ROI (panel **F**) similarity was based on 147 most abundant components. A similarity that corresponded to CDF(similarity) = 0.5 represents its median value.

A similar analysis was performed taking into account 147 most abundant components (Figure 1). In this case, high intra-ROI similarity within NT and FA remained (median similarity of 0.91 and 0.78), yet intra-cancers similarities were markedly reduced (median similarities in the range between 0.65 and 0.62; Figure 3E, Supplementary Table S2). Furthermore, although inter-ROI similarity between NT and FA remained high (median similarity of 0.80), similarities among other types of ROI were markedly reduced when the 147 most abundant components were considered (Supplementary Table S3). Noteworthy, inter-ROI similarities between different types of cancer were comparable (median similarity in the range between 0.62 and 0.58; Figure 3F). These observations suggested collectively that components with a lower intensity markedly contributed to similarities among spectra, while a significant differential potential could be attributed to highly abundant components (with the particular exception of the similarity index between NT and FA).

In the next step, the classification of tissue cores was performed based on the classification of individual spectra that we called "a single-pixel approach." First, binary classification models were tested for comparison of spectra from all possible pairs of seven ROIs (i.e., 21 classifiers were tested); then, individual spectra in a tissue core were analyzed using these multiple models. Classifier features involved either the complete set of components (n = 1536) or the subset of the most abundant components (n = 147). In the former case, the accuracy of binary classifiers was in the range between 0.94 (NT vs. ATC) and 0.63 (FA vs. FTC) with the average 0.84; in the latter case, binary classifiers performed comparably (average accuracy 0.81) (Supplementary Table S4). Subsequently, all individual spectra in all tissue cores were analyzed using all 21 binary classifiers, and then an individual spectrum was considered "classified" only if all 6 binary classifiers including a given ROI showed full concordance. Otherwise, a spectrum was considered "not classified", which provided a high level of confidence in the individual spectrum classification. In general, 7% and 10% of spectra remained not classified when classifiers based 1536 and 147 components were applied, respectively (Supplementary Table S4). The largest and the lowest number of not classified spectra was observed for ATC and NT cores (13% and 3%, respectively, on average), which mirrored the level of intra-ROI heterogeneity established earlier using the similarity index. Finally, tissue cores were classified based on the predicted identity of individual spectra: a core was classified according to the most frequent class in corresponding spectra (including the "not classified" type). In general, the precision of such core classification was higher for spectra classifiers based on all 1536 components than for classifiers based on the most abundant components (67% and 55% of cores were classified properly, respectively) despite overall similar accuracy of binary classifiers. Nevertheless, in both cases, the overall correctness was much higher than approx. 14% expected from random indexing of seven classes. The complete table of classification results is presented in Figure 4A,B. The lowest accuracy of classification was observed for FTC that was frequently confused with FA and PTC-FV (the sensitivity of FTC classification was 0.38 and 0.26 for classifiers based on 1536 and 147 components, respectively). Moreover, low sensitivity (0.33) was observed in the classification of ATC using classifiers based on 147 components. Furthermore, the number of not classified (i.e., "not diagnostic") cores was higher when the classification was based on the most abundant components (16 vs. 11 for classification based on 147 and 1536 components, respectively). Hence, we concluded that the classification of tissue cores performed better using models based on all registered MSI components. Furthermore, we also tested two other classification strategies based on mean spectra computed for separate tissue cores. First, binary classification models were tested using mean spectra, and then cores were classified based on their mean spectra ("a mean spectrum approach"). In this case, binary classifiers performed comparably to binary classifiers based on individual spectra (average accuracy 0.82 for models based on 1536 MSI components; Supplementary Table S4). However, the performance of a final core classification was much lower than that of a single-pixel approach—only 52% of cores were classified correctly (Figure 4C). Second, binary classification models tested using individual spectra were applied for classification of cores based on their mean spectra ("a hybrid approach"). In this case, the overall precision of a final core classification was only slightly lower compared to a single-pixel approach: 66% of cores were classified

correctly (Figure 4D). Notably, cores corresponding to the most homogenous normal thyroid showed even higher precision of classification using this hybrid approach (63/70 vs. 61/70 cores predicted properly). Nevertheless, the number of "not diagnostic" cores was higher when a hybrid approach was implemented (2.9%, 3.7%, and 5.9% of not classified cores using a single-pixel, a hybrid, and a mean spectrum approach, respectively). Hence, we concluded that a single-pixel approach performed best for the classification of analyzed tissue cores.

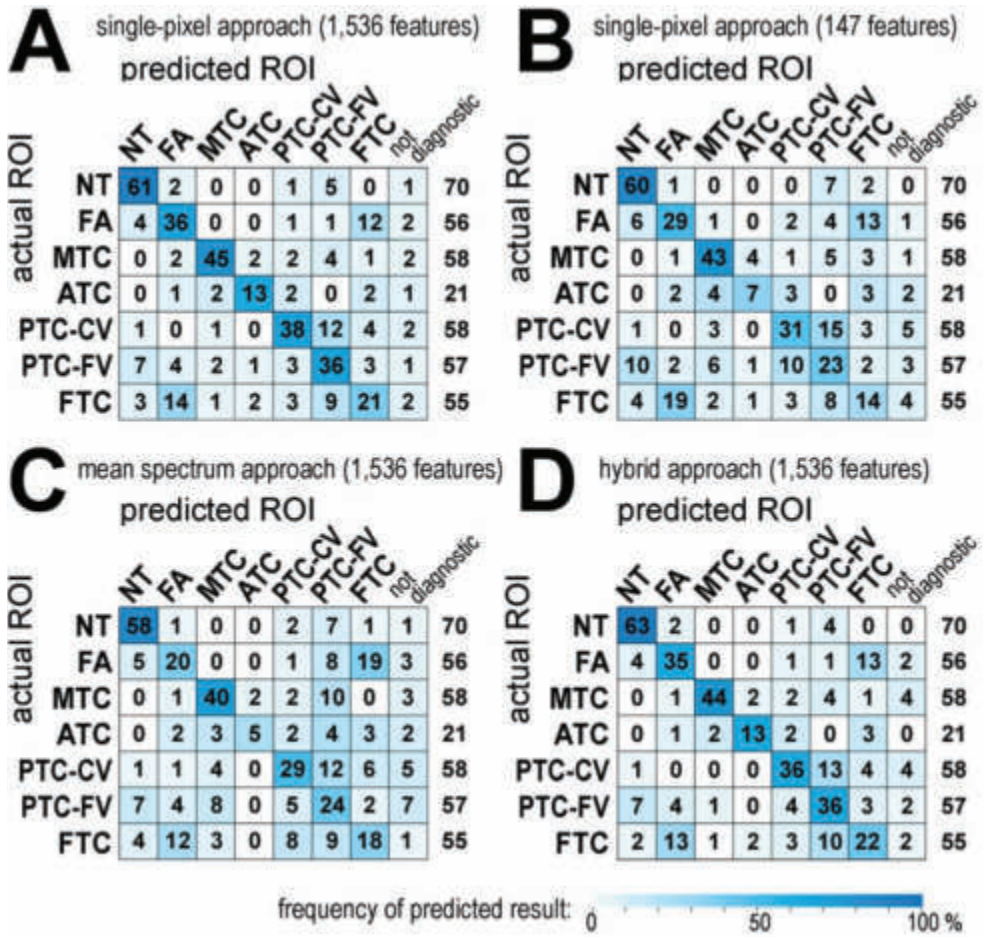


Figure 4. Classification of normal and cancerous thyroid tissue cores based on MSI-registered spectra. Results of a core class prediction are presented as actual numbers (a relative frequency of specific results is color-coded according to the color scale); the last column represents the total number of cores. Presented are: a single-pixel approach for 1536 features (panel **A**) and 147 features (panel **B**), a mean spectrum approach for 1536 features (panel **C**), and a hybrid approach for 1,536 features (panel **D**).

In the last step, molecular components with markedly different abundance between types of thyroid specimens were detected (all spectra from a given ROI were combined for this analysis). Considering the structure of data, the strength of differences was estimated by the effect size factor; Cohen's *d* (absolute) values above 0.5, 0.8, and 1.2 corresponded to medium, large, and very large effects, respectively [26]. The number of components that discriminated pairwise against different ROIs with assumed effect sizes is illustrated in Figure 5 (see details in Supplementary Table S5). First, discriminatory components were detected in the complete set of 1536 molecular components. A substantial number of components with significantly different abundances were detected between normal thyroid and thyroid tumors. The highest number of discriminatory components was observed between NT and ATC (45% of all components showed large or very large effect size). There were fewer discriminatory components between NT and MTC (23% of components showed large or very large effect size), while the number of components markedly discriminating NT from WDTC was

relatively low (about 6% of the registered components showed large effect size). Surprisingly, despite a very high level of general similarity between NT and FA, there was a large number of components whose abundances were markedly different between these two types of ROI, which putatively reflected high homogeneity (i.e., low variability) inside both tissues. On the other hand, there were much fewer components whose abundance was markedly different between benign FA and malignant cancers; notably, components discriminating between FA and WDTC showed only medium effect size. When three major types of thyroid cancer were compared (WDTC, ATC, and MTC), a higher number of components showing large/very large effect was noted between undifferentiated ATC and differentiated cancers (WDTC and MTC). Unexpectedly, we found that the number of discriminatory components between FTC and PTC-CV was lower than the number of discriminatory components between FTC and PTC-FV or between both variants of PTC (Figure 5A). Mass distribution of components discriminating selected types of ROI is depicted in Figure 5B. Moreover, the detection of discriminatory components was performed using 147 most abundant m/z components (Figure 5C). The relative proportion of differentiating components (medium, large, and very large effect size) was similar or even higher in this subset of very abundant components when compared to the whole set of 1536 components (with a marked exception of the reduced contribution of significant differences between NT and FA).

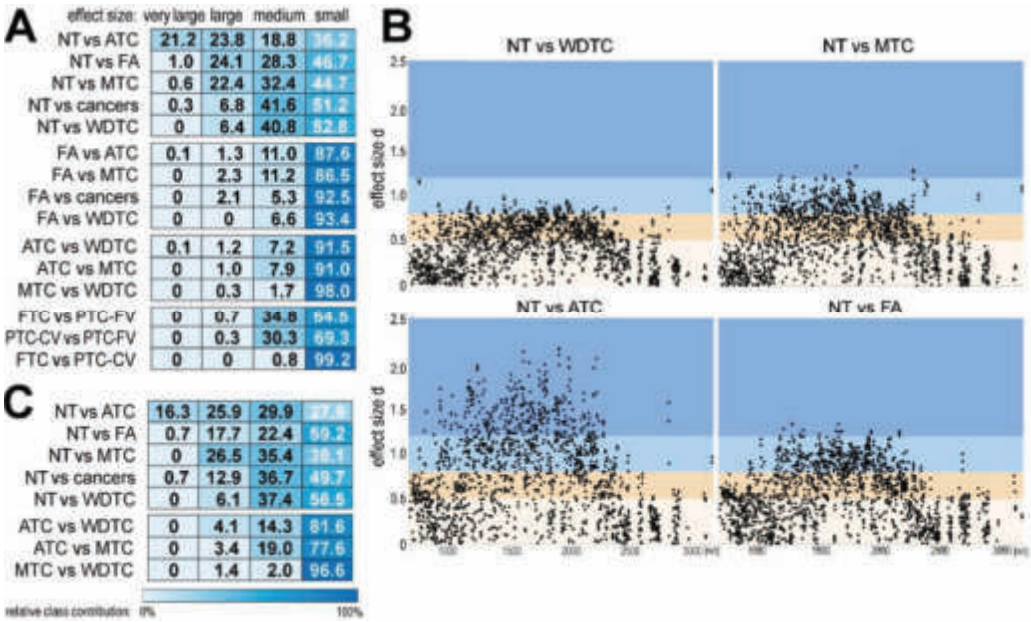


Figure 5. Molecular components that discriminated different types of thyroid tissue. (Panel **A**)—relative contribution (percentage) of components that differentiated pairwise selected ROIs with very large, large, medium, or small effect sizes (all detected components analyzed; n = 1536). (Panel **B**)—distribution of discriminatory components in the m/z axes in four selected pairwise ROI comparisons; very large (>1.2), large (>0.8), and medium (>0.5) effect sized are color-coded (dark blue, light blue, and beige, respectively); each dot corresponds to one spectral component. (Panel **C**)—relative contribution (percentage) of the most abundant components (n = 147) that differentiated pairwise selected ROIs with very large, large, medium, or small effect sizes.

The hypothetical identity of MSI components was established by attributing masses (m/z values) of imaged molecular components (i.e., tryptic peptides) to measured masses of peptides identified by liquid chromatography-tandem mass spectrometry (LC-MS/MS) in tissue lysates. However, though hypothetical identity was attributed to the majority of molecular components detected by MSI, one should be aware that this type of annotation is not unique and more than one identified peptide could be matched to an MSI component due to relatively low resolution of MALDI-TOF MSI (Supplementary Tables S6–S8). Nevertheless, proteins whose tryptic fragments were most frequently

attributed to MSI components markedly upregulated in all types of tumor ROIs (compared to normal thyroid) included thyroglobulin, heat shock proteins, and proteins associated with several gene ontology (GO) terms involved in cancer development: response to growth factors, cytoskeleton organization, extracellular matrix organization, cell-cell adhesion, cell motility, glycolysis and glucose metabolism, regulation of immune response, and inflammatory response (Supplementary Table S9). Moreover, proteins whose tryptic fragments were attributed to MSI components specifically upregulated in MTCs included those involved in their neuroendocrine functions (calcitonin and chromogranin A), and other proteins showed previously to be specifically upregulated in MTC (e.g., APOE, CEACAM5, TTR) [27]. Furthermore, proteins whose tryptic fragments were attributed to MSI components specifically upregulated in ATC included those associated with GO terms involved in cancer progression: regulation of cell migration, regulation of angiogenesis, cytoskeleton organization, regulation of cell death, cell-cell adhesion, and regulation of immune and inflammatory response (Supplementary Table S9). On the other hand, very few components specifically distinguished benign adenoma: when compared to normal thyroid all components upregulated in FA were also upregulated in malignant cancers, while only two components (m/z 1183.6 and 1184.6) markedly upregulated in cancer (large or very large effect size) were not upregulated in FA (small effect size) (components putatively corresponded to a fragment of ribonucleoprotein HNRNPUL2). Furthermore, no components showed marked differences (large or very large effect size) between FA and well-differentiated cancers.

3. Discussion

Classification of thyroid tumors based on pathomorphological features is the primary step in the assessment of prognosis and selection of the treatment. The majority of patients are diagnosed based on the fine needle aspiration cytology (FNAC) of thyroid nodules with further validation and verification based on histopathological examination (either intra- or post-operative) of the resected tissue. Unfortunately, in some cases, cytological and histological patterns are ambiguous, and proper classification is problematic [3–6,28,29]. For example, thyroid tumors with follicular growth pattern include a broad range of lesions that are difficult to distinguish by cytology and could be challenging even in histologic specimens. These lesions include hyperplastic nodules, benign follicular thyroid adenomas (FAs), follicular carcinomas (FTCs), and follicular variant of papillary thyroid carcinomas (PTC-FVs). On the other hand, the most common PTC shares some cytological features with benign lesions (nodular hyperplasia, FTA, Hashimoto's thyroiditis) as well as other malignant lesions (e.g., Hürthle cell carcinoma, FTC, and MTC) [29–31]. Hence, supporting the assessment of morphological patterns with molecular biomarkers would be recommended to improve the classification of thyroid tumors [7,9]. Assuming that the primary diagnosis of thyroid disease is usually based on the fine needle biopsy of thyroid nodules, potential molecular tests would be preferably performed using this type of specimen. However, a small amount of heterogeneous material present in such a biopsy represents a diagnostic challenge for "classical" methods of genomics and proteomics typically based on tissue lysates. A few reports proved already the concept that mass spectrometry imaging could be implemented in the molecular classification of thyroid cancer to overcome these limitations of cytological material [17–22]. Importantly, individual spectra/pixels registered by MSI could be used for sample classification. However, this requires specific approaches to the optimization of data processing [23,24].

Here, tissue microarrays representative for five major types of thyroid malignancies (medullary cancer, undifferentiated anaplastic cancer, and three types of well-differentiated cancers); benign thyroid tumor (adenoma); and normal not cancerous thyroid tissue were analyzed by MALDI-MSI then individual "pixels" of MSI images were used for assessment of differences between tissue types and sample classification to mimic the situation typical for cytological smears. Different numerical approaches were tested to assess molecular similarity within and between types of thyroid tissue (i.e., specific ROIs). The unsupervised PCA approach revealed a large intra- and inter-ROI variability of MSI spectra. Though some separation of normal thyroid from all types of malignancies was visible, specific

types of cancers cannot be separated and benign adenoma overlapped with both normal and cancerous tissue, which defined the major classification problem. To estimate overall intra- and inter-ROIs heterogeneity more specifically, the similarity index was calculated between individual spectra based on the spectral contrast angle approach, which previously proved its high applicability in the analysis of MS data [32]. We found the lowest and the highest intra-ROI heterogeneity for normal thyroid and undifferentiated ATC, respectively, which further affected the performance of the classification of these types of tissue. As expected, normal thyroid was more similar to benign FA than to malignant cancers and more similar to well-differentiated epithelial cancers than to ATC or MTC. Importantly, FA was very similar to well-differentiated cancers with a follicular morphology (FTC and PTC-FV), whose proper discrimination represents the major diagnostic challenge also in the case of classical pathological assessment. The multi-class problem of separation of thyroid tumors was further tested using classification models. Assuming seven classes of thyroid tissue, 21 SVM-based one-versus-one binary classifiers were built to cover all possible combinations of ROIs, then each spectrum was tested with all of them (noteworthy, this strategy outperformed typical multiclass model based on the KNN approach; data not presented). In general, the accuracy of binary classifiers reflected overall similarity between specific ROIs, being high for models comparing normal thyroid with malignant cancers and low for models comparing well-differentiated cancers. An interesting exemption was a high accuracy of the model for the classification of normal thyroid vs. benign adenoma, which likely reflected a high level of intra-ROI homogeneity of both specimens. The discrimination between FA and FTC represented the largest classification problem (discrimination of FA from both variants of PTC performed better). Nevertheless, an individual spectrum was considered "classified" only if decisions were univocal, i.e., all six models featuring a particular type indicated that class; otherwise it remained not classified ("not diagnostic"), which increased credibility of the class prediction. To verify the actual reliability of such classification, the resulting classes of individual spectra were used for the tissue core classification. Importantly, 2/3 of cores were classified properly, while random indexing of seven classes would result in a proper assignment of 14% of samples only. The best performance was observed in the case of normal thyroid and MTC (about 80% of cores classified properly), while the worst performance was observed for FTC (only 38% of cores were classified properly, while 25% were confused with FA). The performance of the classification depended not only on similarities/differences between compared tissues per se but also on the level of intra-tissue heterogeneity. Therefore, undifferentiated ATC that showed very high intra-tumor heterogeneity could represent a classification problem (38% of false-negatives) despite its low overall similarity to other types of thyroid tissue. Additionally, we searched for spectral components that showed the most significant differences in abundance between pairwise compared ROIs using the effect size approach. As expected, a high contribution of differentiating components was usually reversibly associated with the overall similarity and increased accuracy of sample classification. However, a few interesting exemptions from this general rule were noted as mentioned above, which could be explained by the influence of the intra-tissue heterogeneity. Hence, the estimation of this parameter should be always recommended when a comparison between different tissue types is planned. Identification of differentially expressed MSI components based on matching their masses with masses of peptides detected by the classical LC-MS/MS in corresponding tissue lysates is only putative and should be used with caution. Nevertheless, the hypothetical identity of differentiating MSI components fitted general proteome profiles of different thyroid diseases established by classical MS-based proteomics approaches [27], which confirmed the credibility of the tumor classification approach proposed in the current work.

In our basic approach, all molecular components present in MSI data (i.e., 1536 spectral components corresponding to tryptic fragments of proteins) were used in analyses without any preselection. However, we also tested a subset of the 147 most abundant components (putatively corresponding to fragments of more abundant proteins). Interestingly, we found that the overall similarity between the compared tissues was lower when the estimation was based on the latter subset. Similarly, the relative contribution of components showing (very) large effect size of differences between tissue types was

generally higher in the subset of the most abundant ones. Hence, one could conclude that effective differentiation of tissue types could be performed using highly abundant components only which are less sensitive to analytical errors. However, we observed that molecular classifiers based on all components (i.e., including low abundance components) generally performed better than classifiers based on the most abundant components only; hence, the utilization of all registered MSI components for sample classification appeared a credible solution. Artificial intelligence-based methods for the preselection of features for molecular classifiers are frequently in use [33], which approach was validated for MSI-based classification of cytological smears from thyroid tumors [23]. However, a selection of the "best set" of components for sample classification based on the machine learning approaches could be sensitive to specific features present uniquely in a given dataset, which implies the necessity of using "big data" (i.e., very large patients cohorts or sample repositories). Still, we tested here classification models based on spectral components selected using t-test results between ROIs, yet this approach did not improve the performance of classification compared to models based on all components (data not presented). Therefore, using all registered MSI components for the classification based on individual spectra remains a valid alternative when computationally feasible. Nevertheless, using either all components or their pre-selected subset emphasizes the problem of proper spectra generation and data processing to avoid errors and artifacts potentially related to the analysis of low-signal components of MSI spectra. Moreover, we found that a single-pixel approach outperformed a classification model based on the mean spectra of tissue cores. Alternatively, a strategy where binary classification models tested with individual spectra are subsequently used for the sample classification based on its mean spectrum could be considered, yet the implementation of this hybrid approach should be limited to tissue specimens with high homogeneity.

4. Materials and Methods

4.1. Clinical Material

Postoperative tissue was collected during thyroidectomy then stored as formalin-fixed paraffin-embedded (FFPE) material. Samples derived from 134 patients treated at Maria Skłodowska-Curie National Research Institute of Oncology in Gliwice between 2012–2017. The study was approved by the appropriate local Ethics Committee (approval no. KB/430–17/13). Tissue blocks were used to generate tissue microarrays (TMAs) that included 375 individual cores (3 cores from different tissue areas per patient, on average) of 1 mm in diameter. Tissue cores represented seven types of thyroid tissue (Region of Interest, ROI): normal thyroid (NT), follicular adenoma (FA), papillary thyroid carcinoma—classical variant (PTC-CV), papillary thyroid carcinoma—follicular variant (PTC-FV), follicular thyroid carcinoma (FTC), anaplastic thyroid carcinoma (ATC), and medullary thyroid carcinoma (MTC). Normal thyroid was represented by a tissue distant from the cancer region that showed no marks of any pathology. Clinical material included in the study is characterized in the Supplementary Table S1. Tissue cores were distributed randomly into eight arrays, and molecular images were registered in a random sequence.

4.2. Sample Preparation for MALDI-MSI

Tissue microarray blocks were cut into 5 μm sections with the use of a rotary microtome (HM 340E, Thermo Fisher Scientific, Waltham, MA, USA) and placed on ITO glass slides (Bruker Daltonik, Bremen, Germany) coated with poly-L-lysine. Slides were then dried at 37 °C for 18 h. Additional thermal treatment (60 °C, 1 h) was performed directly before paraffin removal. De-waxing was carried out in xylene (2 × 5 min), followed by washing with 99.8% ethanol (5 min), 96% ethanol (5 min), and 50% ethanol (5 min). Finally, glass slides were left to dry on a bench and subjected to heat-induced antigen retrieval in 10 mM Tris-HCl pH 9.0, 95 °C, 20 min (in a water bath). The slides were cooled in the retrieval solution for 20 min at room temperature, then washed with MilliQ water for 1 min and dried on a bench for 10 min, followed by drying in a vacuum desiccator for 15 min. A solution of

sequencing grade modified trypsin (Promega, Madison, WI, USA) (20 μg/mL in 25 mM NH_4HCO_3) was sprayed over each TMA section with the use of SunCollect device (SunChrom, Friedrichsdorf, Germany). The section was then incubated in a humid chamber with MilliQ water for 18 h at 37 °C. After on-tissue digestion, the slide was dried in a vacuum desiccator for 15 min and covered with an α-cyano-4-hydroxycinnamic acid (5 mg/mL w 50% ACN, 0.3% TFA) matrix solution deposited onto TMAs with the use of SunCollect device. Methods of both trypsin and matrix coating were employed according to the reference [34].

4.3. MALDI-MSI Measurements

Spectra of tryptic peptides were acquired using MALDI-TOF mass spectrometry with the use of ultrafleXtreme mass spectrometer (Bruker Daltonik, Bremen, Germany) equipped with smartbeam II™ laser operated at 1 kHz frequency. Ions were accelerated at 25 kV with a PIE time delay of 100 ns. Spectra were recorded in reflectron positive mode within 700–3700 m/z, 400 shots per position, the random walk mode was activated (40 shots at raster spot). Each TMA core was imaged with a raster width of 50 μm (laser setting: 3_medium). After imaging, the matrix was washed off the glass slides with 70% ethanol (two washes, 1 min each), and the sections were stained with hematoxylin and eosin, then scanned and used for image co-registration (using flexImaging software; Bruker Daltonik, Bremen, Germany). Compass for flex 1.4 software package (Bruker Daltonik, Bremen, Germany) was used for spectra acquisition and handling.

4.4. Spectra Processing and Identification of Spectral Components

The MSI spectra dataset was preprocessed by performing mass channels unification, baseline subtraction [35], outlying spectra identification according to TIC (total ion current) value using criterion for skewed and heavy-tailed distributions [36], fast Fourier transform-based peak alignment to reference average spectrum [37], and TIC normalization. Gaussian mixture modeling (GMM) of the average spectrum was applied for peak detection as described in detail elsewhere [38,39]. GMM components of low amplitude and high variance were removed from initial GMM spectra representation; GMM components modeling the same spectrum peak were merged by summing their estimated abundance and setting the location of a dominant component as mass/charge value of a peptide ion. The abundance of the particular component was estimated by pairwise convolution of the GMM components and individual spectra. The resulting dataset featured 1536 components detected in mass/charge range between 700 and 3150 that represent tryptic peptide species imaged by MSI. The average intensities of these 1536 components ranged between 25 and 830,527 arbitrary units, whose distribution could be described by five Gaussian components with the following thresholds: >25, >639, >1507, >4001, and >9610 a.u.; this resulted in five classes, with 386, 328, 332, 343, and 147 (the most abundant) m/z components, respectively (see Figure 1B).

4.5. Statistical Analyses and Sample Classification

Dimensionality reduction was applied to assess the separability of a different region of interest (ROI) types. Principal component analysis (PCA) based on singular value decomposition was performed on the dataset containing all 1536 components. To account for significant differences in variable ranges without losing information on feature covariance, the data were scaled using a pseudo-logarithmic function ($\log_{10}(x + 1)$) and centered. A normalized dot-product of two mass spectra (pairwise similarity index) was calculated to assess the similarity between the compared pair of spectra [40]. Spectra were labeled according to their tissue microarray core location in one of seven ROIs creating seven spectra subsets. The similarity index was calculated in two manners: within particular ROI (intra-ROI similarity) and between different ROIs (inter-ROI similarity) creating all possible combinations of compared spectra pairs. Populations of computed similarity values were plotted as cumulative distribution functions (CDFs) to visualize similarities within and between analyzed ROIs. Sample classification involved two types of classifiers: binary classifiers used to classify individual spectra

and the main core classifier. In the first stage, the dataset was standardized using the Z-score method, then support vector machines (SVM) classifiers with a linear kernel were used with a one-versus-one strategy (21 binary classifiers were tested to cover all possible pairs of seven ROIs). An adapted k-fold method was used for model cross-validation. Each patient was assigned to one of the seven classes based on the ROI type identified in the majority of tissue cores derived from that donor. The patients were then partitioned into five stratified folds. The training and validation sets for each fold were reconstructed from this partition, preserving original tissue type of the cores. Such subsampling of the spectra, when quantitatively unbalanced (contrary to the classic k-fold), ensures independence between folds, preventing data leakage and quality estimation bias. In the second step, the output of the binary classifiers was used to classify the whole tissue core. For each spectrum, all 21 binary classifications were tested and a spectrum was considered classified if the decision was univocal, i.e., all six models featuring a particular type indicated that class (otherwise, the spectrum remained not classified). The final classifier decision for the core was the most frequent of eight classes (seven tissue types ROI or "not diagnostic" if the majority of spectra could not be conclusively classified). Basic classification quality indices were calculated using the one-versus-rest approach. An effect size analysis was applied to indicate discriminatory molecular components. Cohen's *d* value defined here as the difference between the mean abundance of each molecular component between different ROIs divided by pooled standard deviation was calculated [26]. Unlike the t-test statistic, the Cohen's *d* is independent of the sample size, thus avoiding overestimation of the significance of differences between extremely numerous samples (which is typical for MSI data analysis, where large spectra collections are compared); the Cohen's *d* absolute value above 0.5, 0.8, and 1.2 corresponds to medium, large, and very large effects, respectively.

4.6. LC-MALDI MS/MS Analysis and Identification of Molecular Components

Representative samples of the cancerous thyroid gland (ca. 60% of cancer cells, FFPE material) were used for protein identification using the shotgun LC-MS/MS approach. Protein lysates were prepared and subjected to tryptic digestion according to a modified version of a combination of FASP with stage-tip fractionation as described in detail elsewhere [27]. Tryptic peptides were then separated using an EASY-nLC nano-liquid chromatography coupled with PROTEINEER fcII fraction collector (Bruker Daltonik, Bremen, Germany) and analyzed using an ultrafleXtreme mass spectrometer (Bruker Daltonik, Bremen, Germany). A detailed description of the instrumental settings of the LC-MALDI-MS/MS system is given in [27]. Registered MS/MS spectra were exported to ProteinScape 3.1 software (Bruker Daltonik, Bremen, Germany) and analyzed using Mascot Server 2.5.1 (Matrix Science, London, UK); for details, see Gawin et al. [27]. The hypothetical identity of molecular components detected by MSI was determined by the assignment of the mean parameters of MSI components (component location on mass/charge scale) to the measured masses of tryptic peptides identified in LC-MS/MS experiment; the assignment was performed allowing ±0.05% mass tolerance. The MSI molecular component annotations were established based on peptides identified in three major types of thyroid cancer (ATC, MTC, and well-differentiated thyroid cancers).

5. Conclusions

Molecular diagnostics of thyroid disease using the fine-needle cytological biopsy represent a general challenge for "classical" methods of proteomics. This problem could be overcome by mass spectrometry imaging, where sample classification is possible based on a single spectrum that corresponds to small clusters of cells. This approach was recently validated by Capitoli and colleagues who implemented the so-called pixel-by-pixel approach for the bi-state discrimination of malignant and hyperplastic (benign) thyroid nodules using actual diagnostic FNA material analyzed by MALDI-TOF-MSI [19]. Here, we confirmed that a similar approach could be implemented in a more complex situation when multiple classes are to be compared and classified. Our classification model cannot be implemented directly for the analysis of cytological material due to structural differences

between MS spectra registered for tissue microarrays based on FFPE and fresh material in actual FNA. However, the proposed single-pixel approach to multi-state classification problems has potentially general applicability not limited to thyroid tumors.

Supplementary Materials: The following are available online at http://www.mdpi.com/1422-0067/21/17/6289/s1. Figure S1. Average spectra of each type of region of interest (ROI).; Table S1. Included tissue specimens.; Table S2. Intra-ROI similarity of spectral features.; Table S3. Inter-ROI similarity of spectral features.; Table S4. Indices of multicomponent classifiers.; Table S5. Significance of differences in the abundance of molecular components detected by MSI between tissue types shown as the Cohen's *d* effect size.; Table S6. Hypothetical identity of peptide components detected by MSI based on peptides identified in well-differentiated thyroid cancers.; Table S7. Hypothetical identity of peptide components detected by MSI based on peptides identified in medullary thyroid cancers.; Table S8. Hypothetical identity of peptide components detected by mass spectrometry imaging (MSI) based on peptides identified in anaplastic thyroid cancers.; Table S9. Gene ontology (GO) terms associated with proteins that were hypothetically annotated with MSI components upregulated in specific cancer ROIs.

Author Contributions: Conceptualization, M.C., M.P., and P.W.; Data curation, G.M., and K.F. (Katarzyna Frątczak); Formal analysis, A.K., A.W., and K.Ł.; Funding acquisition, P.W.; Investigation, A.K., and M.G.; Methodology, A.K., M.G., K.Ł., J.P., K.F. (Krzysztof Fujarewicz), and M.P.; Project administration, P.W.; Resources, M.C., K.F. (Krzysztof Fujarewicz), and P.W.; Visualization, A.W.; Writing—original draft, P.W.; Writing—review and editing, A.K., M.G., and P.W. All authors have read and agreed to the published version of the manuscript.

Funding: The research leading to these results has received funding from the National Science Centre, Poland, Grant 2016/23/B/NZ4/03901.

Acknowledgments: We thank Ewa Zembala-Nożyńska for help in the retrieving of histopathological data. Calculations were performed using the infrastructure supported by the computer cluster Ziemowit (www.ziemowit.hpc.polsl.pl) funded by the Silesian BIO-FARMA project No. POIG.02.01.00-00-166/08 and expanded in the POIG.02.03.01-00-040/13 in the Computational Biology and Bioinformatics Laboratory of the Biotechnology Centre at the Silesian University of Technology.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

ATC	anaplastic thyroid carcinoma
FA	follicular adenoma
FFPE	formalin-fixed paraffin-embedded
FTC	follicular thyroid carcinoma
KNN	k-nearest neighbors
MALDI-TOF MS	matrix-assisted laser-desorption ionization time-of-flight mass spectrometry
MSI	mass spectrometry imaging
MTC	medullary thyroid carcinoma
NT	normal thyroid
PTC-CV	papillary thyroid carcinoma–classical variant
PTC-FV	papillary thyroid carcinoma–follicular variant
SVM	support vector machine
ROI	region of interest
TMA	tissue microarrays

References

1. Siegel, R.L.; Miller, K.D.; Jemal, A. Cancer statistics, 2015. *CA Cancer J. Clin.* **2015**, *65*, 5–29. [CrossRef] [PubMed]
2. Pellegriti, G.; Frasca, F.; Regalbuto, C.; Squatrito, S.; Vigneri, R. Worldwide increasing incidence of thyroid cancer: Update on epidemiology and risk factors. *J. Cancer Epidemiol.* **2013**, *2013*, 965212. [CrossRef] [PubMed]

3. Haugen, B.R.; Alexander, E.K.; Bible, K.C.; Doherty, G.M.; Mandel, S.J.; Nikiforov, Y.E.; Pacini, F.; Randolph, G.W.; Sawka, A.M.; Schlumberger, M.; et al. 2015 American Thyroid Association Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid Cancer: The American Thyroid Association Guidelines Task Force on Thyroid Nodules and Differentiated Thyroid Cancer. *Thyroid* **2016**, *26*, 1–133. [CrossRef] [PubMed]
4. Sakorafas, G.H. Thyroid nodules; interpretation and importance of fine-needle aspiration (FNA) for the clinician-practical considerations. *Surg. Oncol.* **2010**, *19*, 130–139. [CrossRef] [PubMed]
5. Kakudo, K.; Kameyama, K.; Miyauchi, A.; Nakamura, H. Introducing the reporting system for thyroid fine-needle aspiration cytology according to the new guidelines of the Japan Thyroid Association. *Endocr. J.* **2014**, *61*, 539–552. [CrossRef] [PubMed]
6. Faquin, W.C. The thyroid gland: Recurring problems in histologic and cytologic evaluation. *Arch. Pathol. Lab. Med.* **2008**, *132*, 622–632. [CrossRef] [PubMed]
7. Eszlinger, M.; Paschke, R. Molecular fine-needle aspiration biopsy diagnosis of thyroid nodules by tumor specific mutations and gene expression patterns. *Mol. Cell. Endocrinol.* **2010**, *322*, 29–37. [CrossRef]
8. Aragon Han, P.; Olson, M.T.; Fazeli, R.; Prescott, J.D.; Pai, S.I.; Schneider, E.B.; Tufano, R.P.; Zeiger, M.A. The impact of molecular testing on the surgical management of patients with thyroid nodules. *Ann. Surg. Oncol.* **2014**, *21*, 1862–1869. [CrossRef]
9. Pagni, F.; L'Imperio, V.; Bono, F.; Garancini, M.; Roversi, G.; De Sio, G.; Galli, M.; Smith, A.J.; Chinello, C.; Magni, F. Proteome analysis in thyroid pathology. *Expert Rev. Proteomics.* **2015**, *12*, 375–390. [CrossRef]
10. Cornett, D.S.; Reyzer, M.L.; Chaurand, P.; Caprioli, R.M. MALDI imaging mass spectrometry: Molecular snapshots of biochemical systems. *Nat. Methods* **2007**, *4*, 828–833. [CrossRef]
11. Schwamborn, K.; Caprioli, R.M. Molecular imaging by mass spectrometry-looking beyond classical histology. *Nat. Rev. Cancer* **2010**, *10*, 639–646. [CrossRef] [PubMed]
12. Seeley, E.H.; Caprioli, R.M. MALDI imaging mass spectrometry of human tissue: Method challenges and clinical perspectives. *Trends Biotechnol.* **2011**, *29*, 136–143. [CrossRef] [PubMed]
13. Aichler, M.; Walch, A. MALDI Imaging mass spectrometry: Current frontiers and perspectives in pathology research and practice. *Lab. Investig.* **2015**, *95*, 422–431. [CrossRef] [PubMed]
14. Mainini, V.; Pagni, F.; Garancini, M.; Giardini, V.; De Sio, G.; Cusi, C.; Arosio, C.; Roversi, G.; Chinello, C.; Caria, P.; et al. An alternative approach in endocrine pathology research: MALDI-IMS in papillary thyroid carcinoma. *Endocr. Pathol.* **2013**, *24*, 250–253. [CrossRef] [PubMed]
15. Min, K.W.; Bang, J.Y.; Kim, K.P.; Kim, W.S.; Lee, S.H.; Shanta, S.R.; Lee, J.H.; Hong, J.H.; Lim, S.D.; Yoo, Y.B.; et al. Imaging mass spectrometry in papillary thyroid carcinoma for the identification and validation of biomarker proteins. *J. Korean Med. Sci.* **2014**, *29*, 934–940. [CrossRef]
16. Pietrowska, M.; Diehl, H.C.; Mrukwa, G.; Kalinowska-Herok, M.; Gawin, M.; Chekan, M.; Elm, J.; Drazek, G.; Krawczyk, A.; Lange, D.; et al. Molecular profiles of thyroid cancer subtypes: Classification based on features of tissue revealed by mass spectrometry imaging. *Biochim. Biophys. Acta* **2017**, *1865*, 837–845. [CrossRef]
17. Pagni, F.; De Sio, G.; Garancini, M.; Scardilli, M.; Chinello, C.; Smith, A.J.; Bono, F.; Leni, D.; Magni, F. Proteomics in thyroid cytopathology: Relevance of MALDI-imaging in distinguishing malignant from benign lesions. *Proteomics* **2016**, *16*, 1775–1784. [CrossRef]
18. Mosele, N.; Smith, A.; Galli, M.; Pagni, F.; Magni, F. MALDI-MSI Analysis of Cytological Smears: The Study of Thyroid Cancer. *Methods Mol. Biol.* **2017**, *1618*, 37–47. [CrossRef]
19. Capitoli, G.; Piga, I.; Galimberti, S.; Leni, D.; Pincelli, A.I.; Garancini, M.; Clerici, F.; Mahajneh, A.; Brambilla, V.; Smith, A.; et al. MALDI-MSI as a Complementary Diagnostic Tool in Cytopathology: A Pilot Study for the Characterization of Thyroid Nodules. *Cancers* **2019**, *11*, 1377. [CrossRef]
20. DeHoog, R.J.; Zhang, J.; Alore, E.; Lin, J.Q.; Yu, W.; Woody, S.; Almendariz, C.; Lin, M.; Engelsman, A.F.; Sidhu, S.B.; et al. Preoperative metabolic classification of thyroid nodules using mass spectrometry imaging of fine-needle aspiration biopsies. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 21401–21408. [CrossRef]
21. Piga, I.; Capitoli, G.; Denti, V.; Tettamanti, S.; Smith, A.; Stella, M.; Chinello, C.; Leni, D.; Garancini, M.; Galimberti, S.; et al. The management of haemoglobin interference for the MALDI-MSI proteomics analysis of thyroid fine needle aspiration biopsies. *Anal. Bioanal. Chem.* **2019**, *411*, 5007–5012. [CrossRef] [PubMed]

22. Piga, I.; Capitoli, G.; Tettamanti, S.; Denti, V.; Smith, A.; Chinello, C.; Stella, M.; Leni, D.; Garancini, M.; Galimberti, S.; et al. Feasibility Study for the MALDI-MSI Analysis of Thyroid Fine Needle Aspiration Biopsies: Evaluating the Morphological and Proteomic Stability Over Time. *Proteom. Clin. Appl.* **2019**, *13*, e1700170. [CrossRef] [PubMed]
23. Galli, M.; Zoppis, I.; De Sio, G.; Chinello, C.; Pagni, F.; Magni, F.; Mauri, G. A Support Vector Machine Classification of Thyroid Bioptic Specimens Using MALDI-MSI Data. *Adv. Bioinform.* **2016**, *2016*, 3791214. [CrossRef] [PubMed]
24. Wilk, A.; Gawin, M.; Fratczak, K.; Widlak, P.; Fujarewicz, K. On Stability of Feature Selection Based on MALDI Mass Spectrometry Imaging Data and Simulated Biopsy. In *Current Trends in Biomedical Engineering and Bioimages Analysis. PCBEE 2019. Advances in Intelligent Systems and Computing*; Korbicz, J., Maniewski, R., Patan, K., Kowal, M., Eds.; Springer: Cham, Germany, 2020; Volume 1033, pp. 82–93. [CrossRef]
25. Galli, M.; Pagni, F.; De Sio, G.; Smith, A.; Chinello, C.; Stella, M.; L'Imperio, V.; Manzoni, M.; Garancini, M.; Massimini, D.; et al. Proteomic profiles of thyroid tumors by mass spectrometry-imaging on tissue microarrays. *Biochim. Biophys. Acta Proteins Proteom.* **2017**, *1865*, 817–827. [CrossRef] [PubMed]
26. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed.; Lawrence Erlbaum Associates: Hillsdale, NJ, USA, 1988.
27. Gawin, M.; Wojakowska, A.; Pietrowska, M.; Marczak, Ł.; Chekan, M.; Jelonek, K.; Lange, D.; Jaksik, R.; Gruca, A.; Widłak, P. Proteome profiles of different types of thyroid cancers. *Mol. Cell. Endocrinol.* **2018**, *474*, 68–79. [CrossRef]
28. Salabè, G.B. Pathogenesis of thyroid nodules: Histological classification? *Biomed. Pharmacother.* **2001**, *55*, 39–53. [CrossRef]
29. DeLellis, R.A.; Lloyd, R.V.; Heitz, P.U.; Eng, C. *Pathology and Genetics of Tumours of Endocrine Organs. WHO Classification of Tumours*; IARC Press: Lyon, France, 2004.
30. Schlumberger, M. Papillary and follicular thyroid carcinoma. *Ann. Endocrinol.* **2007**, *68*, 120–128. [CrossRef]
31. Ustun, B.; Chhieng, D.; Prasad, M.L.; Holt, E.; Hammers, L.; Carling, T.; Udelsman, R.; Adeniran, A.J. Follicular Variant of Papillary Thyroid Carcinoma: Accuracy of FNA Diagnosis and Implications for Patient Management. *Endocr. Pathol.* **2014**, *25*, 257–264. [CrossRef]
32. Wan, K.X.; Vidavsky, I.; Gross, M.L. Comparing Similar Spectra: From Similarity Index to Spectral Contrast Angle. *J. Am. Soc. Mass Spectrom.* **2002**, *13*, 85–88. [CrossRef]
33. Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene Selection for Cancer Classification using Support Vector Machines. *Mach. Learn.* **2002**, *46*, 389–422. [CrossRef]
34. Heijs, B.; Carreira, R.J.; Tolner, E.A.; de Ru, A.H.; van den Maagdenberg, A.M.J.M.; van Veelen, P.A.; McDonnell, L.A. Comprehensive Analysis of the Mouse Brain Proteome Sampled in Mass Spectrometry Imaging. *Anal. Chem.* **2015**, *87*, 1867–1876. [CrossRef] [PubMed]
35. Bednarczyk, K.; Gawin, M.; Chekan, M.; Kurczyk, A.; Mrukwa, G.; Pietrowska, M.; Polanska, J.; Widlak, P. Discrimination of normal oral mucosa from oral cancer by mass spectrometry imaging of proteins and lipids. *J. Mol. Histol.* **2019**, *50*, 1–10. [CrossRef] [PubMed]
36. Bruffaerts, C.; Verardi, V.; Vermandele, C. A generalized boxplot for skewed and heavy-tailed distributions. *Stat. Probab. Lett.* **2014**, *95*, 110–117. [CrossRef]
37. Wong, J.W.H.; Durante, C.; Cartwright, H.M. Application of fast Fourier transform cross-correlation for the alignment of large chromatographic and spectral datasets. *Anal. Chem.* **2005**, *77*, 5655–5661. [CrossRef]
38. Polanski, A.; Marczyk, M.; Pietrowska, M.; Widlak, P.; Polanska, J. Signal partitioning algorithm for highly efficient gaussian mixture modeling in mass spectrometry. *PLoS ONE* **2015**, *10*, e0134256. [CrossRef] [PubMed]
39. Polanski, A.; Marczyk, M.; Pietrowska, M.; Widlak, P.; Polanska, J. Initializing EM algorithm for univariate Gaussian, multi-component, heteroscedastic mixture models by dynamic programming partitions. *Int. J. Comput. Methods.* **2018**, *15*, 1850012. [CrossRef]
40. Frank, A.M.; Bandeira, N.; Shen, Z. Clustering millions of tandem mass spectra research articles. *J. Proteome Res.* **2008**, *7*, 113–122. [CrossRef] [PubMed]

Original Article

Radiomic signature accurately predicts the risk of metastatic dissemination in late-stage non-small cell lung cancer

Agata Małgorzata Wilk[1,2#^], Emilia Kozłowska[1#^], Damian Borys[1,3^], Andrea D'Amico[3^], Krzysztof Fujarewicz[1^], Izabela Gorczewska[3^], Iwona Debosz-Suwinska[4^], Rafał Suwinski[5^], Jarosław Smieja[1^], Andrzej Swierniak[1^]

[1]Department of Systems Biology and Engineering, Silesian University of Technology, Gliwice, Poland; [2]Department of Biostatistics and Bioinformatics, Maria Sklodowska-Curie National Research Institute of Oncology, Gliwice Branch, Gliwice, Poland; [3]Department of Nuclear Medicine and Endocrine Oncology, PET Diagnostics Unit, Maria Sklodowska-Curie National Research Institute of Oncology, Gliwice Branch, Gliwice, Poland; [4]Department of Radiotherapy, Maria Sklodowska-Curie National Research Institute of Oncology, Gliwice Branch, Gliwice, Poland; [5]II Radiotherapy and Chemotherapy Clinic, Maria Sklodowska-Curie National Research Institute of Oncology, Gliwice Branch, Gliwice, Poland

Contributions: (I) Conception and design: E Kozłowska, A Swierniak; (II) Administrative support: A Swierniak, K Fujarewicz, J Smieja; (III) Provision of study materials or patients: I Debosz-Suwinska, R Suwinski, A D'Amico; (IV) Collection and assembly of data: I Debosz-Suwinska, D Borys, A D'Amico, I Gorczewska, E Kozłowska; (V) Data analysis and interpretation: AM Wilk, E Kozłowska; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work.

Correspondence to: Professor Andrzej Swierniak, PhD; Emilia Kozłowska, PhD. Department of Systems Biology and Engineering, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland. Email: andrzej.swierniak@polsl.pl; emilia.kozlowska@polsl.pl.

Background: Non-small cell lung cancer (NSCLC) is the most common type of lung cancer, and the median overall survival (OS) is approximately 2–3 years among patients with stage III disease. Furthermore, it is one of the deadliest types of cancer globally due to non-specific symptoms and the lack of a biomarker for early detection. The most important decision that clinicians need to make after a lung cancer diagnosis is the selection of a treatment schedule. This decision is based on, among others factors, the risk of developing metastasis.

Methods: A cohort of 115 NSCLC patients treated using chemotherapy and radiotherapy (RT) with curative intent was retrospectively collated and included patients for whom positron emission tomography/computed tomography (PET/CT) images, acquired before RT, were available. The PET/CT images were used to compute radiomic features extracted from a region of interest (ROI), the primary tumor. Radiomic and clinical features were then classified to stratify the patients into short and long time to metastasis, and regression analysis was used to predict the risk of metastasis.

Results: Classification based on binarized metastasis-free survival (MFS) was applied with moderate success. Indeed, an accuracy of 0.73 was obtained for the selection of features based on the Wilcoxon test and logistic regression model. However, the Cox regression model for metastasis risk prediction performed very well, with a concordance index (C-index) score equal to 0.84.

Conclusions: It is possible to accurately predict the risk of metastasis in NSCLC patients based on radiomic features. The results demonstrate the potential use of features extracted from cancer imaging in predicting the risk of metastasis.

Keywords: Non-small cell lung cancer (NSCLC); metastasis; Cox regression; classification; radiomics

^ ORCID: Agata Małgorzata Wilk, 0000-0001-7554-1803; Emilia Kozłowska, 0000-0002-3069-3085; Damian Borys, 0000-0003-0229-2601; Andrea D'Amico, 0000-0003-4632-2139; Krzysztof Fujarewicz, 0000-0002-1837-6466; Izabela Gorczewska, 0000-0003-1387-5503; Iwona Debosz-Suwinska, 0000-0002-4554-8905; Rafał Suwinski, 0000-0002-3895-7938; Jarosław Smieja, 0000-0002-6120-4424; Andrzej Swierniak, 0000-0002-5698-5721.

Submitted Jan 30, 2023. Accepted for publication Jun 13, 2023. Published online Jul 07, 2023.
doi: 10.21037/tlcr-23-60
View this article at: https://dx.doi.org/10.21037/tlcr-23-60

Introduction

Lung cancer is one of the most frequently diagnosed cancer types worldwide, constituting over 11% of all cancer cases. With 2.2 million new diagnoses in 2020 alone, it was surpassed in incidence only by breast cancer, making the lung the most prevalent cancer site in men (with over 1.43 million diagnoses) and the third most prevalent in women after breast and colorectal cancers, which had 0.77 million diagnoses (1,2). While tobacco smoking is recognized as the primary cause of lung cancer, it can also be attributed to environmental factors such as air pollution, occupational exposure, and genetic predisposition (3-5). It is usually diagnosed at an advanced stage due to non-specific early-stage symptoms, which is reflected in the very high mortality rate. Indeed, the five-year survival rate for lung cancer does not exceed 20% (6-8), thus, it is the leading cause of cancer-related mortality and is responsible for 18% of all deaths from cancer (1).

Diagnosis of lung cancer involves medical imaging, including X-ray and positron emission tomography/computed tomography (PET/CT), which allows for classification according to the tumor node metastasis (TNM) staging system. Detected lesions are sampled by endobronchial ultrasound (EBUS) guided bronchoscopy and undergo histopathological assessment. Management is stage-specific (8), with clinical guidelines divided into early-stage, locally advanced, and metastatic cancer (9,10). In early-stage lung cancer, lobectomy is the preferred treatment option. If the tumor is not initially resectable, neoadjuvant chemotherapy can be implemented to downgrade the tumor, which would eventually allow for surgery. For selected patients with comorbidities, stereotactic ablative radiotherapy (SABR) may also be considered. For locally advanced cancer with lymph node involvement, platinum-based chemotherapy administered concurrently or sequentially with radiotherapy (RT) is the most commonly used curative therapeutic option, and it can be followed by maintenance immunotherapy. For advanced metastatic cancer, immune checkpoint inhibitors, with or without chemotherapy, are a viable therapeutic option. As molecular diagnostics becomes routinely available, targeted therapies aimed at epidermal growth factor receptor (EGFR) (11), fibroblast growth factor receptor (FGFR) (12), anaplastic lymphoma kinase (ALK) (13), or Kirsten rat sarcoma virus (KRAS) (14) are being used to treat mutation carriers.

One of the main reasons for the high mortality seen in lung cancer is its invasiveness, and most patients develop distant metastases. Unfortunately, metastatic tumors are often resistant to treatment, which leads to much shorter survival times for these patients. Although the exact mechanisms of metastasis are still being investigated, it is known that cancer cells can spread by both blood and lymphatic vessels (15,16). Lung cancer metastases are most frequently observed in the brain, bones, liver, lung, and adrenal gland (16). Since the occurrence of distant metastasis is the turning point in the course of the disease, it might be considered an important endpoint in prognostic analysis, along with the standard endpoints. Furthermore, the ability to predict when lung cancer will metastasize could guide clinical decision-making and may be used to indicate the need for therapy intensification in high-risk patients.

The search for accurate prognostic biomarkers in lung cancer is hindered by its high heterogeneity and complexity. Nonetheless, clinical and molecular characteristics have

Highlight box

Key findings

- PET/CT imaging is a routine procedure for radiotherapy planning in NSCLC. NSCLC is a highly metastatic cancer. Prediction of risk of metastasis and a time when metastatic appear is of high interest. Radiomics could be applied for predicting the risk of metastasis.

What is known and what is new?

- Radiomics was applied in the prediction of NSCLC survival before.
- Here, we applied a cohort of heterogenous NSCLC patients for metastasis risk prediction. We extracted a radiomics signature for risk prediction with high prediction performance (concordance index =0.84)

What is the implication, and what should change now?

- We can predict the risk of metastasis based on routinely collected PET/CT images from a primary tumor.

shown some promise in predicting metastasis. Metastasis-associated lung adenocarcinoma transcript 1 (*MALAT-1*), a long non-coding ribonucleic acid (RNA), was demonstrated to be significantly associated with metastasis in non-small cell lung cancer (NSCLC) (8). Meanwhile, cancer antigen 125 (CA125) and neuron-specific enolase (NSE) were found to be indicative of liver metastasis (17).

The radiomics-based approach has been successfully applied for different endpoints in lung cancer, including overall survival (OS) and progression-free survival (PFS) (18). It has also shown promising results for the prediction of distant metastases. Coroller *et al.* (19) selected a radiomic signature based on CT images to predict distant metastasis in lung adenocarcinoma. Wu *et al.* constructed and validated a Cox proportional hazards model using ^{18}F-fluorodeoxyglucose PET (^{18}F-FDG PET) imaging to predict freedom of distant metastasis in early-stage NSCLC patients (20). Fave *et al.* (21) demonstrated that adding pre-treatment radiomic features extracted from CT images could improve the ability of clinical prognostic models to predict distant metastasis (21). Meanwhile, Dou *et al.* (22) focused on locally advanced lung adenocarcinoma and investigated radiomic features from the primary tumor and peritumoral region (22).

In this study, 115 NSCLC patients with various histological subtypes were retrospectively analyzed. The prognostic value of standard clinical features, and radiomic features extracted from PET/CT images acquired for RT planning, were evaluated by determining if they could be used to predict time to distant metastasis. To answer this question, machine learning models were constructed for continuous and categorical metastasis-free survival (MFS) prediction. We present this article in accordance with the TRIPOD reporting checklist (available at https://tlcr.amegroups.com/article/view/10.21037/tlcr-23-60/rc).

Methods

Study design

A cohort of NSCLC patients was collated to investigate if PET/CT imaging routinely performed for RT planning could help in planning the future treatment strategy, with a focus on predicting the risk and time of relapse with distant metastases. MFS was defined as the time elapsed between diagnosis and the detection of distant metastasis or the time of death/last follow-up if distant metastases did not emerge. In addition, classification algorithms were used to predict if MFS would be short or long.

As the prediction of metastasis risk was the focus of the study, the primary lung cancer tumor was the region of interest (ROI). Using the available PET/CT scans, radiomic features were extracted from the ROI and assessed.

The specific clinical question considered in this work was whether or not a radiomic signature could be extracted that would help discriminate between a primary tumor that has the potential to metastasize early from one that metastasizes late or not at all.

The pipeline of our method is presented in *Figure 1*. The method includes processing the clinical and PET/CT data and applying them as predictors (features) in the calibration of survival models (Cox regression, random survival forest) and perform classification.

Study population

Data were collected retrospectively at the Maria Sklodowska-Curie National Research Institute of Oncology, Gliwice Branch (NRIO). The cohort consisted of 115 patients with NSCLC who were treated with curative intent at the Institute between 2009 and 2017. All patients in the cohort had been treated with a combination of chemotherapy and RT. Most of the patients received a platinum-based doublet with vinorelbine. Patients received between one and six cycles (median four), followed by RT with a total dose between 60 and 70 Gray (Gy) in two Gy fractions. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study was approved by the institutional review board of Maria Sklodowska-Curie National Research Institute of Oncology (Gliwice Branch) (No. KB/430-48/23) and individual consent for this retrospective analysis was waived. The clinical data were anonymized before the computational analysis.

All patients underwent PET/CT imaging for RT planning. Only patients with non-detectable distant tumors at the onset of treatment were assessed. However, most patients had locally disseminated tumors to the lymph nodes, as they were diagnosed late due to non-specific symptoms.

In the cohort, 72.2% of patients were males, and 27.8% were female. This is consistent with population data showing that most lung cancer patients are male. The median age of patients in the cohort was 61 years, and over half of the patients had tumors located in the left lung. The most prevalent cancer subtype was squamous cell carcinoma, which constituted two-thirds of all cases, followed by not

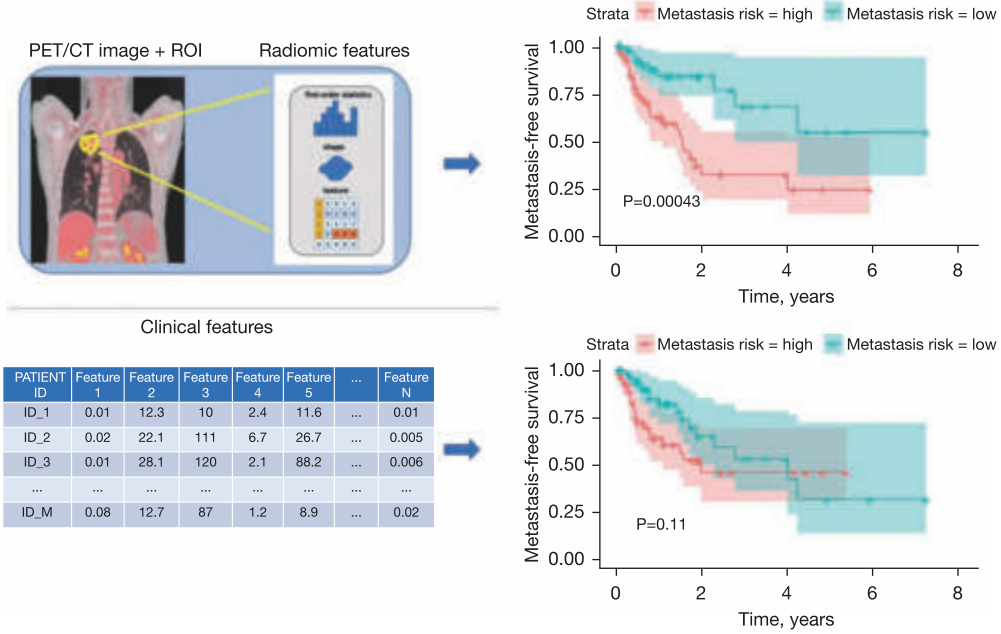


Figure 1 Project workflow. PET/CT images were acquired and radiomic features were extracted from ROI. Integration of clinical and radiomic data led to the prediction of short-term and long-term MFS and the risk of metastasis. The output from the workflow was a radiomic signature, which could be used for the prediction of metastasis risk in newly diagnosed NSCLC patients being treated with platinum-based chemotherapy. PET/CT, positron emission tomography/computed tomography; ROI, regions of interest; MFS, metastasis-free survival; NSCLC, non-small cell lung cancer.

otherwise specified (24.3%) and adenocarcinoma (7.0%). Detailed characteristics of the cohort are presented in *Table 1*.

Extraction of radiomic features

Feature extraction was performed with PyRadiomics version 3.0.1, a Python package designed to increase the reproducibility of radiomic studies (23). Using the PET dataset, 105 standard features were calculated. Radiomic features belong to one of three classes, including first-order statistics such as energy, entropy, and minimum, as well as shape features such as volume, surface area, and sphericity, and texture features including gray level co-occurrence matrix (GLCM), gray level dependence matrix (GLDM), gray level run length matrix (GLRLM), gray level size zone matrix (GLSZM), and neighboring-gray tone difference matrix (NGTDM).

MFS categorization

For the classification, a threshold of one year was used to create two classes, which included patients with MFS below and over this threshold. Due to the presence of censored observations (in the cohort this primarily signified the patient's death), such stratification divided patients into a group who suffered either metastasis or death within a year (66 patients), and those who did not (49 patients). To create subgroups that were more related to the research question, the binary MFS was defined as "short" if the patient developed metastasis within a year (25 patients) and "long" if the patient developed metastasis or was censored after longer than a year (49 patients).

Statistical analysis

Statistical analysis was performed using the R environment (version 4.1.3). For survival analysis, survival (version 3.2-13) was used, caret (version 6.0.93) and RandomForest (version 4.7-1.1) were used for classification, and randomForestSRC (version 3.1.1) was used to perform random survival forest. A heatmap of the radiomic features was created with ComplexHeatmap (version 2.10.0).

Filtering of the radiomic features was applied based on the Pearson correlation coefficient to avoid redundancy,

Table 1 Patient characteristics

Characteristics	Subtypes	Patients (n=115)
Sex, n (%)	Male	83 (72.2)
	Female	32 (27.8)
Age (years)	Median [Q1–Q3]	61 [57–67]
Histopathology, n (%)	Squamous	77 (67.0)
	Adenocarcinoma	8 (7.0)
	Not otherwise specified	28 (24.3)
	Large cell	2 (1.7)
Location, n (%)	Left	65 (56.5)
	Right	50 (43.5)
T, n (%)	1	4 (3.5)
	2	37 (32.2)
	3	37 (32.2)
	4	37 (32.2)
N, n (%)	0	19 (16.5)
	1	6 (5.2)
	2	83 (72.2)
	3	7 (6.1)
M, n (%)	0	115 (100.0)
	1	0 (0)
Zubrod score, n (%)	0	34 (29.6)
	1	80 (69.6)
	2	1 (0.9)

with a cutoff threshold equal to 0.9 (see Table S1). Since the PET images were acquired using two scanners, principal component analysis was applied to determine if there was any grouping of samples due to the scanner used (see Figure S1).

The clinical and radiomic features with potential for event-free survival (EFS) and MFS prediction were assessed (see Table S2). In addition, differences in the values of radiomic and clinical features between 'short' and 'long' MFS patient subgroups were investigated statistically. Fisher's exact test was performed for categorical variables, while the Mann-Whitney *U* test was used for continuous variables (see Table S3). A log-rank test was also conducted for both categorical and continuous features (see Table S4). As the log-rank test assesses if there is a significant difference between two or more survival curves, continuous features were binarized with respect to the median value.

Cross-validation

The value of any type of predictive model lies in its applicability to unknown data, and not just its ability to fit the training data. Cross-validation enables evaluation of the model's ability to generalize by removing part of the data from the cohort and applying them in the estimation of model performance. In addition, data partitioning at the beginning of each iteration prevents information leakage.

For a more consistent comparison between the regression and classification results, modified k-fold data partitioning was applied. Firstly, the data was ordered according to (continuous) MFS values. Then, the observations were assigned consecutive numbers, from one to five, which were used as cross-validation folds. Such partitioning ensures proper stratification of both continuous and binarized MFS.

Classification algorithms

The observed relationships between binary MFS and binary EFS and extracted features (both clinical radiomic) were verified by employing classification models. Firstly, three main feature selection methods were applied, including Student's *t*-test, Wilcoxon test, and a mutual information test. To investigate the impact of a varying number of features on classification quality, between 1 and 10 features were tested. Since only the mutual information method handles both categorical and continuous variables, a hybrid selection was used for the other two methods by applying the main method for continuous variables and Fisher's exact test for categorical variables. The categorical variables that passed the significance threshold equal to 0.1 were added to the model.

The following classification methods were tested: K nearest neighbor (KNN) with different K values (for clarity, only the best one, K=5, is presented), random forest, support vector machines (SVM) with linear and radial kernels, and logistic regression (LogReg). Considering the inconsistent orders of magnitude for radiomic features, a z-score transformation was used to scale the data. In each k-fold iteration, the scaling parameters (mean and standard deviation) were determined from the training set and applied to both the training and test sets. Classification accuracy was then used to assess model performance.

Regression algorithms

For the prediction of continuous MFS, Cox proportional hazards regression (using survival R package) and random survival forest (using randomForestSRC R package) were applied. Variable selection was performed based on univariate analysis, with the Harrell concordance index (C-index) adopted as a ranking metric. The model performance was validated using the k-fold partitioning described above. Again, models containing between 1 and 10 features were tested.

Radiomic-based risk score

Although cross-validation facilitates the estimation of prediction quality, the results and selected features can be different in each iteration due to subsampling. Therefore, all selections were repeated on the entire dataset to obtain conclusive feature rankings. To demonstrate the validity of the obtained signature, the Cox model was chosen, which is the classic approach to survival data analysis with known interpretation. The patients were then divided into high-risk and low-risk groups based on the calculated median risk score, and MFS was compared using Kaplan-Meier curves.

Results

Patient characteristics

The cohort included only NSCLC patients, as it is the most common type of lung cancer. Most patients (67%) had squamous histopathological subtypes, and almost two-thirds had an advanced stage of the primary tumor (T3 or T4). In total, 37 patients eventually developed distant metastases. *Figure 2* shows the radiomics features of a whole cohort and the time-to-metastasis as a survival curve. The median time-to-metastasis was 2.77 years, with a secondary tumor observed most frequently in the second lung, brain, bones, and liver.

None of the clinical features of the cohort were informative in relation to the time to metastasis onset (Table S2). This means that clinicians are unable to predict if a particular patient will develop metastatic cancer, based only on clinical variables at diagnosis. On the other hand, 34 radiomic features were statistically significant against continuous MFS, 36 features against the binarized MFS, and 18 against EFS.

PET/CT data acquisition and segmentation

The PET/CT images were acquired at the NRIO using Philips GeminiGXL 16 (Philips, Amsterdam, Netherlands) (24 patients) and Siemens Biograph mCT 131 (Siemens AG, Munich, Germany) (88 patients) PET/CT scanners. For each patient, the ROI was contoured by the same experienced nuclear medicine specialist using Medical Image Merge (MIM) 7.0.1 software and the PET Edge™ tool (both MIM Software Inc., OH, USA).

Integration of clinical and radiomic data

Figure 2A shows the normalized z-score values of radiomic features for each patient. The patients were divided into short and long EFS groups. As can be seen from the results, the hierarchical clustering correctly divided patients into these two groups. Furthermore, it was observed that the radiomic feature spectrum varied between patients with short and long EFS. This demonstrates that there is potential for the use of radiomic features in predicting EFS.

We also investigated the correlation between radiomic features. High correlations were observed between the radiomic features, which resulted in only 65 of 105 features passing the initial correlation filtering (Figure S2). The highest redundancy was found for the first-order features (6 out of 18 were kept) and the lowest for the GLSZM features (15 out of 16 were kept) (see Table S1).

Classification of advanced NSCLC

As expected, no clinical features were selected by the models. The feature rankings obtained for EFS (*Figure 3*) and MFS (*Figure 4*) prediction differed, which aligns with the different interpretations of these endpoints. While the rankings varied with respect to feature selection and classification methods, there was some consistency among the top features. Indeed, TotalEnergy, ZoneEntropy, and RootMeanSquared favored EFS prediction, while Variance, TotalEnergy, RunLengthNonUniformity (GLRLM), SizeZoneNonUniformityNormalized, and Maximum2DDiameterColumn favored MFS prediction.

The highest accuracy for EFS prediction (approx. 0.65) was achieved using the SVM classifier with linear kernels for the mutual information selection of eight features. The highest accuracy for MFS prediction (approx. 0.73) was

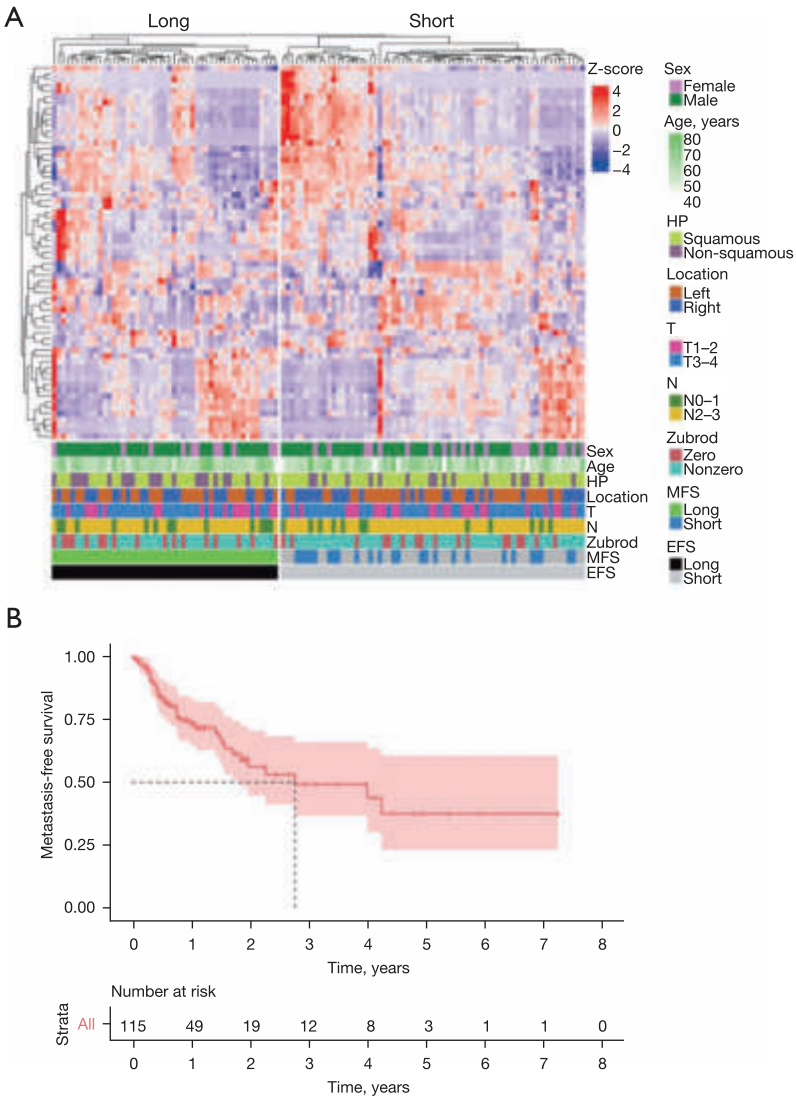


Figure 2 The integration of clinical and radiomic data. (A) Integration of clinical and radiomic data. Patients were split by the binary EFS. (B) Kaplan-Meier plot of MFS for the entire population. HP, histopathology; EFS, event-free survival; MFS, metastasis-free survival.

achieved using the LogReg classifier for the five features selected using the Wilcoxon test. Due to the imbalanced classes, with "long" (treated as the negative class) being the predominant group, the models for MFS tended to yield high specificity and relatively low sensitivity. Most models performed better for a small number of features.

Prediction of risk of metastasis

For regression-based models, the tendency was similar, with the highest predictive ability observed for a small number of features. The highest median C-index across folds was reached for two features (GLRLM and NGTDM Business) in Cox regression and one feature (shapeMinorAxisLength) in the random survival forest (*Figure 5*).

The mean C-index for the best set of features using Cox regression was 0.84, whereas the C-index for the random survival forest was 0.8. The inclusion of more features in the model resulted in a loss of prediction quality due to overfitting. No clinical features were selected for the best models, which is consistent with the preliminary patient cohort analysis.

Feature selection on the entire dataset revealed that the two top features for Cox regression,

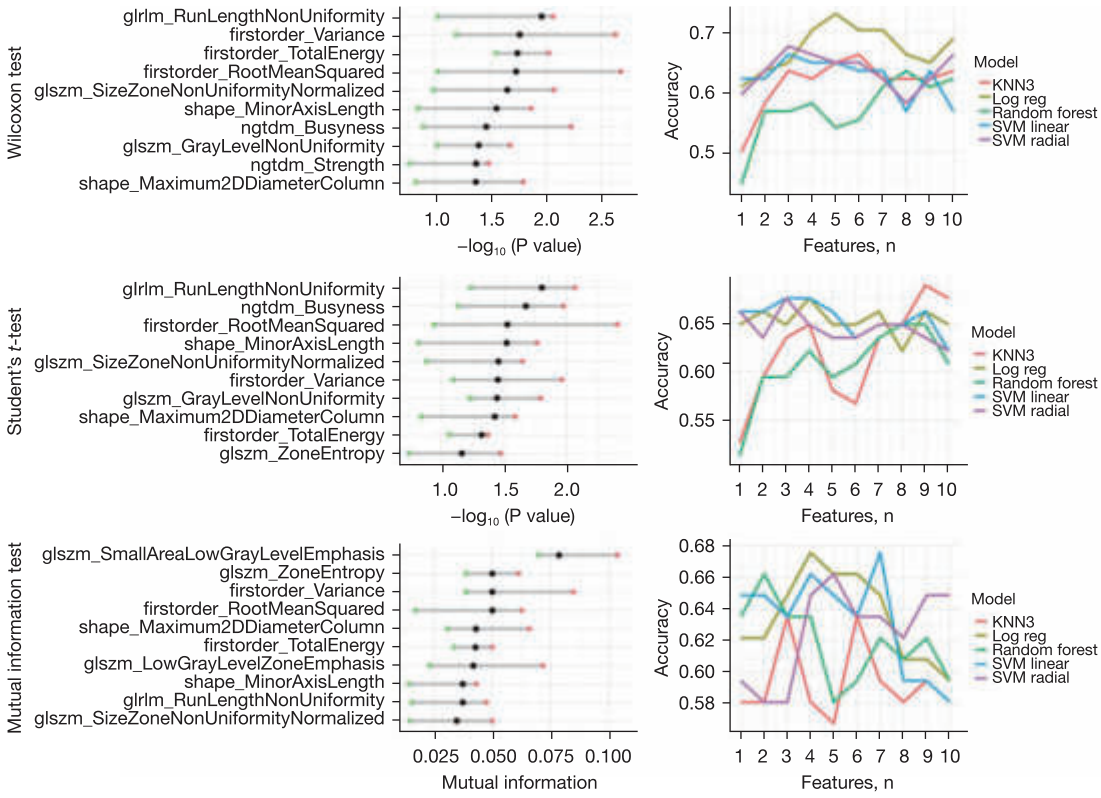


Figure 3 MFS prediction using the classification approach. Top row: Wilcoxon test; middle row: Student's *t*-test; bottom row: mutual information test. Left column: feature selection in a 5-fold cross-validation. Features were ranked according to the $-\log_{10}$ (P value) for the Wilcoxon test and Student's *t*-test selections, and mutual information score for mutual information selection. Black dots indicate the median value across folds, green dots indicate the lowest value across folds, and red dots indicate the highest value. Right column: classification results for the test set in a 5-fold cross-validation for different models, depending on the number of features. KNN, K nearest neighbor; SVM, support vector machine; MFS, metastasis-free survival.

SmallAreaLowGrayLevelEmphasis (GLSZM) and RunLengthNonUniformity (GLRLM), also held high-ranking positions in the classification approach. Also, univariate analysis shows that those two features could discriminate high risk from low-risk patients (Figures S3,S4). Therefore, the Cox model was constructed using those two features. The high-risk and low-risk groups (*Figure 6*) had significantly different MFS, with the log-rank test P<0.001.

Discussion

Lung cancer is the leading cause of cancer-related death worldwide, claiming over 1.7 million lives yearly. It is characterized by high invasiveness, and the occurrence of distant spread significantly influences survival and treatment options. This necessitates the search for prognostic biomarkers that could help determine the time to metastasis onset. With the rapid development of the radiomics field, researchers have turned to medical imaging, which is routinely performed and non-invasive, as a source of information that could shed some light on the tumor dissemination process and aid clinicians in therapy planning.

A cohort of NSCLC patients with different subtypes and stages of the disease was collated. It was concluded that the standard clinical data available for the patients, except for higher metastatic potential exhibited by the squamous subtype, were largely uninformative regarding metastasis occurrence. To assess the potential of radiomics for MFS prediction, we extracted 105 radiomic features from PET/CT scans, using the primary tumor as the ROI.

Correlations between radiomic and clinical features were mostly low, signifying that both datasets carried independent information. Also, the hierarchical clustering of radiomic features did not correspond to any discernible

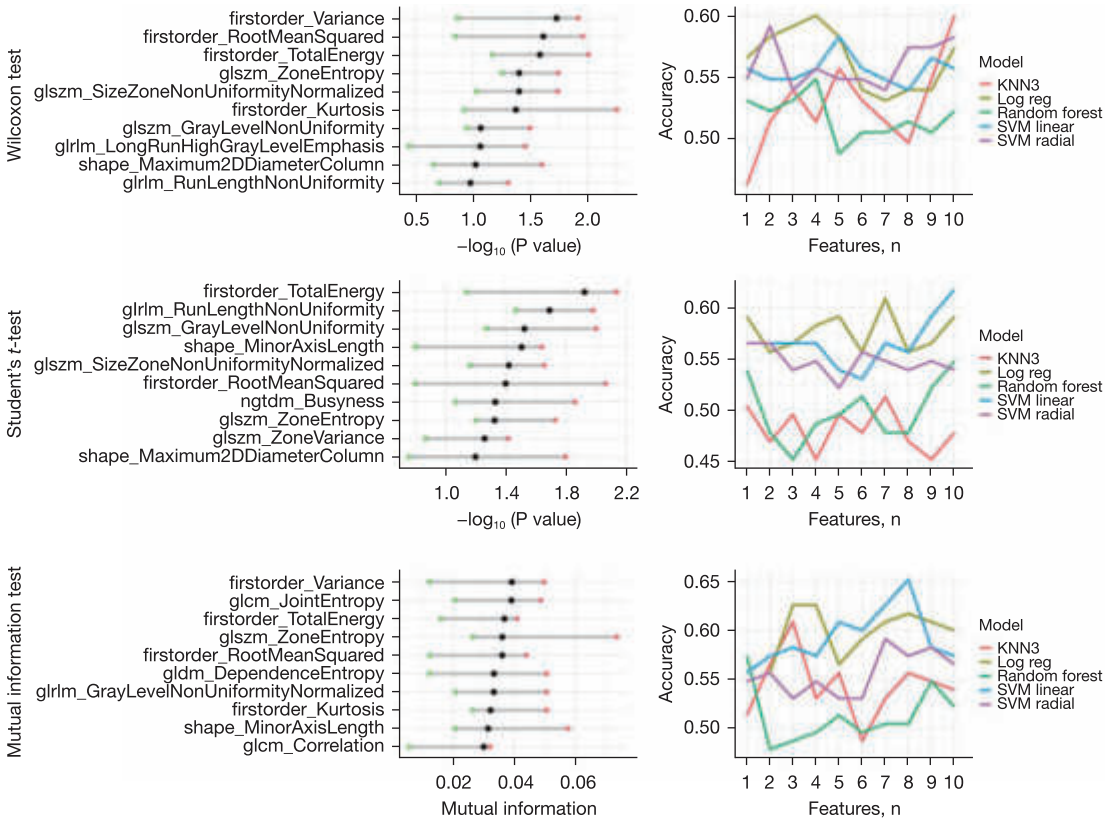


Figure 4 EFS prediction using the classification approach. Top row: Wilcoxon test; middle row: Student's *t*-test; bottom row: mutual information test. Left column: feature selection in a 5-fold cross-validation. Features are ranked according to the $-\log_{10}$ (P value) for the Wilcoxon test and Student's *t*-test selections, and mutual information score for mutual information selection. Black dots indicate the median value across folds, green dots indicate the lowest value across folds, and red dots indicate the highest value across folds. Right column: classification results for the test set in a 5-fold cross-validation for different models, depending on the number of features. KNN, K nearest neighbor; SVM, support vector machine; EFS, event-free survival.

grouping of clinical features (see *Figure 2*).

Regression and machine learning methods were then used to select radiomic signatures that could predict the risk of metastasis and achieved a C-index of 0.84 for the Cox proportional hazards model and 0.8 for the random survival forest, and an accuracy of 0.72 for the KNN classifier. These results confirm that medical images contain information that could be successfully applied to MFS prediction.

Several studies have shown the potential of radiomic features in predicting distant metastasis in lung cancer, with most of them focusing on either a particular subtype or stage (24,25). Coroller *et al.* (19) investigated radiomic features extracted from CT images for predicting distant metastasis in lung adenocarcinoma, which had a C-index of 0.61 on an independent validation set. Fave *et al.* (21) demonstrated that combining pre-treatment radiomic features with clinical information improved the ability of prognostic models to predict distant metastasis in stage III NSCLC patients, reporting a C-index of 0.63 (19). Wu *et al.* used features extracted from PET images to predict the freedom of distant metastasis, with a high C-index of 0.71 in independent validation (20). However, this work only focused on early-stage lung cancer. Dou *et al.* (22) presented an interesting approach, extracting features from both the tumor and tumor rim and achieving a C-index of 0.64 in a cohort of patients with locally advanced lung adenocarcinoma (22). In the current work, significantly better model quality was achieved in a cohort including patients with varying subtypes (squamous cell carcinoma, adenocarcinoma, large cell carcinoma) and stages.

While a regression approach, such as a Cox proportional

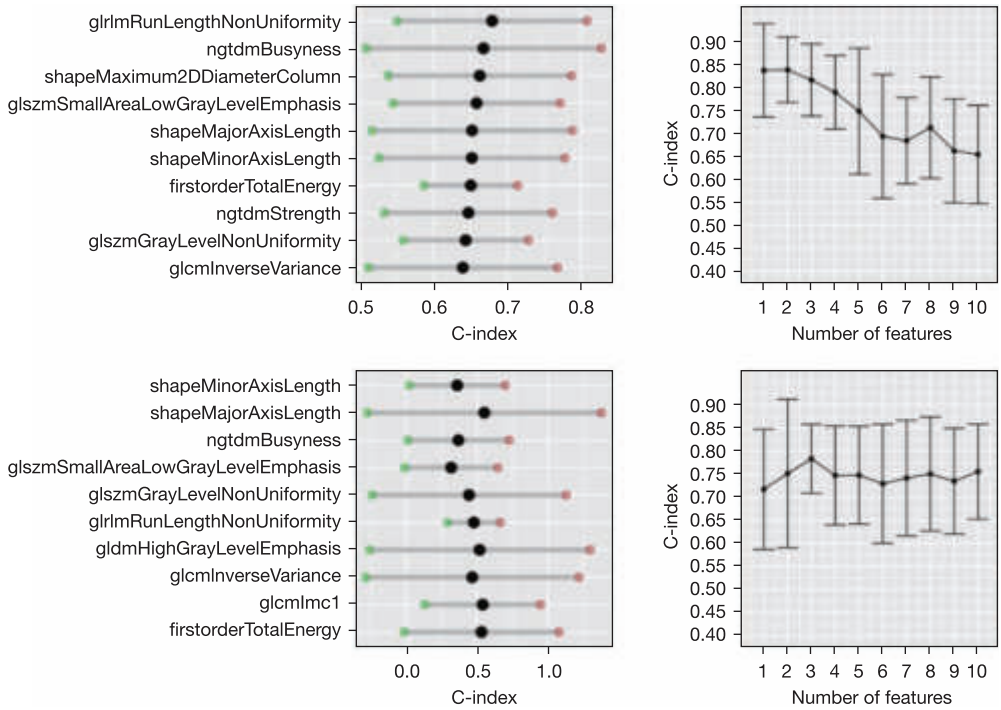


Figure 5 MFS prediction using a regression approach. Top row: Cox regression; bottom row: random survival forest. Left column: feature selection in a 5-fold cross-validation. Features were ranked according to the concordance index value for the univariate model. Right column: prediction results for the test set in a 5-fold cross-validation, depending on the number of features. MFS, metastasis-free survival.

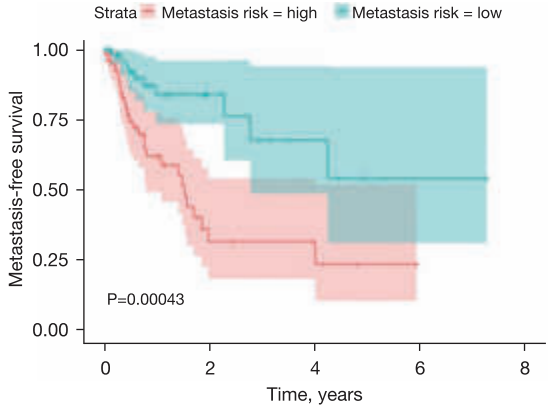


Figure 6 Kaplan-Meier plot of metastasis-free survival for the whole cohort. The patients were divided into high-risk and low-risk groups according to a Cox model constructed using the feature selection and number for which the cross-validation accuracy was highest.

hazards model, is typically used for survival-type analysis, the risk score it yields does not directly translate to the time of event occurrence. The C-index only compares pairs of observations, resulting in a global assessment of whether a higher risk is related to a shorter time-to-event. Therefore, classification was also performed and achieved an accuracy of 0.72 in cross-validation.

After testing several methods and approaches to variable selection, it was observed that similar predictive ability could be achieved for different feature sets, which indicates that even unrelated radiomic features carry equivalent information. Interestingly, the quality dropped drastically with increased feature numbers in all models. This suggests that features with high predictive potential perform much worse when combined than when used in isolation, and emphasizes the importance of selecting algorithms that are sensitive to feature interactions.

Certain variables retained high positions across different selections. These included GLRLM RunLengthNonUniformity, NGTDM Strength, and NGTDM Business. This demonstrates that these radiomic features are important for predicting if and when metastasis will occur in a lung cancer patient.

This analysis was not without limitations. While the study design ensured all images were contoured by one expert, which prevented bias, this did not allow for an

assessment of the reproducibility of radiomic feature extraction. In addition, plans are in place to collect an independent patient cohort to validate the signature. Future work will also investigate tumor growth and dissemination dynamics, to achieve more clinically meaningful predictions.

Conclusions

Based on a cohort comprising 115 NSCLC patients, clinical features routinely collected during diagnostic procedures are not sufficient for the prediction of the risk of metastasis. Medical images (PET/CT scans) were investigated as a potential source of prognostic markers by assessing radiomic features in various classes of predictive models. A model based on two texture features (GLSZM and GLRLM) was constructed, which divided the patient cohort into low-risk and high-risk groups that significantly differed in MFS. The findings of this study have the potential to help clinicians make adjustments to therapy and create a rational basis for the intensification of systemic treatment in high-risk lung cancer patients.

Acknowledgments

Funding: This work was supported by the Polish National Science Centre (No. UMO-2020/37/B/ST6/01959). Calculations were performed on the Ziemowit computer cluster in the Laboratory of Bioinformatics and Computational Biology, created in the EU Innovative Economy Programme POIG.02.01.00-00-166/08 and expanded in the POIG.02.03.01-00-040/13 project. Data analysis was partially carried out using the Biotest Platform developed within project PBS3/B3/32/2015, which was financed by the Polish National Centre of Research and Development (NCBiR). This work was carried out in part by the Silesian University of Technology internal research funding (to K.F., J.S.). The funders have no role in designing the study and writing the manuscript.

Footnote

Reporting Checklist: The authors have completed the TRIPOD reporting checklist. Available at https://tlcr.amegroups.com/article/view/10.21037/tlcr-23-60/rc

Data Sharing Statement: Available at https://tlcr.amegroups.com/article/view/10.21037/tlcr-23-60/dss

Peer Review File: Available at https://tlcr.amegroups.com/article/view/10.21037/tlcr-23-60/prf

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://tlcr.amegroups.com/article/view/10.21037/tlcr-23-60/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work and will ensure that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The study was conducted in accordance with the Declaration of Helsinki (as revised in 2013). The study was approved by the institutional review board of Maria Sklodowska-Curie National Research Institute of Oncology (Gliwice Branch) (No. KB/430-48/23), and individual consent for this retrospective analysis was waived.

References

1. Malvezzi M, Santucci C, Boffetta P, et al. European cancer mortality predictions for the year 2023 with focus on lung cancer. Ann Oncol 2023;34:410-9.
2. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin 2021;71:209-49.
3. Alberg AJ, Samet JM. Epidemiology of lung cancer. Chest 2003;123:21S-49S.
4. Dela Cruz CS, Tanoue LT, Matthay RA. Lung cancer: epidemiology, etiology, and prevention. Clin Chest Med 2011;32:605-44.
5. Barta JA, Powell CA, Wisnivesky JP. Global Epidemiology of Lung Cancer. Ann Glob Health 2019;85:8.
6. Lu T, Yang X, Huang Y, et al. Trends in the incidence, treatment, and survival of patients with lung cancer in the

last four decades. Cancer Manag Res 2019;11:943-53.
7. Torre LA, Siegel RL, Jemal A. Lung Cancer Statistics. Adv Exp Med Biol 2016;893:1-19.
8. Molina JR, Yang P, Cassivi SD, et al. Non-small cell lung cancer: epidemiology, risk factors, treatment, and survivorship. Mayo Clin Proc 2008;83:584-94.
9. Planchard D, Popat S, Kerr K, et al. Metastatic non-small cell lung cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol 2018;29:iv192-237.
10. Postmus PE, Kerr KM, Oudkerk M, et al. Early and locally advanced non-small-cell lung cancer (NSCLC): ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol 2017;28:iv1-iv21.
11. Le T, Gerber DE. Newer-Generation EGFR Inhibitors in Lung Cancer: How Are They Best Used? Cancers (Basel) 2019;11:366.
12. Zhou Z, Liu Z, Ou Q, et al. Targeting FGFR in non-small cell lung cancer: implications from the landscape of clinically actionable aberrations of FGFR kinases. Cancer Biol Med 2021. [Epub ahead of print]. doi: 10.20892/j.issn.2095-3941.2020.0120.
13. Yuan M, Huang LL, Chen JH, et al. The emerging treatment landscape of targeted therapy in non-small-cell lung cancer. Signal Transduct Target Ther 2019;4:61.
14. Salgia R, Pharaon R, Mambetsariev I, et al. The improbable targeted therapy: KRAS as an emerging target in non-small cell lung cancer (NSCLC). Cell Rep Med 2021;2:100186.
15. Popper HH. Progression and metastasis of lung cancer. Cancer Metastasis Rev 2016;35:75-91.
16. Zhu T, Bao X, Chen M, et al. Mechanisms and Future of Non-Small Cell Lung Cancer Metastasis. Front Oncol 2020;10:585284.
17. Wang CF, Peng SJ, Liu RQ, et al. The Combination of CA125 and NSE Is Useful for Predicting Liver Metastasis of Lung Cancer. Dis Markers 2020;2020:8850873.
18. Zhang Y, Oikonomou A, Wong A, et al. Radiomics-based Prognosis Analysis for Non-Small Cell Lung Cancer. Sci Rep 2017;7:46349.
19. Coroller TP, Grossmann P, Hou Y, et al. CT-based radiomic signature predicts distant metastasis in lung adenocarcinoma. Radiother Oncol 2015;114:345-50.
20. Wu J, Aguilera T, Shultz D, et al. Early-Stage Non-Small Cell Lung Cancer: Quantitative Imaging Characteristics of (18)F Fluorodeoxyglucose PET/CT Allow Prediction of Distant Metastasis. Radiology 2016;281:270-8.
21. Fave X, Zhang L, Yang J, et al. Delta-radiomics features for the prediction of patient outcomes in non-small cell lung cancer. Sci Rep 2017;7:588.
22. Dou TH, Coroller TP, van Griethuysen JJM, et al. Peritumoral radiomics features predict distant metastasis in locally advanced NSCLC. PLoS One 2018;13:e0206108.
23. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Res 2017;77:e104-7.
24. Liu RS, Ye J, Yu Y, et al. The predictive accuracy of CT radiomics combined with machine learning in predicting the invasiveness of small nodular lung adenocarcinoma. Transl Lung Cancer Res 2023;12:530-46.
25. Li J, Zhang B, Ge S, et al. Prognostic value of (18)F-FDG PET/CT radiomic model based on primary tumor in patients with non-small cell lung cancer: A large single-center cohort study. Front Oncol 2022;12:1047905.

Cite this article as: Wilk AM, Kozłowska E, Borys D, D'Amico A, Fujarewicz K, Gorczewska I, Debosz-Suwinska I, Suwinski R, Smieja J, Swierniak A. Radiomic signature accurately predicts the risk of metastatic dissemination in late-stage non-small cell lung cancer. Transl Lung Cancer Res 2023;12(7):1372-1383. doi: 10.21037/tlcr-23-60

Improving the predictive ability of radiomics-based regression survival models through incorporating multiple regions of interest

Agata Małgorzata Wilk[1,2], Emilia Kozłowska[1], Damian Borys[1,3], Andrea D'Amico[3], Izabela Gorczewska[3], Iwona Debosz-Suwińska[4], Seweryn Gałecki[1], Krzysztof Fujarewicz[1], Rafał Suwiński[5], and Andrzej Swierniak[1]

[1] Silesian University of Technology, Department of Systems Biology and Engineering, Gliwice 44-102, Poland,
andrzej.swierniak@polsl.pl
[2] Department of Biostatistics and Bioinformatics, Maria Sklodowska-Curie National Research Institute of Oncology, Gliwice Branch, Gliwice 44-102, Poland
[3] Department of Nuclear Medicine and Endocrine Oncology, PET Diagnostics Unit, Maria Sklodowska-Curie National Research Institute of Oncology, Gliwice Branch, Gliwice 44-102, Poland
[4] Department of Radiotherapy, Maria Sklodowska-Curie National Research Institute of Oncology, Gliwice Branch, Gliwice 44-102, Poland
[5] II-nd Radiotherapy and Chemotherapy Clinic and Teaching Hospital, Maria Sklodowska-Curie National Research Institute of Oncology, Gliwice Branch, Gliwice 44-102, Poland

Abstract. Radiomic features, numeric values extracted from a region of interest (ROI) in medical images, can be used to train prognostic models for various types of cancer. However, in locally advanced diseases, more than one lesion may be present. Using the information contained in multiple regions increases the complexity and necessitates additional processing. Here, we tested seven strategies of handling multiple regions in radiomic-based regularized Cox regression for predicting metastasis-free survival using a cohort of 115 non-small cell lung cancer patients. We have found that using all ROIs to fit the model allowed for better results than using only the largest ROI, achieving c-indexes of 0.617 and 0.581, respectively.

Keywords: lung cancer, radiomics, ROI, metastasis free survival, regularized Cox regression

1 Introduction

Medical imaging is a standard procedure used in cancer both as part of diagnostics, and disease management, for example in radiotherapy planning [5] The high-depth images generated during PET/CT, MRI, or other types of scans

contain an abundance of information, invisible to the human eye, but usable in machine learning models. One of the ways of using this information is radiomics.

In radiomics, quantitative features are extracted from an expert-defined region of interest (ROI) in the image, describing various characteristics related to, among others, voxel intensity distributions, shape, or texture. Following successful standardization efforts [11], radiomic biomarkers are widely researched for use in cancer prognostics [4, 6, 10]. Usually, the modeling focuses on the primary tumor as a source of features and thus incorporates only one region of interest per patient. However, especially for locally advanced disease, the image may include multiple lesions, lymph nodes, or other affected regions, disregarded in the one patient — one ROI approach.

One example of cancer that is often already locally disseminated at diagnosis, is lung cancer, with most patients diagnosed at an advanced clinical stage of the disease. Combined with high metastatic potential, the late diagnosis contributes to low survival rates, with an estimated five-year survival of below 20% [3].

In our earlier work, we demonstrated the potential of radiomic features for predicting metastasis-free survival in non-small cell lung cancer (NSCLC) patients [2, 9]. Building upon this research, we investigate whether incorporating multiple ROIs from one patient in the model can improve its predictive ability. In a cross-validation procedure modified to account for the dependence of samples from one patient, we proposed and evaluated seven different methods of handling multiple ROIs. We show that although introducing multiple ROIs is associated with the higher noisiness of the data, using an appropriate strategy allows for more accurate prediction than using only one lesion.

2 Materials and Methods

2.1 Patient characteristics

The cohort consists of 115 NSCLC patients treated in the Maria Sklodowska-Curie National Research Institute of Oncology, Gliwice Branch between 2009 and 2017. The patients received combination treatment, with between one and six cycles of platinum-based chemotherapy followed by a 60–70 Gy total dose of radiotherapy. The primary endpoint of interest was metastasis-free survival (MFS), defined as the time between cancer diagnosis and detection of distant metastases or time to last follow-up otherwise. Therefore, only patients without distant tumor spread detectable at diagnosis were included in the study group.

Typically for lung cancer, the majority of patients were male. The median age at diagnosis was 61 years. Two-thirds of the cohort had squamous cell carcinoma, other represented subtypes were adenocarcinoma (24.3%) and large cell (7%). The left lung was slightly more prevalent as the tumor location. Most patients had advanced disease, with over 60% classified as T3 or T4 in the TNM staging, and over 80% exhibiting various degrees of lymph node spread. However, the general condition of the patients was good, as all but two patients were categorized as 0 or 1 on the Zubrod performance scale. Detailed information about the study group is presented in Table 1.

Table 1: Patient characteristics. For age, median and quartiles are given

Sex	Male	83 (72.2%)
	Female	32 (27.8%)
Age		61 (57-67)
Histopathology	Squamous	77(67.0%)
	Large cell	8 (7.0%)
	Adenocarcinoma	28 (24.3%)
	Other	2 (1.7%)
Location	Left	65 (56.5%)
	Right	50 (43.5%)
T	1	4 (3.5%)
	2	37 (32.2%)
	3	37 (32.2%)
	4	37 (32.2%)
N	0	19 (16.5%)
	1	6 (5.2%)
	2	83 (72.2%)
	3	7 (6.1%)
M	0	115 (100%)
	1	0 (0%)
Zubrod score	0	34 (29.6%)
	1	80 (69.6%)
	2	1 (0.9%)

2.2 Image acquisition

For 24 patients the planning PET/CT images were acquired using Philips GeminiGXL 16 (Philips, Amsterdam, Netherlands) scanner, and for 88 patients using Siemens Biograph mCT 128 (Siemens AG, Munich, Germany) scanner. Images were contoured by an experienced nuclear medicine specialist utilizing MIM 7.0.1 software and the PET Edge™ tool (both MIM Software Inc., OH, USA). In addition to the primary tumor, the ROIs included other tumor lesions in the lung and lymph nodes. Out of the entire cohort, 47 patients had only one ROI contoured, the remaining 68 had between two and eleven ROIs (Figure 1).

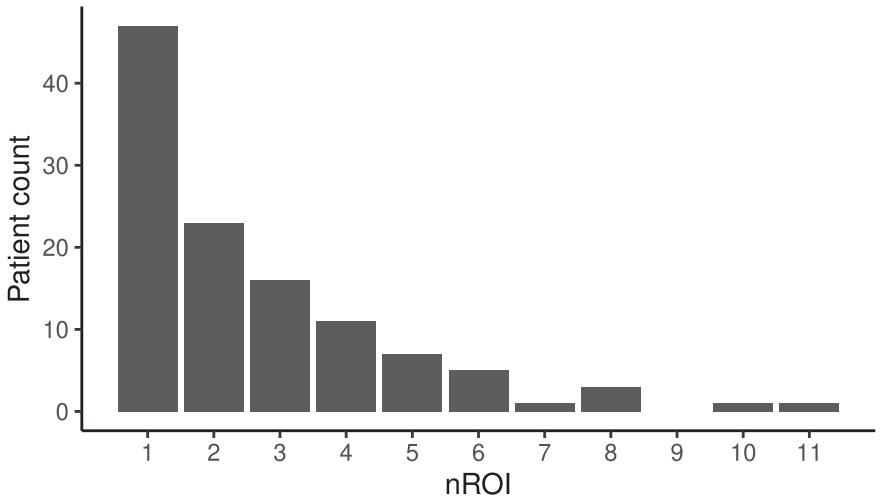


Fig. 1: Number of contoured ROIs per patient

2.3 Radiomic feature extraction

For each of the contoured ROIs, we extracted 105 radiomic features describing their shape, texture etc., using PyRadiomics version 3.0.1 [8]. The features can be divided into seven categories: first-order statistics (18 features), shape (14 features), Gray Level Co-occurrence Matrix (GLCM, 22 features), Gray Level Dependence Matrix (GLDM, 14 features), Gray Level Run Length Matrix (GLRLM, 16 features), Gray Level Size Zone Matrix (GLSZM, 16 features), and Neighboring-Gray Tone Difference Matrix (NGTDM, 5 features).

2.4 Survival analysis — general framework

To test the constructed models, we employed a 1000-iteration Monte Carlo Cross-Validation scheme with fixed subsets. Since the dataset contained multiple observations (ROIs) per patient, the subsets had to be generated in a way preventing information leakage. Therefore, in each iteration, the patients were divided into a training (two-thirds) and test (one-third) groups, stratified according to the metastasis status. The resulting training and test sets contained all ROIs from the assigned patients. While such partitioning ensured that all ROIs of a given patient were used either for training or for testing, and the number of patients in the subsets was always the same, the set cardinalities could be different in each iteration (depending on method) due to varying number of contoured lesions per patient.

To explore the association between radiomic features and the metastasis-free surival we applied the Regularized Cox Regression model (CoxNet) implemented

in the glmnet R package, version 4.1-6 [1, 7]. The optimal λ value was found using a cross-validation procedure. We assessed the predictive ability of the fitted models using the Harrell's c-index calculated for patients in the test set, adopting the median across all iterations as the final quality index.

2.5 Handling multiple ROIs

We focused on two main approaches to handling multiple ROIs:

1. selecting (or constructing) a single ROI per patient and using it to fit the model, or
2. fitting the model on a full dataset (containing all ROIs) and reconstructing patient risk from the risks obtained for the ROIs.

The advantage of the second idea is having a larger dataset to build the model. However, because of the substantial heterogeneity, it may also be more subject to data noise.

Between the two approaches, we tested a total of seven distinct methods of handling multiple ROIs. The overview of the tested strategies is presented in Figure 2.

randomROI In this approach, one ROI is randomly selected to represent a patient. It requires no defined order existing for the ROIs, but the results are largely dependent on chance and change with each realization of the method. Considering the pronounced differences between ROIs from the same patient, the instability makes it largely unsuitable for potential clinical use. Here it will be regarded as the baseline model for comparison.

largestROI Again, one ROI is selected to represent each patient, taking advantage of the intuitive order according to size (here defined as the voxel number of the mask). Selecting the largest ROI ensures that the result is identical each time the method is called. However, it does not utilize any information from the remaining ROIs. Indeed, since the largest of the lesions is usually the primary tumor, this method corresponds to the classic approach adopted in radiomics studies.

arithmeticMeanROI A representative ROI is constructed, where each feature is an arithmetic average calculated across all the patient's ROIs. Although this method incorporates information from all the contoured lesions, it puts into question the interpretability of obtained radiomic features. For example, shape features like major axis length have fairly straightforward interpretation when applied to one ROI, but are no longer meaningful when averaged.

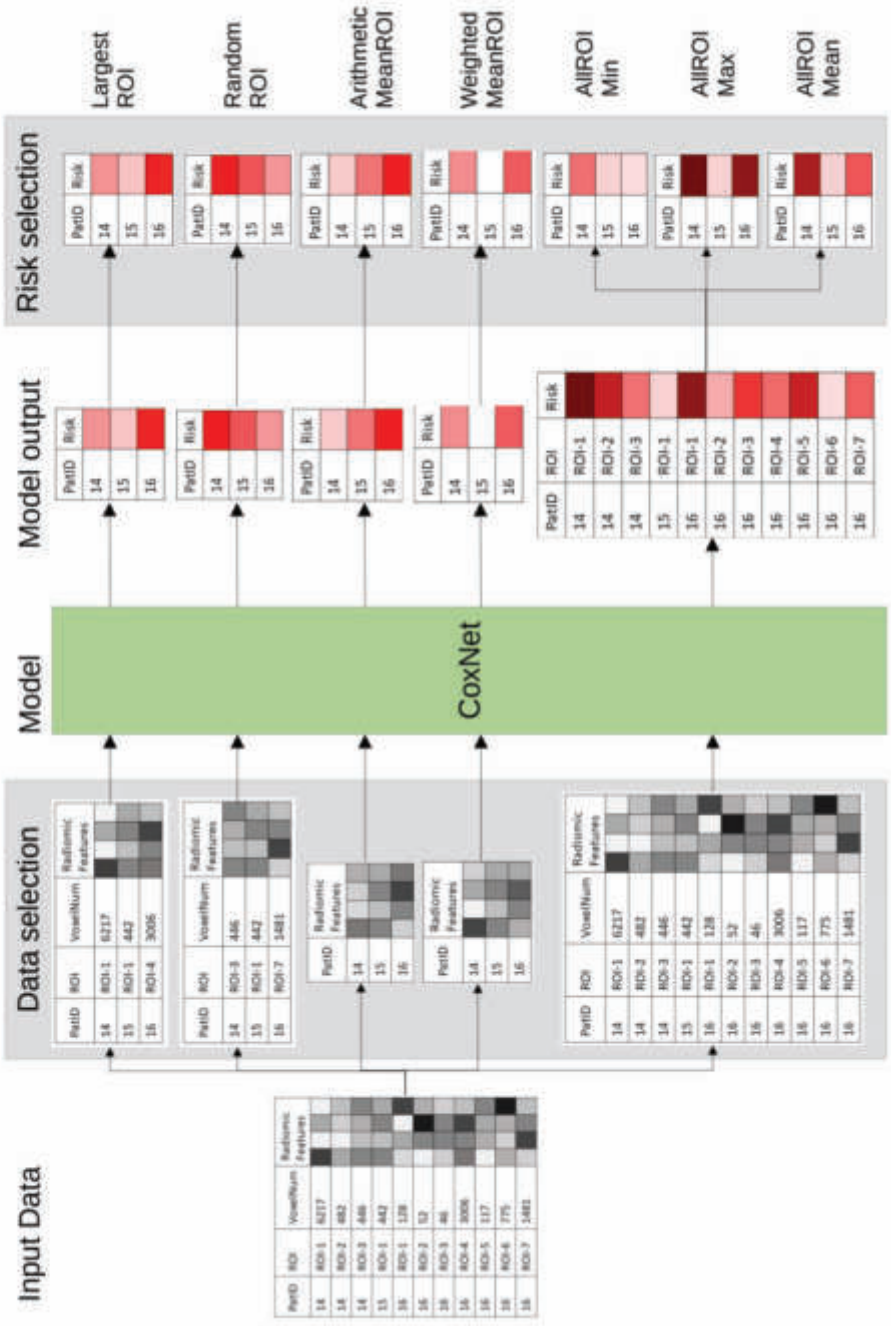


Fig. 2: Methods of handling multiple ROIs

weightedMeanROI Analogous to the previous method. The representative ROI is calculated as a weighted mean of all ROIs with their sizes (voxel numbers) as weights.

allROIMin In the subsequent three strategies, no selection takes place before training. The model is fitted using all available ROIs, resulting in a risk value assigned to each ROI. Thus, the methods require post-processing, in which the risk for the patient is derived. In the first version, the smallest risk is selected. As a consequence of such an approach, patients with multiple ROIs are more likely to be assigned smaller risks.

allROIMax Opposite to the previous method, the highest risk is chosen for the patient, resulting in patients with multiple ROIs generally displaying higher risks. Notably, the highest risk is not necessarily associated with the largest ROI.

allROIMean The risk for the patient is calculated as the mean of the risks of all the ROIs. Here, the risk is not directly dependent on the initial ROI number.

3 Results

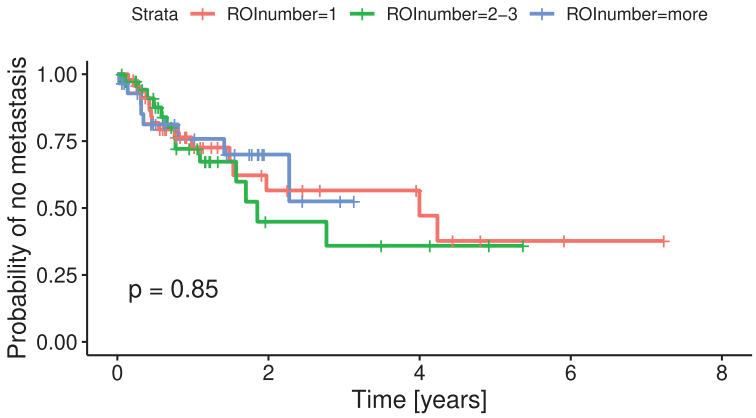
3.1 ROI characteristics

As expected, there is high intra-patient heterogeneity — the contoured ROIs had varying shapes, sizes, and even origin (lung tissue vs lymph nodes), which was reflected in the extracted radiomic features.

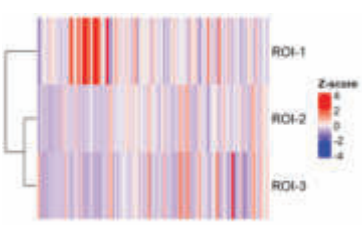
Perhaps surprisingly, there was no direct link between number of ROIs and MFS (Figure 3a). The p-value in the log-rank test between groups with one, two to three, and more contoured lesions was 0.85. However, some differences could be observed in the radiomic profiles between patients with long and short MFS, even with the same number of ROIs, as seen in Figures 3c and 3b. For clarity, features (depicted in columns) were not clustered.

3.2 Prediction of metastasis free survival

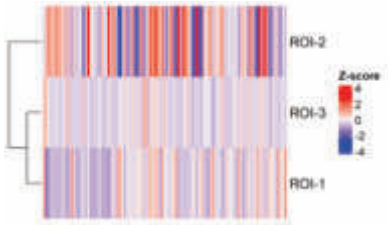
Although the values of the obtained c-indexes were highly dependent on iteration, which can be attributed to the small sample size, distinct trends emerged in performance of the different strategies. Predictably, the worst result (median c-index 0.534) was obtained by the *randomROI* method. The ,,classic" approach using only the largest ROI yielded better results (median c-index 0.581), but was still outperformed by the *weightedMeanROI* method incorporating information from multiple ROI (c-index 0.592). The *arithmeticMeanROI* worked worse (c-index 0.557), confirming that a good strategy for incorporating multiple samples is crucial as the model is susceptible to noise.



(a) Kaplan-Meier plot with patients stratified according to ROI number



(b) Radiomic features for a sample patient with long MFS (over four years) and three contoured ROIs



(c) Radiomic features for a sample patient with short MFS (below three months) and three contoured ROIs

Fig. 3: Characterization of the dataset with respect to ROIs

Among the methods using all available ROIs to fit the model, *allROIMin* had the poorest performance, giving a median c-index of 0.566. The other two methods, *allROIMax*, and *allROIMean* achieved the best overall performance, yielding median c-indexes of 0.617 and 0.616, respectively. Notably, the *allROIMean* approach displayed the lowest variability between subsets, emphasizing the added advantage of a larger training set associated with these approaches.

4 Discussion

Medical images are gaining popularity as a source of prognostic biomarkers in cancer due to the relative simplicity and low invasiveness of their acquisition. With great emphasis and effort toward reproducibility and standardization, radiomics provides a way to extract quantitative features from regions of interest in the image. The extracted variables, incorporated into predictive models such as Cox regression, can be used, for example, to assess the risk of distant metastasis.

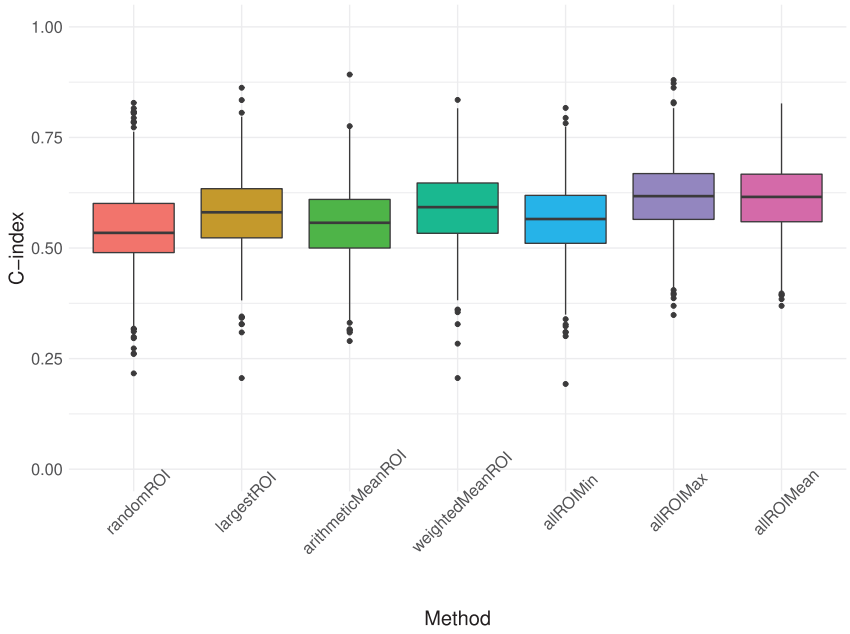


Fig. 4: Benchmark results. Boxplots depict c-indexes obtained in 1000 cross-validation iterations.

However, in malignancies like lung cancer, the disease is often locally disseminated at diagnosis. Hence, apart from the primary tumor, the regions of interest can include other tumor lesions or lymph nodes, resulting in an additional layer of information often overlooked in radiomic studies.

While the idea of utilizing the additional information in modeling is tempting, the exact strategy of handling multiple ROIs is a non-trivial task. It presents a set of unique challenges, from dataset structure, through inference of risk for a particular patient, to validation preventing information leakage. In an attempt to tackle these challenges, we performed a benchmark of seven different approaches, evaluating their performance in a tailored Monte Carlo Cross-Validation scheme.

On a cohort of 115 non-small cell lung cancer patients, we found that training the model on all available ROIs allowed for better prediction than using only the largest one. Setting the patient's risk to the maximum over all ROIs the median c-index in all iterations was 0.617, which was the highest reached quality. Also a representative ROI constructed as a weighted average outperformed the largestROI approach.

There are some limitations to the presented study, the most important being the limited cohort size. In every iteration, the training set contained only 77 patients, further reduced in the internal cross-validation within CoxNet. It resulted in a certain instability that could be observed in the large dispersion

between iterations. Furthermore, we tested only one type of survival model, a limitation we plan to address in the future.

5 Conclusion

In radiomics-based survival modeling, incorporating multiple ROIs per patient can improve the predictive ability, provided that an appropriate strategy is used. For our dataset, the best results were achieved when all ROIs were used for prediction and the risk for a particular patient was derived subsequently.

6 Acknowledgments

This work was supported by the Polish National Science Centre, grant number: UMO-2020/37/B/ST6/01959, and Silesian University of Technology statutory research funds. Calculations were performed on the Ziemowit computer cluster in the Laboratory of Bioinformatics and Computational Biology created in the EU Innovative Economy Programme POIG.02.01.00-00-166/08 and expanded in the POIG.02.03.01-00-040/13 project.

7 Author contributions

A.M.W. designed the computational study and methodology, performed the analysis and visualizations, and wrote the initial draft of the manuscript. D.B. extracted the radiomic features and edited the manuscript. E.K. participated in the study design, conceptualization and methodology, prepared the data for analysis and edited the manuscript. A.D'A. and I.G. assembled and prepared the imaging data. I.D.S. and R.S. provided and assembled the patient database. S.G. participated in the analysis and manuscript writing. K.F. provided administrative support. A.S. participated in the study design and conceptualization and supervised the project. All authors have read and approved the final version of the manuscript.

References

1. Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software **33**(1) (2010) 1–22
2. Fujarewicz, K., Wilk, A., Borys, D., d'Amico, A., Suwiński, R., Świerniak, A.: Machine learning approach to predict metastasis in lung cancer based on radiomic features. In Nguyen, N.T., Tran, T.K., Tukayev, U., Hong, T.P., Trawiński, B., Szczerbicki, E., eds.: Intelligent Information and Database Systems, Cham, Springer Nature Switzerland (2022) 40–50
3. Lu, T., Yang, X., Huang, Y., Zhao, M., Li, M., Ma, K., Yin, J., Zhan, C., Wang, Q.: Trends in the incidence, treatment, and survival of patients with lung cancer in the last four decades. Cancer Management and Research **Volume 11** (January 2019) 943–953

4. Reginelli, A., Nardone, V., Giacobbe, G., Belfiore, M.P., Grassi, R., Schettino, F., Del Canto, M., Grassi, R., Cappabianca, S.: Radiomics as a new frontier of imaging for cancer prognosis: A narrative review. Diagnostics **11**(10) (2021)
5. Saif, W., Tzannou, I., Makrilia, N., Syrigos, K.: Role and cost effectiveness of pet/ct in management of patients with cancer. The Yale journal of biology and medicine **83** (06 2010) 53–65
6. Shen, C., Liu, Z., Guan, M., Song, J., Lian, Y., Wang, S., Tang, Z., Dong, D., Kong, L., Wang, M., Shi, D., Tian, J.: 2d and 3d ct radiomics features prognostic performance comparison in non-small cell lung cancer. Translational Oncology **10**(6) (2017) 886–894
7. Simon, N., Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for cox's proportional hazards model via coordinate descent. Journal of Statistical Software **39**(5) (2011) 1–13
8. van Griethuysen, J.J., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R.G., Fillion-Robin, J.C., Pieper, S., Aerts, H.J.: Computational radiomics system to decode the radiographic phenotype. Cancer Research **77**(21) (October 2017) e104–e107
9. Wilk, A., Borys, D., Fujarewicz, K., d'Amico, A., Suwiński, R., Świerniak, A.: Potential of radiomics features for predicting time to metastasis in nsclc. In Nguyen, N.T., Tran, T.K., Tukayev, U., Hong, T.P., Trawiński, B., Szczerbicki, E., eds.: Intelligent Information and Database Systems, Cham, Springer Nature Switzerland (2022) 64–76
10. Zhang, Y., Oikonomou, A., Wong, A., Haider, M.A., Khalvati, F.: Radiomics-based prognosis analysis for non-small cell lung cancer. Scientific Reports **7**(1) (April 2017)
11. Zwanenburg, A., Vallièe res, M., Abdalah, M.A., Aerts, H.J.W.L., Andrearczyk, V., Apte, A., Ashrafinia, S., Bakas, S., Beukinga, R.J., Boellaard, R., Bogowicz, M., Boldrini, L., Buvat, I., Cook, G.J.R., Davatzikos, C., Depeursinge, A., Desseroit, M.C., Dinapoli, N., Dinh, C.V., Echegaray, S., Naqa, I.E., Fedorov, A.Y., Gatta, R., Gillies, R.J., Goh, V., Götz, M., Guckenberger, M., Ha, S.M., Hatt, M., Isensee, F., Lambin, P., Leger, S., Leijenaar, R.T., Lenkowicz, J., Lippert, F., Losnegård, A., Maier-Hein, K.H., Morin, O., Müller, H., Napel, S., Nioche, C., Orlhac, F., Pati, S., Pfaehler, E.A., Rahmim, A., Rao, A.U., Scherer, J., Siddique, M.M., Sijtsema, N.M., Fernandez, J.S., Spezi, E., Steenbakkers, R.J., Tanadini-Lang, S., Thorwarth, D., Troost, E.G., Upadhaya, T., Valentini, V., van Dijk, L.V., van Griethuysen, J., van Velden, F.H., Whybra, P., Richter, C., Löck, S.: The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping. Radiology **295**(2) (May 2020) 328–338

Towards the use of multiple ROIs for radiomics-based survival modelling: finding a strategy of aggregating lesions

Agata Małgorzata Wilk[a,b], Andrzej Swierniak[a], Andrea d'Amico[c], Rafał Suwiński[d], Krzysztof Fujarewicz[a], Damian Borys[a,*]

[a]*Department of Systems Biology and Engineering, Silesian University of Technology, Akademicka 16, Gliwice 44-100, Poland*
[b]*Department of Biostatistics and Bioinformatics, Maria Sklodowska-Curie National Research Institute of Oncology, Gliwice Branch, Wybrzeze AK 15, Gliwice 44-102, Poland*
[c]*Department of Nuclear Medicine and Endocrine Oncology, PET Diagnostics Unit, Maria Sklodowska-Curie National Research Institute of Oncology, Gliwice Branch, Wybrzeze AK 15, Gliwice 44-102, Poland*
[d]*II-nd Radiotherapy and Chemotherapy Clinic, Maria Sklodowska-Curie National Research Institute of Oncology, Gliwice Branch, Wybrzeze AK 15, Gliwice 44-102, Poland*

Abstract

The main objective of this work is to explore the possibility of incorporating radiomic information from multiple lesions into survival models. We hypothesise that when more lesions are present, their inclusion can improve model performance, and we aim to find an optimal strategy for using multiple distinct regions in modelling.

The idea of using multiple regions of interest (ROIs) to extract radiomic features for predictive models has been implemented in many recent works. However, in almost all studies, analogous regions were segmented according to particular criteria for all patients — for example, the primary tumour and peritumoral area, or subregions of the primary tumour. They can be included in a model in a straightforward way as additional features. A more interesting scenario occurs when multiple distinct ROIs are present, such as multiple lesions in a regionally disseminated cancer. Since the number of such regions may differ between patients, their inclusion in a model is non-trivial and requires additional processing steps.

We proposed several methods of handling multiple ROIs representing either ROI or risk aggregation strategy, compared them to a published one, and evaluated their performance in different classes of survival models in a Monte-Carlo Cross-Validation scheme. We demonstrated the effectiveness of the methods using a cohort of 115 non-small cell lung cancer patients, for whom we predicted the metastasis risk based on features extracted from PET images in original resolution or interpolated to CT image resolution. For both feature sets, incorporating all available lesions, as opposed to a singular ROI representing the primary tumour, allowed for considerable improvement of predictive ability regardless of the model.

Keywords:
multiple ROIs, radiomics, survival models, ROI aggregation, risk aggregation

1. Introduction

Non-invasive, quick to achieve, and accurate imaging is a staple of modern medicine. With various available modalities, it allows for observing the structural, metabolic, and biochemical state of the patient's body,

*Corresponding author: Tel.: +48-32-237-11-59; fax: +48-32-237-16-55;
Email addresses: agata.wilk@polsl.pl (Agata Małgorzata Wilk), damian.borys@polsl.pl (Damian Borys)

proving invaluable for diagnostics, therapy planning and management. In addition to their standard, visual application, resulting digital images can be treated as data used in machine learning models [1]. Although deep learning can be applied directly to raw images, a major obstacle is the cohort size necessary to construct a believable model [2]. A compromise solution is radiomics, where numerical features are first extracted from a specific part of the image. These can be fed into statistical models, achieving better performance compared to clinical or radiological features [3]. The growing interest in the field, related among others to oncology [4, 5] can largely be attributed to consistent efforts to make the extracted features robust and reproducible [6].

One of the challenges in radiomics is the definition of the region of interest (ROI). Typically, in cancer-related studies, an intuitive region is the tumour itself, however, various other ROI delineations have been explored. In [7], radiomic features were extracted from intra-tumour region as well as the peritumoural area and applied for survival risk prediction in early-stage lung cancer. Xu et al. [8] presented a diagnostic approach utilizing five different ROIs in two types of breast ultrasound images, including the whole tumour region, the strongest perfusion region, and the surrounding region. In [9], radiation-induced skin toxicity was investigated using six types of ROIs based on radiation doses. Shan et al. [10] used ROIs corresponding to the lesion and the peritumoural area to predict recurrence in hepatocellular carcinoma. Chen et al. [11] predicted metastasis in rectal cancer based on ROIs delineating tumour area, peritumoural fat, and the largest pelvic lymph node. Chen et al. [12] described a deep learning model for MGMT promoter methylation prediction using radiomic features extracted from the whole tumour region and the tumour core. In [13], three ROIs of different sizes created around the nodule center served for differentiation between benign and malignant thyroid nodules. Han et al. [14] used features extracted from the tumour zone and the tumour-liver interface to predict predominant histopathological growth patterns of colorectal liver metastases. In [15], radiomic features extracted from MR images for two types of ROIs (Facet and Circle) were used to identify axial spondyloarthritis. In [16], axillary lymph node (ALN) metastasis status in breast cancer was predicted using radiomics extracted from tumour and ALN ROIs in MRI and mammography images. Dammak et al. [17] distinguished between lung cancer recurrence and benign radiation-induced injury based on six semi-automatically contoured ROIs (each initialised with a RECIST line drawn by a specialist). In [18], immunotherapy response in non-small cell lung cancer was predicted based on radiomic features extracted from three subregions of the primary tumour obtained through k-means clustering. Hou et al. [19] construct a model predicting neutrophil-lymphocite ratio for lung cancer patients based on CT radiomics from five anatomical regions. A work by Zhang et al. [20] demonstrates the influence of peritumoural margin on the performance of radiomics models. A later study [21] describes a model combining deep learning and conventional radiomics for ROIs comprising the tumour region and peritumoural area. Similarly, in [22], radiomic and deep features are integrated to classify breast lesions. In [23], ROIs resulting from radiotherapy planning are considered — gross tumour volume, planning tumour volume, and the difference between the two. Even in studies focusing on extraction of deep features, multiple ROI delineations are used, such as in [24], where the VOI used to predict nodal metastasis in lung cancer is divided into sections representing tumour core and peritumoral area. Finally, Huang et al [25] predict acute coronary syndrome based on two ROIs — pericoronary adipose tissue and atherosclerotic plaques. These studies show that including other ROIs can improve the predictive ability of models. Still, their inclusion in the model remains relatively straightforward since they can be determined for each of the patients and accordingly labelled, introducing no ambiguity.

However, in many cases the tumour is already locally spread [26], and in addition to the primary tumour there are also secondary lesions and involved lymph nodes. All these structures may constitute separate regions of interest and carry information valuable for modelling. Yet, as the number of ROIs may differ for each patient, there is no trivial way to incorporate them all in one model. Beyond medical imaging, this problem can easily be extended to other aspects of diagnostics and prognostics where multiple measurements per patient might occur, such as blood tests, biopsies or tissue samples. In the work [27] dedicated to mass spectrometry imaging, it was discussed that a "single-pixel" approach (using all available data for training the model) allowed for better classification quality than a "mean spectrum" approach.

Expanding on this idea, we take lung cancer as an example, because it is often diagnosed at a late, locally advanced stage and involves a high probability of multiple ROIs being present. It has already been shown [28, 29, 30, 31] that radiomic features extracted from the primary tumour have the potential for predicting the risk of metastatic dissemination. Here we investigate whether incorporating information from additional available ROIs allows for more accurate prediction. While many works, as mentioned above, present models including different (often multiple) *definitions* of the region of interest, the problem of using actually distinct regions, whose number is inconsistent between patients, is addressed much less frequently. Notably, Zhao et al. [32] proposed a method of ROI aggregation based on "meta histograms" and applied it for PET/CT radiomics in lung adenocarcinoma to train classification models corresponding to 3- and 4-year overall survival, as well as tumour grade and risk. Nevertheless, the method was not compared to a single-ROI approach. We perform the comparison for a time-to-event problem, for the "meta histogram" strategy as well as original ROI or risk aggregation methods. We test several different classes of survival models in a Monte Carlo validation scheme. Moreover, considering the differences in radiomic profiles of patients' ROIs, we hypothesise that the inter-ROI heterogeneity might be a better predictor of early metastasis than simply the number of ROIs itself.

2. Materials and methods

The retrospectively collected study cohort, described in detail elsewhere [29], consists of 115 Polish non-small cell lung cancer patients treated in the Maria Sklodowska-Curie National Research Institute of Oncology Gliwice Branch. The study was approved by the institutional Ethics Committee (KB/430-48/23) and the data were anonymized before the analysis.

The endpoint considered in this study was metastasis-free survival (MFS), defined as the time from diagnosis to detection of distant metastases (event) or the last screening without detectable distant metastases (censored observations). It is important to note, that as implied by this definition, censoring from the MFS point of view does not coincide with the patient's death.

2.1. PET/CT image acquisition and processing

As part of radiotherapy planning, PET/CT images were acquired using Philips GeminiGXL 16 (Philips, Amsterdam, Netherlands) (24 patients) and Siemens Biograph mCT 131 (Siemens AG, Munich, Germany) (91 patients). An experienced nuclear medicine expert contoured all the radiologically changed regions within the lung. Taking into consideration the validity of texture features, we excluded regions smaller than two voxels. Using PyRadiomics v3.1.0 [33] with bin width set to 0.1, we extracted 100 radiomic features for each region of interest (ROI). Before calculating the features we standardized the PET images using body weight (SUVbw) and used the images both in the original resolution (PET dataset) and interpolated to the CT resolution with the nearest neighbour algorithm (PET_CT dataset). None of the additional filters were used in the pre-processing step.

2.2. Inter-ROI heterogeneity

To check whether the number of ROIs is related to MFS, we divided the patients into three groups of similar cardinalities — with only one contoured ROI, with two to three ROIs, and with four or more ROIs. We compared the Kaplan-Meier curves for the resulting groups using the log-rank test.

To assess the level of inter-ROI diversity, several concerns had to be addressed. Firstly, some of the radiomic features can take negative values, which excludes certain popular dissimilarity measures, in particular entropy-based indices. In addition, the number of ROIs was not consistent between patients. We decided to employ several distance- and correlation-based indices, listed below:
Canberra distance. The Canberra distance $d_C(x, y)$ is defined as:

$$d_C(x, y) = \sum_{i=1}^{n} \frac{|x_i - y_i|}{|x_i| + |y_i|} \tag{1}$$

Euclidean distance. The Euclidean distance $d_E(x, y)$ is defined as:

$$d_E(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{2}$$

Minkowski distance. The Minkowski distance $d_M(x, y)$ is defined as:

$$d_M(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}} \tag{3}$$

Kendall distance. The Kendall distance $d_K(x, y)$, based on the Kendall correlation coefficient, is defined as:

$$d_K(x, y) = 1 - |\tau| = 1 - \left| \frac{2}{n(n-1)} \sum_{i<j} sgn(x_i - x_j)\,\mathrm{sgn}(y_i - y_j) \right| \tag{4}$$

Spearman distance. The Spearman distance $d_S(x, y)$, based on the Spearman correlation coefficient, is defined as:

$$d_S(x, y) = 1 - |\rho| = 1 - \left| \frac{cov(R(x), R(y))}{\sigma_{R(x)}\sigma_{R(y)}} \right| \tag{5}$$

where $R(x)$ and $R(y)$ are x and y converted to rank vectors. Due to the vast differences in magnitude levels of the radiomic features, Pearson's correlation coefficient was impractical for this application.

For a particular patient, the heterogeneity index is equal to 0 if there is only one ROI, and to the average value for all distinct pairs of ROIs otherwise. For consistency against the number of ROIs, the patients were again stratified into three groups, according to terciles of the heterogeneity index values. Naturally, in our cohort, the one-ROI group and the low-heterogeneity group for any given index are equal.

2.3. Methods of handling multiple ROIs

Generally, survival-type models take as input a collection of observations (patients, objects), each of them consisting of a feature vector, in our case radiomic features for the ROI, and the response, which is a pair of values denoting time-to-event and status (event or censoring). After fitting, the model can be used for new observations, assigning a risk value corresponding to a new feature vector. Utilizing multiple feature vectors, non-equivalent between patients, requires aggregation of either ROIs before the model training or risks thereafter. A schematic illustration of the methods is shown in Fig. 1. They are listed and described in the following Section.

2.3.1. Aggregation of ROIs

The main idea of ROI aggregation approaches is to select or construct a single ROI to represent each patient. The dataset consists of fewer data points, but the models can be used as is, and the results require no further processing.

largestROI. In the case of medical imaging, the ROIs can be intuitively ordered according to their size, that is the number of voxels. The largest one, usually corresponding to the primary tumour, is a default choice for the modelling, and as such will be treated as the reference method. Although it is a straightforward approach, its major disadvantage is the loss of information from the remaining ROIs.

randomROI. For each patient, a random ROI is selected. Since the result is largely dependent on chance, it may differ for each realization of the method. Taking into account considerable variability between ROIs for a single patient, the instability it introduces in the model makes it practically unsuitable for clinical application. Nevertheless, this method is useful when there exists no intrinsic ordering of the measurements.

largestROIdiversityIndex. For each patient, the heterogeneity indices described above are calculated based on all the ROIs. The five resulting variables are concatenated to the radiomic feature vector from the largest ROI, and can subsequently be selected within the model. This method utilizes, to some extent, information from all the available data points.

randomROIdiversityIndex. Similar to *largestROIdiversityIndex*, except diversity indices are considered together with the radiomic features extracted from a randomly selected ROI.

arithmeticMeanROI. A representative ROI $\hat{x}$ is constructed as the arithmetic average of all corresponding ROIs. Although it uses information from all contoured regions, in this method many radiomic features, particularly shape features, lose their interpretability.

weightedMeanROI. The representative ROI is calculated as the weighted mean of all regions scaled by the volumes of the ROIs.

MetaHistogram. The method proposed by Zhao et al.[32]. Briefly, for each patient a "meta histogram" is constructed — the "bins" are values of a particular radiomic feature for ROIs arranged in descending order of size. Next, multi-lesion features are extracted as characteristics of the "meta histogram": mean, variance, sum, skewness, kurtosis, energy, and entropy. Since this approach results in a large number of features, correlation-based redundancy filtering was first employed with a 0.9 threshold.

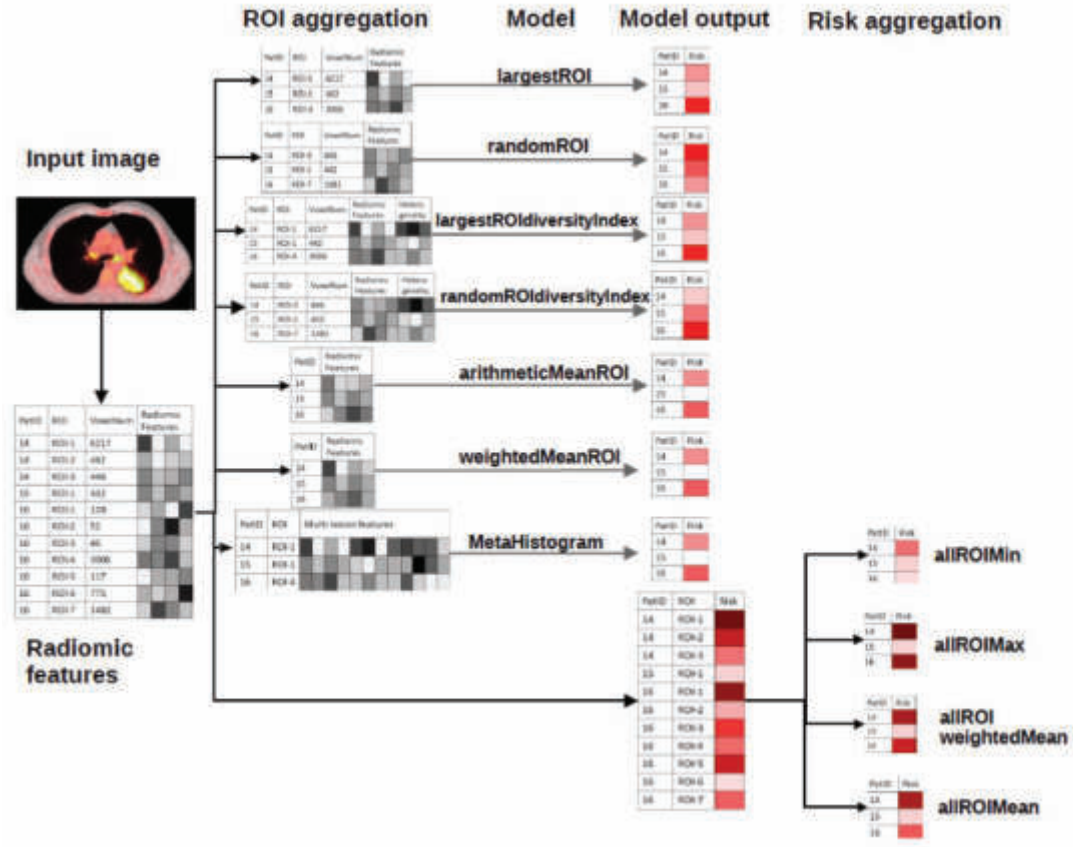


Figure 1: Methods of handling multiple ROIs. Radiomic features (represented by grey rectangles) are extracted from the input image for each region of interest. In the ROI aggregation methods, a representative ROI is constructed. The ROIs serve as input for a survival model, whose output is a risk score (represented by red rectangles). In risk aggregation methods, risk scores are aggregated to obtain a single value for every patient. See a more detailed description in the text.

2.3.2. *Aggregation of risks*

In the following methods, all available patient ROIs are used for fitting. When the trained model is applied for prediction, each ROI is assigned its own risk, resulting in multiple scores per patient. Only then, aggregation is performed on the risk values. Since all ROIs are included in the dataset, more data points are available for training. However, high levels of intra-patient heterogeneity can contribute to noise in the data.

allROIMin. The smallest risk is chosen for each patient.

allROIMax. Opposite to the previous method, the highest risk is chosen.

allROIWeightedMean.The risk for the patient is an average of all the risks for the corresponding ROIs weighted with ROI volumes.

allROIMean. Analogous to the previous method, but the patient's risk is calculated as an arithmetic average of ROI risks.

2.4. *Survival models*

Survival models are used in a setting where the response variable consists of pairs (t_i, δ_i), where δ_i represents the censoring status, and t_i is the time to event if $\delta_i = 1$ or time to last follow-up if $\delta_i = 0$. Typically,

the output of such models can be interpreted as a relative risk — the higher the risk score, the sooner the event should occur. Several survival models, representing different classes, usually coupled with some variable selection strategy, were included in the analysis.

CoxStepAIC. The classic Cox proportional hazards regression, with forward stepwise selection of variables according to the Akaike Information Criterion (AIC). Due to its popularity in survival analysis, it is treated as the reference model.

Coxnet. Regularized Cox regression, with embedded feature selection. The optimal λ value was determined in cross-validation. The R package *glmnet* was used [34, 35].

Weibull. Model-based boosting (using *mboost* package) with Weibull accelerated failure time (AFT) model [36, 37].

Loglog. Model-based boosting with Loglog AFT model.

Lognormal. Model-based boosting with Lognormal AFT model.

randomForest. A random survival forest model grown with 1000 trees. The package *RandomForestSRC* was used [38, 39].

SVMregression. A survival support vector machine [40] with a regression model and an additive kernel function. The package *survivalsvm* was used [41].

SVMvanbelle1. A survival support vector machine with a Van Belle model and an additive kernel function.

2.5. Model quality evaluation

Since for the risk aggregation methods the dataset contained multiple ROIs from the same patient, the validation scheme had to be specifically constructed to prevent information leakage. This would be a situation when ROIs from the same patient were placed in the training set and test set in the same iteration. To avoid it, and to ensure unbiased comparisons between models and aggregation methods, a tailored Monte Carlo cross-validation partitioning was constructed, and fixed for all the schemes.

In each of 1000 iterations, the patients were partitioned into the training group (2/3) and the test group (1/3). Then, depending on the ROI handling method, the training and test sets were constructed to contain either the representative ROIs for all patients from the corresponding group or all ROIs for patients from the corresponding group. Thus, in every iteration and scheme, the same patient groups were used for training and testing, albeit in the risk aggregation models the actual sizes of the training and test sets differ.

3. Statistical analysis

We employed Harrel's c-Index to evaluate the predictive ability of the models [42]. Defined as a ratio of concordant observation pairs (where higher risk corresponds to shorter time-to-event) to all comparable pairs, it can be viewed as a survival analysis counterpart of the area under the ROC curve (AUC) in classification. Indeed, the value 0.5 indicates a completely random model, while 1 is a perfect model. Because of the large sample sizes (1000 iterations constituting a group), we used Cohen's d effect size as an indicator of the strength of the difference against the reference scheme. The standard interpretation was assumed, with $|d| < 0.2$ treated as a negligible effect, $0.2 \leq |d| < 0.5$ meaning small effect, $0.5 \leq |d| < 0.8$ medium effect and otherwise large effect.

For comparison of survival between two or more groups, we used the log-rank test, assuming the standard significance threshold $p = 0.05$.

All statistical analyses were conducted using the R environment for statistical computing, version 4.1.3 (R Core Team, 2022).

4. Results

4.1. Patient characteristics

The clinical characteristics of the study population were consistent with a typical lung cancer cohort, with the majority of patients being male and diagnosed in an advanced stage of the disease (according to the TNM classification). The patients differed in the number of contoured regions of interest, ranging from one to ten (see Fig. 2). In nearly 60% cases, multiple ROIs were present.

4.2. Inter-ROI heterogeneity related to early metastasis

In the analysed cohort, there was no significant difference in metastasis-free survival probability between the groups based on ROI number (Fig. 3a). In contrast, for the PET_CT dataset, the two subgroups with more than one ROI differed significantly for some of the diversity

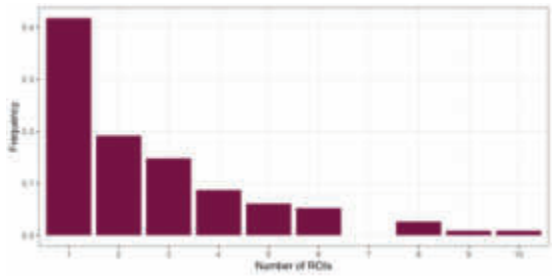


Figure 2: Characterization of the cohort with respect to ROI number

indices, with higher heterogeneity related to early metastasis. The lowest p-value equal to 0.026 was obtained for the Euclidean distance (Fig. 3b). It shows that the diversity indices can be used as variables in a survival model for MFS prediction. However, the Kaplan-Meier curve for the low-heterogeneity group was positioned between the other two, suggesting that the inclusion of other features is necessary for the generality of the model.

4.3. Comparison of approaches to handling multiple ROIs and survival models

For the PET dataset, the model quality varied both between models and ROI handling approaches (Fig. 4a), with median c-Indices between 0.456 for the combination *largestROIdiversityIndex* + *SVMvanbelle1* and 0.632 for *MetaHistogram* + *randomForest*. For the reference scheme, *largestROI* + *CoxStepAIC*, the median c-Index was 0.554. 7 schemes performed better with large effect size, 20 performed better with medium effect size, 37 performed better with small effect size, 3 performed worse with small effect size, 1 performed worse with medium effect size, and 2 performed worse with large effect size. For the rest, the effect size was negligible (Fig. 4b).

randomForest, *Loglog*, and *Weibull* generally yielded high c-Indices. Interestingly, for the *largestROI* method, all but one model were better than the reference (*CoxStepAIC*). Although adding the diversity indices improved the *largestROI* method only for two models (*randomForest* and *SVMregression*), it was beneficial for the *randomROI* method, improving the performance of all but one model.

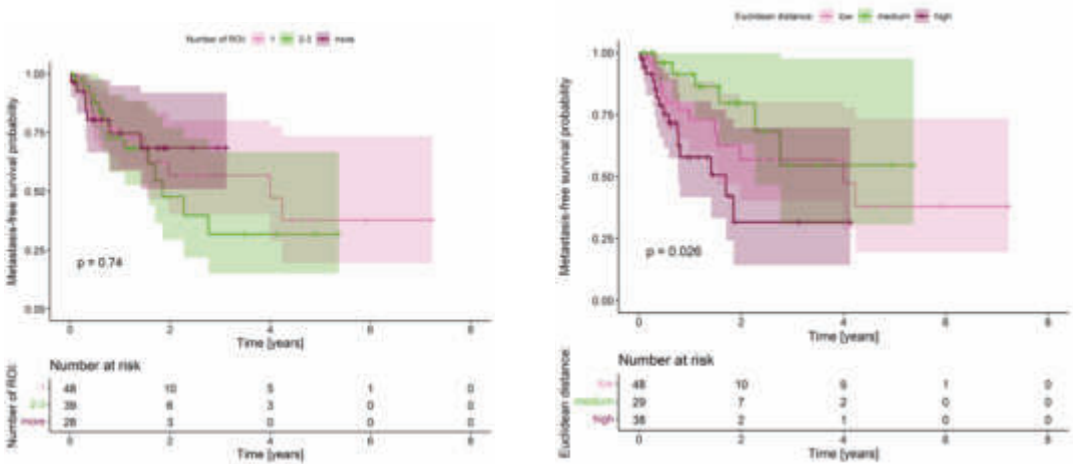
The proposed methods of handling multiple ROIs allowed for improving the predictive ability of the models. Unsurprisingly, *randomROI* worked worse than the *largestROI* approach for every model but *SVMvanbelle1*. Among the ROI aggregation methods, *arithmeticMeanROI* achieved the highest overall median c-indexes — 0.629 for *CoxStepAIC* and 0.630 for *Weibull*. Although not as effective, *weightedMeanROI* method was more consistent across different models, outperforming *largestROI* for all of them. Risk aggregation, with the exception of *allROIMin*, also yielded good results, particularly for the Weibull model — 0.623 for *allROIMean* and 0.625 for *allROIMax*. The latter method achieved higher c-Indices than *largestROI* for all tested models.

The schemes also differed in the distribution of c-Indices across the cross-validation iterations (Fig 4c). The minimum c-Indices were between 0.090 for the scheme *weightedMeanROI* + *SVMvanbelle1* and 0.364 for *allROIMax* + *randomForest*, while the maximum c-Indices ranged from 0.733 for *largestROI* + *SVMvanbelle1* to 0.941 for *randomROIdiversityIndex* + *SVMregression*. Since c-Indices below 0.5 indicate a performance worse than random, such values for the test set likely stem from overfitting. Schemes *MetaHistogram* + *randomForest* *allROIMean* + *Weibull*, *weightedMean* + *randomForest*, and *allROIMax* + *Weibull* were most robust to overfitting, as all of them had c-Index < 0.5 in fewer than 60 out of 1000 iterations. A complete summary of the results is given in the Supplementary Table 2.

For the PET_CT dataset, similar tendencies could be observed with large differences between schemes (Fig. 5a); the median c-Indices varied between 0.498 for the combination *largestROIdiversityIndex* + *SVMvanbelle1* and 0.634 for *allROIMax* + *randomForest*. For the reference scheme, *largestROI* + *CoxStepAIC*, the median c-Index was 0.540. 6 schemes performed better with large effect size, 14 performed better with medium effect size, 35 performed better with small effect size, and 5 performed worse with small effect size. For the rest, the effect size was negligible (Fig. 5b).

The *randomForest* model consistently outperformed others regardless of the ROI handling approach, with the sole exception of the *MetaHistogram* approach. For the *largestROI* method, all but one model was better than the reference (*CoxStepAIC*). Once again, the worst performing model for most ROI handling methods was *SVMvanbelle1*.

Also for this dataset, the results obtained for the *largestROI* method could be improved by incorporating mul-



(a) (b)

Figure 3: Predictive value of the inter-ROI heterogeneity. (a) Kaplan-Meier plot of metastasis-free survival for patients divided according to the number of ROIs. (b) Kaplan-Meier plot of metastasis-free survival for patients divided according to the heterogeneity index of ROIs (euclidean distance, PET resolution)

tiple ROIs. Risk aggregation methods seemed to be more suitable — *allROIMax* achieved better c-Indices for all models, *allROIMean* and *allROIWeightedMean* performed similarly or better than *largestROI*.

The schemes displayed different robustness levels observed in the distribution of c-Indices across the cross-validation iterations (Fig 5c). The minimum c-Indices were between 0.062 for the scheme *weightedMeanROI + SVMvanbelle1* and 0.357 for *allROIMax + randomForest*, while the maximum c-Indices ranged from 0.762 for *weightedMeanROI + SVMvanbelle1* to 0.922 for *allROIMax + SVMregression*. Schemes *allROIMax + randomForest*, *allROIWeightedMean + randomForest*, *allROIMean + randomForest*, and *weightedMeanROI + randomForest* were most robust to overfitting, achieving c-Indices lower than 0.5 in 50, 60, 61 and 61 iterations, respectively. A complete summary of the results is given in the Supplementary Table 3.

5. Discussion

Radiomics is gaining popularity in cancer research as a source of biomarkers, both in diagnostic and prognostic settings [4, 43, 44, 45, 46]. In addition to the imaging data, extraction of radiomic features requires a defined region of interest. In many cases, the cancer is already locally disseminated at diagnosis - for example in the US, from 2011 to 2020, 23.3% of lung cancer cases were diagnosed at a localized stage, 21.1% at a regional stage, and 48.3 % at a distant stage [26]. Even in a regionally advanced cancer, the number of distinct lesions that can be considered regions of interest is greater than one, and can include lymph nodes and secondary tumours.

These ROIs carry information that can prove prognostically useful but is non-trivial to include in the modelling. In this work, we answer the question of whether including all available regions improves the performance of survival models. Since they are distinct foci, not different ways of contouring the primary tumour, their number and nature vary between patients. Therefore, dedicated methods must be developed to account for patient heterogeneity and relationships between ROIs. To explore the possibility of incorporating multiple ROIs into survival-type models, we collected a cohort of 115 non-metastatic non-small cell lung cancer patients, with between one and ten contoured ROIs. We used radiomic features extracted from the ROIs to predict metastasis-free survival, employing different regression models and proposing various strate-

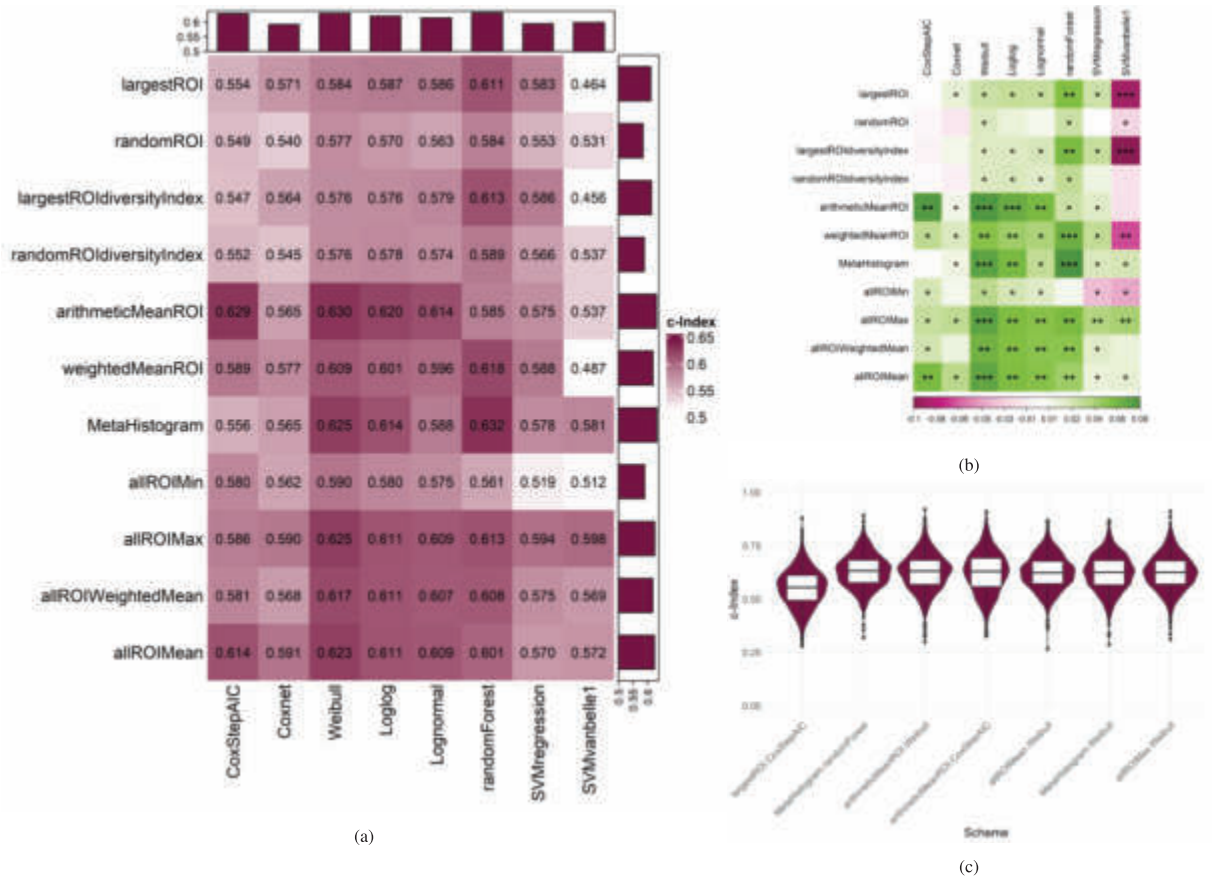


Figure 4: Results of the comparison for the PET dataset. (a) Median c-Indices of all the tested schemes. The top and right annotations show the best median c-Indices for the models and ROI handling methods, respectively. (b) Comparison of the schemes with the reference model (largestROI+CoxStepAIC). Colours represent differences between median c-Indices, and stars show Cohen's d effect size, ( ) meaning negligible, (*) — small, (**) — medium, and (***) — large. (c) c-Indices of the selected schemes across 1000 iterations.

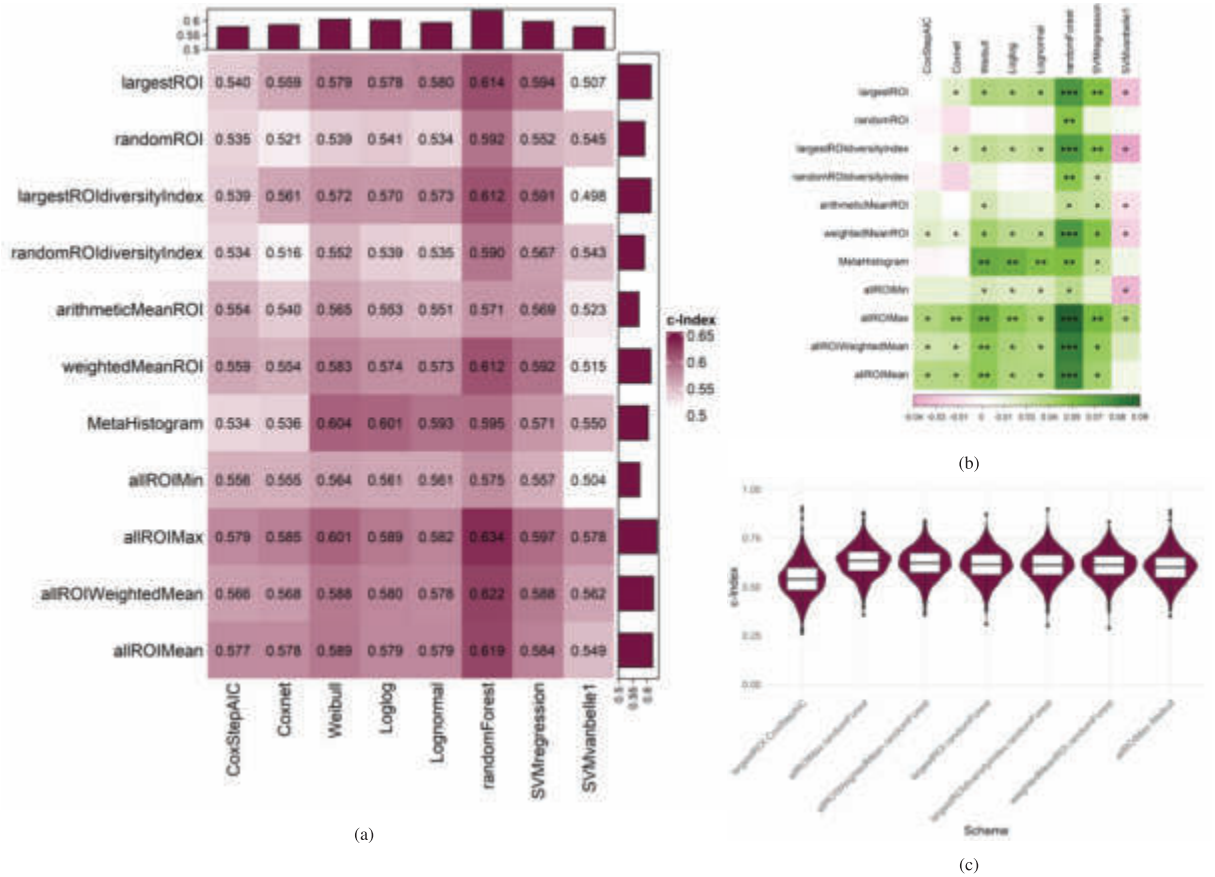


Figure 5: Results of the comparison for the PET_CT dataset. (a) Median c-Indices of all the tested schemes. The top and right annotations show the best median c-Indices for the models and ROI handling methods, respectively. (b) Comparison of the schemes with the reference model (largestROI+CoxStepAIC). Colours represent differences between median c-Indices, and stars show Cohen's d effect size, ( ) meaning negligible, (*) — small, (**) — medium, and (***) — large. (c) c-Indices of the selected schemes across 1000 iterations.

gies of multiple ROI handling. Additionally, we evaluated the relation between inter-ROI heterogeneity and time-to-metastasis.

We have shown that for patients with more than one ROI, their number alone was not significantly related to the metastasis-free survival probability (see Fig. 3a). However, for the PET dataset, the groups created by dividing patients by heterogeneity indices differed in MFS — higher inter-ROI heterogeneity was related to a higher risk of metastasis (Figure 3b). Therefore, heterogeneity indices, synthetic features combining information from all the patient's ROIs, can be used as prognostic variables in the models. Still, the trend is disturbed, as the Kaplan-Meier curve for a group of patients with only one ROI, and consequently low heterogeneity, was between the two others. For a more generally useful model, the heterogeneity indices must be augmented with other radiomic features.

The prediction quality, evaluated in terms of Harrell's c-Index, was highly dependent on the used survival model (Figures 5, 3). While the classic approach, consisting of the Cox proportional hazards regression with stepwise selection based on AIC performed quite well, other models achieved better predictions — random survival forest achieved higher c-Index for each ROI handling method. Ensemble models, such as model-based boosting and the random forest, deserve attention for their performance. They model the data well without overfitting, as seen on the test data, where they rarely yielded c-Indices below 0.5 (Supplementary Tables 2 and 3).

The ROI aggregation methods performed better for the PET dataset. Adding diversity indices to radiomic features allowed for improving some largest-ROI models, and many randomROI models. Even better results were obtained for the *arithmeticMeanROI* and *weightedMeanROI*, which outperformed the *largestROI* method for most models. This confirms the benefit of incorporating information from all lesions, especially since in ROI aggregation methods the improvement cannot be explained by a larger data set used for training.

Very good overall results for both datasets were achieved by the risk aggregation methods except for allROIMin. An intuitive explanation of this poor performance can be found in the inter-ROI heterogeneity analysis. Patients with highly diverse ROIs will likely be assigned both very high and very low risks. After risk aggregation, they will be given the lowest value, contrary to what is presented in Fig. 3b, where the high-heterogeneity group exhibits more and earlier events than the other two. The opposite strategy implemented in the allROIMax method is consistent with the findings from the heterogeneity analysis, resulting in the highest c-Indices of all the methods.

The *MetaHistogram* method, introduced by Zhao and coworkers [32], improved the model performance compared to the largest ROI particularly in the PET dataset. However, since the endpoint of this study was related to metastasis, our cohort was limited to cases where cancer was regionally advanced at most. Consequently, the number of ROIs was relatively smaller than in the original work introducing the concept —- in fact, slightly over 40% of patients had only one contoured lesion. In such cases, the constructed "meta histogram" consists of only one bin, which makes some of the extracted features uninformative.

This study is not without limitations, the greatest being only one, relatively small cohort used for comparing the schemes. A search was conducted for an external imaging dataset to validate the findings. While there are studies providing multiple nodules per patient [47] or survival data [48], to the authors' best knowledge no public dataset exists with both available. Also, different possible ROIs have been explored [8, 9, 10, 2, 11, 13, 14, 17, 15, 16, 18, 20, 21, 23, 19], usually related to either narrowing or extending the tumour area, for instance tumour core, tumour margin, peritumoural area etc. Including such regions in a machine learning model is relatively straightforward, since they can be identified for every patient and accordingly labelled, introducing no ambiguity. In our work, we understand multiple regions of interest as independent uptakes in the PET images, as opposed to the elsewhere adopted idea based on different ways of contouring or defining new types of ROI. In one study also involving multiple actual lesions, the method of extracting multi-lesion radiomic features was used for a classification problem [32]. Nevertheless, we included the described strategy (*MetaHistogram*) in our comparison demonstrating that for certain datasets it can indeed improve the performance of survival models. Overall, the presented results can neither be directly compared to published studies nor verified in an external cohort. It should however be noted, that the aim of the present study is not to present a solution to a particular clinical problem (in

this case the prediction of metastasis risk) but to investigate whether incorporating information from multiple lesions can improve model performance.

6. Conclusion

We proposed several methods of incorporating information from multiple lesions/regions of interest in radiomics-based survival models. We evaluated their performance with different types of models for the prediction of metastasis-free survival, based on PET images from a cohort of 115 non-small cell lung cancer patients. Regardless of the model, the predictive ability (assessed in terms of Harrell's c-Index in 1000 validation iterations) could be improved by using aggregation of ROIs or risks, compared to a model built using only the primary tumour. Furthermore, including information from all lesions improved the robustness of the models.

Using all available feature vectors in a structurally heterogneous dataset to fit survival models can be considered a promising direction not only for imaging studies, but also for molecular research, in particular spatial or single-cell analysis.

Acknowledgments

This work was supported by the Polish National Science Centre, grant number: UMO-2020/37/B/ST6/01959 (AS), Medical Research Agency, grant number: 07/010/ABB23/1037 (DB) and Silesian University of Technology statutory research funds (KF). AMW was also supported by a subsidy for the maintenance and development of research potential BKM. Calculations were performed on the Ziemowit computer cluster in the Laboratory of Bioinformatics and Computational Biology created in the EU Innovative Economy Programme POIG.02.01.00-00-166/08 and expanded in the POIG.02.03.01-00-040/13 project.

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data and code availability

The data used in this work and source code are available upon reasonable request.

Author contributions

Agata Małgorzata Wilk: Conceptualization, Methodology, Software, Formal Analysis, Investigation, Visualization, Writing — Original Draft. **Andrzej Swierniak**: Conceptualization, Supervision, Project administration, Funding acquisition, Writing — Review & editing. **Andrea d'Amico**: Investigation. **Rafał Suwinski**: Resources. **Krzysztof Fujarewicz**: Project administration, Supervision, Writing — Review & editing. **Damian Borys**: Data curation, Software, Investigation, Visualization, Writing — Review & editing.

References

[1] R. J. Gillies, P. E. Kinahan, H. Hricak, Radiomics: Images are more than pictures, they are data, Radiology 278 (2016) 563–577. doi:10.1148/radiol.2015151169, pMID: 26579733.

[2] X. Chen, X. Wang, K. Zhang, et al., Recent advances and clinical applications of deep learning in medical image analysis, Medical Image Analysis 79 (2022) 102444. URL: https://www.sciencedirect.com/science/article/pii/S1361841522000913. doi:https://doi.org/10.1016/j.media.2022.102444.

[3] H. J. Park, B. Park, S. Y. Park, et al., Preoperative prediction of postsurgical outcomes in mass-forming intrahepatic cholangiocarcinoma based on clinical, radiologic, and radiomics features, European Radiology 31 (2021) 8638–8648. URL: https://doi.org/10.1007/s00330-021-07926-6. doi:10.1007/s00330-021-07926-6.

[4] A. Tagliafico, M. Piana, D. Schenone, et al., Overview of radiomics in breast cancer diagnosis and prognostication, The Breast 49 (2020) 74–80. URL: https://www.sciencedirect.com/science/article/pii/S0960977619305922.

doi:https://doi.org/10.1016/j.breast.2019.10.018.

[5] M. R. Chetan, F. V. Gleeson, Radiomics in predicting treatment response in non-small-cell lung cancer: current status, challenges and future perspectives, European Radiology 31 (2020) 1049–1058. URL: https://doi.org/10.1007/s00330-020-07141-9. doi:10.1007/s00330-020-07141-9.

[6] A. Zwanenburg, M. Vallières, M. A. Abdalah, et al., The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping, Radiology 295 (2020) 328–338. doi:10.1148/radiol.2020191145, pMID: 32154773.

[7] T. Wang, Y. She, Y. Yang, et al., Radiomics for survival risk stratification of clinical and pathologic stage ia pure-solid non–small cell lung cancer, Radiology 302 (2022) 425–434. URL: https://doi.org/10.1148/radiol.2021210109. doi:10.1148/radiol.2021210109, pMID: 34726531.

[8] Z. Xu, Y. Wang, M. Chen, et al., Multi-region radiomics for artificially intelligent diagnosis of breast cancer using multimodal ultrasound, Computers in Biology and Medicine 149 (2022) 105920. URL: https://www.sciencedirect.com/science/article/pii/S001048252200662X. doi:https://doi.org/10.1016/j.compbiomed.2022.105920.

[9] H. Feng, H. Wang, L. Xu, et al., Prediction of radiation-induced acute skin toxicity in breast cancer patients using data encapsulation screening and dose-gradient-based multi-region radiomics technique: A multicenter study, Frontiers in Oncology 12 (2022). doi:10.3389/fonc.2022.1017435.

[10] Q. Shan, H. Hu, S. Feng, et al., CT-based peritumoral radiomics signatures to predict early recurrence in hepatocellular carcinoma after curative tumor resection or ablation, Cancer Imaging 19(1) (2019). doi:10.1186/s40644-019-0197-5.

[11] L.-D. Chen, J.-Y. Liang, H. Wu, et al., Multiparametric radiomics improve prediction of lymph node metastasis of rectal cancer compared with conventional radiomics, Life Sciences 208 (2018) 55–63. URL: https://www.sciencedirect.com/science/article/pii/S0024320518303849. doi:https://doi.org/10.1016/j.lfs.2018.07.007.

[12] S. Chen, Y. Xu, M. Ye, et al., Predicting mgmt promoter methylation in diffuse gliomas using deep learning with radiomics, Journal of Clinical Medicine 11 (2022). URL: https://www.mdpi.com/2077-0383/11/12/3445. doi:10.3390/jcm11123445.

[13] H. Zhou, Y. Jin, L. Dai, et al., Differential diagnosis of benign and malignant thyroid nodules using deep learning radiomics of thyroid ultrasound images, European Journal of Radiology 127 (2020) 108992. URL: https://www.sciencedirect.com/science/article/pii/S0720048X20301819. doi:https://doi.org/10.1016/j.ejrad.2020.108992.

[14] Y. Han, F. Chai, J. Wei, et al., Identification of predominant histopathological growth patterns of colorectal liver metastasis by multi-habitat and multi-sequence based radiomics analysis, Frontiers in Oncology 10 (2020). URL: https://www.frontiersin.org/articles/10.3389/fonc.2020.01363. doi:10.3389/fonc.2020.01363.

[15] P. Xin, Q. Wang, R. Yan, et al., Assessment of axial spondyloarthritis activity using a magnetic resonance imaging-based multi-region-of-interest fusion model, Arthritis Research & Therapy 25 (2023). URL: http://dx.doi.org/10.1186/s13075-023-03193-6. doi:10.1186/s13075-023-03193-6.

[16] Q. Wang, Y. Lin, C. Ding, et al., Multi-modality radiomics model predicts axillary lymph node metastasis of breast cancer using mri and mammography, European Radiology (2024). URL: http://dx.doi.org/10.1007/s00330-024-10638-2. doi:10.1007/s00330-024-10638-2.

[17] S. Dammak, S. Gulstene, D. A. Palma, et al., Distinguishing recurrence from radiation-induced lung injury at the time of recist progressive disease on post-sabr ct scans using radiomics, Scientific Reports 14 (2024). URL: http://dx.doi.org/10.1038/s41598-024-52828-4. doi:10.1038/s41598-024-52828-4.

[18] J. Peng, D. Zou, X. Zhang, et al., A novel subregional radiomics model to predict immunotherapy response in non-small cell lung carcinoma, Journal of Translational Medicine 22 (2024). doi:10.1186/s12967-024-04904-6.

[19] R. Hou, W. Xia, C. Zhang, et al., Dosiomics and radiomics improve the prediction of post-radiotherapy neutrophil-lymphocyte ratio in locally advanced non-small cell lung cancer, Medical Physics 51 (2024) 650–661. doi:https://doi.org/10.1002/mp.16829.

[20] X. Zhang, G. Zhang, X. Qiu, et al., Optimizing the Size of Peritumoral Region for Assessing Non-Small Cell Lung Cancer Heterogeneity Using Radiomics, Springer Nature Singapore, 2023, p. 309–320. URL: http://dx.doi.org/10.1007/978-981-99-7108-4_26. doi:10.1007/978-981-99-7108-4_26.

[21] X. Zhang, G. Zhang, X. Qiu, et al., Exploring non-invasive precision treatment in non-small cell lung cancer patients through deep learning radiomics across imaging features and molecular phenotypes, Biomarker Research 12 (2024). URL: http://dx.doi.org/10.1186/s40364-024-00561-5. doi:10.1186/s40364-024-00561-5.

[22] X. Zhang, C. Liang, D. Zeng, et al, Pattern classification for breast lesion on ffdm by integration of radiomics and deep features, Computerized Medical Imaging and Graphics 90 (2021) 101922. URL: https://www.sciencedirect.com/science/article/pii/S0895611121000719. doi:https://doi.org/10.1016/j.compmedimag.2021.101922.

[23] T. Nie, Z. Chen, J. Cai, et al., Integration of dosimetric parameters, clinical factors, and radiomics to predict symptomatic radiation pneumonitis in lung cancer patients undergoing combined immunotherapy and radiotherapy, Radiotherapy and Oncology 190 (2024) 110047. doi:https://doi.org/10.1016/j.radonc.2023.110047.

[24] Y.-W. Wang, C.-J. Chen, H.-C. Huang, et al., Dual energy ct image prediction on primary tumor of lung cancer for nodal metastasis using deep learning, Computerized Medical Imaging and Graphics 91 (2021) 101935. URL: https://www.sciencedirect.com/science/article/pii/S0895611121000847. doi:https://doi.org/10.1016/j.compmedimag.2021.101935.

[25] Y. Huang, J. Yang, Y. Hou, et al., Automatic prediction of acute coronary syndrome based on pericoronary adipose tissue and atherosclerotic plaques, Computerized Medical Imaging and Graphics 108 (2023) 102264. URL: https://www.sciencedirect.com/science/article/pii/S0895611123000824. doi:https://doi.org/10.1016/j.compmedimag.2023.102264.

[26] N. C. I. Surveillance Research Program, Seer*explorer: An interactive website for seer cancer statistics, 2023. URL: https://seer.cancer.gov/statistics-network/explorer/, data source(s): SEER Incidence Data, November 2022 Submission (1975-2020),2023 Apr 19. [cited 2023 Jun 7].

[27] A. Kurczyk, M. Gawin, M. Chekan, et al., Classification of thyroid tumors based on mass spectrometry imaging of tissue microarrays; a single-pixel approach, International Journal of Molecular Sciences 21 (2020). URL: https://www.mdpi.com/1422-0067/21/17/6289. doi:10.3390/ijms21176289.

[28] A. Wilk, D. Borys, K. Fujarewicz, et al., Potential of radiomics features for predicting time to metastasis in nsclc, in: N. T. Nguyen, T. K. Tran, U. Tukayev, T.-P. Hong, B. Trawiński, E. Szczerbicki (Eds.), Intelligent Information and Database

Systems, Springer Nature Switzerland, Cham, 2022, pp. 64–76.

[29] A. M. Wilk, E. Kozłowska, D. Borys, et al., Radiomic signature accurately predicts the risk of metastatic dissemination in late-stage non-small cell lung cancer, Translational Lung Cancer Research 12 (2023) 1372–1383. doi:10.21037/tlcr-23-60.

[30] K. Fujarewicz, A. Wilk, D. Borys, et al., Machine learning approach to predict metastasis in lung cancer based on radiomic features, in: N. T. Nguyen, T. K. Tran, U. Tukayev, T.-P. Hong, B. Trawiński, E. Szczerbicki (Eds.), Intelligent Information and Database Systems, Springer Nature Switzerland, Cham, 2022, pp. 40–50.

[31] T. P. Coroller, P. Grossmann, Y. Hou, et al., Ct-based radiomic signature predicts distant metastasis in lung adenocarcinoma, Radiotherapy and Oncology 114 (2015) 345–350. URL: https://www.sciencedirect.com/science/article/pii/S0167814015001073. doi:https://doi.org/10.1016/j.radonc.2015.02.015.

[32] M. Zhao, K. Kluge, L. Papp, et al., Multi-lesion radiomics of pet/ct for non-invasive survival stratification and histologic tumor risk profiling in patients with lung adenocarcinoma, European Radiology 32 (2022) 7056–7067. URL: http://dx.doi.org/10.1007/s00330-022-08999-7. doi:10.1007/s00330-022-08999-7.

[33] J. J. van Griethuysen, A. Fedorov, C. Parmar, et al., Computational radiomics system to decode the radiographic phenotype, Cancer Research 77 (2017) e104–e107. doi:10.1158/0008-5472.can-17-0339.

[34] J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent, Journal of Statistical Software 33 (2010) 1–22. URL: https://www.jstatsoft.org/v33/i01/. doi:10.18637/jss.v033.i01.

[35] N. Simon, J. Friedman, T. Hastie, et al., Regularization paths for cox's proportional hazards model via coordinate descent, Journal of Statistical Software 39 (2011) 1–13. URL: https://www.jstatsoft.org/v39/i05/. doi:10.18637/jss.v039.i05.

[36] B. Hofner, A. Mayr, N. Robinzonov, et al., Model-based boosting in R: A hands-on tutorial using the R package mboost, Computational Statistics 29 (2014) 3–35.

[37] T. Hothorn, P. Buehlmann, T. Kneib, et al., mboost: Model-Based Boosting, 2022. URL: https://CRAN.R-project.org/package=mboost, R package version 2.9-7.

[38] H. Ishwaran, U. Kogalur, Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC), 2023. URL: https://cran.r-project.org/package=randomForestSRC, r package version 3.2.1.

[39] H. Ishwaran, U. Kogalur, E. Blackstone, et al., Random survival forests, Ann. Appl. Statist. 2 (2008) 841–860. URL: https://arXiv.org/abs/0811.1645v1.

[40] V. Van Belle, K. Pelckmans, S. Van Huffel, et al., Improved performance on high-dimensional survival data by application of survival-SVM, Bioinformatics 27 (2010) 87–94. URL: https://doi.org/10.1093/bioinformatics/btq617. doi:10.1093/bioinformatics/btq617.

[41] C. Fouodo, survivalsvm: Survival Support Vector Analysis, 2018. URL: https://CRAN.R-project.org/package=survivalsvm, r package version 0.0.5.

[42] F. Harrell, K. Lee, D. Mark, Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors, Statistics in Medicine 15 (1996) 361–387. URL: https://doi.org/10.1002/(sici)1097-0258(19960229)15:4<361::aid-sim168>3.0.co;2-4. doi:10.1002/(sici)1097-0258(19960229)15:4<361::aid-sim168>3.0.co;2-4.

[43] G. Wu, Y. Chen, Y. Wang, et al., Sparse representation-based radiomics for the diagnosis of

brain tumors, IEEE Transactions on Medical Imaging 37 (2018) 893–905. doi:10.1109/TMI.2017.2776967.

[44] Z. Liu, S. Wang, D. Dong, et al., The applications of radiomics in precision diagnosis and treatment of oncology: Opportunities and challenges, Theranostics 9 (2019) 1303–1322. URL: https://doi.org/10.7150/thno.30309. doi:10.7150/thno.30309.

[45] A. d'Amico, D. Borys, I. Gorczewska, Radiomics and artificial intelligence for pet imaging analysis, Nuclear Medicine Review 23 (2020) 36 – 39. URL: https://journals.viamedica.pl/nuclear_medicine_review/article/view/NMR.2020.0005. doi:10.5603/NMR.2020.0005.

[46] H. Homayoun, W. Y. Chan, T. Y. Kuzan, et al., Applications of machine-learning algorithms for prediction of benign and malignant breast lesions using ultrasound radiomics signatures: A multi-center study, Biocybernetics and Biomedical Engineering 42 (2022) 921–933. URL: https://www.sciencedirect.com/science/article/pii/S0208521622000687. doi:https://doi.org/10.1016/j.bbe.2022.07.004.

[47] S. G. Armato III, G. McLennan, L. Bidaut, et al., The lung image database consortium (lidc) and image database resource initiative (idri): A completed reference database of lung nodules on ct scans, Medical Physics 38 (2011) 915–931. URL: https://aapm.onlinelibrary.wiley.com/doi/abs/10.1118/1.3528204. doi:https://doi.org/10.1118/1.3528204.

[48] H. J. W. L. Aerts, E. R. Velazquez, R. T. H. Leijenaar, et al., Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach, Nature Communications 5 (2014). URL: https://doi.org/10.1038/ncomms5006. doi:10.1038/ncomms5006.

Oświadczenia Współautorów

Dla prac wieloautorskich wchodzących w skład cyklu przedstawiono oświadczenia podpisane przez ich Autorów korespondencyjnych.

Gdańsk

29.05.2024 dnia .....

OŚWIADCZENIE

O WKŁADZIE AUTORÓW

Tytuł pracy: p38 Mediates Resistance to FGFR Inhibition in Non-Small Cell Lung Cancer

Czasopismo, rok publikacji: Cells. 2021; 10(12):3363. DOI: 10.3390/cells10123363

Wykaz autorów wraz z udziałem procentowym:

Lp.	Imię i nazwisko	Udział procentowy
1	Izabela Zarczyńska	17
2	Monika Górska-Arcisz	8
3	Alexander J Cortez	7
4	Katarzyna A Kujawa	7
5	Agata M Wilk	7
6	Andrzej C Składanowski	2
7	Aleksandra Stanczak	2
8	Monika Skupińska	2
9	Maciej Wieczorek	3
10	Katarzyna M Lisowska	5
11	Rafał Sadej*	15
12	Kamila Kitowska*	25

* Autor (autorzy) korespondencyjny

Wkład merytoryczny doktorantki:

1. Analiza bioinformatyczna zmienności liczby kopii – ekstrakcja cech genomicznych z surowych danych dla macierzy aCGH, preprocessing (normalizacja, korekcja GC i cy3/cy5, centrowanie), estymacja liczby kopii, analiza różnicowa.
2. Analiza bioinformatyczna danych RNAseq – preprocessing (analiza jakości, usuwanie adapterów, mapowanie i annotacja, wyznaczenie macierzy zliczeń), normalizacja, analiza różnicowa.
3. Analiza nienadzorowana i wizualizacja danych z eksperymentów wysokoprzepustowych.
4. Analiza szlaków sygnałowych – GSEA.

Katedra Enzymologii i Onkologii Molekularnej
Gdański Uniwersytet Medyczny
prof. dr hab. Rafał Sadej

Podpisał prof. Rafał Sadej (podpis odręczny)

.....

Podpis Współautora

Podpisała Kamila Kitowska (podpis odręczny)

Ustroń, dnia 18.05.2024

OŚWIADCZENIE
O WKŁADZIE AUTORÓW

Tytuł pracy: Profiling of serum metabolome of breast cancer: multi-cancer features discriminate between healthy women and patients with breast cancer

Czasopismo, rok publikacji: Frontiers in Oncology, 2024, 14:1377373. DOI: 10.3389/fonc.2024.1377373

Wykaz autorów wraz z udziałem procentowym:

Lp.	Imię i nazwisko	Udział procentowy
1	Katarzyna Mrowiec	15
2	Julia Debik	10
3	Karol Jelonek	10
4	Agata Kurczyk	10
5	Lucyna Ponge	5
6	Agata Wilk	10
7	Marcela Krzempek	5
8	Guro F. Giskeødegård	10
9	Tone F. Bathen	10
11	Piotr Widłak*	15

* Autor (autorzy) korespondencyjny

Wkład merytoryczny doktorantki:

1. Uczenie i walidacja modeli uczenia maszynowego, badanie stabilności selekcji cech oraz wpływu hiperparametrów.
2. Opracowanie schematu testowania modeli w celu zbadania generalizowalności dla różnych kohort.
3. Udział w przygotowaniu manuskryptu.

Podpisał prof. Piotr Widłak (podpis odręczny)

Podpis Współautora

Gliwice....., dnia 24.05.2024

OŚWIADCZENIE
O WKŁADZIE AUTORÓW

Tytuł pracy: Impact of government policies on the COVID-19 pandemic unraveled by mathematical modeling.

Czasopismo, rok publikacji: Scientific Reports, 2022, 12, 16987. DOI: 10.1038/s41598-022-21126-2

Wykaz autorów wraz z udziałem procentowym:

Lp.	Imię i nazwisko	Udział procentowy
1	Agata Wilk*	70
2	Krzysztof Łakomiec	15
3	Krzysztof Psiuk-Maksymowicz	5
4	Krzysztof Fujarewicz*	10

* Autor (autorzy) korespondencyjny

Wkład merytoryczny doktorantki:

1. Przygotowanie zbioru danych – pozyskanie z zewnętrznych baz danych (zachorowania, obostrzenia, populacje), odfiltrowanie braków, zapis w formacie odpowiednim do dalszej analizy.
2. Estymacja wpływu obostrzeń – implementacja modelu wspólnego, niezależnych i indywidualizowanych.
3. Walidacja modeli na testowym przedziale czasowym – estymacja liczby zachorowań z modelu SEIR, ocena błędu dopasowania.
4. Wizualizacja wyników.
5. Przygotowanie manuskryptu.

Podpisał prof. Krzysztof Fujarewicz (podpis odręczny)

.....

Podpis Współautora

Ustroń, dnia 18.05.2024

OŚWIADCZENIE
O WKŁADZIE AUTORÓW

Tytuł pracy: Classification of Thyroid Tumors Based on Mass Spectrometry Imaging of Tissue Microarrays; a Single-Pixel Approach

Czasopismo, rok publikacji: International Journal of Molecular Sciences, 2020, 21(17), 6289; DOI: 10.3390/ijms21176289

Wykaz autorów wraz z udziałem procentowym:

Lp.	Imię i nazwisko	Udział procentowy
1	Agata Kurczyk	15
2	Marta Gawin	15
3	Mykola Chekan	5
4	Agata Wilk	10
5	Krzysztof Łakomiec	8
6	Grzegorz Mrukwa	2
7	Katarzyna Frątczak	2
8	Joanna Polańska	5
9	Krzysztof Fujarewicz	10
10	Monika Pietrowska	13
11	Piotr Widłak*	15

* Autor (autorzy) korespondencyjny

Wkład merytoryczny doktorantki:

1. Analiza nienadzorowana (PCA) zbioru danych.
2. Opracowanie schematu walidacji modeli uczenia maszynowego uwzględniającego wielopoziomowe powiązania pomiędzy próbkami.
3. Implementacja podejścia „single-pixel" umożliwiającego wykorzystanie nierównolicznych zbiorów wektorów cech do klasyfikacji.
4. Wizualizacja.

Podpisał prof. Piotr Widłak (podpis odręczny)

Podpis Współautora

Gliwice, dnia 5.06.2024

OŚWIADCZENIE
O WKŁADZIE AUTORÓW

Tytuł pracy: Radiomic signature accurately predicts the risk of metastatic dissemination in late-stage non-small cell lung cancer.

Czasopismo, rok publikacji: Translational lung cancer research, 2023. 12(7), 1372–1383. DOI: 10.21037/tlcr-23-60

Wykaz autorów wraz z udziałem procentowym:

Lp.	Imię i nazwisko	Udział procentowy
1	Agata Wilk	55
2	Emilia Kozłowska*	20
3	Damian Borys	10
4	Andrea D'Amico	2
5	Krzysztof Fujarewicz	2
6	Izabela Gorczewska	1
7	Iwona Dębosz-Suwińska	1
8	Rafał Suwiński	2
9	Jarosław Śmieja	2
10	Andrzej Świerniak*	5

* Autor (autorzy) korespondencyjny

Wkład merytoryczny doktorantki:

1. Analiza nienadzorowana zbioru danych ---- PCA, analiza korelacji.
2. Statystyka opisowa, analiza jednoczynnikowa.
3. Przewdywanie przeżycia wolnego od przerzutów i wolnego od zdarzeń z wykorzystaniem metod klasyfikacji.
4. Konstrukcja finalnego regresyjnego modelu przeżycia.
5. Przygotowanie pierwszej wersji manuskryptu.

Podpisał prof. Andrzej Świerniak (podpis odręczny)

Podpis Współautora

Gliwice, dnia 5.06.2024

OŚWIADCZENIE
O WKŁADZIE AUTORÓW

Tytuł pracy: Improving the Predictive Ability of Radiomics-Based Regression Survival Models Through Incorporating Multiple Regions of Interest.

Czasopismo, rok publikacji: The Latest Developments and Challenges in Biomedical Engineering. PCBEE 2023. Lecture Notes in Networks and Systems, vol 746. Springer, Cham. DOI: 10.1007/978-3-031-38430-1_13

Wykaz autorów wraz z udziałem procentowym:

Lp.	Imię i nazwisko	Udział procentowy
1	Agata Wilk	80
2	Emilia Kozłowska	2
3	Damian Borys	9
4	Andrea D'Amico	1
5	Izabela Gorczewska	1
6	Iwona Dębosz-Suwińska	1
7	Seweryn Gałecki	1
8	Krzysztof Fujarewicz	1
9	Rafał Suwiński	1
10	Andrzej Świerniak*	3

* Autor (autorzy) korespondencyjny

Wkład merytoryczny doktorantki:

1. Konceptualizacja.
2. Metodyka – opracowanie i implementacja metod agregacji.
3. Analiza statystyczna i testowanie modeli przeżycia.
4. Wizualizacja.
5. Przygotowanie manuskryptu

Podpisał prof. Andrzej Świerniak (podpis odręczny)

Podpis Współautora

Gliwice, dnia 7.06.2024

OŚWIADCZENIE
O WKŁADZIE AUTORÓW

Tytuł pracy: Towards the use of multiple ROIs for radiomics-based survival modelling: finding a strategy of aggregating lesions

Czasopismo, rok publikacji: arXiv preprint, 2024. arXiv: 2405.17668 [stat.AP].

Wykaz autorów wraz z udziałem procentowym:

Lp.	Imię i nazwisko	Udział procentowy
1	Agata Wilk	80
2	Andrzej Świerniak	5
3	Andrea d'Amico	3
4	Rafał Suwiński	1
5	Krzysztof Fujarewicz	1
6	Damian Borys*	10

* Autor (autorzy) korespondencyjny

Wkład merytoryczny doktorantki:

1. Konceptualizacja — określenie problemu badawczego, przegląd istniejących rozwiązań, zaprojektowanie badania.
2. Metodologia —opracowanie i implementacja metod agregacji.
3. Analiza statystyczna.
4. Walidacja, testowainie modeli przeżycia.
5. Wizualizacja.
6. Przygotowanie manuskryptu.

Podpisał prof. Damian Borys (podpis odręczny)

Podpis Współautora

Dorobek naukowy Autorki

Artykuły w czasopismach:

1. Karło A, Wilk A, Ziembińska-Buczyńska A, Surmacz-Gorska J. Cultivation Parameters Adjustment for Effective Algal Biomass Production. Rocznik Ochrona Środowiska, 17(cz. 1),275-288, (2015)

2. Kurczyk A, Gawin M, Chekan M, Wilk A, Łakomiec K, Mrukwa G, Frątczak K, Polanska J, Fujarewicz K, Pietrowska M i Widlak P. Classification of Thyroid Tumors Based on Mass Spectrometry Imaging of Tissue Microarrays; a Single-Pixel Approach. International journal of molecular sciences, 21(17), 6289, (2020)

3. Cortez AJ, Kujawa KA, Wilk AM, Sojka DR, Syrkis JP, Olbryt M, Lisowska KM. Evaluation of the Role of ITGBL1 in Ovarian Cancer. Cancers. 12(9):2676. (2020)

4. Swoboda R, Giebel S, Knopińska-Posłuszny W, Chmielowska E, Drozd-Sokołowska J, Paszkiewicz-Kozik E, Kulikowski W, Taszner M, Mendrek W, Najda J, Czerw T, Olszewska-Szopa M, Czyż A, Giza A, Spychałowicz W, Subocz E, Szwedyk P, Krzywon A, Wilk A, Zaucha JM. High efficacy of BGD (bendamustine, gemcitabine, and dexamethasone) in relapsed/refractory Hodgkin Lymphoma. Ann Hematol, 100, 1755–1767 (2021)

5. Zarczynska I, Gorska-Arcisz M, Cortez AJ, Kujawa KA, Wilk AM, Skladanowski AC, Stanczak A, Skupinska M, Wieczorek M, Lisowska KM, Sadej R, Kitowska K. p38 Mediates Resistance to FGFR Inhibition in Non-Small Cell Lung Cancer. Cells. Nov 30;10(12):3363. (2021)

6. Zeman M, Skałba W, Wilk AM, Cortez AJ, Maciejewski A, Czarniecka A. Impact of renin-angiotensin system inhibitors on the survival of patients with rectal cancer. BMC Cancer 22, 815 (2022)

7. Płuciennik A, Płaczek A, Wilk A, Student S, Oczko-Wojciechowska M, Fujarewicz K. Data Integration–Possibilities of Molecular and Clinical Data Fusion on the Example of Thyroid Cancer Diagnostics. International Journal of Molecular Sciences. 23(19):11880 (2022)

8. Wilk AM, Łakomiec K, Psiuk-Maksymowicz K i Fujarewicz K. Impact of government policies on the COVID-19 pandemic unraveled by mathematical modelling. Scientific Reports 12, 16987 (2022)

9. Wilk AM, Kozłowska E, Borys D, D'Amico A, Fujarewicz K, Gorczewska I, Debosz-Suwinska I, Suwinski R, Smieja J, Swierniak A. Radiomic signature accurately predicts the risk of metastatic dissemination in late-stage non-small cell lung cancer. Translational Lung Cancer Research 12(7):1372-1383. (2023)

10. Mrowiec K, Debik J, Jelonek K, Kurczyk A, Ponge L, Wilk A, Krzempek M, Giskeødegård GF, Bathen TF i Widłak P. Profiling of serum metabolome of breast cancer: multi-cancer features discriminate between healthy women and patients with breast cancer. Front. Oncol. 14:1377373. (2024)

11. Sojka DR, Gogler A, Kania D, Vydra N, Wiecha K, Adamiec-Organiściok M, Wilk A, Chumak V, Matyśniak D, Scieglinska D. The human testis-enriched HSPA2 interacts with HIF-1α in epidermal keratinocytes, yet HIF-1α stability and HIF-1-dependent gene expression rely on the HSPA (HSP70) activity. Biochim Biophys Acta Mol Cell Res. 1871(5):119735. (2024)

Rozdziały w monografiach/publikacje konferencyjne:

1. Student S, Płuciennik A, Łakomiec K, Wilk A, Bensz W, Fujarewicz K. Integration Strategies of Cross-Platform Microarray Data Sets in Multiclass Classification Problem. In: Misra S, et al. Computational Science and Its Applications – ICCSA 2019. ICCSA 2019. Lecture Notes in Computer Science, vol 11623 (2019)

2. Wilk A, Gawin M, Frątczak K, Widłak P, Fujarewicz K. On Stability of Feature Selection Based on MALDI Mass Spectrometry Imaging Data and Simulated Biopsy. In: Korbicz J, Maniewski R, Patan K, Kowal M. (eds) Current Trends in Biomedical Engineering and Bioimages Analysis. PCBEE 2019. Advances in Intelligent Systems and Computing, vol 1033 (2020)

3. Cortez AJ, Kujawa KA, Wilk AM, Krzempek MK, Syrkis JP, Olbryt M, Lisowska M. 219 Signaling pathways related with ITGBL1 in ovarian cancer cells International Journal of Gynecologic Cancer 2020;30:A80.

4. Łakomiec K, Wilk A, Psiuk-Maksymowicz K, Fujarewicz K. Finding the Time-Dependent Virus Transmission Intensity via Gradient Method and Adjoint Sensitivity Analysis. In: Pietka E, Badura P, Kawa J, Wieclawek W (eds) Information Technology in Biomedicine. ITIB 2022. Advances in Intelligent Systems and Computing, vol 1429.(2022)

5. Fujarewicz K, Wilk A, Borys D, d'Amico A, Suwiński R, Świerniak A. Machine Learning Approach to Predict Metastasis in Lung Cancer Based on Radiomic Features. In: Nguyen NT, Tran TK, Tukayev U, Hong TP, Trawiński B, Szczerbicki E. (eds) Intelligent Information and Database Systems. ACIIDS 2022. Lecture Notes in Computer Science, vol 13758 (2022)

6. Wilk AM, Kozłowska E, Borys D, D'Amico A, Gorczewska I, Debosz-Suwińska I, Gałecki S, Fujarewicz K, Suwiński R i Świerniak A. Improving the Predictive Ability of Radiomics-Based Regression Survival Models Through Incorporating Multiple Regions of Interest. W: Strumiłło, P., Klepaczko, A., Strzelecki, M., Bociąga, D. (eds) The Latest Developments and Challenges in Biomedical Engineering. PCBEE 2023. Lecture Notes in Networks and Systems, vol 746. (2024)

7. Kozłowska E, Wilk AM, Butkiewicz D, Krześniak M, Gdowicz-Kłosok A, Giglok M, Suwiński R, Świerniak A . Predicting the Risk of Metastatic Dissemination in Non-small Cell Lung Cancer Using Clinical and Genetic Data. In: Strumiłło, P., Klepaczko, A., Strzelecki, M., Bociąga, D. (eds) The Latest Developments and Challenges in Biomedical Engineering. PCBEE 2023. Lecture Notes in Networks and Systems, vol 746. (2024)

8. Gałecki S, Kysiak M, Kozłowska E, Wilk AM, Suwiński R, Świerniak A. Assessing the Prognosis of Patients with Metastatic or Recurrent Non-small Cell Lung Cancer in the Era of Immunotherapy and Targeted Therapy. In: Strumiłło, P., Klepaczko, A., Strzelecki, M., Bociąga, D. (eds) The Latest Developments and Challenges in Biomedical Engineering. PCBEE 2023. Lecture Notes in Networks and Systems, vol 746. (2024).

Udział w konferencjach:

1. Polish Conference on Bioinformatics and Biomedical Engineering — Zielona Góra 2019, Warszawa 2021, Łódź 2021;
2. Gliwice Scientific Meetings — 2019, 2020, 2022, 2023;
3. Krajowa Konferencja Zastosowań Matematyki w Biologii i Medycynie — Iwonicz 2021, Wisła 2022 (jako członek Komitetu Organizacyjnego), Złoty Potok 2023;
4. 9th International Conference on Risk Analysis (ICRA9) — 25-27 maja 2022, Perugia, Włochy;
5. 14th Asian Conference on Intelligent Information and Database Systems — 28-30 listopada 2022, Ho Chi Minh City, Wietnam;
6. Festival of Genomics — 24-25 stycznia 2024, Londyn, Anglia
7. Śląskie Spotkania Naukowe — 17-19 maja 2024, Ustroń

Życiorys Autorki

Mgr inż. Agata Wilk ukończyła studia magisterskie na kierunku Biotechnologia, specjalność Bioinformatyka. Otrzymała medal „Omnium Studiosorum Optimo" przyznawany dla najlepszych absolwentów Politechniki Śląskiej w roku akademickim 2019/2020. W kolejnym roku akademickim rozpoczęła kształcenie w Szkole Doktorów Politechniki Śląskiej w dyscyplinie Inżynieria Biomedyczna.

Aktywnie uczestniczyła w siedmiu projektach badawczych, finansowanych przez Narodowe Centrum Badań i Rozwoju, Narodowe Centrum Nauki oraz Agencję Badań Medycznych, w tym w projekcie „Nowe narzędzia diagnostyki molekularnej i obrazowania w indywidualizowanej terapii raka piersi, tarczycy i gruczołu krokowego" akronim MILESTONE; „Uczenie maszynowe, modelowanie biologiczne i przetwarzanie obrazów medycznych w prognozowaniu przerzutów raka płuc." oraz „Centrum Medycyny Cyfrowej SILESIA". Realizowała również własne projekty, finansowane z subwencji badawczych dla młodych naukowców. Brała udział w szkoleniach i kursach, na przykład szkole wiosennej „Structured Population Models". Odbyła także trzymiesięczny staż w German Cancer Research Center (DKFZ), na który pozyskała finansowanie w programie Helmholtz Information and Data Science Academy.

Współpracowała z wieloma zespołami badawczymi, m.in. ze Śląskiego Uniwersytetu Medycznego, Gdańskiego Uniwersytetu Medycznego, oraz Narodowego Instytutu Onkologii, gdzie od 2020 r. jest pracownikiem Działu Analiz Bioinformatyczno - Biostatystycznych; a także z partnerami komercyjnymi takimi jak Wasko S.A. czy CelonPharma S.A.

Dorobek naukowy doktorantki to m.in. jedenaście artykułów opublikowanych w czasopismach, indeksowanych w bazie Web of Science, kilka artykułów w monografiach pokonferencyjnych, w tym indeksowanych w bazie Scopus, oraz kilkanaście doniesień i plakatów, prezentowanych na konferencjach naukowych w kraju i za granicą, takich jak Festival of Genomics w Londynie, Asian Conference on Intelligent Information and Database Systems w Ho Chi Minh City czy International Conference on Risk Analysis w Perugii. Jest również współautorką trzech europejskich zgłoszeń patentowych, które są aktualnie procedowane przez EPO.

Poza działalnością badawczą, mgr inż. Agata Wilk współorganizowała XXVII Krajową Konferencję Zastosowań Matematyki w Biologii i Medycynie. Była także laureatką

Ogólnopolskiej Olimpiady Języka Angielskiego dla Studentów Wyższych Uczelni Technicznych.

Doktorantka prowadziła zajęcia dydaktyczne, obejmujące ćwiczenia i laboratoria z różnych przedmiotów dotyczących zagadnień sztucznej inteligencji i uczenia maszynowego dla studentów kierunków Informatyka, Automatyka i Robotyka, Data Science (Makrokierunek), Cognitive Technologies oraz Biotechnologia.