Διατμηματικό Πρόγραμμα Μεταπτυχιακών Σπουδών

Πληροφορική Επιστημών Ζωής

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Ανάλυση ωμικών δεδομένων για τον καρκίνο του προστάτη με χρήση μεθόδων μηχανικής μάθησης

Χρήστος Γιαννόπουλος

ΤΡΙΜΕΛΗΣ ΕΠΙΤΡΟΠΗ

Σεφερίνα Μαυρουδή
Επίκουρος Καθηγήτρια, Τμήμα Νοσηλευτικής Πανεπιστημίου Πατρών

Κώστας Θεοφιλάτος
Senior Lecturer, Imperial College

Αναλίζα Δελαστίκ
Αναπληρώτρια Καθηγήτρια, Τμήμα Ιατρικής Πανεπιστημίου Πατρών

ΠΑΤΡΑ, 2026

Ευχαριστίες

Η παρούσα διπλωματική εργασία εκπονήθηκε στο πλαίσιο των μεταπτυχιακών μου σπουδών. Θα ήθελα να ευχαριστήσω θερμά τους επιβλέποντες καθηγητές μου για την καθοδήγηση, τις εύστοχες παρατηρήσεις και την υπομονή της καθ' όλη τη διάρκεια της εκπόνησης, καθώς και για την ελευθερία που μου έδωσε να ακολουθήσω μια μεθοδολογικά αυστηρή προσέγγιση ακόμη και όταν αυτή οδηγούσε σε λιγότερο εντυπωσιακά αποτελέσματα.

Ευχαριστώ επίσης τη διεθνή ερευνητική ομάδα Stand Up To Cancer–Prostate Cancer Foundation (SU2C/PCF) για τη δημόσια διάθεση των δεδομένων της μελέτης prad_su2c_2019 μέσω της πλατφόρμας cBioPortal, καθώς και την κοινότητα ανοικτού λογισμικού, χωρίς την οποία η υπολογιστική ανάλυση της παρούσας εργασίας δεν θα ήταν εφικτή.

Τέλος, ευχαριστώ την οικογένειά μου για την αμέριστη στήριξη.

Περίληψη

Ο μεταστατικός, ανθεκτικός στον ευνουχισμό καρκίνος του προστάτη (mCRPC) είναι η θανατηφόρος μορφή της νόσου. Οι αναστολείς της σηματοδότησης του ανδρογονικού υποδοχέα (ARSI) παρατείνουν την επιβίωση, αλλά η ανταπόκριση διαφέρει πολύ μεταξύ των ασθενών. Το βασικό ερευνητικό ερώτημα της εργασίας είναι αν υπάρχουν γονίδια που προβλέπουν την εξέλιξη του mCRPC σε ασθενείς που λαμβάνουν θεραπεία ARSI, πέρα από όσα προβλέπουν τα κλινικά στοιχεία.

Αναλύθηκαν 128 ασθενείς της κοόρτης SU2C/PCF (76 θάνατοι) που έλαβαν abiraterone ή enzalutamide ως πρώτη θεραπεία ARSI, με έκβαση τη συνολική επιβίωση από την έναρξή της. Από βιοψίες μεταστάσεων με αλληλούχιση όλων των εξωνίων ορίστηκαν μεταλλάξεις, γονιδιακές ενισχύσεις και ομόζυγες απαλοιφές σε 20 από τα συχνότερα αλλοιωμένα γονίδια της νόσου. Από αυτές τις αλλοιώσεις, οι 22 εμφανίζονταν σε τουλάχιστον 5 ασθενείς. Τέσσερις αλγόριθμοι ανάλυσης επιβίωσης (Cox με ποινή ridge, Cox elastic net, random survival forest, gradient boosting) αξιολογήθηκαν σε τέσσερα σύνολα εισόδων με επαναλαμβανόμενη εμφωλευμένη διασταυρωμένη επικύρωση. Επιπλέον, κάθε αλλοίωση εξετάστηκε με μοντέλο Cox και διόρθωση του ρυθμού ψευδών ανακαλύψεων, με επιβίωση με αποκοπή στους 12, 24 και 36 μήνες.

Οι αλλοιώσεις των γονιδίων μόνες τους δεν διέκριναν τους ασθενείς επαρκώς (Harrell C-index 0,488–0,529). Η προσθήκη τους στα κλινικά στοιχεία μείωσε τον δείκτη συμφωνίας σε όλους τους αλγορίθμους (κατά 0,027–0,048) και επιδείνωσε το κριτήριο Brier. Η καλύτερη επίδοση επιτεύχθηκε μόνο με κλινικά στοιχεία (0,614 με gradient boosting και έως 0,694 στους 12 μήνες). Από τους 77 ελέγχους συσχέτισης, 9 είχαν $p < 0,05$. Η ισχυρότερη ένδειξη αφορούσε την ομόζυγη απαλοιφή του RB1 (7 φορείς), με όλους τους θανάτους των φορέων έως τους 21 μήνες.

Στο σύνολο δεδομένων αυτό, κανένα από τα γονίδια δεν προέβλεψε την επιβίωση υπό θεραπεία ARSI μετά από αυστηρή επικύρωση. Μια σπάνια αλλοίωση με μεγάλο λόγο κινδύνων, όπως η ομόζυγη απαλοιφή του RB1, μπορεί να έχει βιολογική σημασία χωρίς να βελτιώνει την πρόβλεψη για το σύνολο των ασθενών. Η επιβεβαίωση απαιτεί μεγαλύτερες κοόρτες, πληρέστερα κλινικά δεδομένα και εξωτερική επικύρωση.

Λέξεις-κλειδιά: καρκίνος του προστάτη, mCRPC, αμπιρατερόνη, ενζαλουταμίδη, ανάλυση επιβίωσης, μηχανική μάθηση, γονιδιωματικές μεταλλάξεις , εμφωλευμένη διασταυρωμένη επικύρωση, RB1

Abstract

Metastatic castration-resistant prostate cancer (mCRPC) is the lethal form of the disease. Androgen receptor signalling inhibitors (ARSI) prolong survival, but the response varies widely between patients. The main research question of this thesis is whether there are genes that predict the course of mCRPC in patients receiving ARSI therapy, beyond what clinical data predict.

We analysed 128 patients of the SU2C/PCF cohort (76 deaths) who received abiraterone or enzalutamide as first-line ARSI, with overall survival from the start of treatment as the outcome. Mutations, amplifications and deep deletions in 20 of the most frequently altered genes of the disease were identified from whole-exome sequencing of metastatic biopsies. Of these alterations, 22 were present in at least five patients. Four survival analysis algorithms (ridge-penalised Cox, Cox elastic net, random survival forest and gradient boosting) were evaluated on four input sets with repeated nested cross-validation. In addition, each alteration was examined with a Cox model and false discovery rate correction, using survival truncated at 12, 24 and 36 months.

The gene alterations alone did not discriminate patients at a satisfactory raten(Harrell's C-index 0.488–0.529). Adding them to the clinical variables reduced the C-index in all algorithms (by 0.027–0.048) and worsened the Brier score. The best performance was achieved with clinical variables only (0.614 with gradient boosting and up to 0.694 at 12 months). Of the 77 association tests, 9 had $p < 0.05$. The strongest signal concerned RB1 deep deletion (7 carriers), with all deaths among carriers occurring within 21 months.

In this dataset, none of the genes predicted survival under ARSI therapy after rigorous validation. A rare alteration with a large hazard ratio, such as RB1 deep deletion, may be biologically relevant without improving prediction for the patient population as a whole. Confirmation requires larger cohorts, more complete clinical data and external validation.

Keywords: prostate cancer, mCRPC, abiraterone, enzalutamide, survival analysis, machine learning, genomic mutations, nested cross-validation, RB1

Πίνακας Περιεχομένων

Ευχαριστίες 2

Περίληψη 3

Abstract 5

Πίνακας Περιεχομένων 6

Πίνακας Εικόνων 10

Κατάλογος Πινάκων 12

1. ΕΙΣΑΓΩΓΗ 13

1.1 Καρκίνος του προστάτη 13

1.1.1 Ιστοπαθολογική ταξινόμηση: από το άθροισμα Gleason στις ομάδες ISUP 13

1.1.2 Ειδικό προστατικό αντιγόνο (PSA) 14

1.2 Μεταστατικός, ανθεκτικός στον ευνουχισμό καρκίνος και θεραπεία ARSI 15

1.3 Γονιδιωματικές αλλοιώσεις στον μεταστατικό καρκίνο του προστάτη 16

1.4 Η κοόρτη SU2C/PCF και η πλατφόρμα cBioPortal 18

1.5 Δομή της εργασίας 18

2. ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ ΚΑΙ ΑΝΑΛΥΣΗ ΕΠΙΒΙΩΣΗΣ 20

2.1 Τεχνητή νοημοσύνη και μηχανική μάθηση: είδη εκπαίδευσης 20

2.1.1 Εποπτευόμενη μάθηση 20

2.1.2 Μη εποπτευόμενη μάθηση 20

2.1.3 Ανάλυση επιβίωσης 20

2.2 Αλγόριθμοι ανάλυσης επιβίωσης 22

2.2.1 Cox proportional hazards 22

2.2.2 Cox elastic net 23

2.2.3 Random survival forest 24

2.2.4 Gradient boosting 25
2.3 Λίγα συμβάντα, πολλές μεταβλητές 26
2.3.1 Επιλογή χαρακτηριστικών 27
2.3.2 Κανονικοποίηση 27
2.3.3 Διαρροή πληροφορίας 27
2.4 Εμφωλευμένη διασταυρωμένη επικύρωση 28
2.5 Δείκτες αξιολόγησης μοντέλων επιβίωσης 29
2.5.1 Δείκτης συμφωνίας 29
2.5.2 Χρονοεξαρτώμενη AUC 30
2.5.3 Κριτήριο Brier 30
2.6 Πολλαπλοί έλεγχοι 30
2.7 Συσχέτιση και πρόβλεψη 31
3. ΒΙΒΛΙΟΓΡΑΦΙΚΗ ΕΠΙΣΚΟΠΗΣΗ 32
3.1 Γονιδιωματικοί προγνωστικοί δείκτες στο mCRPC υπό θεραπεία ARSI 32
3.1.1 Ο ανδρογονικός υποδοχέας 33
3.1.2 Ογκοκατασταλτικά γονίδια και επιδιόρθωση του DNA 34
3.2 Μοντέλα πρόβλεψης της επιβίωσης και μηχανική μάθηση 35
3.3 Ελληνικές μελέτες 36
3.4 Κριτική αποτίμηση 37
4. ΣΚΟΠΟΣ ΚΑΙ ΕΡΕΥΝΗΤΙΚΑ ΕΡΩΤΗΜΑΤΑ 39
4.1 Βασικό ερευνητικό ερώτημα 39
4.2 Επιμέρους ερωτήματα 40
4.3 Ερευνητικές υποθέσεις 40
5. ΥΛΙΚΑ ΚΑΙ ΜΕΘΟΔΟΙ 42
5.1 Πηγή δεδομένων 42
5.2 Κοόρτη και έκβαση 42

5.3 Μεταβλητές 43
5.3.1 Κλινικές μεταβλητές 43
5.3.2 Γονιδιωματικές μεταβλητές: τα συχνότερα αλλοιωμένα γονίδια ..44
5.3.3 Σύνολα εισόδων 45
5.4 Προεπεξεργασία 45
5.5 Αλγόριθμοι και υπερπαράμετροι 45
5.6 Πρωτόκολλο αξιολόγησης 46
5.7 Ανάλυση σε χρονικούς ορίζοντες 47
5.8 Ανάλυση ανά αλλοίωση 48
5.9 Λογισμικό και αναπαραγωγιμότητα 49
6. ΑΠΟΤΕΛΕΣΜΑΤΑ 50
6.1 Περιγραφή της κοόρτης 50
6.1.1 Κλινικά χαρακτηριστικά και επιβίωση 50
6.1.2 Αλλοιώσεις των γονιδίων 51
6.2 Επίδοση των μοντέλων πρόβλεψης 52
6.2.1 Τα γονίδια ως μόνη πηγή πληροφορίας (E1) 54
6.2.2 Ακρίβεια των πιθανοτήτων επιβίωσης 54
6.3 Προστιθέμενη αξία των γονιδίων πέρα από τα κλινικά στοιχεία (E2) 55
6.4 Ανάλυση σε χρονικούς ορίζοντες 57
6.5 Αλλοιώσεις που σχετίζονται με την επιβίωση (E3) 60
6.6 Σύνοψη των αποτελεσμάτων 63
7. ΣΥΖΗΤΗΣΗ 64
7.1 Απάντηση στο ερευνητικό ερώτημα 64
7.2 Ερμηνεία των ευρημάτων 64
7.2.1 Γιατί τα γονίδια δεν βελτίωσαν την πρόβλεψη 64
7.2.2 Η χρονική διάσταση και το RB1 65
7.2.3 Ο ρόλος των κλινικών στοιχείων 66

7.3 Περιορισμοί 67
7.4 Μελλοντικές κατευθύνσεις 68
7.5 Συμπεράσματα 68
8. ΒΙΒΛΙΟΓΡΑΦΙΑ 70
Παράρτημα Α: Πηγαίος κώδικας Python — βασικές ενότητες 77
A.1 src/data.py 77
A.2 src/models.py 80
A.3 src/evaluation.py 81
A.4 src/horizons.py 85
Σημείωμα αναπαραγωγιμότητας 89

Πίνακας Εικόνων

Σχήμα 1. Καμπύλη Kaplan–Meier της συνολικής επιβίωσης από την έναρξη της πρώτης θεραπείας ARSI (128 ασθενείς, 76 θάνατοι). Η σκιασμένη περιοχή είναι το 95% διάστημα εμπιστοσύνης και οι κάθετες γραμμές δείχνουν λογοκριμένους ασθενείς. Κάτω από τον άξονα, οι ασθενείς σε κίνδυνο. 51

Σχήμα 2. Πλήθος ασθενών με κάθε αλλοίωση των 20 γονιδίων (από 128). Εμφανίζονται οι 42 αλλοιώσεις με τουλάχιστον έναν φορέα. Η διακεκομμένη γραμμή δείχνει το φίλτρο των 5 φορέων και με γκρι αριθμό σημειώνονται οι αλλοιώσεις που δεν το πέρασαν. 52

Σχήμα 3. Δείκτης συμφωνίας του Harrell για κάθε σύνολο εισόδων και αλγόριθμο: μέσος όρος 50 διαχωρισμών και διορθωμένο 95% διάστημα εμπιστοσύνης. Η διακεκομμένη γραμμή αντιστοιχεί στην τυχαία κατάταξη (0,5). 54

Σχήμα 4. Ολοκληρωμένο κριτήριο Brier (6–30 μήνες) για κάθε σύνολο εισόδων και αλγόριθμο, με διορθωμένο 95% διάστημα εμπιστοσύνης. Η διακεκομμένη γραμμή είναι η τιμή της καμπύλης Kaplan–Meier (0,197). Μικρότερες τιμές είναι καλύτερες.....55

Σχήμα 5. Ζευγαρωμένες διαφορές «κλινικά + γονίδια» μείον «κλινικά διευρυμένα» στον δείκτη Harrell, στον δείκτη Uno και στο IBS, ανά αλγόριθμο (μέσος όρος και διορθωμένο 95% διάστημα εμπιστοσύνης). Για τους δείκτες συμφωνίας οι θετικές τιμές σημαίνουν βελτίωση, για το IBS οι αρνητικές. 56

Σχήμα 6. Ζευγαρωμένες διαφορές «κλινικά διευρυμένα» μείον «κλινικά βασικά» στον δείκτη Harrell, στον δείκτη Uno και στο IBS, ανά αλγόριθμο (μέσος όρος και διορθωμένο 95% διάστημα εμπιστοσύνης). 57

Σχήμα 7. Δείκτης συμφωνίας του Harrell έως τον ορίζοντα (12, 24 και 36 μήνες) για κάθε σύνολο εισόδων και αλγόριθμο, με διορθωμένο 95% διάστημα εμπιστοσύνης. Η διακεκομμένη γραμμή αντιστοιχεί στην τυχαία κατάταξη (0,5). 58

Σχήμα 8. Ζευγαρωμένες διαφορές στον δείκτη Harrell ανά ορίζοντα: «κλινικά + γονίδια» μείον «κλινικά διευρυμένα» (τετράγωνα) και «κλινικά διευρυμένα» μείον «κλινικά βασικά» (κύκλοι), με διορθωμένο 95% διάστημα εμπιστοσύνης. 59

Σχήμα 9. Λόγοι κινδύνων (λογαριθμική κλίμακα) με 95% διάστημα εμπιστοσύνης για κάθε εκτιμήσιμη αλλοίωση, στην ανάλυση χωρίς προσαρμογή, ανά ορίζοντα. Με πορτοκαλί οι συσχετίσεις με $p < 0,05$ πριν από τη διόρθωση FDR. Καμία δεν είχε $q < 0,05$. 62

Κατάλογος Πινάκων

Πίνακας 1. Αντιστοίχιση αθροίσματος Gleason με τις Ομάδες Βαθμού ISUP 2014 (Epstein et al., 2016). 14

Πίνακας 2. Επιλεγμένες μελέτες γονιδιωματικών και μοριακών δεικτών σε ασθενείς με mCRPC υπό θεραπεία με abiraterone ή enzalutamide.....33

Πίνακας 3. Αρχεία δεδομένων που αξιοποιήθηκαν.....42

Πίνακας 4. Ορισμός των τριών τύπων αλλοίωσης των γονιδίων. 44

Πίνακας 5. Τα τέσσερα σύνολα εισόδων των μοντέλων. 45

Πίνακας 6. Σταθερές ρυθμίσεις και πλέγματα υπερπαραμέτρων των τεσσάρων αλγορίθμων. 46

Πίνακας 7. Επιβίωση με αποκοπή στους τρεις ορίζοντες (128 ασθενείς). ..48

Πίνακας 8. Χαρακτηριστικά των 128 ασθενών της κοόρτης. Συνεχείς μεταβλητές: διάμεσος (ενδοτεταρτημοριακό εύρος). Κατηγορικές: πλήθος (%). 50

Πίνακας 9. Επίδοση των 16 συνδυασμών συνόλου εισόδων και αλγορίθμου στην κύρια ανάλυση (μέσος όρος 50 διαχωρισμών). 53

Πίνακας 10. Ζευγαρωμένες διαφορές μεταξύ συνόλων εισόδων στον ίδιο αλγόριθμο και στους ίδιους 50 διαχωρισμούς. 56

Πίνακας 11. Δείκτης συμφωνίας του Harrell έως κάθε ορίζοντα (μέσος όρος 50 διαχωρισμών και διορθωμένο 95% διάστημα εμπιστοσύνης). 57

Πίνακας 12. Ζευγαρωμένες διαφορές στον δείκτη συμφωνίας του Harrell ανά ορίζοντα. 59

Πίνακας 13. Αριθμός εκτιμήσιμων συσχετίσεων αλλοίωσης–επιβίωσης και πόσες είχαν $p < 0,05$ ή $q < 0,05$, ανά ανάλυση και ορίζοντα (22 αλλοιώσεις). ..60

Πίνακας 14. Συσχετίσεις αλλοίωσης–επιβίωσης με $p < 0,05$ (μοντέλο Cox ανά αλλοίωση). 61

Πίνακας 15. Σύνοψη των απαντήσεων στα επιμέρους ερευνητικά ερωτήματα.....63

Πίνακας 16. Οι ενότητες κώδικα που παρατίθενται στο Παράρτημα Α. 77

1. ΕΙΣΑΓΩΓΗ

1.1 Καρκίνος του προστάτη

Ο καρκίνος του προστάτη αποτελεί μία από τις σημαντικότερες προκλήσεις της σύγχρονης ογκολογίας. Σύμφωνα με τις εκτιμήσεις GLOBOCAN για το 2022, καταγράφηκαν παγκοσμίως 1.467.854 νέα κρούσματα, ποσοστό που αντιστοιχεί στο 7,3% του συνόλου των νεοδιαγνωσθεισών κακοηθειών (Bray et al., 2024). Στις χώρες υψηλού δείκτη ανθρώπινης ανάπτυξης η τυποποιημένη κατά ηλικία επίπτωση ανέρχεται σε 35,5 ανά 100.000 άνδρες, καθιστώντας τη νόσο τη δεύτερη συχνότερη μετά τον καρκίνο του πνεύμονα, ενώ στις χώρες χαμηλότερου δείκτη κατέχει την πρώτη θέση παρά τη χαμηλότερη απόλυτη επίπτωση (12,6 ανά 100.000). Η διαφορά αυτή αντανακλά τόσο πραγματικές βιολογικές και περιβαλλοντικές διαφοροποιήσεις όσο και ανισότητες στην πρόσβαση σε διαγνωστικό έλεγχο.

Χαρακτηριστικό γνώρισμα της νόσου είναι η εξαιρετική ετερογένεια της κλινικής της πορείας. Ένα σημαντικό ποσοστό των όγκων παραμένει ενδοπροστατικό και βραδέως εξελισσόμενο για δεκαετίες, χωρίς ποτέ να απειλήσει τη ζωή του ασθενούς. Ένα άλλο, μικρότερο αλλά κρίσιμο ποσοστό εμφανίζει επιθετική συμπεριφορά με ταχεία τοπική επέκταση και μεταστατική διασπορά. Η διάκριση των δύο αυτών φαινοτύπων κατά τη διάγνωση συνιστά το κεντρικό άλυτο πρόβλημα της κλινικής διαχείρισης, καθώς καθορίζει την επιλογή μεταξύ ενεργού παρακολούθησης και ριζικής θεραπείας — με το κόστος της υπερθεραπείας αφενός και της απώλειας θεραπευτικού παραθύρου αφετέρου.

1.1.1 Ιστοπαθολογική ταξινόμηση: από το άθροισμα Gleason στις ομάδες ISUP

Η βαθμονόμηση κατά Gleason, που εισήχθη το 1966, παραμένει ο ισχυρότερος μεμονωμένος προγνωστικός παράγοντας (Gleason, 1966). Το σύστημα αποδίδει βαθμό από 1 έως 5 στα δύο επικρατέστερα αρχιτεκτονικά πρότυπα του όγκου και τα αθροίζει. Η σειρά των δύο συνιστωσών έχει σημασία, καθώς το πρότυπο 4+3 συνεπάγεται σαφώς δυσμενέστερη πρόγνωση από το 3+4 παρά το ίδιο άθροισμα.

Η συναινετική διάσκεψη της Διεθνούς Εταιρείας Ουρολογικής Παθολογίας (International Society of Urological Pathology, ISUP) το 2014 αναδιοργάνωσε το σύστημα σε πέντε προγνωστικά διακριτές Ομάδες Βαθμού (Grade Groups), με στόχο την απλοποίηση της επικοινωνίας και τη μείωση της υπερθεραπείας (Epstein et al., 2016). Η αντιστοίχιση παρουσιάζεται στον Πίνακα που ακολουθεί.

Πίνακας 1. Αντιστοίχιση αθροίσματος Gleason με τις Ομάδες Βαθμού ISUP 2014 (Epstein et al., 2016).

Ομάδα ISUP	Άθροισμα Gleason	Πρόγνωση
1	≤ 6 (3+3)	Άριστη, υποψήφιοι για ενεργό παρακολούθηση
2	7 (3+4)	Ευνοϊκή ενδιάμεση
3	7 (4+3)	Δυσμενής ενδιάμεση
4	8 (4+4, 3+5, 5+3)	Υψηλού κινδύνου
5	9–10	Πολύ υψηλού κινδύνου, υψηλό μεταστατικό δυναμικό

Η ομαδοποίηση ISUP επικυρώθηκε σε πληθυσμιακές μελέτες και ενσωματώθηκε στις κατευθυντήριες οδηγίες. Στα δεδομένα της παρούσας εργασίας είναι διαθέσιμο μόνο το άθροισμα Gleason κατά τη διάγνωση, χωρίς τις επιμέρους συνιστώσες. Η ομάδα ISUP δεν μπορεί επομένως να υπολογιστεί και το άθροισμα χρησιμοποιείται απευθείας ως κλινική μεταβλητή.

1.1.2 Ειδικό προστατικό αντιγόνο (PSA)

Το ειδικό προστατικό αντιγόνο είναι γλυκοπρωτεΐνη που παράγεται από το αδενικό επιθήλιο του προστάτη. Η μέτρησή του στον ορό αποτελεί τον ακρογωνιαίο λίθο της διάγνωσης και της παρακολούθησης, παρά τους γνωστούς περιορισμούς του. το PSA είναι ειδικό για το όργανο αλλά όχι για τη νεοπλασία, με αποτέλεσμα η καλοήθης υπερπλασία, η προστατίτιδα και μηχανικοί χειρισμοί να προκαλούν ψευδώς θετικά αποτελέσματα.

Στο πλαίσιο της πρόγνωσης, η τιμή PSA κατά τη διάγνωση αποτελεί ανεξάρτητο προγνωστικό παράγοντα της εντοπισμένης νόσου και συμπεριλαμβάνεται στα καθιερωμένα νομογράμματα κινδύνου. Στην παρούσα εργασία είναι διαθέσιμη μόνο αυτή η τιμή, η οποία προηγείται κατά πολύ της έναρξης της θεραπείας για τη μεταστατική νόσο. Αποτυπώνει επομένως την αρχική εικόνα της νόσου και όχι την κατάσταση του ασθενούς κατά την έναρξη της θεραπείας ARSI. Λόγω της έντονα δεξιά ασύμμετρης κατανομής της, η

μεταβλητή μετασχηματίζεται συνήθως λογαριθμικά πριν από τη στατιστική μοντελοποίηση — πρακτική που ακολουθείται και εδώ.

1.2 Μεταστατικός, ανθεκτικός στον ευνουχισμό καρκίνος και θεραπεία ARSI

Η ανάπτυξη του καρκίνου του προστάτη εξαρτάται σε μεγάλο βαθμό από τη σηματοδότηση του ανδρογονικού υποδοχέα (androgen receptor, AR). Γι' αυτό η βασική θεραπεία της προχωρημένης νόσου είναι η ανδρογονική αποστέρηση (androgen deprivation therapy, ADT), δηλαδή ο φαρμακευτικός ή χειρουργικός ευνουχισμός. Η αρχική ανταπόκριση είναι συνήθως καλή, αλλά με τον χρόνο η νόσος εξελίσσεται παρά τα χαμηλά επίπεδα τεστοστερόνης. Η κατάσταση αυτή ονομάζεται ανθεκτικός στον ευνουχισμό καρκίνος του προστάτη (castration-resistant prostate cancer, CRPC) και, όταν υπάρχουν μεταστάσεις, mCRPC (Sandhu et al., 2021).

Η ανθεκτικότητα στον ευνουχισμό δεν σημαίνει ότι ο όγκος έπαψε να εξαρτάται από τον AR. Προκλινικές μελέτες έδειξαν ότι η διατηρούμενη σηματοδότηση του υποδοχέα παραμένει ο κύριος οδηγός του CRPC, και αυτό οδήγησε στην ανάπτυξη νέων φαρμάκων που στοχεύουν τον άξονα του AR (Watson et al., 2015). Τα φάρμακα αυτά αναφέρονται συλλογικά ως αναστολείς της σηματοδότησης του ανδρογονικού υποδοχέα (androgen receptor signaling inhibitors, ARSI). Οι δύο πρώτοι που καθιερώθηκαν είναι το abiraterone, που αναστέλλει τη βιοσύνθεση των ανδρογόνων και εκτός των όρχεων (de Bono et al., 2011), και το enzalutamide, που αναστέλλει πολλαπλά βήματα της σηματοδότησης του ίδιου του υποδοχέα (Scher et al., 2012).

Και τα δύο φάρμακα παρέτειναν την επιβίωση σε μεγάλες τυχαιοποιημένες μελέτες φάσης III. Σε ασθενείς που είχαν ήδη λάβει χημειοθεραπεία με docetaxel, το abiraterone αύξησε τη διάμεση συνολική επιβίωση από 10,9 σε 14,8 μήνες (HR 0,65, de Bono et al., 2011) και το enzalutamide από 13,6 σε 18,4 μήνες (HR 0,63, Scher et al., 2012). Σε ασθενείς χωρίς προηγούμενη χημειοθεραπεία, το abiraterone σχεδόν διπλασίασε τη διάμεση επιβίωση χωρίς ακτινολογική εξέλιξη, από 8,3 σε 16,5 μήνες, με τάση βελτίωσης της συνολικής επιβίωσης (Ryan et al., 2013), ενώ το enzalutamide μείωσε τον κίνδυνο θανάτου κατά 29% (HR 0,71, Beer et al., 2014). Έτσι οι ARSI έγιναν θεραπεία πρώτης γραμμής του mCRPC. Στην Ελλάδα, σε πολυκεντρική μελέτη

149 ασθενών, το 78,5% έλαβε ARSI ως θεραπεία πρώτης γραμμής για το mCRPC (Liontos et al., 2024).

Παρά το όφελος σε επίπεδο πληθυσμού, η ανταπόκριση στους ARSI διαφέρει πολύ από ασθενή σε ασθενή. Σε μελέτη 60 ασθενών με οστικές μεταστάσεις, το 37% εμφάνισε πρωτογενή αντίσταση στο enzalutamide και διέκοψε τη θεραπεία μέσα σε τέσσερις μήνες (Efstathiou et al., 2015). Οι μηχανισμοί της αντίστασης ταξινομούνται σε τρεις κατηγορίες: αποκατάσταση της σηματοδότησης του AR, παράκαμψή της μέσω άλλων οδών και πλήρη ανεξαρτησία από τον υποδοχέα (Watson et al., 2015). Για τον λόγο αυτό η ομάδα εργασίας PCWG3 συνιστά τη σειριακή βιολογική ανάλυση δειγμάτων όγκου, αίματος ή απεικόνισης, ώστε να κατανοηθούν οι μηχανισμοί αντίστασης και να εντοπιστούν δείκτες που προβλέπουν την ανταπόκριση (Scher et al., 2016). Η αναζήτηση τέτοιων δεικτών στο DNA του όγκου αποτελεί το κλινικό πλαίσιο της παρούσας εργασίας.

1.3 Γονιδιωματικές αλλοιώσεις στον μεταστατικό καρκίνο του προστάτη

Ο καρκίνος αναπτύσσεται μέσα από σωματικές αλλοιώσεις του DNA, δηλαδή αλλοιώσεις που αποκτώνται στα καρκινικά κύτταρα και δεν κληρονομούνται. Δύο κατηγορίες είναι κεντρικές για την παρούσα εργασία. Οι σημειακές μεταλλάξεις (mutations) και οι μικρές ενθέσεις ή ελλείψεις αλλάζουν την αλληλουχία ενός γονιδίου και, όταν είναι μη συνώνυμες, την πρωτεΐνη που αυτό κωδικοποιεί. Οι μεταβολές του αριθμού αντιγράφων αλλάζουν τη «δόση» του γονιδίου: η γονιδιακή ενίσχυση (amplification) πολλαπλασιάζει τα αντίγραφα ενός γονιδίου και συνήθως αυξάνει την έκφρασή του, ενώ η απαλοιφή αφαιρεί ένα ή και τα δύο αντίγραφα.

Για τα ογκοκατασταλτικά γονίδια ισχύει η υπόθεση των «δύο χτυπημάτων», σύμφωνα με την οποία η λειτουργία του γονιδίου χάνεται όταν αδρανοποιηθούν και τα δύο αλληλόμορφα (Knudson, 1971). Η ομόζυγη απαλοιφή (deep deletion), δηλαδή η απώλεια και των δύο αντιγράφων, αδρανοποιεί το γονίδιο με ένα μόνο γεγονός. Αντίθετα, η ημιζυγής απαλοιφή (shallow deletion) αφαιρεί ένα αντίγραφο και οδηγεί σε πλήρη απώλεια της λειτουργίας μόνο αν συνδυαστεί με αδρανοποίηση του άλλου αλληλομόρφου, για παράδειγμα με

μετάλλαξη. Γι’ αυτό, στα δεδομένα αριθμού αντιγράφων η ομόζυγη απαλοιφή είναι η σαφέστερη ένδειξη απώλειας ενός ογκοκατασταλτικού γονιδίου.

Μεγάλες μελέτες αλληλούχισης μεταστάσεων περιέγραψαν το γονιδιωματικό τοπίο του mCRPC. Ο συνολικός ρυθμός μεταλλάξεων είναι χαμηλός, περίπου 2 μεταλλάξεις ανά εκατομμύριο βάσεις, ακόμη και σε όγκους που έχουν λάβει πολλές θεραπείες (Grasso et al., 2012). Στους πρώτους 150 ασθενείς της κοόρτης SU2C/PCF, αλλοιώσεις των AR, ETS, TP53 και PTEN εμφανίζονταν στο 40–60% των περιπτώσεων. Οι αλλοιώσεις των TP53 και AR ήταν συχνότερες απ’ ό,τι στον πρωτοπαθή όγκο, ενώ αλλοιώσεις των BRCA2, BRCA1 και ATM βρέθηκαν στο 19,3% των ασθενών (Robinson et al., 2015). Η ενιαία ανάλυση 1.013 καρκίνων του προστάτη έδειξε ότι, πέρα από τα συχνά γονίδια, υπάρχει μια «μακριά ουρά» γονιδίων που μεταλλάσσονται σε λιγότερο από 3% των ασθενών (Armenia et al., 2018). Τέλος, η αλληλούχιση ολόκληρου του γονιδιώματος αποκάλυψε δομικές αλλοιώσεις που δεν ανιχνεύονται με την αλληλούχιση των εξωνίων, όπως η ενίσχυση μιας ενισχυτικής περιοχής ανοδικά του AR στο 81% των ασθενών (Quigley et al., 2018).

Τα 20 γονίδια της εργασίας καλύπτουν τις κύριες λειτουργικές κατηγορίες αυτών των αλλοιώσεων (Grasso et al., 2012, Robinson et al., 2015 και Quigley et al., 2018). Περιλαμβάνουν τον ίδιο τον AR και τον συνεργαζόμενο μεταγραφικό παράγοντα FOXA1, καθώς και το SPOP, του οποίου οι μεταλλάξεις ορίζουν έναν από τους μοριακούς υποτύπους του πρωτοπαθούς καρκίνου (The Cancer Genome Atlas Research Network, 2015). Περιλαμβάνουν επίσης τα ογκοκατασταλτικά γονίδια TP53, RB1 και PTEN, γονίδια της επιδιόρθωσης του DNA (BRCA1, BRCA2, ATM, CDK12), γονίδια των μονοπατιών PI3K (PIK3CA, PIK3CB, AKT1) και WNT (APC, CTNNB1), ρυθμιστές του κυτταρικού κύκλου (CCND1, CDKN1B), το ογκογονίδιο MYC και γονίδια που ρυθμίζουν τη χρωματίνη (CHD1, KMT2D).

Ένα πρακτικό ερώτημα σε κάθε μελέτη βιοψιών είναι αν μία μετάσταση αντιπροσωπεύει τη νόσο του ασθενούς. Η σύγκριση πολλαπλών μεταστάσεων του ίδιου ασθενούς έδειξε μεγάλη ετερογένεια μεταξύ ασθενών αλλά περιορισμένη μέσα στον ίδιο ασθενή. Η ανάλυση μίας μετάστασης δίνει επομένως εύλογη εικόνα των κύριων αλλοιώσεων-οδηγών, με ορισμένες εξαιρέσεις (Kumar et al., 2016).

1.4 Η κοόρτη SU2C/PCF και η πλατφόρμα cBioPortal

Η διεθνής ερευνητική ομάδα Stand Up To Cancer–Prostate Cancer Foundation (SU2C/PCF) δημιούργησε μια πολυκεντρική υποδομή για την προοπτική αλληλούχιση όλων των εξωνίων και του μεταγραφώματος σε βιοψίες μεταστάσεων σε οστά και μαλακά μόρια ασθενών με mCRPC (Robinson et al., 2015). Στόχος ήταν να γίνει η γονιδιωματική πληροφορία διαθέσιμη για κλινικές αποφάσεις: στους πρώτους 150 ασθενείς, το 89% είχε τουλάχιστον μία αλλοίωση με πιθανή θεραπευτική σημασία. Η κοόρτη επεκτάθηκε αργότερα σε 429 ασθενείς με γονιδιωματική και μεταγραφωματική ανάλυση και κλινικά δεδομένα παρακολούθησης. Από αυτούς, 128 έλαβαν abiraterone ή enzalutamide ως πρώτη θεραπεία ARSI (Abida et al., 2019). Αυτοί οι 128 ασθενείς αποτελούν τον πληθυσμό της παρούσας εργασίας.

Το cBioPortal for Cancer Genomics αποτελεί τη διαδικτυακή πλατφόρμα που καθιστά τα δεδομένα μεγάλων γονιδιωματικών μελετών προσβάσιμα σε ερευνητές χωρίς εξειδικευμένη υπολογιστική υποδομή (Cerami et al., 2012 και Gao et al., 2013). Παρέχει ενοποιημένη μορφή αρχείων, οπτικοποίηση και δυνατότητα μεταφόρτωσης ολόκληρων μελετών.

Η τυποποίηση αυτή διευκολύνει την ανάλυση αλλά εισάγει και έναν περιορισμό: το δημόσιο εξαγόμενο μιας μελέτης δεν ταυτίζεται απαραίτητα με το σύνολο των δεδομένων που ανέλυσε η αρχική δημοσίευση. Για τον λόγο αυτό, στην παρούσα εργασία η επιλογή του δείγματος κάθε ασθενούς και ο ορισμός της έκβασης διασταυρώθηκαν με το συμπληρωματικό υλικό της δημοσίευσης (Κεφάλαιο 5). Η εργασία συνιστά δευτερογενή ανάλυση δημόσιων δεδομένων με δικό της ερευνητικό ερώτημα.

1.5 Δομή της εργασίας

Η εργασία εξετάζει αν γονιδιωματικές αλλοιώσεις του όγκου μπορούν να προβλέψουν την εξέλιξη ασθενών με mCRPC που λαμβάνουν θεραπεία ARSI, πέρα από τα κλινικά στοιχεία. Το Κεφάλαιο 2 παρουσιάζει τις μεθόδους μηχανικής μάθησης και ανάλυσης επιβίωσης που χρησιμοποιούνται και το Κεφάλαιο 3 τη σχετική βιβλιογραφία. Το Κεφάλαιο 4 διατυπώνει το ερευνητικό ερώτημα και τις υποθέσεις. Το Κεφάλαιο 5 περιγράφει τα δεδομένα και τη

μεθοδολογία, το Κεφάλαιο 6 τα αποτελέσματα και το Κεφάλαιο 7 τη συζήτησή τους και τα συμπεράσματα.

2. ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ ΚΑΙ ΑΝΑΛΥΣΗ ΕΠΙΒΙΩΣΗΣ

2.1 Τεχνητή νοημοσύνη και μηχανική μάθηση: είδη εκπαίδευσης

Η μηχανική μάθηση αποτελεί υποπεδίο της τεχνητής νοημοσύνης που μελετά αλγορίθμους ικανούς να βελτιώνουν την επίδοσή τους σε μια εργασία μέσω της εμπειρίας, χωρίς ρητό προγραμματισμό των κανόνων. Στη βιοϊατρική, η προσέγγιση αυτή είναι ιδιαίτερα ελκυστική επειδή οι υποκείμενοι μοριακοί μηχανισμοί είναι συχνά υπερβολικά πολύπλοκοι για να διατυπωθούν ως κανόνες.

2.1.1 Εποπτευόμενη μάθηση

Στην εποπτευόμενη μάθηση κάθε παρατήρηση συνοδεύεται από ετικέτα και ο αλγόριθμος καλείται να μάθει την απεικόνιση από τα χαρακτηριστικά στην ετικέτα. Όταν η ετικέτα είναι διακριτή, το πρόβλημα είναι ταξινόμηση, ενώ όταν είναι συνεχής, παλινδρόμηση. Η πρόβλεψη της επιβίωσης ανήκει στην εποπτευόμενη μάθηση, με την ιδιαιτερότητα ότι για μέρος των ασθενών η έκβαση δεν είναι πλήρως παρατηρημένη (ενότητα 2.1.3).

2.1.2 Μη εποπτευόμενη μάθηση

Απουσία ετικετών, ο στόχος μετατοπίζεται στην ανακάλυψη δομής: ομαδοποίηση, μείωση διαστατικότητας, ανίχνευση ανωμαλιών. Στη μοριακή ογκολογία η μη εποπτευόμενη ομαδοποίηση αποτελεί το βασικό εργαλείο ανακάλυψης μοριακών υποτύπων της νόσου.

Η διάκριση έχει πρακτική σημασία για την αποφυγή διαρροής πληροφορίας: τα μη εποπτευόμενα βήματα προεπεξεργασίας, όπως το φιλτράρισμα των γονιδιωματικών αλλοιώσεων με βάση τη συχνότητά τους, δεν «βλέπουν» την ετικέτα και επομένως μπορούν να εκτελεστούν μία φορά πριν από τη διασταυρωμένη επικύρωση. Αντίθετα, κάθε βήμα που χρησιμοποιεί την έκβαση πρέπει να επαναλαμβάνεται μέσα σε κάθε διαχωρισμό, μόνο στα δεδομένα εκπαίδευσης (Hastie et al., 2009, ενότητα 2.3.3).

2.1.3 Ανάλυση επιβίωσης

Ιδιαίτερη κατηγορία αποτελεί η ανάλυση επιβίωσης, όπου η έκβαση είναι ζεύγος χρόνου και δείκτη συμβάντος, με μέρος των παρατηρήσεων

λογοκριμένο. Η δυαδικοποίηση μιας λογοκριμένης έκβασης απορρίπτει την πληροφορία του χρόνου και εισάγει συστηματικό σφάλμα: ασθενής με σύντομη παρακολούθηση και ετικέτα «χωρίς συμβάν» δεν είναι ισοδύναμος με ασθενή που παρακολουθήθηκε επί δεκαετία. Το μοντέλο αναλογικών κινδύνων του Cox (1972) και ο εκτιμητής Kaplan–Meier (1958) αποτελούν τα θεμελιώδη εργαλεία.

Η λογοκρισία (censoring) προκύπτει όταν η παρακολούθηση ενός ασθενούς τελειώνει πριν από το συμβάν, για παράδειγμα επειδή η συλλογή των δεδομένων ολοκληρώθηκε ενώ ο ασθενής ζούσε. Για τον ασθενή αυτόν είναι γνωστό μόνο ότι επιβίωσε τουλάχιστον έως τον τελευταίο χρόνο παρακολούθησης (λογοκρισία από δεξιά). Οι μέθοδοι ανάλυσης επιβίωσης αξιοποιούν αυτή τη μερική πληροφορία, με την παραδοχή ότι η λογοκρισία είναι μη πληροφοριακή (non-informative): οι λογοκριμένοι ασθενείς δεν έχουν συστηματικά διαφορετική πρόγνωση από όσους παραμένουν υπό παρακολούθηση.

Η κατανομή του χρόνου επιβίωσης T περιγράφεται με δύο ισοδύναμες συναρτήσεις. Η συνάρτηση επιβίωσης $S(t) = P(T > t)$ δίνει την πιθανότητα επιβίωσης πέρα από τον χρόνο t. Η συνάρτηση κινδύνου $h(t)$ εκφράζει τον στιγμιαίο ρυθμό εμφάνισης του συμβάντος τη χρονική στιγμή t, για όσους έχουν επιβιώσει έως τότε. Ο αθροιστικός κίνδυνος $H(t)$ είναι το ολοκλήρωμα της $h(t)$ από το 0 έως το t, και οι δύο περιγραφές συνδέονται με τη σχέση $S(t) = \exp[-H(t)]$.

Ο εκτιμητής Kaplan–Meier εκτιμά τη συνάρτηση επιβίωσης χωρίς παραμετρικές υποθέσεις (Kaplan & Meier, 1958). Σε κάθε χρόνο θανάτου t_j υπολογίζεται το ποσοστό των ασθενών σε κίνδυνο που επιβίωσαν, $1 - d_j/n_j$, όπου d_j οι θάνατοι και n_j οι ασθενείς που παρακολουθούνταν ακόμη εκείνη τη στιγμή. Η εκτίμηση της $S(t)$ είναι το γινόμενο αυτών των ποσοστών για όλους τους χρόνους θανάτου έως τον t. Οι λογοκριμένοι ασθενείς συμμετέχουν στο σύνολο κινδύνου έως τον χρόνο λογοκρισίας τους. Η διάμεση επιβίωση εκτιμάται ως ο χρόνος στον οποίο η καμπύλη πέφτει στο 0,5. Η απλή διάμεσος των παρατηρημένων χρόνων θα την υποεκτιμούσε, διότι οι χρόνοι των λογοκριμένων ασθενών είναι μόνο κάτω όρια.

Μια χρήσιμη παραλλαγή είναι η επιβίωση με αποκοπή σε χρονικό ορίζοντα h. Όσοι ασθενείς ζουν πέρα από τον h θεωρούνται λογοκριμένοι στον h, και οι θάνατοι μετά τον h δεν προσμετρώνται ως συμβάντα. Ένα μοντέλο που

εκπαιδεύεται σε αυτά τα δεδομένα μαθαίνει τον κίνδυνο θανάτου έως τον h. Η σύγκριση διαφορετικών οριζόντων δείχνει αν η επίδραση μιας μεταβλητής αφορά κυρίως τους πρώιμους ή τους όψιμους θανάτους, χωρίς να αποκλείεται κανένας ασθενής από την ανάλυση.

2.2 Αλγόριθμοι ανάλυσης επιβίωσης

Στην παρούσα εργασία αξιολογούνται τέσσερις αλγόριθμοι ανάλυσης επιβίωσης, οργανωμένοι σε δύο οικογένειες. Η πρώτη περιλαμβάνει δύο γραμμικά μοντέλα: το μοντέλο αναλογικών κινδύνων του Cox με ποινή ridge και το μοντέλο Cox με ποινή elastic net. Και τα δύο υποθέτουν ότι ο λογάριθμος του κινδύνου είναι γραμμικός συνδυασμός των χαρακτηριστικών. Η δεύτερη περιλαμβάνει δύο σύνολα δέντρων απόφασης, το random survival forest και το gradient boosting, τα οποία μαθαίνουν μη γραμμικές σχέσεις και αλληλεπιδράσεις χωρίς να χρειάζεται να οριστούν εκ των προτέρων. Η επιλογή καλύπτει έτσι δύο βασικές αντιθέσεις: γραμμικές έναντι μη γραμμικών υποθέσεων και μεμονωμένα μοντέλα έναντι συνόλων.

Και οι τέσσερις αλγόριθμοι χρησιμοποιούν απευθείας τη λογοκριμένη έκβαση, χωρίς δυαδικοποίηση (ενότητα 2.1.3). Για κάθε ασθενή παράγουν δύο προβλέψεις. Η πρώτη είναι ένας βαθμός κινδύνου, που κατατάσσει τους ασθενείς από τον υψηλότερο στον χαμηλότερο κίνδυνο και αξιολογείται με τον δείκτη συμφωνίας (C-index). Η δεύτερη είναι μια εκτιμώμενη συνάρτηση επιβίωσης $S(t \mid x)$, δηλαδή η πιθανότητα ο ασθενής να επιβιώσει πέρα από τον χρόνο t. Αυτή απαιτείται για τους δείκτες που εξαρτώνται από τον χρόνο, όπως το χρονοεξαρτώμενο AUC και το κριτήριο Brier.

2.2.1 Cox proportional hazards

Το μοντέλο του Cox (1972) εκφράζει τον κίνδυνο (hazard) ενός ασθενούς με χαρακτηριστικά x τη χρονική στιγμή t ως $h(t \mid x) = h_0(t) \cdot \exp(\beta^T x)$. Ο βασικός κίνδυνος $h_0(t)$ είναι κοινός για όλους τους ασθενείς και δεν προσδιορίζεται παραμετρικά. Γι' αυτό το μοντέλο χαρακτηρίζεται ημιπαραμετρικό. Κάθε συντελεστής β_j είναι λογάριθμος λόγου κινδύνων (hazard ratio, HR): η αύξηση του χαρακτηριστικού x_j κατά μία μονάδα πολλαπλασιάζει τον κίνδυνο επί $\exp(\beta_j)$, σε κάθε χρονική στιγμή. Η άμεση αυτή ερμηνεία καθιστά το μοντέλο

πρότυπο της κλινικής ανάλυσης επιβίωσης και φυσική γραμμή αναφοράς για κάθε πιο σύνθετο αλγόριθμο.

Οι συντελεστές εκτιμώνται με μεγιστοποίηση της μερικής πιθανότητας (partial likelihood). Σε κάθε χρόνο θανάτου, η μερική πιθανότητα συγκρίνει τον ασθενή που πέθανε με όσους βρίσκονταν ακόμη υπό παρακολούθηση εκείνη τη στιγμή (σύνολο κινδύνου). Ο βασικός κίνδυνος απαλείφεται από τον λόγο και δεν χρειάζεται για τον υπολογισμό των συντελεστών. Όταν περισσότεροι από ένας θάνατοι καταγράφονται στον ίδιο χρόνο (ties), εφαρμόζεται η προσέγγιση του Efron (1977). Αφού εκτιμηθούν οι συντελεστές, ο βασικός αθροιστικός κίνδυνος $H_0(t)$ εκτιμάται με τη μέθοδο του Breslow (1974) και η συνάρτηση επιβίωσης κάθε ασθενούς προκύπτει ως $S(t \mid x) = \exp[-H_0(t) \cdot \exp(\beta^T x)]$.

Σε μικρά δείγματα με πολλές ή σπάνιες δυαδικές μεταβλητές η μεγιστοποίηση μπορεί να μη συγκλίνει. Όταν, για παράδειγμα, κανείς φορέας μιας σπάνιας αλλοίωσης δεν καταλήγει, ή όλοι οι φορείς καταλήγουν νωρίτερα από κάθε μη φορέα, η πιθανότητα αυξάνεται μονότονα και ο αντίστοιχος συντελεστής αποκλίνει προς το άπειρο. Για τον λόγο αυτό προστίθεται στην αρνητική λογαριθμική μερική πιθανότητα μια ποινή ridge (L2), $(\alpha/2) \cdot \|\beta\|^2$, που συρρικνώνει τους συντελεστές προς το μηδέν χωρίς να τους μηδενίζει (Verweij & Van Houwelingen, 1994). Στην παρούσα εργασία η ποινή είναι σταθερή ($\alpha = 1$) και δεν ρυθμίζεται με βάση την επίδοση. Επειδή τα χαρακτηριστικά είναι τυποποιημένα, η ποινή αυτή ισοδυναμεί με κανονική εκ των προτέρων κατανομή $N(0, 1)$ για τον λογάριθμο του HR ανά τυπική απόκλιση: σταθεροποιεί την εκτίμηση χωρίς να επιλέγει μεταβλητές.

Το μοντέλο στηρίζεται σε δύο υποθέσεις. Η πρώτη είναι ότι κάθε χαρακτηριστικό επιδρά γραμμικά και προσθετικά στον λογάριθμο του κινδύνου. Η δεύτερη είναι ότι οι λόγοι κινδύνων μένουν σταθεροί στον χρόνο (υπόθεση αναλογικών κινδύνων), κάτι που ελέγχεται με τα υπόλοιπα Schoenfeld (Schoenfeld, 1982). Τη γραμμικότητα χαλαρώνουν και οι δύο αλγόριθμοι δέντρων των ενοτήτων 2.2.3 και 2.2.4, ενώ την υπόθεση αναλογικών κινδύνων μόνο το random survival forest.

2.2.2 Cox elastic net

Όταν τα χαρακτηριστικά είναι πολλά σε σχέση με τα συμβάντα, η ποινή μπορεί να αναλάβει και την επιλογή μεταβλητών. Η ποινή lasso (L1) προτάθηκε

για τη γραμμική παλινδρόμηση (Tibshirani, 1996) και επεκτάθηκε στο μοντέλο Cox (Tibshirani, 1997). Σε αντίθεση με την ridge, μηδενίζει ακριβώς τους συντελεστές των λιγότερο χρήσιμων μεταβλητών και παράγει έτσι αραιό μοντέλο. Μειονέκτημά της είναι ότι, μεταξύ ισχυρά συσχετισμένων μεταβλητών, τείνει να κρατά αυθαίρετα μία και να απορρίπτει τις υπόλοιπες.

Το elastic net (Zou & Hastie, 2005) συνδυάζει τις δύο ποινές: $\alpha \cdot [\rho \cdot \|\beta\|_1 + (1 - \rho)/2 \cdot \|\beta\|^2]$. Η παράμετρος α ορίζει τη συνολική ένταση της ποινής και η ρ (l1_ratio) το μερίδιο της L1. Για $\rho = 1$ προκύπτει το καθαρό lasso, ενώ για $\rho < 1$ το τμήμα ridge συγκρατεί ομάδες συσχετισμένων μεταβλητών αντί να επιλέγει μία από αυτές. Η ιδιότητα αυτή έχει σημασία στα γονιδιωματικά δεδομένα, όπου αλλοιώσεις γονιδίων του ίδιου μονοπατιού ή διαφορετικές αλλοιώσεις του ίδιου γονιδίου μπορεί να συσχετίζονται. Η εκτίμηση γίνεται με κατάβαση συντεταγμένων (coordinate descent), μέθοδο που χειρίζεται αποδοτικά τη μη παραγωγίσιμη ποινή L1 (Simon et al., 2011).

Οι α και ρ επιλέγονται με εσωτερική διασταυρωμένη επικύρωση (ενότητα 2.4), με τον κανόνα του ενός τυπικού σφάλματος (Hastie et al., 2009): από τους συνδυασμούς των οποίων η επίδοση απέχει το πολύ ένα τυπικό σφάλμα από τη βέλτιστη, επιλέγεται εκείνος με την ισχυρότερη ποινή. Ο κανόνας προτιμά το απλούστερο μοντέλο όταν οι διαφορές επίδοσης δεν ξεχωρίζουν από τον θόρυβο, κάτι συνηθισμένο σε μικρά δείγματα.

2.2.3 Random survival forest

Το random survival forest (RSF) μεταφέρει τη λογική του random forest (Breiman, 2001) στα λογοκριμένα δεδομένα (Ishwaran et al., 2008). Κάθε δέντρο εκπαιδεύεται σε ένα δείγμα bootstrap των ασθενών. Σε κάθε κόμβο εξετάζεται ένα τυχαίο υποσύνολο των χαρακτηριστικών και επιλέγεται ο διαχωρισμός που μεγιστοποιεί τη διαφορά επιβίωσης ανάμεσα στους δύο θυγατρικούς κόμβους, όπως τη μετρά το στατιστικό log-rank. Σε κάθε φύλλο, η αθροιστική συνάρτηση κινδύνου εκτιμάται με τον εκτιμητή Nelson–Aalen από τους ασθενείς εκπαίδευσης που ανήκουν σε αυτό. Η πρόβλεψη του δάσους για έναν νέο ασθενή είναι ο μέσος όρος των εκτιμήσεων όλων των δέντρων. Ως βαθμός κινδύνου χρησιμοποιείται η αναμενόμενη θνησιμότητα (ensemble mortality), δηλαδή το άθροισμα του αθροιστικού κινδύνου στους παρατηρημένους χρόνους συμβάντων.

Επειδή δεν υποθέτει ούτε γραμμικότητα ούτε αναλογικούς κινδύνους, το RSF μπορεί να αναδείξει μη γραμμικές επιδράσεις και αλληλεπιδράσεις χωρίς να οριστούν εκ των προτέρων. Η διπλή τυχαιότητα, σε ασθενείς και σε χαρακτηριστικά, κάνει τα δέντρα να διαφέρουν μεταξύ τους, ώστε ο μέσος όρος τους να έχει πολύ μικρότερη διακύμανση από ένα μεμονωμένο δέντρο. Οι κρίσιμες υπερπαράμετροι είναι δύο. Το ελάχιστο πλήθος ασθενών ανά φύλλο καθορίζει πόσο λεπτομερείς γίνονται οι εκτιμήσεις: ένα φύλλο με λίγους ασθενείς και ελάχιστα συμβάντα δίνει ασταθή καμπύλη επιβίωσης, γι' αυτό σε μικρά δείγματα χρειάζονται μεγαλύτερα φύλλα. Το πλήθος των χαρακτηριστικών που εξετάζονται σε κάθε διαχωρισμό ρυθμίζει πόσο διαφέρουν τα δέντρα μεταξύ τους.

Το τίμημα της ευελιξίας είναι η ερμηνευσιμότητα. Το RSF δεν παράγει συντελεστές ή λόγους κινδύνων, οπότε η συμβολή κάθε μεταβλητής εκτιμάται έμμεσα, για παράδειγμα με τη σημαντικότητα μετάθεσης (permutation importance): πόσο πέφτει ο δείκτης συμφωνίας όταν οι τιμές της μεταβλητής ανακατεύονται τυχαία στο σύνολο ελέγχου.

2.2.4 Gradient boosting

Το gradient boosting χτίζει διαδοχικά μικρά δέντρα παλινδρόμησης. Κάθε νέο δέντρο προσαρμόζεται στο αρνητικό της κλίσης της συνάρτησης απώλειας ως προς τις τρέχουσες προβλέψεις, δηλαδή στο μέρος της έκβασης που τα προηγούμενα δέντρα δεν έχουν ακόμη εξηγήσει (Friedman, 2001). Για δεδομένα επιβίωσης, ως απώλεια χρησιμοποιείται η αρνητική λογαριθμική μερική πιθανότητα του Cox (Ridgeway, 1999). Το αποτέλεσμα είναι μοντέλο αναλογικών κινδύνων της μορφής $h(t \mid x) = h_0(t) \cdot \exp(f(x))$, όπου ο γραμμικός όρος $\beta^T x$ έχει αντικατασταθεί από το άθροισμα των δέντρων $f(x)$. Η συνάρτηση επιβίωσης προκύπτει, όπως και στο μοντέλο Cox, με τον εκτιμητή Breslow.

Τρεις υπερπαράμετροι ελέγχουν την πολυπλοκότητα του μοντέλου. Το βάθος των δέντρων καθορίζει τον βαθμό των αλληλεπιδράσεων: δέντρα βάθους 1 (κολοβώματα, stumps) δίνουν προσθετικό μοντέλο, με μη γραμμική επίδραση για κάθε μεταβλητή χωριστά, ενώ δέντρα βάθους 2 επιτρέπουν αλληλεπιδράσεις ανά ζεύγη μεταβλητών. Ο ρυθμός μάθησης πολλαπλασιάζει τη συμβολή κάθε δέντρου και, μαζί με το πλήθος των δέντρων, καθορίζει πόσο στενά προσαρμόζεται το μοντέλο στα δεδομένα εκπαίδευσης. Επιπλέον, κάθε

δέντρο εκπαιδεύεται σε τυχαίο υποσύνολο των ασθενών (στην παρούσα εργασία στο 80%), τεχνική γνωστή ως stochastic gradient boosting (Friedman, 2002), που μειώνει τη διακύμανση και λειτουργεί ως πρόσθετη κανονικοποίηση.

Οι δύο αλγόριθμοι δέντρων ακολουθούν αντίθετες στρατηγικές. Το random survival forest συνδυάζει βαθιά, ανεξάρτητα δέντρα και ο μέσος όρος τους μειώνει κυρίως τη διακύμανση. Η προσθήκη περισσότερων δέντρων δεν οδηγεί σε υπερπροσαρμογή. Το gradient boosting συνδυάζει ρηχά δέντρα, καθένα εξαρτημένο από τα προηγούμενα, και μειώνει κυρίως τη μεροληψία. Εδώ το πλήθος των δέντρων πρέπει να περιορίζεται, διότι με πάρα πολλά δέντρα το μοντέλο αρχίζει να προσαρμόζεται στον θόρυβο.

2.3 Λίγα συμβάντα, πολλές μεταβλητές

Στα ωμικά δεδομένα το πλήθος των χαρακτηριστικών p υπερβαίνει συχνά κατά τάξεις μεγέθους το πλήθος των δειγμάτων n. Σε τέτοιο καθεστώς, οποιοδήποτε σύνολο ετικετών μπορεί να διαχωριστεί τέλεια από κάποιον γραμμικό συνδυασμό χαρακτηριστικών, ακόμη και αν οι ετικέτες είναι εντελώς τυχαίες.

Στην ανάλυση επιβίωσης το κρίσιμο μέγεθος δεν είναι μόνο το πλήθος των ασθενών αλλά και το πλήθος των συμβάντων, διότι η πληροφορία για τον κίνδυνο προέρχεται κυρίως από τους θανάτους. Οι Peduzzi et al. (1995) έδειξαν με προσομοιώσεις ότι, όταν τα συμβάντα ανά μεταβλητή (events per variable, EPV) μειώνονται κάτω από περίπου δέκα, οι συντελεστές του μοντέλου Cox γίνονται μεροληπτικοί και τα διαστήματα εμπιστοσύνης χάνουν την ονομαστική τους κάλυψη. Νεότερες προσεγγίσεις αντικαθιστούν τον σταθερό εμπειρικό κανόνα με υπολογισμό του ελάχιστου μεγέθους δείγματος, ώστε η αναμενόμενη υπερπροσαρμογή να είναι μικρή (Riley et al., 2019).

Στην παρούσα εργασία, με 76 θανάτους συνολικά, κάθε σύνολο εκπαίδευσης περιέχει περίπου 61 θανάτους, ενώ το πληρέστερο σύνολο εισόδων περιλαμβάνει έως 33 μεταβλητές μαζί με τους δείκτες έλλειψης (ενότητα 5.3). Αντιστοιχούν δηλαδή λιγότεροι από δύο θάνατοι σε κάθε μεταβλητή. Το πρόβλημα δεν είναι $p \gg n$ με την αυστηρή έννοια, έχει όμως τις ίδιες συνέπειες: χωρίς κανονικοποίηση, ένα μοντέλο προσαρμόζεται εύκολα στον θόρυβο του συνόλου εκπαίδευσης.

2.3.1 Επιλογή χαρακτηριστικών

Οι μέθοδοι επιλογής διακρίνονται σε τρεις οικογένειες (Guyon & Elisseeff, 2003). Οι μέθοδοι φίλτρου αξιολογούν κάθε χαρακτηριστικό ανεξάρτητα από το μοντέλο, με στατιστικά κριτήρια όπως ο έλεγχος F ή η αμοιβαία πληροφορία. Είναι ταχύτατες αλλά αγνοούν τις αλληλεπιδράσεις. Οι ενσωματωμένες μέθοδοι εκτελούν την επιλογή κατά την εκπαίδευση, όπως το lasso ή η σημαντικότητα των δέντρων. Οι μέθοδοι περιτύλιξης αξιολογούν υποσύνολα χαρακτηριστικών εκπαιδεύοντας επανειλημμένα το μοντέλο, με χαρακτηριστικότερο παράδειγμα την αναδρομική απαλοιφή. Είναι οι πλέον δαπανηρές υπολογιστικά.

Στην παρούσα εργασία εφαρμόζονται δύο μορφές επιλογής. Η πρώτη είναι ένα φίλτρο συχνότητας, που αποκλείει τις πολύ σπάνιες γονιδιωματικές αλλοιώσεις χωρίς να εξετάζει την έκβαση (ενότητα 5.3.2). Η δεύτερη είναι η ενσωματωμένη επιλογή της ποινής L1 του μοντέλου Cox με ποινή elastic net (ενότητα 2.2.2).

2.3.2 Κανονικοποίηση

Η κανονικοποίηση περιορίζει την πολυπλοκότητα του μοντέλου προσθέτοντας όρο ποινής στη συνάρτηση απώλειας. Λειτουργεί ως ρητή ενσωμάτωση της παραδοχής ότι οι περισσότεροι συντελεστές είναι μικροί ή μηδενικοί — παραδοχή βιολογικά εύλογη, καθώς ελάχιστα γονίδια σχετίζονται πραγματικά με μια δεδομένη έκβαση.

Στο μοντέλο Cox η κανονικοποίηση έχει τη μορφή των ποινών ridge, lasso και elastic net της ενότητας 2.2. Στα σύνολα δέντρων τον ίδιο ρόλο έχουν το ελάχιστο πλήθος ασθενών ανά φύλλο, το βάθος των δέντρων, ο ρυθμός μάθησης και η εκπαίδευση κάθε δέντρου σε υποσύνολο των ασθενών.

2.3.3 Διαρροή πληροφορίας

Η διαρροή πληροφορίας (data leakage) αποτελεί το συχνότερο και σοβαρότερο σφάλμα σε εργασίες μηχανικής μάθησης επί ωμικών δεδομένων. Συμβαίνει όποτε πληροφορία από το σύνολο ελέγχου επηρεάζει την εκπαίδευση.

Η κλασική μορφή της είναι η επιλογή χαρακτηριστικών εκτός της διασταυρωμένης επικύρωσης. Οι Ambroise και McLachlan (2002) κατέδειξαν ότι η πρακτική αυτή παράγει σχεδόν μηδενικά ποσοστά σφάλματος ακόμη και σε

δεδομένα με τυχαία ανακατεμένες ετικέτες. Το σφάλμα είναι ιδιαίτερα ύπουλο διότι δεν παράγει καμία προειδοποίηση: το μοντέλο φαίνεται απλώς εξαιρετικό.

Στην παρούσα εργασία κάθε βήμα που εκτιμά παραμέτρους από τα δεδομένα —η συμπλήρωση ελλειπουσών τιμών, η τυποποίηση και η επιλογή υπερπαραμέτρων— εκτελείται αποκλειστικά εντός του σωλήνα επεξεργασίας (pipeline) και προσαρμόζεται μόνο στο εκάστοτε σύνολο εκπαίδευσης (ενότητα 5.4).

2.4 Εμφωλευμένη διασταυρωμένη επικύρωση

Η απλή διασταυρωμένη επικύρωση επαρκεί για την εκτίμηση της επίδοσης ενός μοντέλου με προκαθορισμένες υπερπαραμέτρους. Όταν όμως οι υπερπαράμετροι επιλέγονται με βάση την ίδια διαδικασία, η προκύπτουσα εκτίμηση είναι αισιόδοξα μεροληπτική, καθώς το σύνολο ελέγχου έχει ήδη χρησιμοποιηθεί για επιλογή (Varma & Simon, 2006 και Cawley & Talbot, 2010).

Η εμφωλευμένη διασταυρωμένη επικύρωση διαχωρίζει τις δύο λειτουργίες σε δύο επίπεδα. Ο εξωτερικός βρόχος εκτιμά αμερόληπτα την επίδοση, ενώ ο εσωτερικός, εκτελούμενος αποκλειστικά εντός του εκπαιδευτικού τμήματος κάθε εξωτερικής επανάληψης, επιλέγει τις υπερπαραμέτρους. Το σύνολο ελέγχου του εξωτερικού βρόχου παραμένει έτσι πραγματικά αχρησιμοποίητο.

Επιπλέον, με 76 θανάτους ένα μοναδικό σύνολο ελέγχου 20% θα περιείχε περίπου δεκαπέντε θανάτους, και το διάστημα εμπιστοσύνης του δείκτη συμφωνίας θα ήταν τόσο ευρύ ώστε καμία σύγκριση να μην είναι ουσιαστική. Η επανάληψη ολόκληρης της διαδικασίας με διαφορετικούς διαχωρισμούς αποτυπώνει τη μεταβλητότητα που οφείλεται στην τύχη του διαχωρισμού. Στα δεδομένα επιβίωσης οι διαχωρισμοί γίνονται με διαστρωμάτωση ως προς τον δείκτη συμβάντος, ώστε κάθε υποσύνολο να περιέχει παρόμοια αναλογία θανάτων και λογοκριμένων ασθενών.

Για τη σύγκριση δύο μοντέλων απαιτείται και εκτίμηση της αβεβαιότητας. Οι τιμές ενός δείκτη στα επιμέρους υποσύνολα της επαναλαμβανόμενης διασταυρωμένης επικύρωσης δεν είναι ανεξάρτητες, διότι τα σύνολα εκπαίδευσης επικαλύπτονται κατά μεγάλο μέρος. Ένα απλό t-test θα υποεκτιμούσε τη διασπορά και θα έδινε υπερβολικά στενά διαστήματα. Οι Nadeau και Bengio (2003) πρότειναν διόρθωση της διασποράς του μέσου

όρου: η δειγματική διασπορά s^2 των k τιμών πολλαπλασιάζεται με $(1/k + n_2/n_1)$ αντί για 1/k, όπου n_1 και n_2 τα μεγέθη των συνόλων εκπαίδευσης και ελέγχου. Σε πενταπλή διασταυρωμένη επικύρωση ο όρος n_2/n_1 ισούται με 1/4. Στην παρούσα εργασία η διόρθωση εφαρμόζεται τόσο στα διαστήματα εμπιστοσύνης των δεικτών όσο και στις ζευγαρωμένες διαφορές μεταξύ μοντέλων που αξιολογήθηκαν στα ίδια υποσύνολα.

2.5 Δείκτες αξιολόγησης μοντέλων επιβίωσης

Ένα προγνωστικό μοντέλο επιβίωσης αξιολογείται σε δύο άξονες (Steyerberg et al., 2010). Ο πρώτος είναι η διάκριση (discrimination): αν το μοντέλο κατατάσσει σωστά τους ασθενείς από τον υψηλότερο στον χαμηλότερο κίνδυνο. Ο δεύτερος είναι η βαθμονόμηση (calibration): αν οι προβλεπόμενες πιθανότητες επιβίωσης αντιστοιχούν στις συχνότητες που παρατηρούνται. Οι δείκτες της δυαδικής ταξινόμησης, όπως η ακρίβεια, δεν εφαρμόζονται άμεσα, διότι για τους λογοκριμένους ασθενείς η κατάσταση σε έναν μεταγενέστερο χρόνο είναι άγνωστη.

2.5.1 Δείκτης συμφωνίας

Ο δείκτης συμφωνίας (concordance index, C-index) του Harrell εξετάζει τα συγκρίσιμα ζεύγη ασθενών, δηλαδή τα ζεύγη στα οποία ο ασθενής με τον μικρότερο παρατηρημένο χρόνο πέθανε (Harrell et al., 1996). Ένα ζεύγος είναι σύμφωνο όταν ο ασθενής που πέθανε πρώτος είχε και τον μεγαλύτερο προβλεπόμενο κίνδυνο. Ο δείκτης ισούται με το ποσοστό των σύμφωνων ζευγών: η τιμή 0,5 αντιστοιχεί σε τυχαία κατάταξη και η τιμή 1 σε τέλεια. Επειδή κρίνει μόνο τη σειρά των ασθενών, ο δείκτης δεν επηρεάζεται από τη βαθμονόμηση των προβλέψεων.

Η εκτίμηση του Harrell εξαρτάται από την κατανομή της λογοκρισίας στο συγκεκριμένο δείγμα. Οι Uno et al. (2011) πρότειναν παραλλαγή στην οποία κάθε ζεύγος σταθμίζεται με το αντίστροφο της πιθανότητας να μην έχει λογοκριθεί ο ασθενής (inverse probability of censoring weighting, IPCW). Ο δείκτης του Uno εκτιμά έναν δείκτη συμφωνίας που δεν εξαρτάται από τη λογοκρισία, έως έναν προκαθορισμένο χρόνο τ.

2.5.2 Χρονοεξαρτώμενη AUC

Η χρονοεξαρτώμενη AUC μεταφέρει την ανάλυση ROC σε έναν σταθερό χρόνο t. Στον αθροιστικό/δυναμικό ορισμό, οι ασθενείς που πέθαναν έως τον t αποτελούν τις «περιπτώσεις» και όσοι ζουν πέρα από τον t τους «μάρτυρες». Η AUC(t) είναι η πιθανότητα ένας τυχαίος ασθενής της πρώτης ομάδας να έχει μεγαλύτερο προβλεπόμενο κίνδυνο από έναν τυχαίο ασθενή της δεύτερης (Heagerty & Zheng, 2005). Οι ασθενείς που λογοκρίθηκαν πριν από τον t λαμβάνονται υπόψη με στάθμιση IPCW. Ο δείκτης απαντά σε ένα κλινικά συγκεκριμένο ερώτημα, για παράδειγμα πόσο καλά ξεχωρίζει το μοντέλο τους ασθενείς που θα πεθάνουν μέσα στον πρώτο χρόνο.

2.5.3 Κριτήριο Brier

Το κριτήριο Brier σε χρόνο t είναι το μέσο τετραγωνικό σφάλμα μεταξύ της προβλεπόμενης πιθανότητας επιβίωσης S(t | x) και της παρατηρούμενης κατάστασης του ασθενούς, η οποία ισούται με 1 αν ο ασθενής ζει στον t και με 0 αν έχει πεθάνει. Η λογοκρισία αντιμετωπίζεται με στάθμιση IPCW (Graf et al., 1999). Το κριτήριο αξιολογεί ταυτόχρονα διάκριση και βαθμονόμηση, και μικρότερες τιμές είναι καλύτερες. Το ολοκληρωμένο κριτήριο Brier (integrated Brier score, IBS) είναι ο μέσος όρος του κριτηρίου σε ένα διάστημα χρόνων.

Ως σημείο αναφοράς χρησιμοποιείται η καμπύλη Kaplan–Meier του συνόλου εκπαίδευσης, η οποία αποδίδει την ίδια πιθανότητα επιβίωσης σε όλους τους ασθενείς. Ένα μοντέλο με κριτήριο Brier ίσο ή μεγαλύτερο από εκείνο της καμπύλης Kaplan–Meier δεν προσφέρει χρήσιμη εξατομικευμένη πληροφορία, ακόμη κι αν κατατάσσει τους ασθενείς καλύτερα από την τύχη.

2.6 Πολλαπλοί έλεγχοι

Όταν ελέγχονται πολλές υποθέσεις ταυτόχρονα, όπως η συσχέτιση καθεμίας από δεκάδες γονιδιωματικές αλλοιώσεις με την επιβίωση, το κατώφλι $p < 0,05$ χάνει τη συνήθη ερμηνεία του. Με 20 ανεξάρτητους ελέγχους χωρίς καμία πραγματική συσχέτιση αναμένεται κατά μέσο όρο ένα «σημαντικό» αποτέλεσμα από τύχη. Η διόρθωση Bonferroni ελέγχει την πιθανότητα να προκύψει έστω ένα ψευδώς θετικό εύρημα, αλλά γίνεται πολύ συντηρητική όταν οι έλεγχοι είναι πολλοί.

Οι Benjamini και Hochberg (1995) πρότειναν τον έλεγχο του ρυθμού ψευδών ανακαλύψεων (false discovery rate, FDR), δηλαδή του αναμενόμενου ποσοστού ψευδώς θετικών μεταξύ των ευρημάτων που δηλώνονται σημαντικά. Οι m τιμές p κατατάσσονται σε αύξουσα σειρά, $p_{(1)} \leq \ldots \leq p_{(m)}$, και απορρίπτονται οι μηδενικές υποθέσεις έως τη μεγαλύτερη θέση i για την οποία $p_{(i)} \leq (i/m) \cdot \alpha$. Ισοδύναμα, για κάθε έλεγχο υπολογίζεται μια προσαρμοσμένη τιμή q, η οποία συγκρίνεται απευθείας με το επιθυμητό επίπεδο FDR. Στην παρούσα εργασία οι τιμές p αναφέρονται μαζί με τις τιμές q, ώστε να διακρίνεται ένα ονομαστικά σημαντικό εύρημα ($p < 0,05$) από ένα εύρημα που παραμένει σημαντικό μετά τη διόρθωση ($q < 0,05$).

2.7 Συσχέτιση και πρόβλεψη

Ένας προγνωστικός παράγοντας και ένα προγνωστικό μοντέλο απαντούν σε διαφορετικά ερωτήματα. Η συσχέτιση μιας μεταβλητής με την έκβαση εκφράζεται με ένα μέτρο επίδρασης, όπως ο λόγος κινδύνων, και εκτιμάται σε ολόκληρο το δείγμα. Η προγνωστική ικανότητα ενός μοντέλου εκφράζεται με δείκτες όπως ο δείκτης συμφωνίας και αξιολογείται σε ασθενείς που δεν χρησιμοποιήθηκαν στην εκπαίδευσή του.

Οι δύο έννοιες δεν ταυτίζονται. Οι Pepe et al. (2004) έδειξαν ότι ένας δείκτης με λόγο πιθανοτήτων ίσο με 3, τιμή που συνήθως θεωρείται ισχυρή συσχέτιση, μπορεί να είναι πολύ φτωχό εργαλείο ταξινόμησης. Ανάλογο φαινόμενο αναμένεται στις σπάνιες γονιδιωματικές αλλοιώσεις: μια αλλοίωση με μεγάλο λόγο κινδύνων αφορά λίγους ασθενείς, οπότε αλλάζει τη σειρά λίγων μόνο ζευγών και ελάχιστα τον δείκτη συμφωνίας ολόκληρης της κοόρτης. Για τον λόγο αυτό η παρούσα εργασία εξετάζει χωριστά την πρόβλεψη (ενότητες 6.2–6.4) και τη συσχέτιση (ενότητα 6.5).

3. ΒΙΒΛΙΟΓΡΑΦΙΚΗ ΕΠΙΣΚΟΠΗΣΗ

Το κεφάλαιο επισκοπεί τη βιβλιογραφία σε τρεις άξονες: τους γονιδιωματικούς και μοριακούς δείκτες που έχουν συνδεθεί με την έκβαση ασθενών με mCRPC υπό θεραπεία ARSI, τα μοντέλα πρόβλεψης της επιβίωσης και τη θέση της μηχανικής μάθησης σε αυτά, και τη σχετική ελληνική έρευνα. Κλείνει με μια κριτική αποτίμηση που τοποθετεί την παρούσα εργασία σε αυτό το πλαίσιο.

3.1 Γονιδιωματικοί προγνωστικοί δείκτες στο mCRPC υπό θεραπεία ARSI

Οι δείκτες που έχουν μελετηθεί προέρχονται είτε από βιοψίες ιστού είτε από την υγρή βιοψία, δηλαδή από κυκλοφορούντα καρκινικά κύτταρα (circulating tumor cells, CTC) ή από DNA του όγκου που κυκλοφορεί στο πλάσμα (circulating tumor DNA, ctDNA). Η υγρή βιοψία επιτρέπει επαναλαμβανόμενες μετρήσεις χωρίς επεμβατική λήψη ιστού. Ο Πίνακας 2 συνοψίζει τις μελέτες που παρουσιάζονται στη συνέχεια.

Πίνακας 2. Επιλεγμένες μελέτες γονιδιωματικών και μοριακών δεικτών σε ασθενείς με mCRPC υπό θεραπεία με abiraterone ή enzalutamide.

Μελέτη	Ασθενείς και υλικό	Δείκτης	Κύριο εύρημα
Antonarakis et al. (2014)	62 (enzalutamide ή abiraterone), CTC	AR-V7	AR-V7: καμία ανταπόκριση PSA στο enzalutamide (0% έναντι 53%), βραχύτερη PFS και OS
Efstathiou et al. (2015)	60 με οστικές μεταστάσεις, μυελός οστών	AR, AR-V7, CYP17	AR-V7: πρωτογενής αντίσταση στο enzalutamide. Υψηλή έκφραση AR και CYP17: όφελος
Ferraldeschi et al. (2015)	144 (abiraterone μετά από docetaxel), ιστός (IHC)	Απώλεια PTEN	OS 14 έναντι 21 μηνών (HR 1,75)
Romanel et al. (2015)	97 (abiraterone), πλάσμα	Αύξηση αντιγράφων και μεταλλάξεις AR	Χειρότερη OS (HR 7,33) και PFS
Wyatt et al. (2016)	65 (enzalutamide), cfDNA	Ενίσχυση AR, μεταλλάξεις AR, απώλεια RB1	Βραχύτερη PFS (HR 2,92, 3,94 και 4,46 αντίστοιχα)
Conteduca et al. (2017)	171 + 94 (enzalutamide ή abiraterone), πλάσμα	Αύξηση αντιγράφων AR	Χειρότερη OS (HR 3,98 πριν και 3,81 μετά τη χημειοθεραπεία)
Annala et al. (2018)	202 (πρώτη γραμμή), ctDNA	BRCA2, ATM, TP53, AR	BRCA2/ATM και TP53: χειρότερη έκβαση ανεξάρτητα από κλινικούς παράγοντες. AR: όχι
De Laere et al. (2019)	168 (ARSI), CTC και ctDNA	AR-V, AR, TP53	Μόνο η αδρανοποίηση του TP53 ανεξάρτητη (HR 1,88 για PFS)
Chen et al. (2019)	101, αλληλούχιση ολικού γονιδιώματος	RB1, CTNNB1	Δύο αλλοιώσεις RB1: OS 14,1 έναντι 42,0 μηνών

OS: συνολική επιβίωση, PFS: επιβίωση χωρίς εξέλιξη, CTC: κυκλοφορούντα καρκινικά κύτταρα, ctDNA/cfDNA: DNA του όγκου/ελεύθερο DNA στο πλάσμα, IHC: ανοσοϊστοχημεία.

3.1.1 Ο ανδρογονικός υποδοχέας

Επειδή οι ARSI στοχεύουν τον άξονα του AR, οι αλλοιώσεις του ίδιου του υποδοχέα ήταν οι πρώτοι υποψήφιοι δείκτες αντίστασης. Η παραλλαγή ματίσματος AR-V7 δεν έχει την περιοχή σύνδεσης των ανδρογόνων, αλλά παραμένει μόνιμα ενεργή ως μεταγραφικός παράγοντας. Σε 62 ασθενείς που ξεκινούσαν enzalutamide ή abiraterone, κανένας ασθενής με ανιχνεύσιμο AR-V7 στα CTC δεν είχε ανταπόκριση PSA στο enzalutamide, έναντι 53% όσων δεν το είχαν, και η επιβίωση χωρίς εξέλιξη και η συνολική επιβίωση ήταν βραχύτερες (Antonarakis et al., 2014). Ανάλογα, σε βιοψίες μυελού των οστών η παρουσία του AR-V7 συσχετίστηκε με πρωτογενή αντίσταση στο enzalutamide, ενώ η υψηλή πυρηνική έκφραση του AR και η έκφραση του CYP17 με όφελος (Efstathiou et al., 2015).

Στο επίπεδο του DNA, οι αλλοιώσεις του AR είναι σπάνιες πριν από την ορμονοθεραπεία αλλά εμφανίζονται με την ανθεκτικότητα στον ευνουχισμό. Σε 97 ασθενείς που έλαβαν abiraterone, η αύξηση των αντιγράφων του AR ή οι

μεταλλάξεις T878A και L702H στο πλάσμα πριν από τη θεραπεία συσχετίστηκαν με πολύ χειρότερη συνολική επιβίωση (HR 7,33, Romanel et al., 2015). Σε πολυκεντρική μελέτη με ψηφιακή PCR, η αύξηση των αντιγράφων του AR στο πλάσμα βρέθηκε στο 14% των ασθενών πριν από τη χημειοθεραπεία και στο 34% μετά από αυτήν, και συσχετίστηκε με χειρότερη συνολική επιβίωση υπό enzalutamide ή abiraterone (HR 3,98 και 3,81 αντίστοιχα, Conteduca et al., 2017). Σε 65 ασθενείς που έλαβαν enzalutamide, η ενίσχυση του AR, οι πολλαπλές μεταλλάξεις του και η απώλεια του RB1 στο ελεύθερο DNA του πλάσματος συσχετίστηκαν με βραχύτερη επιβίωση χωρίς εξέλιξη (Wyatt et al., 2016).

3.1.2 Ογκοκατασταλτικά γονίδια και επιδιόρθωση του DNA

Η μεγαλύτερη μελέτη ctDNA σε θεραπεία πρώτης γραμμής τυχαιοποίησε 202 ασθενείς σε abiraterone ή enzalutamide και ανέλυσε το ctDNA πριν από την έναρξη. Ελλείμματα των BRCA2 και ATM συσχετίστηκαν ισχυρά με χειρότερη έκβαση, ανεξάρτητα από τους κλινικούς προγνωστικούς παράγοντες και από την ποσότητα του ctDNA, ενώ οι αλλοιώσεις του TP53 συσχετίστηκαν ανεξάρτητα με ταχεία αντίσταση. Αντίθετα, η ανίχνευση ενίσχυσης του AR δεν υπερείχε των καθιερωμένων προγνωστικών δεικτών (Annala et al., 2018). Σε 168 ασθενείς από 10 ευρωπαϊκά κέντρα, καμία μεμονωμένη αλλοίωση του AR δεν παρέμεινε συσχετισμένη με την έκβαση μετά από πολυμεταβλητή ανάλυση. Σημαντικοί ήταν οι δείκτες φορτίου της νόσου και η αδρανοποίηση του TP53 (HR 1,88 για την επιβίωση χωρίς εξέλιξη, De Laere et al., 2019).

Στους ιστούς, η απώλεια της πρωτεΐνης PTEN βρέθηκε στο 40% από 144 ασθενείς που έλαβαν abiraterone μετά από docetaxel και συσχετίστηκε με βραχύτερη συνολική επιβίωση (14 έναντι 21 μηνών, HR 1,75) και βραχύτερη διάρκεια θεραπείας (Ferraldeschi et al., 2015). Σε 285 δείγματα από όλα τα στάδια της νόσου, οι αλλοιώσεις των TP53, PTEN και RB1 γίνονταν συχνότερες όσο προχωρούσε η νόσος και έφταναν το 92% στο μεταστατικό CRPC. Ο συνδυασμός περισσότερων αλλοιώσεων συνδεόταν με χειρότερη έκβαση, αλλά η συσχέτιση δεν παρέμεινε ανεξάρτητη μετά από προσαρμογή για το συνολικό φορτίο μεταλλάξεων και μεταβολών του αριθμού αντιγράφων (Hamid et al., 2019). Σε 101 ασθενείς με mCRPC και αλληλούχιση ολικού γονιδιώματος, η ύπαρξη δύο αλλοιώσεων του RB1 προέβλεπε ανεξάρτητα βραχύτερη συνολική

επιβίωση (διάμεση 14,1 έναντι 42,0 μηνών), όπως και οι μεταλλάξεις της β-κατενίνης (CTNNB1), που βρέθηκαν μόνο σε ασθενείς ανθεκτικούς στο enzalutamide (Chen et al., 2019).

Ο ρόλος των TP53 και RB1 έχει και μηχανιστική τεκμηρίωση. Σε μοντέλα ποντικών, η απώλεια του Rb1 διευκόλυνε την πλαστικότητα της κυτταρικής ταυτότητας (lineage plasticity) και τις μεταστάσεις, ενώ η επιπλέον απώλεια του Trp53 προκάλεσε αντίσταση στην αντιανδρογονική θεραπεία (Ku et al., 2017). Σε ανθρώπινα μοντέλα, η απώλεια της λειτουργίας των TP53 και RB1 επέτρεψε στα καρκινικά κύτταρα να μεταπέσουν σε φαινότυπο ανεξάρτητο από τον AR, μέσω του μεταγραφικού παράγοντα SOX2, και να γίνουν ανθεκτικά στο enzalutamide (Mu et al., 2017). Σε 410 βιοψίες μεταστάσεων, μια υπογραφή έκφρασης που αντανακλά τη συνδυασμένη απώλεια των TP53 και RB1 συσχετίστηκε με μικρότερη ανταπόκριση στους ανταγωνιστές του AR και βραχύτερη επιβίωση, αν και το 40% αυτών των όγκων παρέμενε αδενοκαρκίνωμα με ενεργό AR (Nyquist et al., 2020).

3.2 Μοντέλα πρόβλεψης της επιβίωσης και μηχανική μάθηση

Τα πιο τεκμηριωμένα προγνωστικά μοντέλα στο mCRPC βασίζονται σε κλινικά και εργαστηριακά στοιχεία. Το μοντέλο των Halabi et al. (2014) αναπτύχθηκε σε 1.050 ασθενείς που έλαβαν χημειοθεραπεία πρώτης γραμμής. Με προσαρμοστική ποινή lasso επέλεξε οκτώ παράγοντες: τη γενική κατάσταση κατά ECOG, τη θέση της νόσου, τη γαλακτική αφυδρογονάση, τη χρήση οπιοειδών, τη λευκωματίνη, την αιμοσφαιρίνη, το PSA και την αλκαλική φωσφατάση. Πέτυχε χρονοεξαρτώμενη AUC 0,73 στο σύνολο ελέγχου και 0,76 σε ανεξάρτητο σύνολο επικύρωσης.

Ο διαγωνισμός Prostate Cancer DREAM Challenge διέθεσε ανοικτά τα δεδομένα των ομάδων ελέγχου τεσσάρων μελετών φάσης III για το mCRPC: περίπου 2.000 ασθενείς υπό docetaxel, με περισσότερες από 150 κλινικές μεταβλητές. Από τις 50 μεθόδους που υποβλήθηκαν, η καλύτερη ήταν ένα σύνολο μοντέλων Cox με ποινή (ePCR), με ολοκληρωμένη χρονοεξαρτώμενη AUC 0,791 έναντι 0,743 του μοντέλου αναφοράς με οκτώ κλινικές μεταβλητές (Guinney et al., 2017). Το κέρδος, δηλαδή, ήταν πραγματικό αλλά μέτριο, παρά τον μεγάλο όγκο δεδομένων. Πρόσφατα, ένα απλό σκορ που ενσωματώνει το

φορτίο των μεταστάσεων σε οστά και μαλακά μόρια έδειξε υψηλότερη διάκριση από την κατάταξη Halabi (δείκτης συμφωνίας 0,755 έναντι 0,646) σε 266 ασθενείς. Το εύρημα όμως είναι διερευνητικό και χρειάζεται εξωτερική επικύρωση (Uccello et al., 2026).

Για τα ωμικά δεδομένα, μια μεγάλη συγκριτική μελέτη εφάρμοσε έντεκα μεθόδους, βασισμένες στο boosting, στην παλινδρόμηση με ποινή και στα random forests, σε 18 πολυωμικά σύνολα δεδομένων καρκίνου του TCGA, με επαναλαμβανόμενη πενταπλή διασταυρωμένη επικύρωση και δείκτες τον δείκτη συμφωνίας του Uno και το ολοκληρωμένο κριτήριο Brier. Μόνο μία μέθοδος ξεπέρασε κατά μέσο όρο το μοντέλο Cox με κλινικές μεταβλητές, και μάλιστα οριακά. Η χρησιμότητα των ωμικών δεδομένων για πρόβλεψη ήταν επομένως περιορισμένη (Herrmann et al., 2021). Σε ευρύτερο επίπεδο, συστηματική ανασκόπηση 71 μελετών κλινικής πρόβλεψης δεν βρήκε υπεροχή της μηχανικής μάθησης έναντι της λογιστικής παλινδρόμησης όταν η επικύρωση ήταν αμερόληπτη. Επιπλέον, το 68% των μελετών είχε πιθανή μεροληψία στη διαδικασία επικύρωσης και το 79% δεν αξιολόγησε τη βαθμονόμηση (Christodoulou et al., 2019).

Ανασκοπήσεις των εφαρμογών της μηχανικής μάθησης στην πρόγνωση του καρκίνου καταλήγουν σε ανάλογο συμπέρασμα: οι μέθοδοι αυτές μπορούν να αναδείξουν πληροφορία από σύνθετα δεδομένα, αλλά χρειάζονται κατάλληλη επικύρωση, διαφάνεια και ερμηνευσιμότητα πριν χρησιμοποιηθούν στην κλινική πράξη (Kourou et al., 2015, 2021).

3.3 Ελληνικές μελέτες

Στην Ελλάδα, η πολυκεντρική αναδρομική μελέτη PROSPECT κατέγραψε τη θεραπευτική πρακτική σε 149 ασθενείς με mCRPC από επτά μεγάλα ογκολογικά κέντρα. Το 78,5% έλαβε ARSI ως θεραπεία πρώτης γραμμής (enzalutamide 52,3%, abiraterone 26,2%), και η διάμεση διάρκεια της θεραπείας ήταν 9,8 μήνες στην πρώτη γραμμή, 4,4 στη δεύτερη και 3,7 στην τρίτη (Liontos et al., 2024). Ο πληθυσμός της παρούσας εργασίας, ασθενείς με ARSI πρώτης γραμμής, αντιστοιχεί επομένως στη συχνότερη περίπτωση της ελληνικής κλινικής πρακτικής.

Ελληνικές ερευνητικές ομάδες έχουν συμβάλει και στη μελέτη βιοδεικτών. Σε 208 ασθενείς με mCRPC από πολυκεντρική προοπτική κοόρτη με δεδομένα RNA-seq, η υψηλή έκφραση της πρωτεΐνης ενεργοποίησης ινοβλαστών (fibroblast activation protein, FAP), που εμφάνιζε το 10% των ασθενών, συνδεόταν με νευροενδοκρινικά χαρακτηριστικά και με βραχύτερη επιβίωση (διάμεση 12 έναντι 28 μηνών, Vlachostergios et al., 2022). Με ελληνική συμμετοχή αναπτύχθηκε και το σκορ MIRA, που ενσωματώνει το φορτίο των μεταστάσεων στην πρόγνωση (Uccello et al., 2026), ενώ οι Efstathiou et al. (2015) μελέτησαν σε βιοψίες μυελού των οστών τα μοριακά χαρακτηριστικά που σχετίζονται με την ανταπόκριση στο enzalutamide. Ανασκοπήσεις ελληνικών ομάδων συνοψίζουν τους προγνωστικούς και προβλεπτικούς βιοδείκτες του καρκίνου του προστάτη (Kontos et al., 2015 και Adamaki & Zoumpourlis, 2021).

Στο μεθοδολογικό πεδίο, ερευνητές του Πανεπιστημίου Ιωαννίνων ανασκόπησαν τις εφαρμογές της μηχανικής μάθησης στη διάγνωση και την πρόγνωση του καρκίνου (Kourou et al., 2015, 2021). Πρόσφατα αναπτύχθηκαν επίσης εργαλεία για ωμικά δεδομένα με λίγα δείγματα και πολλές μεταβλητές: το Omics-CNN, που συνδυάζει συνελικτικά νευρωνικά δίκτυα με μείωση της διαστατικότητας και τεχνικές ερμηνευσιμότητας (Zompola et al., 2023), και το MEvA-X, που συνδυάζει έναν πολυκριτηριακό εξελικτικό αλγόριθμο με τον ταξινομητή XGBoost για την επιλογή βιοδεικτών (Panagiotopoulos et al., 2023).

Στην αναζήτηση που έγινε για την παρούσα εργασία δεν εντοπίστηκε ελληνική μελέτη που να εφαρμόζει μηχανική μάθηση σε γονιδιωματικά δεδομένα όγκου για την πρόβλεψη της επιβίωσης ασθενών με mCRPC υπό θεραπεία ARSI.

3.4 Κριτική αποτίμηση

Από την επισκόπηση προκύπτουν τέσσερις παρατηρήσεις. Πρώτον, οι περισσότεροι γονιδιωματικοί δείκτες μελετήθηκαν ως προγνωστικοί παράγοντες: αναφέρεται ένας λόγος κινδύνων σε μία κοόρτη, συνήθως χωρίς αξιολόγηση της ικανότητας πρόβλεψης σε νέους ασθενείς. Όπως εξηγήθηκε στην ενότητα 2.7, μια ισχυρή συσχέτιση δεν συνεπάγεται χρήσιμη πρόβλεψη (Pepe et al., 2004).

Δεύτερον, όταν οι δείκτες εξετάστηκαν μαζί με κλινικούς παράγοντες ή με το συνολικό γονιδιωματικό φορτίο, αρκετοί δεν διατήρησαν ανεξάρτητη αξία (De Laere et al., 2019 και Hamid et al., 2019), ενώ το φορτίο της νόσου αναδείχθηκε επανειλημμένα σημαντικό (De Laere et al., 2019 και Uccello et al., 2026). Το ερώτημα επομένως δεν είναι μόνο αν ένα γονίδιο σχετίζεται με την επιβίωση, αλλά και αν προσθέτει πληροφορία πέρα από όσα είναι ήδη γνωστά.

Τρίτον, πολλές από τις αλλοιώσεις με τη μεγαλύτερη βιολογική σημασία είναι σπάνιες. Με λίγους φορείς οι εκτιμήσεις είναι ασταθείς, και όταν εξετάζονται πολλές αλλοιώσεις ταυτόχρονα, ο κίνδυνος ψευδώς θετικών ευρημάτων αυξάνεται αν δεν γίνει διόρθωση για πολλαπλούς ελέγχους (ενότητα 2.6).

Τέταρτον, τα κλινικά μοντέλα του mCRPC φτάνουν χρονοεξαρτώμενη AUC περίπου 0,73 έως 0,79 με εργαστηριακές μεταβλητές (Halabi et al., 2014 και Guinney et al., 2017), ενώ οι συγκριτικές μελέτες δείχνουν περιορισμένο κέρδος από τα ωμικά δεδομένα και από τους πιο σύνθετους αλγορίθμους όταν η επικύρωση είναι αυστηρή (Herrmann et al., 2021 και Christodoulou et al., 2019).

Η παρούσα εργασία εξετάζει τα ζητήματα αυτά σε μία κοόρτη με γονιδιωματική ανάλυση ιστού και ομοιογενή θεραπεία, ARSI πρώτης γραμμής. Αξιολογεί χωριστά την πρόβλεψη, με εμφωλευμένη διασταυρωμένη επικύρωση και ζευγαρωμένη σύγκριση με τα κλινικά στοιχεία, και τη συσχέτιση, με διόρθωση για πολλαπλούς ελέγχους, και προσθέτει τη χρονική διάσταση μέσω της ανάλυσης σε ορίζοντες (Κεφάλαιο 4).

4. ΣΚΟΠΟΣ ΚΑΙ ΕΡΕΥΝΗΤΙΚΑ ΕΡΩΤΗΜΑΤΑ

4.1 Βασικό ερευνητικό ερώτημα

Σκοπός της εργασίας είναι να διερευνήσει τη σχέση των γονιδιωματικών αλλοιώσεων του όγκου με την εξέλιξη του μεταστατικού, ανθεκτικού στον ευνουχισμό καρκίνου του προστάτη. Η διερεύνηση ξεκίνησε από τη συσχέτιση των αλλοιώσεων με την επιβίωση. Επειδή όμως η συσχέτιση δεν εξασφαλίζει πρόβλεψη σε νέους ασθενείς (ενότητα 2.7), το κύριο ζητούμενο της εργασίας είναι η πρόβλεψη.

> Το βασικό ερευνητικό ερώτημα της παρούσας εργασίας είναι **αν υπάρχουν γονίδια που προβλέπουν την εξέλιξη του μεταστατικού, ανθεκτικού στον ευνουχισμό καρκίνου του προστάτη, πέρα από όσα προβλέπουν τα κλινικά στοιχεία**.

Για να απαντηθεί το ερώτημα με τα διαθέσιμα δεδομένα, οι όροι του ορίζονται ως εξής:

- **Γονίδια:** τα γονίδια που αλλοιώνονται συχνότερα στον μεταστατικό καρκίνο του προστάτη, σύμφωνα με τη βιβλιογραφία (ενότητα 1.3) και με την κοόρτη της εργασίας. Από τη βιβλιογραφία προέκυψαν 20 γονίδια και από την κοόρτη κρατήθηκαν οι αλλοιώσεις τους που εμφανίζονται σε τουλάχιστον 5 ασθενείς (ενότητα 5.3.2). Εξετάζονται οι σωματικές αλλοιώσεις του DNA του όγκου: μεταλλάξεις (mutations), γονιδιακές ενισχύσεις (amplifications) και ομόζυγες απαλοιφές (deep deletions).
- **Ασθενείς:** 128 ασθενείς με μεταστατικό, ανθεκτικό στον ευνουχισμό καρκίνο του προστάτη (metastatic castration-resistant prostate cancer, mCRPC) από την κοόρτη SU2C/PCF, οι οποίοι έλαβαν abiraterone ή enzalutamide, δηλαδή αναστολέα της σηματοδότησης του ανδρογονικού υποδοχέα (androgen receptor signaling inhibitor, ARSI), ως πρώτη θεραπεία αυτής της κατηγορίας. Η έναρξη της θεραπείας ορίζει πληθυσμό με προχωρημένη νόσο και δίνει σε όλους τους ασθενείς κοινό σημείο έναρξης της παρακολούθησης.
- **Εξέλιξη:** η πορεία της νόσου έως την έκβασή της. Μετριέται με τη συνολική επιβίωση από την έναρξη της πρώτης θεραπείας ARSI έως τον θάνατο από οποιαδήποτε αιτία.
- **Πρόβλεψη:** η ικανότητα ενός μοντέλου, εκπαιδευμένου σε ορισμένους ασθενείς, να κατατάσσει σωστά κατά κίνδυνο ασθενείς που δεν χρησιμοποιήθηκαν στην εκπαίδευσή του. Εκτιμάται με αλγορίθμους μηχανικής μάθησης για δεδομένα επιβίωσης, από μοντέλα Cox με ποινή έως σύνολα

δέντρων απόφασης, και με εμφωλευμένη διασταυρωμένη επικύρωση (ενότητες 2.2 και 2.4).

Εκτός του πεδίου της εργασίας μένουν:

- **Μεταγραφωματικά δεδομένα:** η έκφραση των γονιδίων (RNA-seq) και οι βαθμολογίες δραστηριότητας του AR και νευροενδοκρινικής υπογραφής της κοόρτης.
- **Χρόνος σε θεραπεία:** η συνέχιση ή η διακοπή της θεραπείας ARSI δεν εξετάζεται ως έκβαση.
- **Ανταπόκριση στη θεραπεία:** όλοι οι ασθενείς έλαβαν ARSI και δεν υπάρχει ομάδα σύγκρισης. Η εργασία εξετάζει την πρόγνωση υπό ARSI και όχι το ποιοι ωφελούνται περισσότερο από αυτήν.
- **Άλλες μορφές αλλοιώσεων:** οι ημιζυγείς απαλοιφές, οι γαμετικές μεταλλάξεις, οι δομικές αλλοιώσεις που δεν ανιχνεύονται με την αλληλούχιση των εξωνίων και ο συνδυασμός μετάλλαξης και απώλειας στο ίδιο γονίδιο.

4.2 Επιμέρους ερωτήματα

Το βασικό ερώτημα αναλύεται σε τρία επιμέρους. Τα δύο πρώτα αφορούν την πρόβλεψη και το τρίτο τη συσχέτιση μεμονωμένων αλλοιώσεων με την επιβίωση (ενότητα 2.7):

- **Ε1.** Προβλέπουν οι αλλοιώσεις των συχνότερα αλλοιωμένων γονιδίων, μόνες τους, την εξέλιξη σε νέους ασθενείς; Απαντάται με μοντέλα που χρησιμοποιούν μόνο τις αλλοιώσεις (ενότητα 6.2).
- **Ε2.** Προσθέτουν οι αλλοιώσεις πληροφορία πέρα από τα κλινικά στοιχεία; Απαντάται με τη ζευγαρωμένη σύγκριση μοντέλων με και χωρίς τις αλλοιώσεις στους ίδιους ασθενείς (ενότητα 6.3).
- **Ε3.** Ποιες αλλοιώσεις σχετίζονται με την επιβίωση και σε ποιον χρονικό ορίζοντα (12, 24 ή 36 μήνες); Απαντάται με ανάλυση αναλογικών κινδύνων ανά αλλοίωση σε κάθε ορίζοντα (ενότητα 6.5).

Τα Ε1 και Ε2 εξετάζονται σε όλη την παρακολούθηση και, διερευνητικά, σε ορίζοντες 12, 24 και 36 μηνών (ενότητα 6.4), ώστε να φανεί αν η πληροφορία αφορά τους πρώιμους ή τους όψιμους θανάτους. Το Ε3 εξετάζεται στους ίδιους ορίζοντες.

4.3 Ερευνητικές υποθέσεις

Σε κάθε επιμέρους ερώτημα αντιστοιχεί μια μηδενική υπόθεση και ένα κριτήριο απόφασης, σύμφωνα με το σχέδιο ανάλυσης του Κεφαλαίου 5:

- **Y1 (E1).** Τα μοντέλα με μόνο τις αλλοιώσεις των γονιδίων δεν διακρίνουν τους ασθενείς καλύτερα από την τύχη (δείκτης συμφωνίας 0,5). Η υπόθεση απορρίπτεται αν το διορθωμένο 95% διάστημα εμπιστοσύνης του δείκτη συμφωνίας βρίσκεται ολόκληρο πάνω από το 0,5.
- **Y2 (E2).** Η προσθήκη των αλλοιώσεων στα κλινικά στοιχεία δεν μεταβάλλει τον δείκτη συμφωνίας ούτε το ολοκληρωμένο κριτήριο Brier. Η υπόθεση απορρίπτεται υπέρ της προστιθέμενης αξίας αν το 95% διάστημα εμπιστοσύνης της ζευγαρωμένης διαφοράς αποκλείει το μηδέν προς την κατεύθυνση της βελτίωσης.
- **Y3 (E3).** Καμία αλλοίωση δεν σχετίζεται με την επιβίωση (λόγος κινδύνων ίσος με 1) σε κανέναν ορίζοντα. Η υπόθεση απορρίπτεται για μια αλλοίωση αν $q < 0,05$ μετά από διόρθωση Benjamini–Hochberg. Το E3 έχει και διερευνητικό ρόλο: οι αλλοιώσεις με $p < 0,05$ αλλά $q \geq 0,05$ καταγράφονται ως υποψήφιες για περαιτέρω διερεύνηση, ακόμη κι αν δεν βελτιώνουν την πρόβλεψη για το σύνολο των ασθενών.

Η ανάλυση σε χρονικούς ορίζοντες σχεδιάστηκε μετά την κύρια ανάλυση και έχει διερευνητικό χαρακτήρα. Τα ευρήματά της ερμηνεύονται ως ενδείξεις που χρειάζονται επιβεβαίωση σε ανεξάρτητα δεδομένα.

5. ΥΛΙΚΑ ΚΑΙ ΜΕΘΟΔΟΙ

5.1 Πηγή δεδομένων

Τα δεδομένα προέρχονται από τη μελέτη της διεθνούς ερευνητικής ομάδας Stand Up To Cancer–Prostate Cancer Foundation (SU2C/PCF) για τον μεταστατικό, ανθεκτικό στον ευνουχισμό καρκίνο του προστάτη (Abida et al., 2019). Η μελέτη συγκέντρωσε βιοψίες μεταστάσεων από 429 ασθενείς, με γονιδιωματική και μεταγραφωματική ανάλυση και κλινικά δεδομένα παρακολούθησης. Τα δεδομένα διατίθενται δημόσια στην πλατφόρμα cBioPortal με το αναγνωριστικό prad_su2c_2019 (Cerami et al., 2012 και Gao et al., 2013). Στην παρούσα εργασία η δημοσίευση χρησιμοποιείται ως πηγή των δεδομένων και του ορισμού της έκβασης.

Από το εξαγόμενο του cBioPortal αξιοποιήθηκαν τα αρχεία του Πίνακα 3. Χρησιμοποιήθηκε επίσης το συμπληρωματικό Dataset S1 της δημοσίευσης, για την επιλογή του δείγματος κάθε ασθενούς και για τη διασταύρωση της έκβασης και του PSA.

Πίνακας 3. Αρχεία δεδομένων που αξιοποιήθηκαν.

Αρχείο	Περιεχόμενο	Χρήση
data_clinical_patient.txt	Στοιχεία ανά ασθενή: συνολική επιβίωση (OS_MONTHS, OS_STATUS), ηλικία και PSA κατά τη διάγνωση	Έκβαση, PSA
data_clinical_sample.txt	Στοιχεία ανά δείγμα: ηλικία κατά τη βιοψία, άθροισμα Gleason, θέση βιοψίας, ιστολογία, χρονική σχέση βιοψίας και ARSI	Κλινικές μεταβλητές
data_mutations.txt	Σωματικές μεταλλάξεις σε μορφή MAF	Μεταλλάξεις
data_cna.txt	Διακριτές τιμές αριθμού αντιγράφων GISTIC ανά γονίδιο	Γονιδιακές ενισχύσεις, ομόζυγες απαλοιφές
data_gene_panel_matrix.txt	Τεχνολογία αλληλούχισης ανά δείγμα	Έλεγχος ότι όλα τα δείγματα έχουν WXS
Dataset S1 (Abida et al., 2019)	Δείγμα αναφοράς ανά ασθενή, επιβίωση από την έναρξη της ARSI, PSA κατά τη διάγνωση	Επιλογή δείγματος, διασταύρωση

5.2 Κοόρτη και έκβαση

Το εξαγόμενο περιλαμβάνει 429 ασθενείς. Στην ανάλυση εντάχθηκαν όσοι είχαν θετικό χρόνο επιβίωσης (OS_MONTHS) και γνωστή κατάσταση στην τελευταία παρακολούθηση (OS_STATUS): 128 ασθενείς, από τους οποίους 76 πέθαναν και 52 ήταν ζωντανοί στην τελευταία παρακολούθηση (λογοκριμένοι).

Πρόκειται για τους ασθενείς που έλαβαν abiraterone ή enzalutamide ως πρώτη θεραπεία ARSI και για τους οποίους υπάρχουν δεδομένα έκβασης.

Η έκβαση είναι η συνολική επιβίωση από την έναρξη της πρώτης θεραπείας ARSI. Ο χρόνος και η κατάσταση του cBioPortal συμφωνούν για όλους τους ασθενείς με τις στήλες «OS from first-line ARSI start» και «Death status» του Dataset S1, γεγονός που επιβεβαιώνει την αφετηρία του χρόνου.

Για τους 128 ασθενείς υπάρχουν 136 δείγματα, επειδή ορισμένοι είχαν περισσότερες από μία βιοψίες. Για κάθε ασθενή κρατήθηκε το δείγμα που όρισαν οι συγγραφείς στο Dataset S1, ώστε η επιλογή να μην εξαρτάται ούτε από την έκβαση ούτε από τα γονιδιωματικά δεδομένα. Από τις 128 βιοψίες, οι 114 (89%) ελήφθησαν πριν από την έναρξη της θεραπείας ARSI, 10 κατά τη διάρκειά της και 4 μετά από προηγούμενη έκθεση σε αυτήν. Για αυτούς τους 14 ασθενείς το γονιδιωματικό προφίλ δεν αντιστοιχεί ακριβώς στη χρονική στιγμή έναρξης της θεραπείας.

5.3 Μεταβλητές

5.3.1 Κλινικές μεταβλητές

Χρησιμοποιήθηκαν οι κλινικές μεταβλητές που είναι διαθέσιμες για τη μεγάλη πλειονότητα των ασθενών:

- Ηλικία κατά τη βιοψία (έτη). Σε έναν ασθενή η ηλικία βιοψίας ήταν μικρότερη από την ηλικία διάγνωσης και, επειδή δεν είναι γνωστό ποια τιμή είναι λανθασμένη, θεωρήθηκε άγνωστη.
- Άθροισμα Gleason κατά τη διάγνωση (6–10), ως αριθμητική μεταβλητή.
- PSA κατά τη διάγνωση, με λογαριθμικό μετασχηματισμό log(1 + PSA). Η τιμή του cBioPortal συμφωνεί με τη στήλη «PSA at Dx» του Dataset S1.
- Θέση της βιοψίας, με τρεις δυαδικές μεταβλητές (λεμφαδένας, ήπαρ, άλλη θέση) και κατηγορία αναφοράς το οστό.
- Ιστολογία, με δύο δυαδικές μεταβλητές: μικροκυτταρικό καρκίνωμα ή αδενοκαρκίνωμα με νευροενδοκρινικά χαρακτηριστικά, και άγνωστη ιστολογία (μη διαθέσιμη ή ανεπαρκής για διάγνωση). Κατηγορία αναφοράς το αδενοκαρκίνωμα.

Ελλείψεις υπήρχαν στην ηλικία βιοψίας (2 ασθενείς), στο άθροισμα Gleason (21) και στο PSA (26). Κανένας ασθενής δεν αποκλείστηκε λόγω ελλείψεων. Οι τιμές που λείπουν συμπληρώνονται μέσα σε κάθε σύνολο εκπαίδευσης (ενότητα 5.4).

5.3.2 Γονιδιωματικές μεταβλητές: τα συχνότερα αλλοιωμένα γονίδια

Οι γονιδιωματικές μεταβλητές αφορούν 20 γονίδια με επαναλαμβανόμενες αλλοιώσεις στον μεταστατικό καρκίνο του προστάτη (ενότητα 1.3): AR, TP53, RB1, PTEN, BRCA1, BRCA2, ATM, CDK12, SPOP, FOXA1, APC, CTNNB1, PIK3CA, PIK3CB, AKT1, MYC, CCND1, CDKN1B, CHD1 και KMT2D. Όλα τα δείγματα έχουν αλληλούχιση όλων των εξωνίων (whole-exome sequencing, WXS) τόσο για τις μεταλλάξεις όσο και για τον αριθμό αντιγράφων, οπότε η απουσία καταγεγραμμένης μετάλλαξης σημαίνει ότι μετάλλαξη δεν ανιχνεύθηκε.

Για κάθε γονίδιο ορίστηκαν τρεις δυαδικές μεταβλητές (Πίνακας 4): μετάλλαξη (mutation), γονιδιακή ενίσχυση (amplification) και ομόζυγη απαλοιφή (deep deletion). Οι δύο τελευταίες προκύπτουν από τις διακριτές τιμές αριθμού αντιγράφων του αλγορίθμου GISTIC (Mermel et al., 2011). Η τιμή −2 σημαίνει ομόζυγη απαλοιφή, δηλαδή απώλεια και των δύο αντιγράφων του γονιδίου, και η −1 ημιζυγή απαλοιφή (shallow deletion), δηλαδή απώλεια ενός αντιγράφου. Η τιμή 0 αντιστοιχεί σε φυσιολογικό αριθμό αντιγράφων, η +1 σε χαμηλού επιπέδου κέρδος (gain) και η +2 σε γονιδιακή ενίσχυση. Χρησιμοποιήθηκαν μόνο οι τιμές −2 και +2, δηλαδή οι μεταβολές υψηλού επιπέδου.

Πίνακας 4. Ορισμός των τριών τύπων αλλοίωσης των γονιδίων.

Τύπος αλλοίωσης	Όρος cBioPortal	Ορισμός	Αρχείο
Μετάλλαξη	Mutation	Τουλάχιστον μία μη συνώνυμη σωματική μετάλλαξη (missense, nonsense, μετατόπιση πλαισίου ανάγνωσης, ένθεση ή έλλειψη εντός πλαισίου, θέση ματίσματος, nonstop, θέση έναρξης της μετάφρασης)	data_mutations.txt
Γονιδιακή ενίσχυση	Amplification	Τιμή GISTIC +2	data_cna.txt
Ομόζυγη απαλοιφή	Deep Deletion	Τιμή GISTIC −2: απώλεια και των δύο αντιγράφων	data_cna.txt

Οι μεταλλάξεις περιοχής ματίσματος (splice region) εξαιρέθηκαν, επειδή κατά κανόνα δεν μεταβάλλουν την πρωτεΐνη. Για το KMT2D δεν υπάρχουν τιμές αριθμού αντιγράφων στο εξαγόμενο, οπότε από τις 60 μεταβλητές μετρήθηκαν οι 58.

Μια αλλοίωση που εμφανίζεται σε ελάχιστους ασθενείς δεν επιτρέπει την εκτίμηση της επίδρασής της. Για τον λόγο αυτό κρατήθηκαν μόνο οι μεταβλητές με τουλάχιστον 5 φορείς και 5 μη φορείς στους 128 ασθενείς: 22 από τις 60. Το

φίλτρο δεν χρησιμοποιεί την έκβαση και επομένως εφαρμόστηκε μία φορά σε όλη την κοόρτη χωρίς διαρροή πληροφορίας (Hastie et al., 2009). Οι συχνότητες των αλλοιώσεων παρουσιάζονται στην ενότητα 6.1.2.

5.3.3 Σύνολα εισόδων

Τα μοντέλα εκπαιδεύτηκαν σε τέσσερα σύνολα εισόδων (Πίνακας 5), ώστε να φαίνεται τι προσθέτει κάθε επίπεδο πληροφορίας. Το σύνολο «Γονίδια» απαντά στο ερώτημα Ε1 και η σύγκριση «Κλινικά + γονίδια» έναντι «Κλινικά διευρυμένα» στο ερώτημα Ε2. Η σύγκριση των δύο κλινικών συνόλων δείχνει πόση πληροφορία προσθέτουν το PSA, η θέση της βιοψίας και η ιστολογία στα δύο κλασικά στοιχεία, την ηλικία και το άθροισμα Gleason.

Πίνακας 5. Τα τέσσερα σύνολα εισόδων των μοντέλων.

Σύνολο εισόδων	Μεταβλητές	Πλήθος
Κλινικά βασικά	Ηλικία βιοψίας, άθροισμα Gleason	2
Κλινικά διευρυμένα	Τα βασικά, log(1 + PSA διάγνωσης), θέση βιοψίας (3 δυαδικές), ιστολογία (2 δυαδικές)	8
Γονίδια	Οι 22 αλλοιώσεις των γονιδίων που πέρασαν το φίλτρο συχνότητας	22
Κλινικά + γονίδια	Κλινικά διευρυμένα και γονίδια	30

Σε κάθε σύνολο εκπαίδευσης προστίθεται δείκτης έλλειψης για κάθε μεταβλητή με ελλείπουσες τιμές (έως 3 επιπλέον μεταβλητές).

5.4 Προεπεξεργασία

Η προεπεξεργασία αποτελεί το πρώτο βήμα του σωλήνα επεξεργασίας κάθε μοντέλου και προσαρμόζεται μόνο στους ασθενείς εκπαίδευσης κάθε διαχωρισμού. Περιλαμβάνει τέσσερα βήματα: (α) συμπλήρωση κάθε ελλείπουσας τιμής με τη διάμεσο του συνόλου εκπαίδευσης, (β) προσθήκη δυαδικού δείκτη έλλειψης για κάθε μεταβλητή με ελλείψεις στο σύνολο εκπαίδευσης, ώστε το μοντέλο να μπορεί να μάθει αν η ίδια η απουσία μέτρησης σχετίζεται με τον κίνδυνο, (γ) αφαίρεση μεταβλητών χωρίς διακύμανση στο σύνολο εκπαίδευσης και (δ) τυποποίηση με τον μέσο όρο και την τυπική απόκλιση του συνόλου εκπαίδευσης. Οι ίδιες παράμετροι εφαρμόζονται στη συνέχεια στους ασθενείς ελέγχου.

5.5 Αλγόριθμοι και υπερπαράμετροι

Αξιολογήθηκαν οι τέσσερις αλγόριθμοι της ενότητας 2.2, με την υλοποίηση της βιβλιοθήκης scikit-survival (Pölsterl, 2020). Στους πίνακες και στα σχήματα

αναφέρονται με τα αγγλικά τους ονόματα: Cox PH (ridge), Cox Elastic Net, Random Survival Forest και Gradient Boosting. Οι σταθερές ρυθμίσεις και τα πλέγματα υπερπαραμέτρων συνοψίζονται στον Πίνακα 6. Οι υπερπαράμετροι επιλέγονται στον εσωτερικό βρόχο της διασταυρωμένης επικύρωσης με κριτήριο τον δείκτη συμφωνίας του Harrell. Στο Cox Elastic Net εφαρμόζεται ο κανόνας του ενός τυπικού σφάλματος (ενότητα 2.2.2). Ποινές μικρότερες από 0,01 δεν συμπεριλήφθηκαν στο πλέγμα, επειδή αποτύγχαναν αριθμητικά στα μικρά εσωτερικά σύνολα εκπαίδευσης.

Πίνακας 6. Σταθερές ρυθμίσεις και πλέγματα υπερπαραμέτρων των τεσσάρων αλγορίθμων.

Αλγόριθμος	Σταθερές ρυθμίσεις	Πλέγμα (εσωτερικός βρόχος)
Cox PH (ridge)	Ποινή ridge α = 1, ισοβαθμίες κατά Efron	Καμία (σταθερή ποινή)
Cox Elastic Net	Εκτίμηση βασικού κινδύνου για τις καμπύλες επιβίωσης	α ∈ {0,01, 0,02, 0,05, 0,1, 0,2, 0,4}, ρ ∈ {0,5, 1}, κανόνας 1-SE
Random Survival Forest	300 δέντρα, τουλάχιστον 10 ασθενείς για διαχωρισμό κόμβου	Ασθενείς ανά φύλλο ∈ {5, 10, 15}, μεταβλητές ανά διαχωρισμό ∈ {√p, p/2}
Gradient Boosting	Απώλεια μερικής πιθανότητας Cox, κάθε δέντρο σε 80% των ασθενών	Ρυθμός μάθησης ∈ {0,05, 0,1}, δέντρα ∈ {50, 150}, βάθος ∈ {1, 2}

α: ένταση της ποινής, ρ: μερίδιο της ποινής L1 (l1_ratio), p: πλήθος μεταβλητών.

5.6 Πρωτόκολλο αξιολόγησης

Η επίδοση εκτιμήθηκε με επαναλαμβανόμενη εμφωλευμένη διασταυρωμένη επικύρωση (ενότητα 2.4). Ο εξωτερικός βρόχος είναι πενταπλή διασταυρωμένη επικύρωση με διαστρωμάτωση ως προς τον θάνατο, επαναλαμβανόμενη 10 φορές με διαφορετικό τυχαίο ανακάτεμα (50 διαχωρισμοί). Κάθε ασθενής αξιολογείται μία φορά σε κάθε επανάληψη, από μοντέλο που δεν τον είδε κατά την εκπαίδευση. Ο εσωτερικός βρόχος, για την επιλογή υπερπαραμέτρων, εκτελείται μόνο στο σύνολο εκπαίδευσης με τριπλή διασταυρωμένη επικύρωση επαναλαμβανόμενη τρεις φορές. Η καλύτερη ρύθμιση εκπαιδεύεται στη συνέχεια σε ολόκληρο το σύνολο εκπαίδευσης. Όλοι οι αλγόριθμοι και όλα τα σύνολα εισόδων αξιολογήθηκαν στους ίδιους 50 διαχωρισμούς: συνολικά 800 εξωτερικές εκπαιδεύσεις (4 σύνολα × 4 αλγόριθμοι × 50).

Σε κάθε σύνολο ελέγχου υπολογίστηκαν τέσσερις δείκτες. Ο κύριος δείκτης ήταν ο δείκτης συμφωνίας του Harrell. Υπολογίστηκαν επίσης ο δείκτης συμφωνίας του Uno έως τους 24 μήνες, η χρονοεξαρτώμενη AUC στους 12 και

στους 24 μήνες, με βαθμό κινδύνου 1 − S(t | x), και το ολοκληρωμένο κριτήριο Brier στο διάστημα 6–30 μηνών, σε μηνιαίο πλέγμα. Οι χρόνοι αυτοί επιλέχθηκαν επειδή στους 36 μήνες παραμένουν σε κίνδυνο μόνο 11 ασθενείς. Για την AUC και το κριτήριο Brier, οι χρόνοι παρακολούθησης πέρα από τους 31 μήνες αποκόπηκαν: η στάθμιση IPCW απαιτεί την κατανομή λογοκρισίας του συνόλου εκπαίδευσης στον χρόνο κάθε ασθενούς ελέγχου, η οποία δεν ορίζεται πέρα από τη μέγιστη παρακολούθηση του συνόλου εκπαίδευσης, ενώ η αποκοπή δεν μεταβάλλει τους δείκτες έως τους 30 μήνες. Ως σημείο αναφοράς για το κριτήριο Brier υπολογίστηκε, στους ίδιους διαχωρισμούς, η καμπύλη Kaplan–Meier του συνόλου εκπαίδευσης.

Για κάθε συνδυασμό συνόλου εισόδων και αλγορίθμου αναφέρεται ο μέσος όρος των 50 διαχωρισμών με διορθωμένο 95% διάστημα εμπιστοσύνης (Nadeau & Bengio, 2003). Οι δύο συγκρίσεις ενδιαφέροντος, «κλινικά + γονίδια» έναντι «κλινικά διευρυμένα» και «κλινικά διευρυμένα» έναντι «κλινικά βασικά», έγιναν ως ζευγαρωμένες διαφορές στον ίδιο αλγόριθμο και στον ίδιο διαχωρισμό, με το διορθωμένο t-test. Αναφέρεται επίσης σε πόσες από τις 10 επαναλήψεις η μέση διαφορά είχε την κατεύθυνση της βελτίωσης. Οι τιμές p των ζευγαρωμένων συγκρίσεων δεν διορθώθηκαν για πολλαπλούς ελέγχους.

Ρυθμίσεις του Cox Elastic Net που απέτυχαν αριθμητικά σε κάποιον εσωτερικό διαχωρισμό εξαιρέθηκαν από την επιλογή (100 από τις 2.400 αξιολογήσεις ρυθμίσεων). Κάθε ασθενής αξιολογήθηκε ακριβώς μία φορά ανά επανάληψη και δεν υπήρξαν κενές τιμές δεικτών.

5.7 Ανάλυση σε χρονικούς ορίζοντες

Για να εξεταστεί αν η προγνωστική πληροφορία αφορά κυρίως τους πρώιμους ή τους όψιμους θανάτους, η ανάλυση επαναλήφθηκε με επιβίωση με αποκοπή (ενότητα 2.1.3) στους 12, 24 και 36 μήνες. Σε κάθε ορίζοντα h, όσοι ζούσαν πέρα από τον h λογοκρίθηκαν ελάχιστα μετά τον h (h + 0,001 μήνες), ώστε οι δείκτες να ορίζονται ακριβώς στον h. Οι θάνατοι μετά τον h δεν μετρήθηκαν ως συμβάντα και όσοι είχαν λογοκριθεί νωρίτερα κράτησαν τον χρόνο τους. Έτσι παρέμειναν στην ανάλυση και οι 128 ασθενείς (Πίνακας 7).

Πίνακας 7. Επιβίωση με αποκοπή στους τρεις ορίζοντες (128 ασθενείς).

Ορίζοντας	Θάνατοι έως τον ορίζοντα	Ζωντανοί πέρα από τον ορίζοντα	Λογοκριμένοι νωρίτερα
12 μήνες	19	91	18
24 μήνες	50	47	31
36 μήνες	68	11	49

Σε κάθε ορίζοντα εφαρμόστηκε το ίδιο πρωτόκολλο εμφωλευμένης διασταυρωμένης επικύρωσης, με τα ίδια τέσσερα σύνολα εισόδων και τους ίδιους αλγορίθμους και με διαστρωμάτωση ως προς τον θάνατο έως τον ορίζοντα: συνολικά 2.400 εξωτερικές εκπαιδεύσεις (3 ορίζοντες × 4 σύνολα × 4 αλγόριθμοι × 50), χωρίς καμία αποτυχία. Υπολογίστηκαν ο δείκτης του Harrell έως τον ορίζοντα, ο δείκτης του Uno με τ ίσο με τον ορίζοντα, η AUC και το κριτήριο Brier στον ορίζοντα. Η AUC υπολογίστηκε μόνο σε διαχωρισμούς με τουλάχιστον τρεις θανάτους και τρεις επιζώντες στο σύνολο ελέγχου. Στους 36 μήνες αυτό ίσχυε σε 17 από τους 50 διαχωρισμούς. Οι ζευγαρωμένες συγκρίσεις της κύριας ανάλυσης επαναλήφθηκαν σε κάθε ορίζοντα.

5.8 Ανάλυση ανά αλλοίωση

Για να εντοπιστούν οι μεμονωμένες αλλοιώσεις που σχετίζονται με την επιβίωση (ερώτημα Ε3), καθεμία από τις 22 αλλοιώσεις των γονιδίων εξετάστηκε με μοντέλο αναλογικών κινδύνων Cox σε δύο αναλύσεις: χωρίς προσαρμογή, με μόνη μεταβλητή την αλλοίωση (128 ασθενείς), και με προσαρμογή για την ηλικία βιοψίας και το άθροισμα Gleason, στους 106 ασθενείς με πλήρη στοιχεία. Οι αναλύσεις έγιναν σε ολόκληρη την κοόρτη, σε κάθε έναν από τους τρεις ορίζοντες με την επιβίωση με αποκοπή της ενότητας 5.7. Πρόκειται για ανάλυση συσχέτισης και όχι πρόβλεψης (ενότητα 2.7).

Ένας λόγος κινδύνων εκτιμήθηκε μόνο όταν, στα δεδομένα της ανάλυσης και έως τον ορίζοντα, υπήρχαν τουλάχιστον 5 φορείς, 5 μη φορείς και 3 θάνατοι σε κάθε ομάδα. Χωρίς τον κανόνα αυτό, μια αλλοίωση χωρίς θανάτους στους φορείς θα έδινε λόγο κινδύνων 0 με διάστημα εμπιστοσύνης από 0 έως άπειρο. Για κάθε εκτιμήσιμη συσχέτιση αναφέρονται ο λόγος κινδύνων (HR) με 95% διάστημα εμπιστοσύνης και η τιμή p του ελέγχου Wald. Η υπόθεση των αναλογικών κινδύνων ελέγχθηκε με τα υπόλοιπα Schoenfeld (Schoenfeld,

1982). Οι τιμές p διορθώθηκαν με τη μέθοδο Benjamini–Hochberg (ενότητα 2.6), χωριστά για κάθε ανάλυση και κάθε ορίζοντα.

5.9 Λογισμικό και αναπαραγωγιμότητα

Η ανάλυση υλοποιήθηκε σε Python 3.12.10 με τις βιβλιοθήκες scikit-survival 0.25.0 (Pölsterl, 2020), scikit-learn 1.7.2 (Pedregosa et al., 2011), lifelines 0.30.0 (Davidson-Pilon, 2019), pandas 2.3.3, NumPy 2.2.6, SciPy 1.15.3 και Matplotlib 3.10.6. Όλα τα τυχαία βήματα χρησιμοποιούν καθορισμένο σπόρο (2026), ώστε η επανεκτέλεση να δίνει τα ίδια αποτελέσματα. Η κύρια ανάλυση (800 εκπαιδεύσεις) διήρκεσε 8,2 λεπτά και η ανάλυση των οριζόντων (2.400 εκπαιδεύσεις) 20,6 λεπτά, με παράλληλη εκτέλεση σε φορητό υπολογιστή με επεξεργαστή AMD Ryzen 7 5800H (8 πυρήνες, 16 νήματα) και Windows 11. Οι βασικές ενότητες κώδικα παρατίθενται στο Παράρτημα Α.

6. ΑΠΟΤΕΛΕΣΜΑΤΑ

6.1 Περιγραφή της κοόρτης

6.1.1 Κλινικά χαρακτηριστικά και επιβίωση

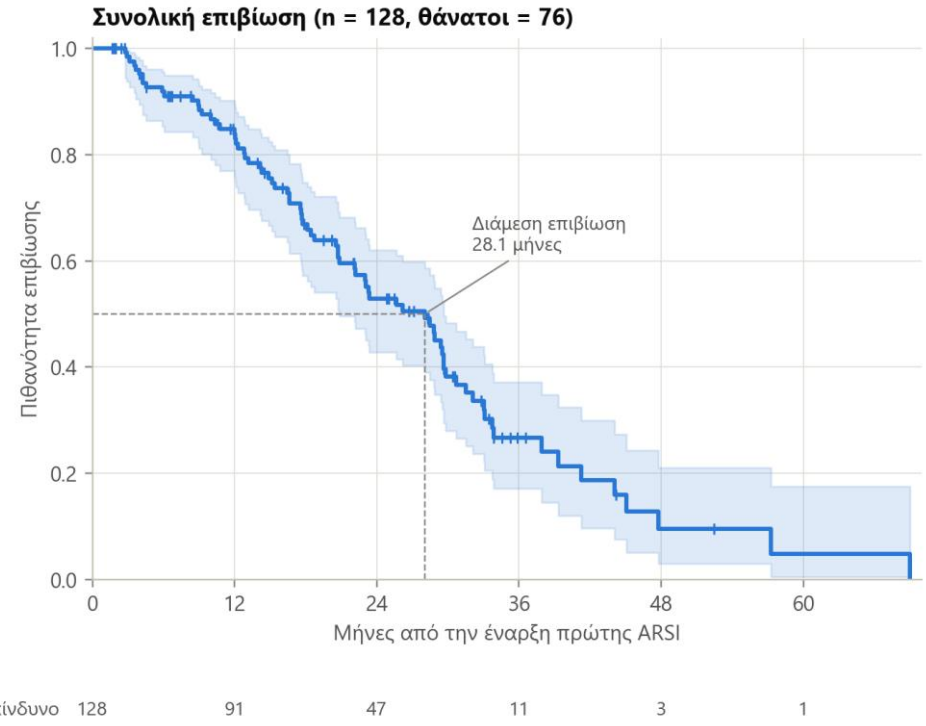
Τα χαρακτηριστικά των 128 ασθενών παρουσιάζονται στον Πίνακα 8. Η διάμεση ηλικία κατά τη βιοψία ήταν 67,3 έτη και το διάμεσο PSA κατά τη διάγνωση 18,4 ng/mL. Το 48% των ασθενών είχε άθροισμα Gleason 8 έως 10, ενώ για το 16% το άθροισμα ήταν άγνωστο. Οι βιοψίες προέρχονταν κυρίως από λεμφαδένες (44%) και οστά (42%), και η ιστολογία ήταν αδενοκαρκίνωμα στο 84% των ασθενών.

Πίνακας 8. Χαρακτηριστικά των 128 ασθενών της κοόρτης. Συνεχείς μεταβλητές: διάμεσος (ενδοτεταρτημοριακό εύρος). Κατηγορικές: πλήθος (%).

Χαρακτηριστικό	Τιμή	Διαθέσιμα
Θάνατοι κατά την παρακολούθηση	76 (59%)	128
Ηλικία κατά τη βιοψία (έτη)	67,3 (63,0–73,0)	126
PSA κατά τη διάγνωση (ng/mL)	18,4 (5,2–47,6)	102
Άθροισμα Gleason κατά τη διάγνωση		
6	10 (8%)	
7	36 (28%)	
8	18 (14%)	
9	35 (27%)	
10	8 (6%)	
άγνωστο	21 (16%)	
Θέση βιοψίας		
Λεμφαδένας	56 (44%)	
Οστό	54 (42%)	
Ήπαρ	8 (6%)	
Άλλος μαλακός ιστός	5 (4%)	
Προστάτης	4 (3%)	
Πνεύμονας	1 (1%)	
Ιστολογία		
Αδενοκαρκίνωμα	107 (84%)	
Ανεπαρκές για διάγνωση	10 (8%)	
Μη διαθέσιμο	6 (5%)	
Μικροκυτταρικό	3 (2%)	
Αδενοκαρκίνωμα με NE χαρακτηριστικά	2 (2%)	
Χρονική σχέση βιοψίας και θεραπείας ARSI		
Πριν την ARSI	114 (89%)	
Κατά τη θεραπεία	10 (8%)	

Χαρακτηριστικό	Τιμή	Διαθέσιμα
Μετά από έκθεση	4 (3%)	

Κατά την παρακολούθηση πέθαναν 76 ασθενείς (59%). Η διάμεση συνολική επιβίωση από την έναρξη της θεραπείας ARSI, κατά Kaplan–Meier, ήταν 28,1 μήνες (95% ΔΕ: 20,9–29,7). Η πιθανότητα επιβίωσης ήταν 0,84 στους 12 μήνες, 0,53 στους 24 και 0,27 στους 36, με 91, 47 και 11 ασθενείς σε κίνδυνο αντίστοιχα (Σχήμα 1). Ο μικρός αριθμός ασθενών σε κίνδυνο στους 36 μήνες είναι ο λόγος για τον οποίο οι δείκτες σε σταθερό χρόνο της κύριας ανάλυσης περιορίστηκαν στους 12 και 24 μήνες.

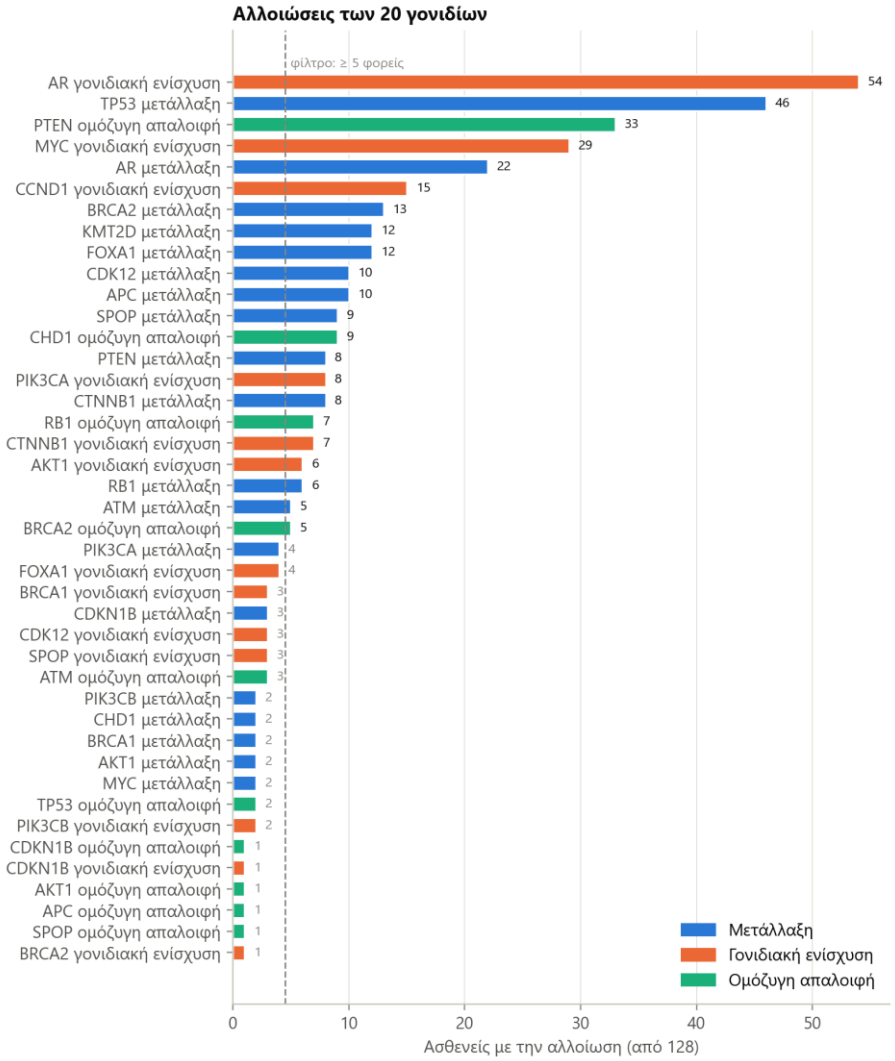


Σχήμα 1. Καμπύλη Kaplan–Meier της συνολικής επιβίωσης από την έναρξη της πρώτης θεραπείας ARSI (128 ασθενείς, 76 θάνατοι). Η σκιασμένη περιοχή είναι το 95% διάστημα εμπιστοσύνης και οι κάθετες γραμμές δείχνουν λογοκριμένους ασθενείς. Κάτω από τον άξονα, οι ασθενείς σε κίνδυνο.

6.1.2 Αλλοιώσεις των γονιδίων

Από τις 58 μετρήσιμες μεταβλητές των γονιδίων, οι 42 εμφανίστηκαν σε τουλάχιστον έναν ασθενή και οι 22 πέρασαν το φίλτρο συχνότητας (Σχήμα 2). Οι συχνότερες αλλοιώσεις ήταν η γονιδιακή ενίσχυση του AR (54 ασθενείς, 42%), η μετάλλαξη του TP53 (46, 36%), η ομόζυγη απαλοιφή του PTEN (33, 26%), η γονιδιακή ενίσχυση του MYC (29, 23%) και η μετάλλαξη του AR (22,

17%). Οι περισσότερες από τις υπόλοιπες ήταν σπάνιες: 17 από τις 22 αλλοιώσεις που κρατήθηκαν εμφανίζονταν σε 5 έως 15 ασθενείς, όπως η ομόζυγη απαλοιφή του RB1 (7 ασθενείς) και η γονιδιακή ενίσχυση του AKT1 (6).



Σχήμα 2. Πλήθος ασθενών με κάθε αλλοίωση των 20 γονιδίων (από 128). Εμφανίζονται οι 42 αλλοιώσεις με τουλάχιστον έναν φορέα. Η διακεκομμένη γραμμή δείχνει το φίλτρο των 5 φορέων και με γκρι αριθμό σημειώνονται οι αλλοιώσεις που δεν το πέρασαν.

6.2 Επίδοση των μοντέλων πρόβλεψης

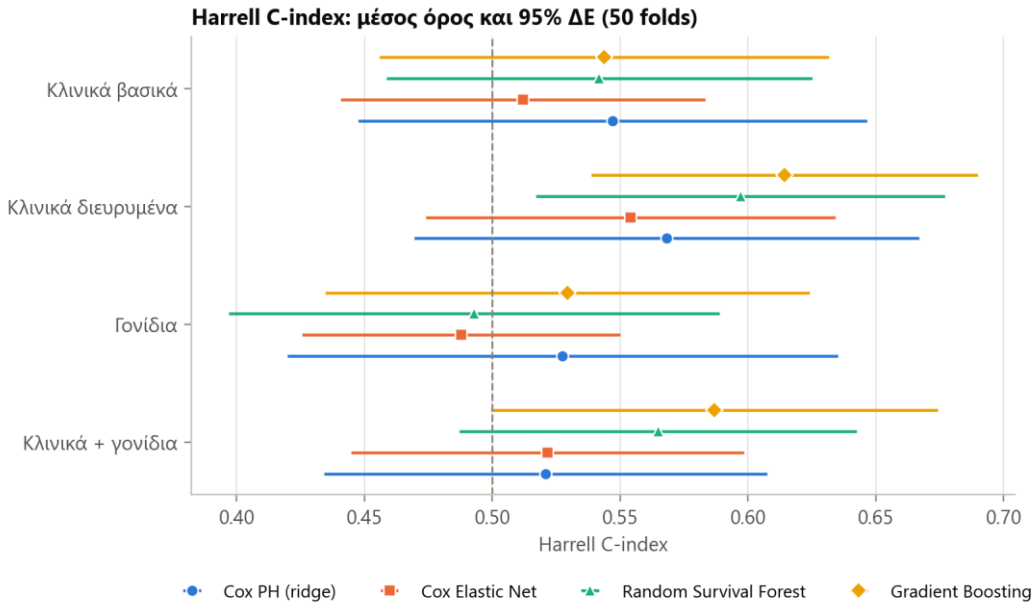
Ο Πίνακας 9 συνοψίζει την επίδοση των 16 συνδυασμών συνόλου εισόδων και αλγορίθμου στην κύρια ανάλυση και το Σχήμα 3 απεικονίζει τον δείκτη συμφωνίας του Harrell. Η διακριτική ικανότητα ήταν μέτρια σε όλα τα μοντέλα. Ο υψηλότερος δείκτης παρατηρήθηκε για το Gradient Boosting με τα διευρυμένα κλινικά στοιχεία, 0,614 (95% ΔΕ: 0,539–0,690), και ακολούθησε το Random

Survival Forest με το ίδιο σύνολο, 0,597 (0,517–0,677). Τα διαστήματα των συνδυασμών επικαλύπτονται εκτεταμένα, οπότε με 128 ασθενείς δεν μπορούν να διακριθούν αξιόπιστα μοντέλα που διαφέρουν κατά 0,02 έως 0,05.

Πίνακας 9. Επίδοση των 16 συνδυασμών συνόλου εισόδων και αλγορίθμου στην κύρια ανάλυση (μέσος όρος 50 διαχωρισμών).

Σύνολο εισόδων	Αλγόριθμος	Harrell C (95% ΔΕ)	Uno C	AUC 12 μ.	AUC 24 μ.	IBS (95% ΔΕ)
Κλινικά βασικά	Cox PH (ridge)	0,547 (0,448–0,647)	0,536	0,593	0,538	0,202 (0,175–0,228)
	Cox Elastic Net	0,512 (0,441–0,583)	0,503	0,548	0,499	0,200 (0,180–0,220)
	Random Survival Forest	0,542 (0,458–0,625)	0,543	0,543	0,588	0,197 (0,173–0,222)
	Gradient Boosting	0,544 (0,456–0,632)	0,534	0,541	0,553	0,200 (0,176–0,224)
Κλινικά διευρυμένα	Cox PH (ridge)	0,568 (0,470–0,667)	0,557	0,585	0,590	0,204 (0,176–0,232)
	Cox Elastic Net	0,554 (0,474–0,634)	0,553	0,558	0,582	0,196 (0,175–0,218)
	Random Survival Forest	0,597 (0,517–0,677)	0,597	0,568	0,660	0,187 (0,166–0,209)
	Gradient Boosting	0,614 (0,539–0,690)	0,614	0,605	0,658	0,192 (0,162–0,222)
Γονίδια	Cox PH (ridge)	0,527 (0,420–0,635)	0,540	0,536	0,569	0,223 (0,180–0,267)
	Cox Elastic Net	0,488 (0,426–0,550)	0,493	0,480	0,498	0,201 (0,179–0,223)
	Random Survival Forest	0,493 (0,397–0,589)	0,502	0,438	0,524	0,203 (0,183–0,223)
	Gradient Boosting	0,529 (0,435–0,624)	0,545	0,496	0,578	0,203 (0,172–0,233)
Κλινικά + γονίδια	Cox PH (ridge)	0,521 (0,434–0,607)	0,525	0,548	0,554	0,256 (0,209–0,303)
	Cox Elastic Net	0,522 (0,445–0,598)	0,530	0,514	0,562	0,204 (0,175–0,234)
	Random Survival Forest	0,565 (0,487–0,643)	0,567	0,511	0,625	0,194 (0,171–0,218)
	Gradient Boosting	0,587 (0,499–0,674)	0,583	0,567	0,622	0,203 (0,167–0,238)
Αναφορά	Kaplan–Meier	—	—	—	—	0,197 (0,179–0,215)

Σε παρένθεση το διορθωμένο 95% διάστημα εμπιστοσύνης (Nadeau & Bengio, 2003). Uno C: δείκτης συμφωνίας του Uno έως τους 24 μήνες. AUC: χρονοεξαρτώμενη AUC στους 12 και 24 μήνες. IBS: ολοκληρωμένο κριτήριο Brier στο διάστημα 6–30 μηνών (μικρότερο = καλύτερο).



Σχήμα 3. Δείκτης συμφωνίας του Harrell για κάθε σύνολο εισόδων και αλγόριθμο: μέσος όρος 50 διαχωρισμών και διορθωμένο 95% διάστημα εμπιστοσύνης. Η διακεκομμένη γραμμή αντιστοιχεί στην τυχαία κατάταξη (0,5).

6.2.1 Τα γονίδια ως μόνη πηγή πληροφορίας (E1)

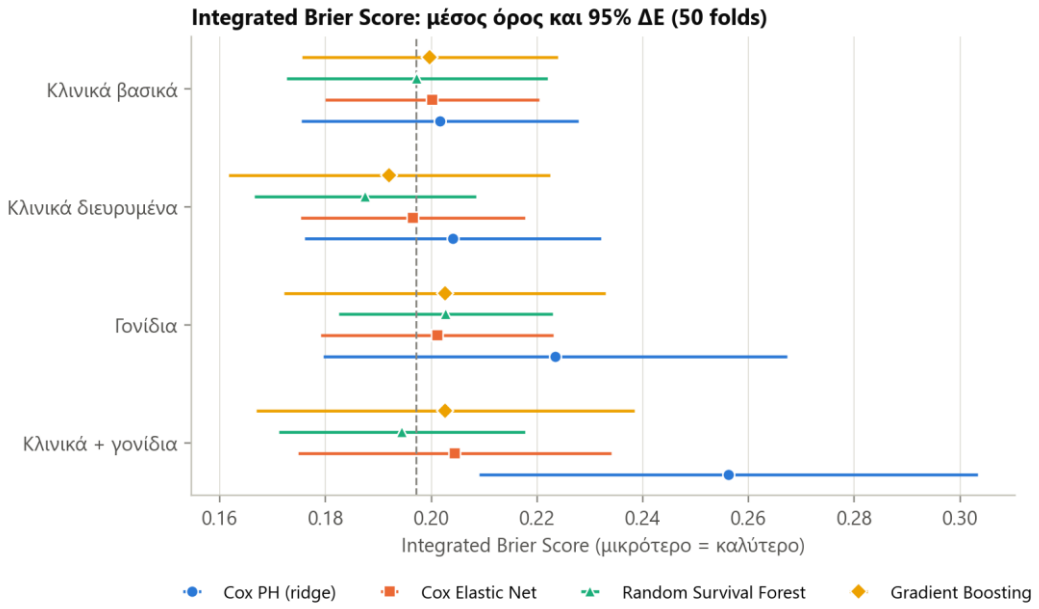
Τα μοντέλα που χρησιμοποίησαν μόνο τις 22 αλλοιώσεις των γονιδίων είχαν δείκτη συμφωνίας από 0,488 (Cox Elastic Net) έως 0,529 (Gradient Boosting), πρακτικά στο επίπεδο της τύχης, και όλα τα διαστήματα εμπιστοσύνης περιείχαν το 0,5. Οι υπόλοιποι δείκτες διάκρισης έδωσαν την ίδια εικόνα (Uno 0,493–0,545, AUC 12 μηνών 0,438–0,536, AUC 24 μηνών 0,498–0,578), και κανένα μοντέλο με μόνο τα γονίδια δεν είχε μικρότερο IBS από την καμπύλη Kaplan–Meier (0,201–0,223 έναντι 0,197).

Απάντηση στο E1: στην κοόρτη αυτή, οι αλλοιώσεις των γονιδίων μόνες τους δεν προέβλεψαν την επιβίωση νέων ασθενών. Η μηδενική υπόθεση Y1 δεν απορρίπτεται.

6.2.2 Ακρίβεια των πιθανοτήτων επιβίωσης

Το ολοκληρωμένο κριτήριο Brier αξιολογεί και την ακρίβεια των προβλεπόμενων πιθανοτήτων (Σχήμα 4). Κανένα μοντέλο δεν βελτίωσε ουσιαστικά την καμπύλη Kaplan–Meier, που αποδίδει την ίδια πιθανότητα επιβίωσης σε όλους (IBS 0,197). Το μικρότερο IBS είχε το Random Survival Forest με τα διευρυμένα κλινικά στοιχεία (0,187). Αντίθετα, το Cox PH (ridge) με

το σύνολο «κλινικά + γονίδια» είχε IBS 0,256 (95% ΔΕ: 0,209–0,303), σαφώς χειρότερο από την καμπύλη Kaplan–Meier: με έως 33 μεταβλητές και περίπου δύο θανάτους ανά μεταβλητή, η μικρή σταθερή ποινή ridge δεν αρκεί για να αποτρέψει την υπερπροσαρμογή. Το Cox Elastic Net, του οποίου η ποινή ρυθμίζεται στον εσωτερικό βρόχο, κράτησε κατά διάμεσο 8,5 από τις 33 μεταβλητές και είχε IBS 0,204, κοντά στο σημείο αναφοράς.



Σχήμα 4. Ολοκληρωμένο κριτήριο Brier (6–30 μήνες) για κάθε σύνολο εισόδων και αλγόριθμο, με διορθωμένο 95% διάστημα εμπιστοσύνης. Η διακεκομμένη γραμμή είναι η τιμή της καμπύλης Kaplan–Meier (0,197). Μικρότερες τιμές είναι καλύτερες.

6.3 Προστιθέμενη αξία των γονιδίων πέρα από τα κλινικά στοιχεία (E2)

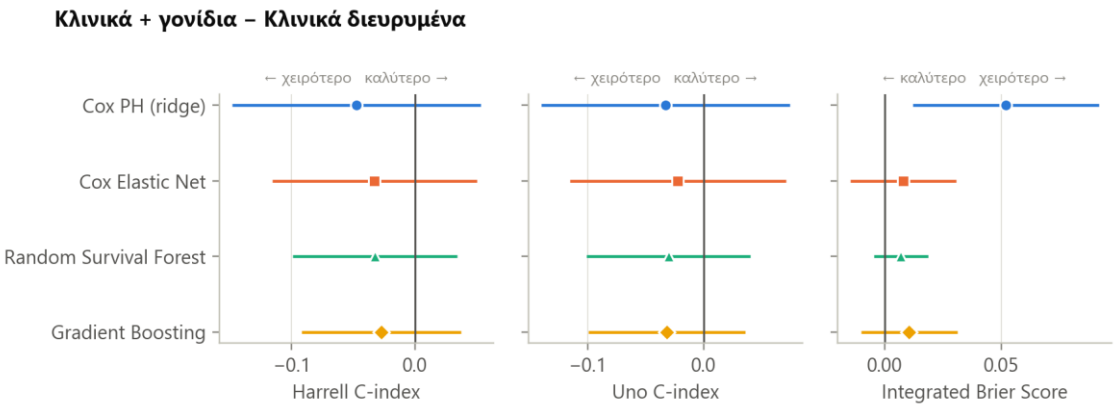
Ο Πίνακας 10 και τα Σχήματα 5 και 6 παρουσιάζουν τις ζευγαρωμένες διαφορές, για κάθε αλγόριθμο, στους ίδιους 50 διαχωρισμούς. Η προσθήκη των γονιδίων στα διευρυμένα κλινικά στοιχεία μείωσε τον δείκτη συμφωνίας σε όλους τους αλγορίθμους, από −0,027 (Gradient Boosting) έως −0,048 (Cox PH), και η μείωση εμφανίστηκε σε 9 έως 10 από τις 10 επαναλήψεις. Τα διαστήματα εμπιστοσύνης περιείχαν το μηδέν, οπότε η επιδείνωση δεν τεκμηριώνεται στατιστικά. Σε καμία περίπτωση όμως δεν παρατηρήθηκε βελτίωση. Το IBS αυξήθηκε επίσης σε όλους τους αλγορίθμους, και για το Cox PH η αύξηση ήταν στατιστικά σημαντική (+0,052, 95% ΔΕ: 0,012–0,092, p = 0,012).

Πίνακας 10. Ζευγαρωμένες διαφορές μεταξύ συνόλων εισόδων στον ίδιο αλγόριθμο και στους ίδιους 50 διαχωρισμούς.

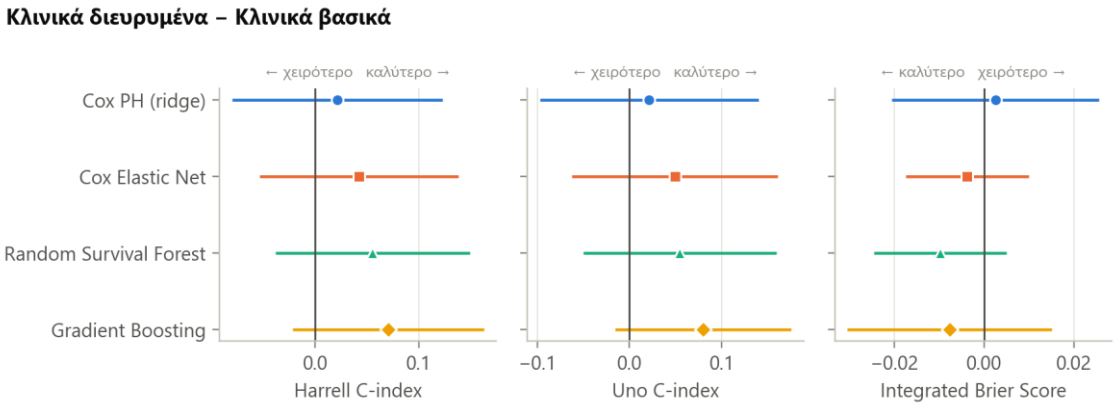
Αλγόριθμος	Δ Harrell C (95% ΔΕ)	p	Επαναλήψεις με βελτίωση	Δ IBS (95% ΔΕ)	p
Κλινικά + γονίδια − κλινικά διευρυμένα					
Cox PH (ridge)	−0,048 (−0,149 έως +0,054)	0,350	0/10	+0,052 (+0,012 έως +0,092)	0,012
Cox Elastic Net	−0,033 (−0,116 έως +0,051)	0,434	0/10	+0,008 (−0,015 έως +0,031)	0,486
Random Survival Forest	−0,032 (−0,099 έως +0,035)	0,337	0/10	+0,007 (−0,005 έως +0,019)	0,237
Gradient Boosting	−0,027 (−0,092 έως +0,037)	0,399	1/10	+0,011 (−0,010 έως +0,031)	0,307
Κλινικά διευρυμένα − κλινικά βασικά					
Cox PH (ridge)	+0,021 (−0,080 έως +0,123)	0,676	10/10	+0,003 (−0,021 έως +0,026)	0,827
Cox Elastic Net	+0,042 (−0,054 έως +0,138)	0,381	9/10	−0,004 (−0,017 έως +0,010)	0,586
Random Survival Forest	+0,055 (−0,039 έως +0,149)	0,242	10/10	−0,010 (−0,025 έως +0,005)	0,188
Gradient Boosting	+0,070 (−0,022 έως +0,162)	0,131	10/10	−0,008 (−0,030 έως +0,015)	0,501

Θετική διαφορά στον δείκτη Harrell και αρνητική στο IBS σημαίνουν βελτίωση. Επαναλήψεις με βελτίωση: σε πόσες από τις 10 επαναλήψεις η μέση διαφορά στον δείκτη Harrell είχε την κατεύθυνση της βελτίωσης. Τιμές p από το διορθωμένο t-test, χωρίς διόρθωση για πολλαπλούς ελέγχους.

Αντίθετα, η προσθήκη του PSA, της θέσης βιοψίας και της ιστολογίας στην ηλικία και στο άθροισμα Gleason αύξησε τον δείκτη συμφωνίας σε όλους τους αλγορίθμους, κατά +0,021 (Cox PH) έως +0,070 (Gradient Boosting), σε 9 έως 10 από τις 10 επαναλήψεις. Και εδώ κανένα διάστημα δεν απέκλεισε το μηδέν (για παράδειγμα Gradient Boosting: +0,070, 95% ΔΕ: −0,022 έως +0,162), οπότε πρόκειται για συνεπή τάση και όχι για τεκμηριωμένη βελτίωση.



Σχήμα 5. Ζευγαρωμένες διαφορές «κλινικά + γονίδια» μείον «κλινικά διευρυμένα» στον δείκτη Harrell, στον δείκτη Uno και στο IBS, ανά αλγόριθμο (μέσος όρος και διορθωμένο 95% διάστημα εμπιστοσύνης). Για τους δείκτες συμφωνίας οι θετικές τιμές σημαίνουν βελτίωση, για το IBS οι αρνητικές.



Σχήμα 6. Ζευγαρωμένες διαφορές «κλινικά διευρυμένα» μείον «κλινικά βασικά» στον δείκτη Harrell, στον δείκτη Uno και στο IBS, ανά αλγόριθμο (μέσος όρος και διορθωμένο 95% διάστημα εμπιστοσύνης).

Απάντηση στο Ε2: τα γονίδια δεν πρόσθεσαν προγνωστική αξία στα κλινικά στοιχεία. Η μηδενική υπόθεση Υ2 δεν απορρίπτεται και η κατεύθυνση των διαφορών ήταν σταθερά προς τη χειροτέρευση.

6.4 Ανάλυση σε χρονικούς ορίζοντες

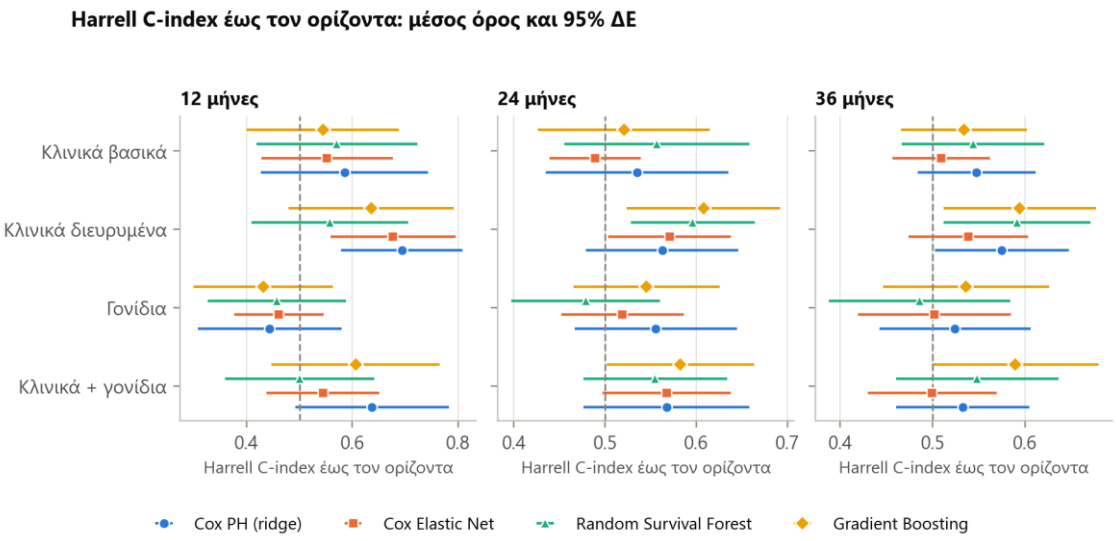
Ο Πίνακας 11 και το Σχήμα 7 παρουσιάζουν τον δείκτη συμφωνίας του Harrell σε κάθε ορίζοντα. Τα διευρυμένα κλινικά στοιχεία είχαν την καλύτερη επίδοση στους 12 μήνες, όπου το Cox PH έφτασε 0,694 (95% ΔΕ: 0,579–0,809) και το Cox Elastic Net 0,677 (0,558–0,796), με διαστήματα πάνω από το 0,5. Στους 24 και 36 μήνες οι τιμές τους ήταν χαμηλότερες (0,538–0,608).

Πίνακας 11. Δείκτης συμφωνίας του Harrell έως κάθε ορίζοντα (μέσος όρος 50 διαχωρισμών και διορθωμένο 95% διάστημα εμπιστοσύνης).

Σύνολο εισόδων	Αλγόριθμος	12 μήνες	24 μήνες	36 μήνες
Κλινικά βασικά	Cox PH (ridge)	0,585 (0,427–0,743)	0,535 (0,435–0,635)	0,548 (0,484–0,612)
	Cox Elastic Net	0,552 (0,428–0,676)	0,489 (0,439–0,539)	0,509 (0,457–0,562)
	Random Survival Forest	0,571 (0,419–0,723)	0,557 (0,455–0,658)	0,544 (0,467–0,621)
	Gradient Boosting	0,544 (0,400–0,688)	0,521 (0,426–0,615)	0,534 (0,466–0,602)
Κλινικά διευρυμένα	Cox PH (ridge)	0,694 (0,579–0,809)	0,563 (0,479–0,646)	0,575 (0,503–0,647)
	Cox Elastic Net	0,677 (0,558–0,796)	0,571 (0,504–0,638)	0,538 (0,474–0,603)
	Random Survival Forest	0,557 (0,409–0,705)	0,596 (0,528–0,665)	0,591 (0,512–0,671)
	Gradient Boosting	0,636 (0,480–0,793)	0,608 (0,524–0,692)	0,594 (0,512–0,677)
Γονίδια	Cox PH (ridge)	0,444 (0,307–0,580)	0,556 (0,467–0,644)	0,524 (0,442–0,606)

Σύνολο εισόδων	Αλγόριθμος	12 μήνες	24 μήνες	36 μήνες
	Cox Elastic Net	0,461 (0,376–0,546)	0,519 (0,452–0,587)	0,502 (0,419–0,585)
	Random Survival Forest	0,457 (0,326–0,588)	0,479 (0,397–0,560)	0,486 (0,388–0,584)
	Gradient Boosting	0,431 (0,299–0,564)	0,546 (0,466–0,625)	0,536 (0,446–0,626)
Κλινικά + γονίδια	Cox PH (ridge)	0,638 (0,492–0,783)	0,567 (0,477–0,658)	0,532 (0,461–0,604)
	Cox Elastic Net	0,545 (0,438–0,651)	0,568 (0,498–0,638)	0,499 (0,430–0,569)
	Random Survival Forest	0,501 (0,359–0,642)	0,555 (0,476–0,634)	0,548 (0,460–0,636)
	Gradient Boosting	0,606 (0,447–0,766)	0,583 (0,502–0,664)	0,590 (0,500–0,679)

Τα γονίδια μόνα τους δεν προέβλεψαν τους πρώιμους θανάτους: στους 12 μήνες ο δείκτης ήταν κάτω από 0,5 και στους τέσσερις αλγορίθμους (0,431–0,461), με διαστήματα που περιείχαν το 0,5. Τιμές ελαφρά κάτω από 0,5 είναι αναμενόμενες όταν δεν υπάρχει σήμα, διότι σε διασταυρωμένη επικύρωση με μικρά δείγματα και ασθενές σήμα οι δείκτες κατάταξης εμφανίζουν αρνητική μεροληψία (Parker et al., 2007). Στους 24 και 36 μήνες τα γονίδια έδωσαν τιμές 0,479–0,556.



Σχήμα 7. Δείκτης συμφωνίας του Harrell έως τον ορίζοντα (12, 24 και 36 μήνες) για κάθε σύνολο εισόδων και αλγόριθμο, με διορθωμένο 95% διάστημα εμπιστοσύνης. Η διακεκομμένη γραμμή αντιστοιχεί στην τυχαία κατάταξη (0,5).

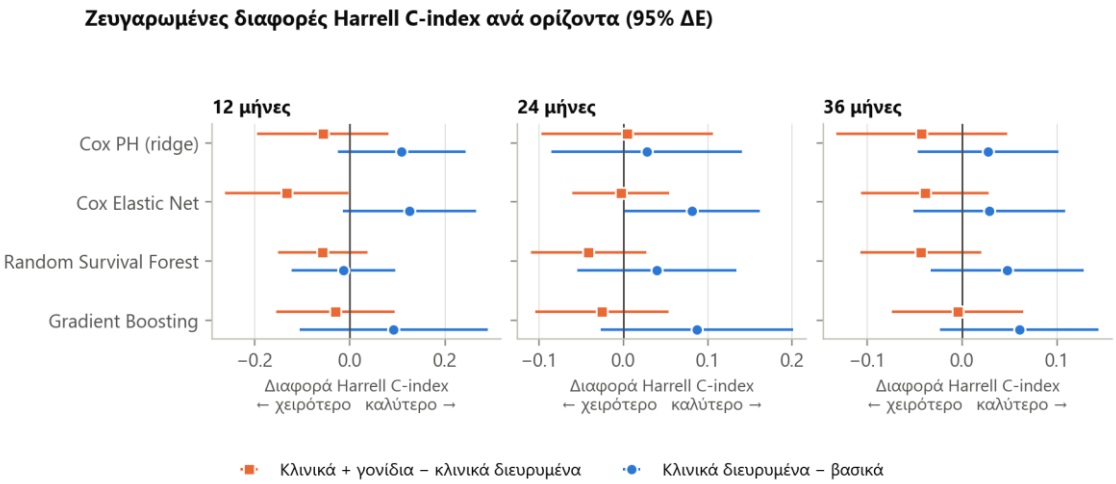
Οι ζευγαρωμένες διαφορές ανά ορίζοντα (Πίνακας 12, Σχήμα 8) επιβεβαιώνουν την εικόνα της κύριας ανάλυσης. Η προσθήκη των γονιδίων στα κλινικά μείωσε τον δείκτη συμφωνίας σε 11 από τους 12 συνδυασμούς ορίζοντα

και αλγορίθμου. Μόνο για το Cox Elastic Net στους 12 μήνες το διάστημα απέκλεισε το μηδέν (−0,132, 95% ΔΕ: −0,262 έως −0,003). Αντίθετα, τα διευρυμένα κλινικά στοιχεία αύξησαν τον δείκτη σε σχέση με τα βασικά σε 11 από τους 12 συνδυασμούς, περισσότερο στους 12 μήνες (+0,092 έως +0,125 για Cox PH, Cox Elastic Net και Gradient Boosting) και λιγότερο στους 24 και 36 μήνες (+0,027 έως +0,087). Μόνο για το Cox Elastic Net στους 24 μήνες το διάστημα απέκλεισε το μηδέν (+0,081, 95% ΔΕ: 0,001–0,162).

Πίνακας 12. Ζευγαρωμένες διαφορές στον δείκτη συμφωνίας του Harrell ανά ορίζοντα.

Αλγόριθμος	12 μήνες	24 μήνες	36 μήνες
Κλινικά + γονίδια − κλινικά διευρυμένα			
Cox PH (ridge)	−0,057 (−0,195 έως +0,082)	+0,005 (−0,097 έως +0,106)	−0,042 (−0,132 έως +0,047)
Cox Elastic Net	**−0,132 (−0,262 έως −0,003)**	−0,003 (−0,060 έως +0,055)	−0,039 (−0,106 έως +0,028)
Random Survival Forest	−0,057 (−0,151 έως +0,037)	−0,041 (−0,110 έως +0,027)	−0,043 (−0,107 έως +0,020)
Gradient Boosting	−0,030 (−0,154 έως +0,094)	−0,025 (−0,104 έως +0,053)	−0,005 (−0,073 έως +0,064)
Κλινικά διευρυμένα − κλινικά βασικά			
Cox PH (ridge)	+0,109 (−0,025 έως +0,243)	+0,028 (−0,085 έως +0,140)	+0,027 (−0,047 έως +0,101)
Cox Elastic Net	+0,125 (−0,015 έως +0,265)	**+0,081 (+0,001 έως +0,162)**	+0,029 (−0,051 έως +0,109)
Random Survival Forest	−0,014 (−0,122 έως +0,095)	+0,040 (−0,055 έως +0,134)	+0,048 (−0,033 έως +0,128)
Gradient Boosting	+0,092 (−0,105 έως +0,289)	+0,087 (−0,027 έως +0,201)	+0,060 (−0,023 έως +0,144)

Μέση διαφορά και διορθωμένο 95% διάστημα εμπιστοσύνης. Με έντονα γράμματα οι διαφορές με διάστημα που αποκλείει το μηδέν.



Σχήμα 8. Ζευγαρωμένες διαφορές στον δείκτη Harrell ανά ορίζοντα: «κλινικά + γονίδια» μείον «κλινικά διευρυμένα» (τετράγωνα) και «κλινικά διευρυμένα» μείον «κλινικά βασικά» (κύκλοι), με διορθωμένο 95% διάστημα εμπιστοσύνης.

Το Cox PH (ridge) με τα γονίδια είχε σε όλους τους ορίζοντες κριτήριο Brier χειρότερο από την καμπύλη Kaplan–Meier. Για παράδειγμα, με το σύνολο «κλινικά + γονίδια» 0,318 έναντι 0,251 στους 24 μήνες. Συνολικά, από τις 96 ζευγαρωμένες συγκρίσεις όλων των δεικτών, 6 είχαν διάστημα που απέκλειε το μηδέν, αριθμός κοντά σε όσες αναμένονται από τύχη με 96 ελέγχους (περίπου 5).

6.5 Αλλοιώσεις που σχετίζονται με την επιβίωση (E3)

Ο αριθμός των εκτιμήσιμων συσχετίσεων αυξάνεται με τον ορίζοντα, επειδή αυξάνονται οι θάνατοι (Πίνακας 13). Στους 12 μήνες, με 19 θανάτους, εκτιμήθηκαν μόνο 11 από τις 22 αλλοιώσεις στην ανάλυση χωρίς προσαρμογή και 6 στην ανάλυση με προσαρμογή. Από τους 77 εκτιμήσιμους ελέγχους, 9 είχαν $p < 0,05$ και κανένας $q < 0,05$ (Πίνακας 14, Σχήμα 9).

Πίνακας 13. Αριθμός εκτιμήσιμων συσχετίσεων αλλοίωσης–επιβίωσης και πόσες είχαν $p < 0,05$ ή $q < 0,05$, ανά ανάλυση και ορίζοντα (22 αλλοιώσεις).

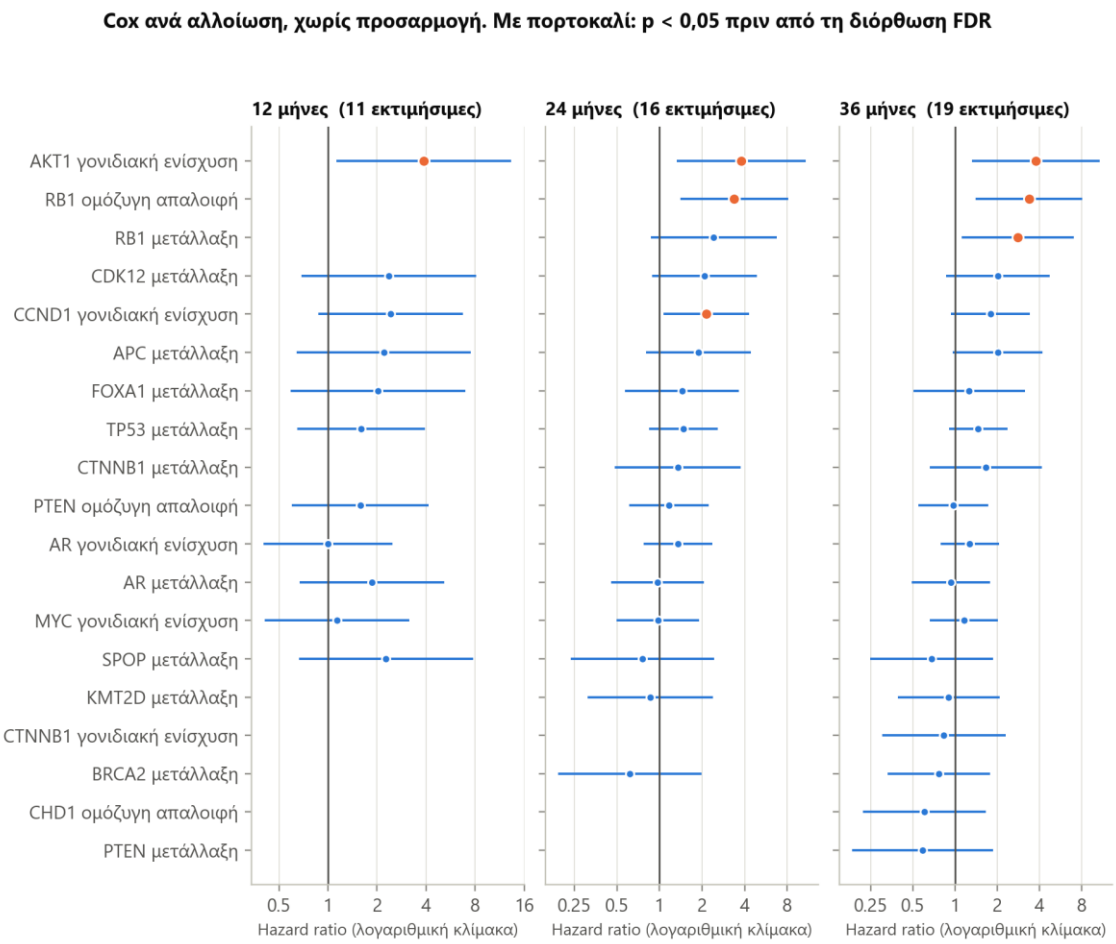
Ανάλυση	Ορίζοντας	Εκτιμήσιμες	$p < 0,05$	$q < 0,05$
Χωρίς προσαρμογή	12 μήνες	11	1	0
	24 μήνες	16	3	0
	36 μήνες	19	3	0
Με προσαρμογή	12 μήνες	6	1	0
	24 μήνες	10	0	0
	36 μήνες	15	1	0
Σύνολο		**77**	**9**	**0**

Ανάλυση με προσαρμογή: για ηλικία βιοψίας και άθροισμα Gleason, στους 106 ασθενείς με πλήρη στοιχεία. Εκτιμήσιμη: τουλάχιστον 5 φορείς, 5 μη φορείς και 3 θάνατοι σε κάθε ομάδα έως τον ορίζοντα. q: τιμή μετά από διόρθωση Benjamini–Hochberg.

Πίνακας 14. Συσχετίσεις αλλοίωσης–επιβίωσης με p < 0,05 (μοντέλο Cox ανά αλλοίωση).

Ορίζοντας	Αλλοίωση	Φορείς (θάνατοι)	HR (95% ΔΕ)	p	q
Χωρίς προσαρμογή (128 ασθενείς)					
12 μήνες	AKT1 γονιδιακή ενίσχυση	6 (3)	3,87 (1,12–13,34)	0,032	0,35
24 μήνες	RB1 ομόζυγη απαλοιφή	7 (6)	3,38 (1,41–8,11)	0,006	0,10
24 μήνες	AKT1 γονιδιακή ενίσχυση	6 (4)	3,79 (1,32–10,84)	0,013	0,10
24 μήνες	CCND1 γονιδιακή ενίσχυση	15 (10)	2,14 (1,07–4,28)	0,032	0,17
36 μήνες	RB1 ομόζυγη απαλοιφή	7 (6)	3,38 (1,41–8,11)	0,006	0,12
36 μήνες	AKT1 γονιδιακή ενίσχυση	6 (4)	3,79 (1,32–10,84)	0,013	0,12
36 μήνες	RB1 μετάλλαξη	6 (5)	2,81 (1,11–7,08)	0,028	0,18
Με προσαρμογή για ηλικία βιοψίας και άθροισμα Gleason (106 ασθενείς)					
12 μήνες	PTEN ομόζυγη απαλοιφή	28 (6)	3,08 (1,04–9,18)	0,043	0,26
36 μήνες	TP53 μετάλλαξη	41 (27)	1,71 (1,00–2,93)	0,049	0,39

Φορείς (θάνατοι): ασθενείς με την αλλοίωση και θάνατοι μεταξύ τους έως τον ορίζοντα. HR: λόγος κινδύνων φορέων έναντι μη φορέων. Καμία συσχέτιση δεν είχε q < 0,05.



Σχήμα 9. Λόγοι κινδύνων (λογαριθμική κλίμακα) με 95% διάστημα εμπιστοσύνης για κάθε εκτιμήσιμη αλλοίωση, στην ανάλυση χωρίς προσαρμογή, ανά ορίζοντα. Με πορτοκαλί οι συσχετίσεις με p < 0,05 πριν από τη διόρθωση FDR. Καμία δεν είχε q < 0,05.

Η ισχυρότερη και σταθερότερη ένδειξη αφορά το RB1. Η ομόζυγη απαλοιφή του RB1 (7 φορείς) συσχετίστηκε με περίπου τριπλάσιο κίνδυνο θανάτου στους 24 μήνες (HR 3,38, 95% ΔΕ: 1,41–8,11, p = 0,006, q = 0,10), με ακριβώς την ίδια εκτίμηση στους 36 μήνες. Η ταύτιση εξηγείται από τη χρονική κατανομή των θανάτων: και οι 6 θάνατοι των φορέων συνέβησαν έως τους 21 μήνες, οι 5 από αυτούς μεταξύ 12 και 21 μηνών. Μετά τους 24 μήνες δεν έμενε κανένας φορέας σε κίνδυνο, οπότε οι μεταγενέστεροι θάνατοι δεν μεταβάλλουν την εκτίμηση. Στους 12 μήνες η συσχέτιση δεν εκτιμήθηκε, διότι έως τότε είχε πεθάνει μόνο ένας φορέας. Η μετάλλαξη του RB1 (6 φορείς) είχε p < 0,05 μόνο στους 36 μήνες (HR 2,81), επειδή ένας από τους 5 θανάτους των φορέων συνέβη στους 29 μήνες.

Η γονιδιακή ενίσχυση του AKT1 (6 φορείς) ήταν η μόνη αλλοίωση με p < 0,05 στους 12 μήνες (HR 3,87, p = 0,032), καθώς 3 από τους 6 φορείς πέθαναν έως τότε, και είχε την ίδια εκτίμηση στους 24 και 36 μήνες (HR 3,79,

p = 0,013). Η εκτίμηση στηρίζεται σε ελάχιστους ασθενείς και είναι εύθραυστη. Με p < 0,05 σε έναν μόνο ορίζοντα εμφανίστηκαν επίσης η γονιδιακή ενίσχυση του CCND1 στους 24 μήνες (HR 2,14, p = 0,032) και, στην ανάλυση με προσαρμογή για ηλικία και Gleason, η ομόζυγη απαλοιφή του PTEN στους 12 μήνες (HR 3,08, p = 0,043) και η μετάλλαξη του TP53, οριακά, στους 36 μήνες (HR 1,71, p = 0,049). Ο έλεγχος των αναλογικών κινδύνων είχε p < 0,05 σε 5 από τους 77 ελέγχους, όλους στους 12 μήνες (γονιδιακές ενισχύσεις των AKT1, CCND1 και AR).

Απάντηση στο E3: καμία αλλοίωση δεν συσχετίστηκε με την επιβίωση μετά τη διόρθωση για πολλαπλούς ελέγχους, οπότε η μηδενική υπόθεση Y3 δεν απορρίπτεται. Οι διερευνητικές ενδείξεις αφορούν κυρίως το RB1, του οποίου η ομόζυγη απαλοιφή σχετίστηκε με θανάτους μέσα στα δύο πρώτα χρόνια από την έναρξη της θεραπείας ARSI. Με 7 φορείς, η στατιστική ισχύς δεν επαρκεί για επιβεβαίωση.

6.6 Σύνοψη των αποτελεσμάτων

Ο Πίνακας 15 συνοψίζει την απάντηση σε κάθε επιμέρους ερώτημα. Η ερμηνεία τους σε σχέση με το βασικό ερευνητικό ερώτημα γίνεται στο Κεφάλαιο 7.

Πίνακας 15. Σύνοψη των απαντήσεων στα επιμέρους ερευνητικά ερωτήματα.

Ερώτημα	Κύριο αποτέλεσμα	Υπόθεση
E1. Προβλέπουν οι αλλοιώσεις μόνες τους την εξέλιξη σε νέους ασθενείς;	Harrell C 0,488–0,529 στην κύρια ανάλυση και 0,431–0,556 στους ορίζοντες. Όλα τα διαστήματα περιέχουν το 0,5	Η Y1 δεν απορρίπτεται
E2. Προσθέτουν πληροφορία πέρα από τα κλινικά στοιχεία;	Μείωση του Harrell C κατά 0,027–0,048 σε όλους τους αλγορίθμους (διαστήματα που περιέχουν το 0), χειρότερο IBS και μείωση σε 11 από τους 12 συνδυασμούς ορίζοντα και αλγορίθμου	Η Y2 δεν απορρίπτεται
E3. Ποιες αλλοιώσεις σχετίζονται με την επιβίωση και σε ποιον ορίζοντα;	9 από 77 ελέγχους με p < 0,05, κανένας με q < 0,05. Ισχυρότερη ένδειξη η ομόζυγη απαλοιφή του RB1 (HR 3,38 στους 24 και 36 μήνες, q = 0,10–0,12), με όλους τους θανάτους των φορέων έως τους 21 μήνες	Η Y3 δεν απορρίπτεται. Διερευνητική ένδειξη για το RB1

7. ΣΥΖΗΤΗΣΗ

7.1 Απάντηση στο ερευνητικό ερώτημα

Το βασικό ερώτημα της εργασίας ήταν αν υπάρχουν γονίδια που προβλέπουν την εξέλιξη του μεταστατικού, ανθεκτικού στον ευνουχισμό καρκίνου του προστάτη, πέρα από όσα προβλέπουν τα κλινικά στοιχεία. Στους 128 ασθενείς της κοόρτης SU2C/PCF, η απάντηση είναι αρνητική για την πρόβλεψη και επιφυλακτική για τη συσχέτιση:

- **Πρόβλεψη (E1):** τα μοντέλα με μόνο τις αλλοιώσεις των γονιδίων δεν διέκριναν τους ασθενείς καλύτερα από την τύχη (δείκτης συμφωνίας 0,488–0,529).
- **Προστιθέμενη αξία (E2):** η προσθήκη των γονιδίων στα κλινικά στοιχεία δεν βελτίωσε κανένα μοντέλο. Ο δείκτης συμφωνίας μειώθηκε σε όλους τους αλγορίθμους και το κριτήριο Brier χειροτέρεψε.
- **Συσχέτιση (E3):** καμία αλλοίωση δεν παρέμεινε σημαντική μετά τη διόρθωση για πολλαπλούς ελέγχους. Η ισχυρότερη ένδειξη αφορούσε την ομόζυγη απαλοιφή του RB1, με όλους τους θανάτους των φορέων μέσα στους πρώτους 21 μήνες.

Το αρνητικό αποτέλεσμα δεν σημαίνει ότι τα γονίδια αυτά στερούνται βιολογικής σημασίας. Σημαίνει ότι, με τα διαθέσιμα δεδομένα, η πληροφορία τους δεν μεταφράζεται σε πρόβλεψη της επιβίωσης για τον μέσο ασθενή. Οι ενότητες που ακολουθούν εξετάζουν γιατί συμβαίνει αυτό και τι σημαίνει.

7.2 Ερμηνεία των ευρημάτων

7.2.1 Γιατί τα γονίδια δεν βελτίωσαν την πρόβλεψη

Ο πρώτος λόγος είναι η σπανιότητα των αλλοιώσεων. Από τις 22 αλλοιώσεις των γονιδίων, οι 17 εμφανίζονταν σε 5 έως 15 ασθενείς. Μια αλλοίωση με μεγάλο λόγο κινδύνων αλλά λίγους φορείς αλλάζει τη σειρά λίγων ζευγών ασθενών και επηρεάζει ελάχιστα τον δείκτη συμφωνίας ολόκληρης της κοόρτης. Η ομόζυγη απαλοιφή του RB1 είναι χαρακτηριστικό παράδειγμα: με λόγο κινδύνων περίπου 3,4 αφορούσε μόλις 7 από τους 128 ασθενείς. Το φαινόμενο είναι γνωστό από τη μεθοδολογική βιβλιογραφία, όπου ένας δείκτης με λόγο πιθανοτήτων 3 μπορεί να είναι φτωχό εργαλείο ταξινόμησης (Pepe et al., 2004).

Ο δεύτερος λόγος είναι η σχέση συμβάντων και μεταβλητών. Με περίπου 61 θανάτους σε κάθε σύνολο εκπαίδευσης και έως 33 μεταβλητές, σε κάθε μεταβλητή αντιστοιχούσαν λιγότεροι από δύο θάνατοι, πολύ κάτω από τα επίπεδα που θεωρούνται ασφαλή (Peduzzi et al., 1995 και Riley et al., 2019). Αυτό φάνηκε καθαρά στο μοντέλο Cox με τη σταθερή ποινή ridge, το οποίο με τα γονίδια είχε κριτήριο Brier χειρότερο από την καμπύλη Kaplan–Meier, δηλαδή υπερπροσαρμόστηκε. Το Cox Elastic Net, του οποίου η ποινή ρυθμίζεται με επικύρωση, κράτησε λίγες μεταβλητές και έμεινε κοντά στο σημείο αναφοράς: η κανονικοποίηση προστάτευσε το μοντέλο από τον θόρυβο, αλλά δεν μπορούσε να δημιουργήσει σήμα που δεν υπήρχε.

Τα ευρήματα συμφωνούν με τη μεγάλη συγκριτική μελέτη των Herrmann et al. (2021), όπου τα ωμικά δεδομένα πρόσθεσαν ελάχιστα στην πρόβλεψη της επιβίωσης πέρα από τις κλινικές μεταβλητές, και με τη διαπίστωση ότι οι πιο σύνθετοι αλγόριθμοι δεν υπερέχουν συστηματικά των απλούστερων όταν η επικύρωση είναι αμερόληπτη (Christodoulou et al., 2019). Στην παρούσα εργασία, τα σύνολα δέντρων είχαν ελαφρά υψηλότερο δείκτη συμφωνίας με τα κλινικά στοιχεία, αλλά η διαφορά από τα μοντέλα Cox ήταν μικρή σε σχέση με την αβεβαιότητα.

7.2.2 Η χρονική διάσταση και το RB1

Η ανάλυση σε ορίζοντες πρόσθεσε μια χρονική οπτική. Η συσχέτιση της ομόζυγης απαλοιφής του RB1 με την επιβίωση οφειλόταν εξ ολοκλήρου σε θανάτους μέσα στα δύο πρώτα χρόνια μετά την έναρξη της θεραπείας ARSI. Αυτό είναι συμβατό με τη βιολογία του γονιδίου: η απώλεια του RB1, ιδίως μαζί με την απώλεια του TP53, διευκολύνει τη μετάπτωση των καρκινικών κυττάρων σε φαινότυπο ανεξάρτητο από τον AR και την αντίσταση στην αντιανδρογονική θεραπεία (Ku et al., 2017, Mu et al., 2017 και Nyquist et al., 2020). Ασθενείς των οποίων ο όγκος δεν εξαρτάται πλέον από τον AR αναμένεται να ωφεληθούν λιγότερο από τους ARSI και να εξελιχθούν νωρίς. Η κατεύθυνση του ευρήματος συμφωνεί επίσης με τη συσχέτιση των διπλών αλλοιώσεων του RB1 με βραχύτερη επιβίωση σε ανεξάρτητη κοόρτη mCRPC με αλληλούχιση ολικού γονιδιώματος (Chen et al., 2019).

Οι υπόλοιπες ονομαστικά σημαντικές συσχετίσεις χρειάζονται ακόμη μεγαλύτερη επιφύλαξη. Η γονιδιακή ενίσχυση του AKT1 στηριζόταν σε 6 φορείς

και 3 πρώιμους θανάτους, και ο έλεγχος αναλογικών κινδύνων υποδήλωσε επίδραση που δεν είναι σταθερή στον χρόνο. Βιολογικά είναι εύλογη, αφού η ενεργοποίηση του μονοπατιού PI3K/AKT έχει συνδεθεί με αντίσταση στην ορμονοθεραπεία. Η απώλεια της πρωτεΐνης PTEN, που ενεργοποιεί το ίδιο μονοπάτι, συσχετίστηκε με βραχύτερη επιβίωση σε ασθενείς υπό abiraterone (Ferraldeschi et al., 2015). Η οριακή συσχέτιση της μετάλλαξης του TP53 στους 36 μήνες είναι ασθενέστερη από όσο θα περίμενε κανείς με βάση τις μελέτες ctDNA, όπου οι αλλοιώσεις του TP53 συνδέθηκαν ανεξάρτητα με ταχύτερη αντίσταση (Annala et al., 2018 και De Laere et al., 2019). Μια πιθανή εξήγηση είναι ότι εδώ ως αλλοίωση του TP53 μετρήθηκε μόνο η μετάλλαξη, ενώ οι μελέτες αυτές εξέτασαν τη συνολική αδρανοποίηση του γονιδίου, συνδυάζοντας μεταλλάξεις και απώλειες αντιγράφων.

Αξιοσημείωτο είναι ότι η γονιδιακή ενίσχυση του AR, η συχνότερη αλλοίωση της κοόρτης (42%), δεν συσχετίστηκε με την επιβίωση σε κανέναν ορίζοντα, ενώ η αύξηση των αντιγράφων του AR στο πλάσμα έχει συνδεθεί με χειρότερη έκβαση υπό ARSI (Romanel et al., 2015 και Conteduca et al., 2017). Η διαφορά μπορεί να οφείλεται στη μέτρηση: στο πλάσμα η ενίσχυση του AR αποτυπώνει εν μέρει και την ποσότητα του DNA του όγκου που κυκλοφορεί, η οποία σχετίζεται με το φορτίο της νόσου, ενώ στη βιοψία ιστού μετριέται σε ένα μόνο σημείο της νόσου. Σε συμφωνία με αυτό, σε τυχαιοποιημένη μελέτη πρώτης γραμμής η ενίσχυση του AR στο ctDNA δεν υπερείχε των καθιερωμένων προγνωστικών δεικτών (Annala et al., 2018).

7.2.3 Ο ρόλος των κλινικών στοιχείων

Η καλύτερη επίδοση παρατηρήθηκε με τα διευρυμένα κλινικά στοιχεία, κυρίως για τους πρώιμους θανάτους (δείκτης συμφωνίας έως 0,694 στους 12 μήνες). Ακόμη και αυτά όμως έδωσαν μόνο μέτρια διάκριση στο σύνολο της παρακολούθησης (έως 0,614). Ένας λόγος είναι ότι το δημόσιο σύνολο δεδομένων δεν περιλαμβάνει τους κλινικούς παράγοντες με την ισχυρότερη τεκμηριωμένη προγνωστική αξία στο mCRPC: τη γενική κατάσταση του ασθενούς, τη γαλακτική αφυδρογονάση, την αλκαλική φωσφατάση, τη λευκωματίνη, την αιμοσφαιρίνη και το φορτίο των μεταστάσεων (Halabi et al., 2014, Guinney et al., 2017 και Uccello et al., 2026). Επιπλέον, το PSA ήταν

διαθέσιμο μόνο κατά την αρχική διάγνωση και όχι κατά την έναρξη της θεραπείας ARSI.

Αυτό έχει δύο συνέπειες. Πρώτον, η κλινική γραμμή αναφοράς της εργασίας είναι ασθενέστερη από όσο θα ήταν στην κλινική πράξη, οπότε η σύγκριση έγινε υπό συνθήκες ευνοϊκές για τα γονίδια. Παρ' όλα αυτά, τα γονίδια δεν πρόσθεσαν αξία. Δεύτερον, το ανώτατο όριο της πρόβλεψης που μπορεί να επιτευχθεί με αυτά τα δεδομένα είναι χαμηλό, και μια δίκαιη αξιολόγηση της προστιθέμενης αξίας των γονιδίων θα απαιτούσε πληρέστερα κλινικά στοιχεία.

7.3 Περιορισμοί

- **Μέγεθος δείγματος.** Με 128 ασθενείς και 76 θανάτους τα διαστήματα εμπιστοσύνης είναι ευρεία, και η εργασία δεν έχει την ισχύ να ανιχνεύσει μέτριες επιδράσεις σπάνιων αλλοιώσεων. Η ομόζυγη απαλοιφή του RB1, για παράδειγμα, αφορούσε 7 ασθενείς.
- **Χρονική αντιστοίχιση και ελλείψεις.** Σε 14 ασθενείς η βιοψία ελήφθη κατά τη διάρκεια ή μετά από θεραπεία ARSI, οπότε το γονιδιωματικό τους προφίλ δεν αντιστοιχεί ακριβώς στην έναρξη της θεραπείας. Ελλείψεις υπήρχαν στο άθροισμα Gleason (21 ασθενείς) και στο PSA (26). Οι δείκτες έλλειψης που χρησιμοποιήθηκαν μπορεί να αποτυπώνουν διαφορές στην πρακτική καταγραφής μεταξύ των κέντρων και όχι βιολογία.
- **Ελλιπή κλινικά στοιχεία.** Λείπουν οι καθιερωμένοι κλινικοί και εργαστηριακοί προγνωστικοί παράγοντες του mCRPC (ενότητα 7.2.3).
- **Κωδικοποίηση των αλλοιώσεων.** Χρησιμοποιήθηκαν δυαδικές μεταβλητές για τις μεταλλάξεις και για τις μεταβολές υψηλού επιπέδου του αριθμού αντιγράφων. Δεν εξετάστηκαν οι ημιζυγείς απαλοιφές, ο συνδυασμός μετάλλαξης και απώλειας στο ίδιο γονίδιο, οι δομικές αλλοιώσεις που δεν ανιχνεύονται με την αλληλούχιση των εξωνίων (Quigley et al., 2018) και οι γαμετικές μεταλλάξεις.
- **Επιλογή γονιδίων και δείγματος.** Τα 20 γονίδια ορίστηκαν εκ των προτέρων και δεν καλύπτουν τη «μακριά ουρά» των σπάνιων γονιδίων-οδηγών (Armenia et al., 2018). Κάθε ασθενής αντιπροσωπεύεται από μία βιοψία. Μία μετάσταση δίνει συνήθως εύλογη εικόνα των κύριων αλλοιώσεων, αλλά υπάρχουν εξαιρέσεις (Kumar et al., 2016).
- **Επικύρωση.** Η επικύρωση ήταν εσωτερική, με επαναλαμβανόμενη εμφωλευμένη διασταυρωμένη επικύρωση. Δεν υπήρξε ανεξάρτητη εξωτερική κοόρτη. Η ανάλυση σε ορίζοντες σχεδιάστηκε μετά την κύρια ανάλυση και έχει διερευνητικό χαρακτήρα.

- **Έκβαση.** Η συνολική επιβίωση επηρεάζεται και από τις θεραπείες που ακολούθησαν την πρώτη ARSI, για τις οποίες δεν υπάρχουν στοιχεία στο σύνολο δεδομένων.

7.4 Μελλοντικές κατευθύνσεις

Τα ευρήματα υποδεικνύουν πέντε κατευθύνσεις:

- **Μεγαλύτερες κοόρτες και εξωτερική επικύρωση.** Η ένδειξη για το RB1 χρειάζεται επιβεβαίωση σε ανεξάρτητες κοόρτες με ARSI πρώτης γραμμής, ιδανικά με τους καθιερωμένους κλινικούς προγνωστικούς παράγοντες, ώστε η προστιθέμενη αξία των γονιδίων να κριθεί απέναντι σε μια ισχυρή κλινική γραμμή αναφοράς.
- **Σειριακή υγρή βιοψία.** Οι αναλύσεις ctDNA επιτρέπουν επαναλαμβανόμενες μετρήσεις κατά τη διάρκεια της θεραπείας, όπως συνιστά η ομάδα PCWG3 (Scher et al., 2016), και έχουν ήδη αναδείξει δείκτες ταχείας αντίστασης (Wyatt et al., 2016 και Annala et al., 2018).
- **Πυκνότερη αναπαράσταση των αλλοιώσεων.** Η συνένωση των αλλοιώσεων σε επίπεδο γονιδίου (για παράδειγμα μετάλλαξη ή απώλεια του TP53) ή βιολογικού μονοπατιού μειώνει τον αριθμό των σπάνιων μεταβλητών και μπορεί να αυξήσει τη στατιστική ισχύ. Μέθοδοι που λαμβάνουν υπόψη τη δομή των ομάδων μεταβλητών προστατεύουν επίσης την πληροφορία των κλινικών στοιχείων (Herrmann et al., 2021).
- **Μεταγραφωματικά δεδομένα.** Υπογραφές έκφρασης που αποτυπώνουν τη δραστηριότητα του AR, τον νευροενδοκρινικό φαινότυπο ή τη συνδυασμένη απώλεια των TP53 και RB1 έχουν συνδεθεί με την επιβίωση (Nyquist et al., 2020 και Vlachostergios et al., 2022) και θα μπορούσαν να συμπληρώσουν τα δεδομένα του DNA στην ίδια κοόρτη.
- **Χρονικά μεταβαλλόμενες επιδράσεις.** Η συγκέντρωση της επίδρασης του RB1 στα δύο πρώτα χρόνια υποδεικνύει ότι μοντέλα που επιτρέπουν λόγους κινδύνων μεταβαλλόμενους στον χρόνο μπορεί να περιγράφουν την πρόγνωση καλύτερα από το μοντέλο αναλογικών κινδύνων.

7.5 Συμπεράσματα

Στην κοόρτη SU2C/PCF, οι αλλοιώσεις σε 20 από τα συχνότερα αλλοιωμένα γονίδια του καρκίνου του προστάτη δεν προέβλεψαν την επιβίωση ασθενών με mCRPC υπό θεραπεία ARSI πρώτης γραμμής, ούτε μόνες τους ούτε μαζί με τα διαθέσιμα κλινικά στοιχεία. Η μόνη συνεπής ένδειξη, η ομόζυγη απαλοιφή του RB1, αφορά λίγους ασθενείς, εκδηλώνεται μέσα στα δύο πρώτα

χρόνια της θεραπείας και δεν επιβεβαιώνεται μετά τη διόρθωση για πολλαπλούς ελέγχους. Η εργασία δείχνει ότι μια αλλοίωση μπορεί να είναι βιολογικά σημαντική χωρίς να βελτιώνει την πρόβλεψη για το σύνολο των ασθενών, και ότι σε μικρές κοόρτες η αξιολόγηση της προγνωστικής αξίας των γονιδίων απαιτεί αυστηρή επικύρωση, σύγκριση με τα κλινικά στοιχεία και έλεγχο για πολλαπλές συγκρίσεις.

8. ΒΙΒΛΙΟΓΡΑΦΙΑ

Οι αναφορές παρατίθενται κατά το πρότυπο APA, 7η έκδοση, σε αλφαβητική σειρά.

Abida, W., Cyrta, J., Heller, G., Prandi, D., Armenia, J., Coleman, I., … Sawyers, C. L. (2019). Genomic correlates of clinical outcome in advanced prostate cancer. Proceedings of the National Academy of Sciences, 116(23), 11428–11436. https://doi.org/10.1073/pnas.1902651116

Adamaki, M., & Zoumpourlis, V. (2021). Prostate cancer biomarkers: From diagnosis to prognosis and precision-guided therapeutics. Pharmacology & Therapeutics, 228, 107932. https://doi.org/10.1016/j.pharmthera.2021.107932

Ambroise, C., & McLachlan, G. J. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. Proceedings of the National Academy of Sciences, 99(10), 6562–6566. https://doi.org/10.1073/pnas.102102699

Annala, M., Vandekerkhove, G., Khalaf, D., Taavitsainen, S., Beja, K., Warner, E. W., … Chi, K. N. (2018). Circulating tumor DNA genomics correlate with resistance to abiraterone and enzalutamide in prostate cancer. Cancer Discovery, 8(4), 444–457. https://doi.org/10.1158/2159-8290.CD-17-0937

Antonarakis, E. S., Lu, C., Wang, H., Luber, B., Nakazawa, M., Roeser, J. C., … Luo, J. (2014). AR-V7 and resistance to enzalutamide and abiraterone in prostate cancer. New England Journal of Medicine, 371(11), 1028–1038. https://doi.org/10.1056/NEJMoa1315815

Armenia, J., Wankowicz, S. A. M., Liu, D., Gao, J., Kundra, R., Reznik, E., … Van Allen, E. M. (2018). The long tail of oncogenic drivers in prostate cancer. Nature Genetics, 50(5), 645–651. https://doi.org/10.1038/s41588-018-0078-z

Beer, T. M., Armstrong, A. J., Rathkopf, D. E., Loriot, Y., Sternberg, C. N., Higano, C. S., … PREVAIL Investigators. (2014). Enzalutamide in metastatic prostate cancer before chemotherapy. New England Journal of Medicine, 371(5), 424–433. https://doi.org/10.1056/NEJMoa1405095

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B, 57(1), 289–300.

Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., & Jemal, A. (2024). Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians, 74(3), 229–263. https://doi.org/10.3322/caac.21834

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breslow, N. E. (1974). Covariance analysis of censored survival data. Biometrics, 30(1), 89–99.

Cawley, G. C., & Talbot, N. L. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. Journal of Machine Learning Research, 11, 2079–2107.

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., … Schultz, N. (2012). The cBio Cancer Genomics Portal: An open platform for exploring multidimensional cancer genomics data. Cancer Discovery, 2(5), 401–404. https://doi.org/10.1158/2159-8290.CD-12-0095

Chen, W. S., Aggarwal, R., Zhang, L., Zhao, S. G., Thomas, G. V., Beer, T. M., … West Coast Prostate Cancer Dream Team. (2019). Genomic drivers of poor prognosis and enzalutamide resistance in metastatic castration-resistant prostate cancer. European Urology, 76(5), 562–571. https://doi.org/10.1016/j.eururo.2019.03.020

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. Journal of Clinical Epidemiology, 110, 12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004

Conteduca, V., Wetterskog, D., Sharabiani, M. T. A., Grande, E., Fernandez-Perez, M. P., Jayaram, A., … Attard, G. (2017). Androgen receptor gene status in plasma DNA associates with worse outcome on enzalutamide or abiraterone for castration-resistant prostate cancer: A multi-institution correlative biomarker study. Annals of Oncology, 28(7), 1508–1516. https://doi.org/10.1093/annonc/mdx155

Cox, D. R. (1972). Regression models and life-tables. Journal of the Royal Statistical Society: Series B, 34(2), 187–220.

Davidson-Pilon, C. (2019). lifelines: Survival analysis in Python. Journal of Open Source Software, 4(40), 1317. https://doi.org/10.21105/joss.01317

de Bono, J. S., Logothetis, C. J., Molina, A., Fizazi, K., North, S., Chu, L., … COU-AA-301 Investigators. (2011). Abiraterone and increased survival in metastatic prostate cancer. New England Journal of Medicine, 364(21), 1995–2005. https://doi.org/10.1056/NEJMoa1014618

De Laere, B., Oeyen, S., Mayrhofer, M., Whitington, T., van Dam, P. J., Van Oyen, P., … Dirix, L. Y. (2019). TP53 outperforms other androgen receptor biomarkers to predict abiraterone or enzalutamide outcome in metastatic castration-resistant prostate cancer. Clinical Cancer Research, 25(6), 1766–1773. https://doi.org/10.1158/1078-0432.CCR-18-1943

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. Journal of the American Statistical Association, 72(359), 557–565.

Efstathiou, E., Titus, M., Wen, S., Hoang, A., Karlou, M., Ashe, R., … Logothetis, C. J. (2015). Molecular characterization of enzalutamide-treated bone metastatic castration-resistant prostate cancer. European Urology, 67(1), 53–60. https://doi.org/10.1016/j.eururo.2014.05.005

Epstein, J. I., Egevad, L., Amin, M. B., Delahunt, B., Srigley, J. R., & Humphrey, P. A. (2016). The 2014 International Society of Urological Pathology (ISUP) Consensus Conference on Gleason grading of prostatic carcinoma. American Journal of Surgical Pathology, 40(2), 244–252. https://doi.org/10.1097/PAS.0000000000000530

Ferraldeschi, R., Nava Rodrigues, D., Riisnaes, R., Miranda, S., Figueiredo, I., Rescigno, P., … de Bono, J. (2015). PTEN protein loss and clinical outcome from castration-resistant prostate cancer treated with abiraterone acetate. European Urology, 67(4), 795–802. https://doi.org/10.1016/j.eururo.2014.10.027

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), 1189–1232.

Friedman, J. H. (2002). Stochastic gradient boosting. Computational Statistics & Data Analysis, 38(4), 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., … Schultz, N. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Science Signaling, 6(269), pl1. https://doi.org/10.1126/scisignal.2004088

Gleason, D. F. (1966). Classification of prostatic carcinomas. Cancer Chemotherapy Reports, 50(3), 125–128.

Graf, E., Schmoor, C., Sauerbrei, W., & Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. Statistics in Medicine, 18(17–18), 2529–2545. https://doi.org/10.1002/(SICI)1097-0258(19990915/30)18:17/18<2529::AID-SIM274>3.0.CO;2-5

Grasso, C. S., Wu, Y. M., Robinson, D. R., Cao, X., Dhanasekaran, S. M., Khan, A. P., … Tomlins, S. A. (2012). The mutational landscape of lethal castration-resistant prostate cancer. Nature, 487(7406), 239–243. https://doi.org/10.1038/nature11125

Guinney, J., Wang, T., Laajala, T. D., Winner, K. K., Bare, J. C., Neto, E. C., … Prostate Cancer Challenge DREAM Community. (2017). Prediction of overall survival for patients with metastatic castration-resistant prostate cancer: Development of a prognostic model through a crowdsourced challenge with open clinical trial data. The Lancet Oncology, 18(1), 132–142. https://doi.org/10.1016/S1470-2045(16)30560-5

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, 3, 1157–1182.

Halabi, S., Lin, C. Y., Kelly, W. K., Fizazi, K. S., Moul, J. W., Kaplan, E. B., … Small, E. J. (2014). Updated prognostic model for predicting overall survival in first-line chemotherapy for patients with metastatic castration-resistant prostate cancer. Journal of Clinical Oncology, 32(7), 671–677. https://doi.org/10.1200/JCO.2013.52.3696

Hamid, A. A., Gray, K. P., Shaw, G., MacConaill, L. E., Evan, C., Bernard, B., … Sweeney, C. J. (2019). Compound genomic alterations of TP53, PTEN, and RB1

tumor suppressors in localized and metastatic prostate cancer. European Urology, 76(1), 89–97. https://doi.org/10.1016/j.eururo.2018.11.045

Harrell, F. E., Lee, K. L., & Mark, D. B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Statistics in Medicine, 15(4), 361–387.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.

Heagerty, P. J., & Zheng, Y. (2005). Survival model predictive accuracy and ROC curves. Biometrics, 61(1), 92–105. https://doi.org/10.1111/j.0006-341X.2005.030814.x

Herrmann, M., Probst, P., Hornung, R., Jurinovic, V., & Boulesteix, A.-L. (2021). Large-scale benchmark study of survival prediction methods using multi-omics data. Briefings in Bioinformatics, 22(3), bbaa167. https://doi.org/10.1093/bib/bbaa167

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. The Annals of Applied Statistics, 2(3), 841–860. https://doi.org/10.1214/08-AOAS169

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. Journal of the American Statistical Association, 53(282), 457–481.

Knudson, A. G. (1971). Mutation and cancer: Statistical study of retinoblastoma. Proceedings of the National Academy of Sciences, 68(4), 820–823. https://doi.org/10.1073/pnas.68.4.820

Kontos, C. K., Adamopoulos, P. G., & Scorilas, A. (2015). Prognostic and predictive biomarkers in prostate cancer. Expert Review of Molecular Diagnostics, 15(12), 1567–1576. https://doi.org/10.1586/14737159.2015.1110022

Kourou, K., Exarchos, K. P., Papaloukas, C., Sakaloglou, P., Exarchos, T., & Fotiadis, D. I. (2021). Applied machine learning in cancer research: A systematic review for patient diagnosis, classification and prognosis. Computational and Structural Biotechnology Journal, 19, 5546–5555. https://doi.org/10.1016/j.csbj.2021.10.006

Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I. (2015). Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal, 13, 8–17. https://doi.org/10.1016/j.csbj.2014.11.005

Ku, S. Y., Rosario, S., Wang, Y., Mu, P., Seshadri, M., Goodrich, Z. W., … Goodrich, D. W. (2017). Rb1 and Trp53 cooperate to suppress prostate cancer lineage plasticity, metastasis, and antiandrogen resistance. Science, 355(6320), 78–83. https://doi.org/10.1126/science.aah4199

Kumar, A., Coleman, I., Morrissey, C., Zhang, X., True, L. D., Gulati, R., … Nelson, P. S. (2016). Substantial interindividual and limited intraindividual genomic diversity among tumors from men with metastatic prostate cancer. Nature Medicine, 22(4), 369–378. https://doi.org/10.1038/nm.4053

Liontos, M., Bournakis, E., Bournakis, A., Kostouros, E., Zolota, V., Papatheodoridi, A. P., … Bamias, A. (2024). Real-world treatment patterns in patients with metastatic castration-resistant prostate cancer in Greece: The PROSPECT study. Clinical Genitourinary Cancer, 22(6), 102170. https://doi.org/10.1016/j.clgc.2024.102170

Mermel, C. H., Schumacher, S. E., Hill, B., Meyerson, M. L., Beroukhim, R., & Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biology, 12(4), R41. https://doi.org/10.1186/gb-2011-12-4-r41

Mu, P., Zhang, Z., Benelli, M., Karthaus, W. R., Hoover, E., Chen, C. C., … Sawyers, C. L. (2017). SOX2 promotes lineage plasticity and antiandrogen resistance in TP53- and RB1-deficient prostate cancer. Science, 355(6320), 84–88. https://doi.org/10.1126/science.aah4307

Nadeau, C., & Bengio, Y. (2003). Inference for the generalization error. Machine Learning, 52(3), 239–281. https://doi.org/10.1023/A:1024068626366

Nyquist, M. D., Corella, A., Coleman, I., De Sarkar, N., Kaipainen, A., Ha, G., … Nelson, P. S. (2020). Combined TP53 and RB1 loss promotes prostate cancer resistance to a spectrum of therapeutics and confers vulnerability to replication stress. Cell Reports, 31(8), 107669. https://doi.org/10.1016/j.celrep.2020.107669

Panagiotopoulos, K., Korfiati, A., Theofilatos, K., Hurwitz, P., Deriu, M. A., & Mavroudi, S. (2023). MEvA-X: A hybrid multiobjective evolutionary tool using an XGBoost classifier for biomarkers discovery on biomedical datasets. Bioinformatics, 39(7), btad384. https://doi.org/10.1093/bioinformatics/btad384

Parker, B. J., Günter, S., & Bedo, J. (2007). Stratification bias in low signal microarray studies. BMC Bioinformatics, 8, 326. https://doi.org/10.1186/1471-2105-8-326

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Peduzzi, P., Concato, J., Feinstein, A. R., & Holford, T. R. (1995). Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. Journal of Clinical Epidemiology, 48(12), 1503–1510. https://doi.org/10.1016/0895-4356(95)00048-8

Pepe, M. S., Janes, H., Longton, G., Leisenring, W., & Newcomb, P. (2004). Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. American Journal of Epidemiology, 159(9), 882–890. https://doi.org/10.1093/aje/kwh101

Pölsterl, S. (2020). scikit-survival: A library for time-to-event analysis built on top of scikit-learn. Journal of Machine Learning Research, 21(212), 1–6.

Quigley, D. A., Dang, H. X., Zhao, S. G., Lloyd, P., Aggarwal, R., Alumkal, J. J., … Feng, F. Y. (2018). Genomic hallmarks and structural variation in metastatic prostate cancer. Cell, 174(3), 758–769.e9. https://doi.org/10.1016/j.cell.2018.06.039

Ridgeway, G. (1999). The state of boosting. Computing Science and Statistics, 31, 172–181.

Riley, R. D., Snell, K. I., Ensor, J., Burke, D. L., Harrell, F. E., Jr., Moons, K. G., & Collins, G. S. (2019). Minimum sample size for developing a multivariable prediction model: Part II — binary and time-to-event outcomes. Statistics in Medicine, 38(7), 1276–1296. https://doi.org/10.1002/sim.7992

Robinson, D., Van Allen, E. M., Wu, Y. M., Schultz, N., Lonigro, R. J., Mosquera, J. M., … Chinnaiyan, A. M. (2015). Integrative clinical genomics of advanced prostate cancer. Cell, 161(5), 1215–1228. https://doi.org/10.1016/j.cell.2015.05.001

Romanel, A., Gasi Tandefelt, D., Conteduca, V., Jayaram, A., Casiraghi, N., Wetterskog, D., … Attard, G. (2015). Plasma AR and abiraterone-resistant prostate cancer. Science Translational Medicine, 7(312), 312re10. https://doi.org/10.1126/scitranslmed.aac9511

Ryan, C. J., Smith, M. R., de Bono, J. S., Molina, A., Logothetis, C. J., de Souza, P., … COU-AA-302 Investigators. (2013). Abiraterone in metastatic prostate cancer without previous chemotherapy. New England Journal of Medicine, 368(2), 138–148. https://doi.org/10.1056/NEJMoa1209096

Sandhu, S., Moore, C. M., Chiong, E., Beltran, H., Bristow, R. G., & Williams, S. G. (2021). Prostate cancer. The Lancet, 398(10305), 1075–1090. https://doi.org/10.1016/S0140-6736(21)00950-8

Scher, H. I., Fizazi, K., Saad, F., Taplin, M. E., Sternberg, C. N., Miller, K., … AFFIRM Investigators. (2012). Increased survival with enzalutamide in prostate cancer after chemotherapy. New England Journal of Medicine, 367(13), 1187–1197. https://doi.org/10.1056/NEJMoa1207506

Scher, H. I., Morris, M. J., Stadler, W. M., Higano, C., Basch, E., Fizazi, K., … Prostate Cancer Clinical Trials Working Group 3. (2016). Trial design and objectives for castration-resistant prostate cancer: Updated recommendations from the Prostate Cancer Clinical Trials Working Group 3. Journal of Clinical Oncology, 34(12), 1402–1418. https://doi.org/10.1200/JCO.2015.64.2702

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. Biometrika, 69(1), 239–241. https://doi.org/10.1093/biomet/69.1.239

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. Journal of Statistical Software, 39(5), 1–13. https://doi.org/10.18637/jss.v039.i05

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., … Kattan, M. W. (2010). Assessing the performance of prediction models: A framework for traditional and novel measures. Epidemiology, 21(1), 128–138. https://doi.org/10.1097/EDE.0b013e3181c30fb2

The Cancer Genome Atlas Research Network. (2015). The molecular taxonomy of primary prostate cancer. Cell, 163(4), 1011–1025. https://doi.org/10.1016/j.cell.2015.10.025

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B, 58(1), 267–288.

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. Statistics in Medicine, 16(4), 385–395.

Uccello, M., Vlachostergios, P. J., Ghose, A., & Boussios, S. (2026). Incorporating metastatic disease burden into prognostic assessment in metastatic castration-resistant prostate cancer: The MIRA score. Cancers, 18(16), 2660. https://doi.org/10.3390/cancers18162660

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., & Wei, L. J. (2011). On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. Statistics in Medicine, 30(10), 1105–1117. https://doi.org/10.1002/sim.4154

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics, 7, 91. https://doi.org/10.1186/1471-2105-7-91

Verweij, P. J. M., & Van Houwelingen, H. C. (1994). Penalized likelihood in Cox regression. Statistics in Medicine, 13(23–24), 2427–2436.

Vlachostergios, P. J., Karathanasis, A., & Tzortzis, V. (2022). Expression of fibroblast activation protein is enriched in neuroendocrine prostate cancer and predicts worse survival. Genes, 13(1), 135. https://doi.org/10.3390/genes13010135

Watson, P. A., Arora, V. K., & Sawyers, C. L. (2015). Emerging mechanisms of resistance to androgen receptor inhibitors in prostate cancer. Nature Reviews Cancer, 15(12), 701–711. https://doi.org/10.1038/nrc4016

Wyatt, A. W., Azad, A. A., Volik, S. V., Annala, M., Beja, K., McConeghy, B., … Chi, K. N. (2016). Genomic alterations in cell-free DNA and enzalutamide resistance in castration-resistant prostate cancer. JAMA Oncology, 2(12), 1598–1606. https://doi.org/10.1001/jamaoncol.2016.0494

Zompola, A., Korfiati, A., Theofilatos, K., & Mavroudi, S. (2023). Omics-CNN: A comprehensive pipeline for predictive analytics in quantitative omics using one-dimensional convolutional neural networks. Heliyon, 9(11), e21165. https://doi.org/10.1016/j.heliyon.2023.e21165

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B, 67(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

Παράρτημα Α: Πηγαίος κώδικας Python — βασικές ενότητες

Στο παρόν παράρτημα παρατίθεται ο πλήρης πηγαίος κώδικας των τεσσάρων ενοτήτων που υλοποιούν την ανάλυση του Κεφαλαίου 5 (Πίνακας 16). Ο κώδικας παρατίθεται αυτούσιος, όπως εκτελέστηκε για την παραγωγή των αποτελεσμάτων του Κεφαλαίου 6. Τα σχόλια και οι τεκμηριώσεις των συναρτήσεων είναι στα ελληνικά και εξηγούν τις μεθοδολογικές επιλογές στο σημείο όπου υλοποιούνται. Στα σχόλια η ομόζυγη απαλοιφή αναφέρεται ως «βαθιά διαγραφή» (DEL) και η γονιδιακή ενίσχυση ως «ενίσχυση» (AMP).

Οι ρυθμίσεις της ανάλυσης (config.py), η ανάλυση ανά αλλοίωση (associations.py), τα γραφήματα (plots.py), τα σενάρια εκτέλεσης και τα σημειωματάρια Jupyter συνοδεύουν τον κώδικα του έργου.

Πίνακας 16. Οι ενότητες κώδικα που παρατίθενται στο Παράρτημα Α.

Αρχείο	Περιεχόμενο	Γραμμές
src/data.py	Φόρτωση των αρχείων cBioPortal, επιλογή της κοόρτης, κλινικές μεταβλητές και γονίδια	163
src/models.py	Προεπεξεργασία και οι τέσσερις αλγόριθμοι επιβίωσης	69
src/evaluation.py	Επαναλαμβανόμενη εμφωλευμένη διασταυρωμένη επικύρωση, δείκτες αξιολόγησης και διορθωμένο t-test	241
src/horizons.py	Ανάλυση σε χρονικούς ορίζοντες με επιβίωση με αποκοπή	203

A.1 src/data.py

Φόρτωση δεδομένων, επιλογή κοόρτης και κατασκευή χαρακτηριστικών

```
"""Φόρτωση δεδομένων, επιλογή κοόρτης και κατασκευή χαρακτηριστικών.

Κανένα βήμα εδώ δεν χρησιμοποιεί την έκβαση για να επιλέξει ασθενείς,
δείγματα ή μεταβλητές. Η έκβαση μόνο διαβάζεται και ελέγχεται.
"""
from dataclasses import dataclass

import numpy as np
import pandas as pd
from sksurv.util import Surv

from .config import DATA, MIN_CARRIERS, NONSYNONYMOUS, PANEL, SUPPLEMENT


def read_cbioportal(name):
    """Αρχείο cBioPortal ως κείμενο. Οι γραμμές με # είναι περιγραφές στηλών."""
    return pd.read_csv(DATA / name, sep="\t", comment="#", dtype=str, keep_default_na=False)
```

```
def load_cohort():
    """Οι 128 ασθενείς με έκβαση, ένα δείγμα ο καθένας.

    Όταν ένας ασθενής έχει πολλά δείγματα, κρατάμε αυτό που όρισαν οι
    συγγραφείς στο Dataset S1 (Sheet2). Ελέγχουμε επίσης ότι χρόνος,
    κατάσταση και PSA συμφωνούν με το πρωτογενές συμπλήρωμα.
    """
    patient = read_cbioportal("data_clinical_patient.txt")
    sample = read_cbioportal("data_clinical_sample.txt")
    months = pd.to_numeric(patient.OS_MONTHS, errors="coerce")
    eligible = patient.loc[patient.OS_STATUS.isin(["0:LIVING", "1:DECEASED"]) &
    months.gt(0)].copy()
    eligible["time"] = months[eligible.index]
    eligible["event"] = eligible.OS_STATUS.str.startswith("1").astype(int)

    outcome = pd.read_excel(SUPPLEMENT, sheet_name="Sheet2", engine="openpyxl")
    assert list(outcome.columns[:4]) == ["Subject ID", "Sample ID", "Death status",
                                         "OS from first-line ARSI start"]
    outcome["PATIENT_ID"] = outcome["Subject ID"].astype("int64").astype(str)
    assert set(outcome.PATIENT_ID) == set(eligible.PATIENT_ID), "Διαφορετικοί ασθενείς από το
    Dataset S1"

    cohort = eligible.merge(outcome, on="PATIENT_ID", validate="one_to_one")
    assert np.allclose(cohort.time, cohort["OS from first-line ARSI start"], atol=1e-7)
    assert cohort.event.eq(cohort["Death status"]).all()
    cohort = cohort.rename(columns={"Sample ID": "SAMPLE_ID"})
    cohort = cohort.merge(sample.drop(columns="PATIENT_ID"), on="SAMPLE_ID",
    validate="one_to_one")

    # PSA: το τοπικό πεδίο πρέπει να είναι το «PSA at Dx» του Sheet1.
    source = pd.read_excel(SUPPLEMENT, sheet_name="Sheet1", engine="openpyxl")
    source = source.rename(columns={"Sample ID": "SAMPLE_ID", "PSA at Dx": "PSA_S1"})
    cohort = cohort.merge(source[["SAMPLE_ID", "PSA_S1"]], on="SAMPLE_ID",
    validate="one_to_one")
    local, s1 = pd.to_numeric(cohort.PSA, errors="coerce"), pd.to_numeric(cohort.PSA_S1,
    errors="coerce")
    assert ((local.isna() & s1.isna()) | np.isclose(local, s1, atol=1e-7)).all(), "Το PSA δεν
    συμφωνεί"

    keep = ["PATIENT_ID", "SAMPLE_ID", "time", "event", "AGE_AT_PROCUREMENT",
    "AGE_AT_DIAGNOSIS",
            "GLEASON_SCORE", "PSA", "TISSUE_SITE", "PATHOLOGY_CLASSIFICATION",
            "ABI_ENZA_EXPOSURE_STATUS", "RACE", "TMB_NONSYNONYMOUS"]
    return cohort[keep].sort_values("PATIENT_ID").set_index("PATIENT_ID")



def make_clinical(cohort):
    """Κλινικές μεταβλητές. Τα κενά μένουν κενά· συμπληρώνονται μέσα σε κάθε fold."""
    X = pd.DataFrame(index=cohort.index)
    X["age_biopsy"] = pd.to_numeric(cohort.AGE_AT_PROCUREMENT, errors="coerce")
    X["gleason"] = pd.to_numeric(cohort.GLEASON_SCORE, errors="coerce")
    psa = pd.to_numeric(cohort.PSA, errors="coerce")
    X["log_psa_dx"] = np.log1p(psa.where(psa >= 0))

    # Ένας ασθενής έχει ηλικία βιοψίας μικρότερη από την ηλικία διάγνωσης.
    # Δεν ξέρουμε ποια τιμή είναι λάθος, οπότε θεωρούμε την ηλικία βιοψίας άγνωστη.
    age_dx = pd.to_numeric(cohort.AGE_AT_DIAGNOSIS, errors="coerce")
    X.loc[X.age_biopsy < age_dx, "age_biopsy"] = np.nan

    # Θέση βιοψίας: αναφορά το οστό.
    site = cohort.TISSUE_SITE
    X["site_lymph_node"] = site.eq("LN").astype(float)
    X["site_liver"] = site.eq("Liver").astype(float)
    X["site_other"] = (~site.isin(["Bone", "LN", "Liver"])).astype(float)

    # Ιστολογία: αναφορά το αδενοκαρκίνωμα. Η άγνωστη ιστολογία είναι δική της κατηγορία.
    histology = cohort.PATHOLOGY_CLASSIFICATION
```

```
    X["histology_ne"] = histology.isin(["Small cell", "Adenocarcinoma with NE
    features"]).astype(float)
    X["histology_unknown"] = histology.isin(["Not available", "Inadequate for
    diagnosis"]).astype(float)
    return X



def make_panel(cohort, genes=PANEL):
    """Πάνελ γονιδίων: μετάλλαξη (MUT), ενίσχυση (AMP, CNA=+2), βαθιά διαγραφή (DEL, CNA=−2).


    Επιστρέφει 3 στήλες ανά γονίδιο (60 για το πάνελ). Στήλη χωρίς μέτρηση (π.χ. το
    KMT2D λείπει από τον πίνακα CNA) μένει κενή και αποκλείεται από το φίλτρο συχνότητας.
    Με άλλη λίστα `genes` χρησιμοποιείται και για τα μονοπάτια (src/pathways.py).
    """
    genes = list(genes)
    samples = cohort.SAMPLE_ID
    coverage = read_cbioportal("data_gene_panel_matrix.txt").set_index("SAMPLE_ID")
    assert coverage.loc[samples, ["mutations", "gistic"]].eq("WXS").all().all()


    maf = pd.read_csv(DATA / "data_mutations.txt", sep="\t", low_memory=False,
                      usecols=["Hugo_Symbol", "Tumor_Sample_Barcode",
    "Variant_Classification"])
    maf = maf[maf.Tumor_Sample_Barcode.isin(samples) &
    maf.Variant_Classification.isin(NONSYNONYMOUS)
              & maf.Hugo_Symbol.isin(genes)]
    # Όλα τα δείγματα έχουν WXS, άρα απουσία εγγραφής σημαίνει «χωρίς μετάλλαξη».
    mut = pd.crosstab(maf.Tumor_Sample_Barcode, maf.Hugo_Symbol).clip(upper=1)
    mut = mut.reindex(index=samples, columns=genes, fill_value=0).astype(float)


    wanted = {"Hugo_Symbol", *samples}
    cna = pd.read_csv(DATA / "data_cna.txt", sep="\t", usecols=lambda c: c in wanted)
    cna = cna[cna.Hugo_Symbol.isin(genes)]
    # Διπλό σύμβολο γονιδίου: κρατάμε την τιμή μόνο όπου οι γραμμές συμφωνούν.
    agree = cna.groupby("Hugo_Symbol").nunique(dropna=False).le(1)
    cna = cna.groupby("Hugo_Symbol").first().where(agree)
    cna = cna.reindex(index=genes, columns=samples).T.astype(float)
    assert set(np.unique(cna.stack().dropna())) <= {-2, -1, 0, 1, 2}


    panel = {}
    for gene in genes:
        panel[f"MUT_{gene}"] = mut[gene].to_numpy()
        panel[f"AMP_{gene}"] =
    cna[gene].eq(2).astype(float).where(cna[gene].notna()).to_numpy()
        panel[f"DEL_{gene}"] = cna[gene].eq(-
    2).astype(float).where(cna[gene].notna()).to_numpy()
    return pd.DataFrame(panel, index=cohort.index)



def panel_counts(panel):
    """Φορείς ανά αλλοίωση και αν περνά το φίλτρο συχνότητας."""
    counts = pd.DataFrame({"carriers": panel.eq(1).sum().astype(int),
                           "measured": panel.notna().sum().astype(int)})
    counts["non_carriers"] = counts.measured - counts.carriers
    counts["kept"] = counts.carriers.ge(MIN_CARRIERS) & counts.non_carriers.ge(MIN_CARRIERS)
    counts.insert(0, "gene", [c.split("_", 1)[1] for c in counts.index])
    counts.insert(1, "type", [c.split("_", 1)[0] for c in counts.index])
    return counts



@dataclass
class Dataset:
    cohort: pd.DataFrame
    clinical: pd.DataFrame
    panel_all: pd.DataFrame
    counts: pd.DataFrame
    feature_sets: dict
    y: np.ndarray
```

```
def load_dataset():
    """Όλα όσα χρειάζονται τα notebooks, με ίδια σειρά ασθενών παντού."""
    cohort = load_cohort()
    clinical = make_clinical(cohort)
    panel_all = make_panel(cohort)
    counts = panel_counts(panel_all)
    panel = panel_all[counts.index[counts.kept]]

    basic = ["age_biopsy", "gleason"]
    feature_sets = {
        "clinical_basic": clinical[basic],
        "clinical": clinical,
        "genomic": panel,
        "clinical_genomic": pd.concat([clinical, panel], axis=1),
    }
    y = Surv.from_arrays(event=cohort.event.astype(bool), time=cohort.time.astype(float))

    assert len(cohort) == 128 and cohort.event.sum() == 76
    for X in feature_sets.values():
        assert X.index.equals(cohort.index)
    return Dataset(cohort, clinical, panel_all, counts, feature_sets, y)
```

A.2 src/models.py

Προεπεξεργασία και οι τέσσερις αλγόριθμοι επιβίωσης

```
"""Προεπεξεργασία και οι τέσσερις αλγόριθμοι επιβίωσης.

Κάθε μοντέλο είναι Pipeline(προεπεξεργασία → εκτιμητής), ώστε διάμεσοι,
μέσοι όροι και τυπικές αποκλίσεις να υπολογίζονται μόνο από τους ασθενείς
εκπαίδευσης του κάθε fold.
"""
import numpy as np
import pandas as pd
from sklearn.base import BaseEstimator, TransformerMixin
from sklearn.pipeline import Pipeline
from sksurv.ensemble import GradientBoostingSurvivalAnalysis, RandomSurvivalForest
from sksurv.linear_model import CoxnetSurvivalAnalysis, CoxPHSurvivalAnalysis

from .config import COX_RIDGE_ALPHA, GRIDS, RSF_TREES



class Preprocessor(BaseEstimator, TransformerMixin):
    """Συμπλήρωση με διάμεσο, δείκτες έλλειψης, αφαίρεση σταθερών, τυποποίηση.

    Δείκτης έλλειψης προστίθεται μόνο για στήλες με κενά στο training set,
    ώστε το μοντέλο να μπορεί να μάθει αν η ίδια η έλλειψη σχετίζεται με τον κίνδυνο.
    """


    def fit(self, X, y=None):
        X = pd.DataFrame(X)
        self.medians_ = X.median()
        self.missing_columns_ = [c for c in X.columns if X[c].isna().any()]
        Z = self._fill(X)
        std = Z.std(ddof=0)
        self.columns_ = list(std.index[std > 1e-12])
        self.mean_ = Z[self.columns_].mean()
        self.scale_ = std[self.columns_]
        return self


    def _fill(self, X):
        Z = X.fillna(self.medians_)
        for c in self.missing_columns_:
            Z[f"{c}_missing"] = X[c].isna().astype(float)
        return Z


    def transform(self, X):
```

```
        Z = self._fill(pd.DataFrame(X))[self.columns_]
        return ((Z - self.mean_) / self.scale_).to_numpy(dtype=float)

    def get_feature_names_out(self, input_features=None):
        return np.array(self.columns_, dtype=object)


def make_model(name, seed):
    """Επιστρέφει (pipeline, πλέγμα υπερπαραμέτρων). Κενό πλέγμα = χωρίς εσωτερικό CV."""
    if name == "cox":
        model = CoxPHSurvivalAnalysis(alpha=COX_RIDGE_ALPHA, ties="efron", n_iter=300)
    elif name == "coxnet":
        model = CoxnetSurvivalAnalysis(fit_baseline_model=True, max_iter=100_000)
    elif name == "rsf":
        model = RandomSurvivalForest(n_estimators=RSF_TREES, min_samples_split=10,
                                     random_state=seed, n_jobs=1)
    elif name == "gbs":
        model = GradientBoostingSurvivalAnalysis(loss="coxph", subsample=0.8,
    random_state=seed)
    else:
        raise ValueError(name)
    grid = {f"model__{k}": v for k, v in GRIDS.get(name, {}).items()}
    return Pipeline([("prep", Preprocessor()), ("model", model)]), grid


def survival_at(pipeline, X, times):
    """Πιθανότητα επιβίωσης S(t) κάθε ασθενούς στους χρόνους times (n × len(times))."""
    functions = pipeline.predict_survival_function(X)
    return np.vstack([[fn(t) for t in times] for fn in functions])
```

A.3 src/evaluation.py

Επαναλαμβανόμενη εμφωλευμένη διασταυρωμένη επικύρωση και δείκτες αξιολόγησης

```
"""Επαναλαμβανόμενο nested cross-validation και δείκτες αξιολόγησης.

Εξωτερικός βρόχος: 5 folds × 10 επαναλήψεις, διαστρωμάτωση ως προς θάνατο/λογοκρισία.
Κάθε ασθενής αξιολογείται μία φορά ανά επανάληψη από μοντέλο που δεν τον είδε.
Εσωτερικός βρόχος: 3 folds × 3 επαναλήψεις μόνο μέσα στο training set, για τις
    υπερπαραμέτρους.
Όλοι οι αλγόριθμοι και όλα τα σύνολα εισόδων χρησιμοποιούν τους ίδιους διαχωρισμούς.
"""
import json
import time
import warnings
from datetime import datetime, timezone

import numpy as np
import pandas as pd
from joblib import Parallel, delayed
from scipy import stats
from sklearn.inspection import permutation_importance
from sklearn.model_selection import GridSearchCV, RepeatedStratifiedKFold
from sksurv.metrics import (concordance_index_censored, concordance_index_ipcw,
                            cumulative_dynamic_auc, integrated_brier_score)
from sksurv.nonparametric import kaplan_meier_estimator

from .config import (AUC_TIMES, IBS_GRID, INNER_FOLDS, INNER_REPEATS, OUTER_FOLDS, REPEATS,
                     RESULTS, SEED, TABLES, TAU)
from .models import make_model, survival_at

METRICS = ["harrell", "uno", "auc_12", "auc_24", "ibs"]
IMPORTANCE_SET = "clinical_genomic"
```

```
def outer_splits(y):
    """Οι κοινές διαιρέσεις (επανάληψη, fold, train, test) για όλα τα μοντέλα."""
    cv = RepeatedStratifiedKFold(n_splits=OUTER_FOLDS, n_repeats=REPEATS, random_state=SEED)
    for i, (train, test) in enumerate(cv.split(np.zeros(len(y)), y["event"])):
        yield i // OUTER_FOLDS, i % OUTER_FOLDS, train, test



def ibs_times(y_train, y_test):
    """Χρόνοι για το Brier score: μέσα στο IBS_GRID και στο εύρος παρακολούθησης του test."""
    grid = np.arange(IBS_GRID[0], IBS_GRID[1] + 1e-9, 1.0)
    upper = min(y_test["time"].max(), y_train["time"].max())
    return grid[(grid >= y_test["time"].min()) & (grid < upper)]



def truncate(y, horizon=IBS_GRID[1] + 1.0):
    """Διοικητική λογοκρισία στον ορίζοντα, μόνο για AUC/IBS.

    Τα βάρη IPCW χρειάζονται την κατανομή λογοκρισίας του training στον χρόνο κάθε
    ασθενούς test· πέρα από το μέγιστο follow-up του training δεν ορίζεται. Ένας
    ασθενής ζωντανός μετά τους 31 μήνες συνεισφέρει ακριβώς το ίδιο στους δείκτες
    έως 30 μήνες, είτε λογοκριθεί στους 31 είτε όχι.
    """
    y = y.copy()
    late = y["time"] > horizon
    y["time"][late] = horizon
    y["event"][late] = False
    return y



def score_fold(y_train, y_test, risk, surv_fn):
    """Όλοι οι δείκτες σε ένα test fold. surv_fn(times) → S(t) πίνακας n × len(times)."""
    out = {"harrell": concordance_index_censored(y_test["event"], y_test["time"], risk)[0]}
    out["uno"] = concordance_index_ipcw(y_train, y_test, risk, tau=TAU)[0]
    y_test = truncate(y_test)
    # AUC στον χρόνο t με κίνδυνο 1 − S(t): ίδια κατάταξη με το risk για τα Cox μοντέλα.
    for t in AUC_TIMES:
        try:
            out[f"auc_{int(t)}"] = cumulative_dynamic_auc(y_train, y_test, 1 - surv_fn([t])[:,
    0], [t])[0][0]
        except ValueError:
            out[f"auc_{int(t)}"] = np.nan
    times = ibs_times(y_train, y_test)
    out["ibs"] = integrated_brier_score(y_train, y_test, surv_fn(times), times)
    return out



def one_se_rule(cv_results):
    """Κανόνας 1-SE (όπως το lambda.1se του glmnet) για το Elastic Net.

    Από τις ρυθμίσεις με μέσο C-index μέσα σε ένα τυπικό σφάλμα από την καλύτερη,
    κρατά την ισχυρότερη ποινή (μεγαλύτερο α). Η διαφορά τους δεν ξεχωρίζει από τον
    θόρυβο, οπότε προτιμάμε το απλούστερο μοντέλο.
    """
    mean = np.asarray(cv_results["mean_test_score"], dtype=float)
    se = np.asarray(cv_results["std_test_score"], dtype=float) / np.sqrt(INNER_FOLDS *
    INNER_REPEATS)
    best = np.nanargmax(mean)
    ok = np.flatnonzero(mean >= mean[best] - se[best])
    alpha = np.array([a[0] for a in cv_results["param_model__alphas"]])
    strongest = ok[alpha[ok] == alpha[ok].max()]
    return int(strongest[np.argmax(mean[strongest])])



def run_fold(X, y, set_name, model_name, repeat, fold, train, test, with_importance):
    """Εκπαίδευση και αξιολόγηση ενός (συνόλου, αλγορίθμου) σε ένα εξωτερικό fold."""
    seed = SEED + 1000 * repeat + fold
    X_train, X_test, y_train, y_test = X.iloc[train], X.iloc[test], y[train], y[test]
```

```
    pipeline, grid = make_model(model_name, seed)
    record = {"feature_set": set_name, "model": model_name, "repeat": repeat, "fold": fold}
    started = time.perf_counter()
    with warnings.catch_warnings(record=True) as caught:
        warnings.simplefilter("always")
        if grid:
            inner = RepeatedStratifiedKFold(n_splits=INNER_FOLDS, n_repeats=INNER_REPEATS,
    random_state=seed)
            search = GridSearchCV(pipeline, grid, cv=list(inner.split(X_train,
    y_train["event"])),
                                  refit=one_se_rule if model_name == "coxnet" else True,
                                  error_score=np.nan, n_jobs=1)
            search.fit(X_train, y_train)
            pipeline = search.best_estimator_
            record["params"] = json.dumps({k.replace("model__", ""): v for k, v in
    search.best_params_.items()})
            record["inner_c"] = search.cv_results_["mean_test_score"][search.best_index_]
            record["failed_configs"] =
    int(np.isnan(search.cv_results_["mean_test_score"]).sum())
        else:
            pipeline.fit(X_train, y_train)
            record["params"] = "{}"
    record["warnings"] = len(caught)
    record["seconds"] = time.perf_counter() - started


    risk = pipeline.predict(X_test)
    record.update(score_fold(y_train, y_test, risk, lambda t: survival_at(pipeline, X_test,
    t)))
    model = pipeline.named_steps["model"]
    names = pipeline.named_steps["prep"].get_feature_names_out()
    record["n_features"] = len(names)
    if model_name in ("cox", "coxnet"):
        coef = np.ravel(model.coef_)
        record["n_nonzero"] = int(np.sum(np.abs(coef) > 1e-10))

    predictions = pd.DataFrame({"feature_set": set_name, "model": model_name, "repeat":
    repeat,
                                "fold": fold, "PATIENT_ID": X_test.index, "risk": risk,
                                "surv_24": survival_at(pipeline, X_test, [TAU])[:, 0]})
    importance = []
    if with_importance:
        if model_name in ("cox", "coxnet"):
            importance += [{"feature": n, "kind": "coef", "value": c} for n, c in zip(names,
    coef)]
        perm = permutation_importance(pipeline, X_test, y_test, n_repeats=5,
    random_state=seed)
        importance += [{"feature": n, "kind": "permutation", "value": v}
                       for n, v in zip(X.columns, perm.importances_mean)]
        importance = [dict(i, feature_set=set_name, model=model_name, repeat=repeat,
    fold=fold)
                      for i in importance]
    return record, predictions, importance



def km_reference(y):
    """IBS του Kaplan–Meier (ίδια καμπύλη για όλους): σημείο αναφοράς «χωρίς μεταβλητές»."""
    rows = []
    for repeat, fold, train, test in outer_splits(y):
        y_test = truncate(y[test])
        times_km, surv_km = kaplan_meier_estimator(y[train]["event"], y[train]["time"])
        times = ibs_times(y[train], y_test)
        idx = np.searchsorted(times_km, times, side="right") - 1  # βηματική συνάρτηση
        s = np.where(idx >= 0, surv_km[np.maximum(idx, 0)], 1.0)
        rows.append({"repeat": repeat, "fold": fold, "ibs": integrated_brier_score(
            y[train], y_test, np.tile(s, (len(test), 1)), times)})
    return pd.DataFrame(rows)



def run_cv(dataset, models, n_jobs=-1, verbose=5):
    """Τρέχει όλους τους συνδυασμούς και αποθηκεύει τα αποτελέσματα στο results/tables/."""
    splits = list(outer_splits(dataset.y))
```

```
    tasks = [(s, m, *split) for s in dataset.feature_sets for m in models for split in splits]
    started = time.perf_counter()
    output = Parallel(n_jobs=n_jobs, verbose=verbose)(
        delayed(run_fold)(dataset.feature_sets[s], dataset.y, s, m, r, f, tr, te, s ==
    IMPORTANCE_SET)
        for s, m, r, f, tr, te in tasks)
    metrics = pd.DataFrame([o[0] for o in output])
    predictions = pd.concat([o[1] for o in output], ignore_index=True)
    importance = pd.DataFrame([i for o in output for i in o[2]])

    TABLES.mkdir(parents=True, exist_ok=True)
    metrics.to_csv(TABLES / "cv_metrics.csv", index=False)
    predictions.to_csv(TABLES / "cv_predictions.csv", index=False)
    importance.to_csv(TABLES / "cv_importance.csv", index=False)
    km_reference(dataset.y).to_csv(TABLES / "cv_km_reference.csv", index=False)
    (RESULTS / "run_info.json").write_text(json.dumps({
        "finished_utc": datetime.now(timezone.utc).isoformat(timespec="seconds"),
        "minutes": round((time.perf_counter() - started) / 60, 1),
        "seed": SEED, "outer_folds": OUTER_FOLDS, "repeats": REPEATS, "inner_folds":
    INNER_FOLDS,
        "inner_repeats": INNER_REPEATS, "coxnet_selection": "1-SE rule",
        "fits": len(tasks), "models": models, "feature_sets": list(dataset.feature_sets),
    }, indent=2, ensure_ascii=False), encoding="utf-8")
    return load_cv()



def load_cv():
    """Φορτώνει τα αποθηκευμένα αποτελέσματα του run_cv."""
    return {name: pd.read_csv(TABLES / f"cv_{name}.csv")
            for name in ["metrics", "predictions", "importance", "km_reference"]}



def corrected_ttest(differences, n_folds=OUTER_FOLDS):
    """Διορθωμένο t-test επαναλαμβανόμενου CV (Nadeau & Bengio, 2003).

    Τα folds μοιράζονται ασθενείς εκπαίδευσης, άρα δεν είναι ανεξάρτητα. Το απλό t-test
    υποεκτιμά τη διασπορά· η διόρθωση προσθέτει τον όρο n_test / n_train = 1 / (K − 1).
    Επιστρέφει μέσο, 95% διάστημα εμπιστοσύνης και p-value.
    """
    d = np.asarray(differences, dtype=float)
    d = d[np.isfinite(d)]
    k = len(d)
    variance = (1 / k + 1 / (n_folds - 1)) * d.var(ddof=1)
    se = np.sqrt(variance)
    t_crit = stats.t.ppf(0.975, k - 1)
    p = 2 * stats.t.sf(abs(d.mean() / se), k - 1) if se > 0 else np.nan
    return d.mean(), d.mean() - t_crit * se, d.mean() + t_crit * se, p



def summarize(metrics):
    """Μέσος όρος 50 folds, διορθωμένο 95% ΔΕ και SD μεταξύ επαναλήψεων για κάθε δείκτη."""
    rows = []
    for (s, m), g in metrics.groupby(["feature_set", "model"], sort=False):
        row = {"feature_set": s, "model": m}
        for metric in METRICS:
            mean, low, high, _ = corrected_ttest(g[metric])
            row[metric] = mean
            row[f"{metric}_low"], row[f"{metric}_high"] = low, high
            row[f"{metric}_sd_repeats"] = g.groupby("repeat")[metric].mean().std()
        rows.append(row)
    return pd.DataFrame(rows)



def paired_differences(metrics, comparisons):
    """Διαφορά δύο συνόλων εισόδων στον ίδιο αλγόριθμο και στο ίδιο fold.

    comparisons: λίστα (νέο, αναφορά). Για C-index/AUC θετική διαφορά = βελτίωση,
    για IBS αρνητική διαφορά = βελτίωση.
    """
```

```
    wide = metrics.set_index(["feature_set", "model", "repeat", "fold"]).sort_index()
    rows = []
    for new, ref in comparisons:
        for model in metrics.model.unique():
            for metric in METRICS:
                try:
                    d = wide.loc[(new, model), metric] - wide.loc[(ref, model), metric]
                except KeyError:
                    continue
                mean, low, high, p = corrected_ttest(d)
                better = d.groupby(level="repeat").mean()
                better = (better > 0) if metric != "ibs" else (better < 0)
                rows.append({"comparison": f"{new} − {ref}", "new": new, "reference": ref,
                             "model": model, "metric": metric, "difference": mean,
                             "low": low, "high": high, "p": p,
                             "repeats_improved": f"{int(better.sum())}/{len(better)}"})
    return pd.DataFrame(rows)
```

A.4 src/horizons.py

Ανάλυση σε χρονικούς ορίζοντες

```
"""Ανάλυση σε χρονικούς ορίζοντες: επιβίωση με αποκοπή στους 12, 24 και 36 μήνες.

Για κάθε ορίζοντα h:
  • Αποκοπή: όποιος ζει πέρα από τον h λογοκρίνεται στο h (ελάχιστα μετά, h + 0,001 μήνες,
    ώστε οι δείκτες να ορίζονται ακριβώς στο h). Θάνατοι μετά τον h δεν μετράνε ως θάνατοι.
    Μένουν και οι 128 ασθενείς· όσοι λογοκρίθηκαν πριν τον h κρατούν τον χρόνο τους.
  • Τα μοντέλα εκπαιδεύονται στα αποκομμένα δεδομένα: μαθαίνουν τον κίνδυνο έως τον h.
  • Ίδιο nested CV με την κύρια ανάλυση (εξωτερικά 5 × 10, εσωτερικά 3 × 3, 1-SE για Elastic
    Net).
    Η διαστρωμάτωση γίνεται ως προς τον θάνατο έως τον h, ώστε κάθε fold να έχει θανάτους.
  • Δείκτες: Harrell C-index έως τον h, Uno C-index (τ = h), AUC στον h, Brier score στον h.
    Η AUC υπολογίζεται μόνο σε folds με ≥ 3 θανάτους έως τον h και ≥ 3 ζωντανούς μετά.

Τα αποτελέσματα γράφονται στο results/horizons/ και δεν επηρεάζουν την κύρια ανάλυση.
"""
import json
import time
import warnings
from datetime import datetime, timezone


import numpy as np
import pandas as pd
from joblib import Parallel, delayed
from sklearn.model_selection import GridSearchCV, RepeatedStratifiedKFold
from sksurv.metrics import brier_score, concordance_index_censored, concordance_index_ipcw,
    cumulative_dynamic_auc
from sksurv.nonparametric import kaplan_meier_estimator


from .config import HORIZONS, INNER_FOLDS, INNER_REPEATS, OUTER_FOLDS, REPEATS, RESULTS, SEED
from .evaluation import corrected_ttest, one_se_rule
from .models import make_model, survival_at


OUT = RESULTS / "horizons"
TABLES_H = OUT / "tables"
FIGURES_H = OUT / "figures"
OUT_PANEL = RESULTS / "horizons_panel"  # notebook 05: τα 4 σύνολα της αρχικής ανάλυσης ανά
    ορίζοντα
DELTA = 1e-3
METRICS_H = ["harrell", "uno", "auc", "brier"]
MIN_AUC_GROUP = 3



def truncate_at(y, h):
    """Επιβίωση με αποκοπή στον ορίζοντα h."""
    y = y.copy()
    late = y["time"] > h
```

```
    y["time"][late] = h + DELTA
    y["event"][late] = False
    return y


def horizon_counts(y):
    """Θάνατοι έως τον h, ζωντανοί πέρα από τον h, λογοκριμένοι πριν τον h."""
    rows = []
    for h in HORIZONS:
        rows.append({"horizon": h, "deaths_by_h": int(((y["time"] <= h) & y["event"]).sum()),
                     "alive_after_h": int((y["time"] > h).sum()),
                     "censored_before_h": int(((y["time"] <= h) & ~y["event"]).sum())})
    return pd.DataFrame(rows)


def splits(y_h):
    cv = RepeatedStratifiedKFold(n_splits=OUTER_FOLDS, n_repeats=REPEATS, random_state=SEED)
    for i, (train, test) in enumerate(cv.split(np.zeros(len(y_h)), y_h["event"])):
        yield i // OUTER_FOLDS, i % OUTER_FOLDS, train, test


def _safe(f):
    """Ένας δείκτης που δεν ορίζεται σε ένα fold (π.χ. κανείς ζωντανός πέρα από τον h) γίνεται
    κενός."""
    try:
        return float(f())
    except (ValueError, ZeroDivisionError):
        return np.nan


def score(y_train, y_test, risk, surv_h, h):
    # Η AUC στον h συγκρίνει όσους πέθαναν έως τον h με όσους ζουν μετά. Με 1–2 ασθενείς στη
    μία
    # ομάδα η τιμή είναι τυχαία (π.χ. 0,06), οπότε απαιτούμε τουλάχιστον MIN_AUC_GROUP σε
    καθεμία.
    enough = min(int(y_test["event"].sum()), int((y_test["time"] > h).sum())) >= MIN_AUC_GROUP
    return {
        "harrell": _safe(lambda: concordance_index_censored(y_test["event"], y_test["time"],
    risk)[0]),
        "uno": _safe(lambda: concordance_index_ipcw(y_train, y_test, risk, tau=h)[0]),
        "auc": _safe(lambda: cumulative_dynamic_auc(y_train, y_test, 1 - surv_h, [h])[0][0])
    if enough else np.nan,
        "brier": _safe(lambda: brier_score(y_train, y_test, surv_h.reshape(-1, 1),
    [h])[1][0]),
    }


def run_fold(X, y_h, h, set_name, model_name, repeat, fold, train, test):
    seed = SEED + 1000 * repeat + fold
    X_train, X_test, y_train, y_test = X.iloc[train], X.iloc[test], y_h[train], y_h[test]
    pipeline, grid = make_model(model_name, seed)
    record = {"horizon": h, "feature_set": set_name, "model": model_name, "repeat": repeat,
    "fold": fold,
              "test_deaths": int(y_test["event"].sum()), "test_alive_after_h":
    int((y_test["time"] > h).sum())}
    started = time.perf_counter()
    try:
        with warnings.catch_warnings(record=True) as caught:
            warnings.simplefilter("always")
            if grid:
                inner = RepeatedStratifiedKFold(n_splits=INNER_FOLDS, n_repeats=INNER_REPEATS,
    random_state=seed)
                search = GridSearchCV(pipeline, grid, cv=list(inner.split(X_train,
    y_train["event"])),
                                      refit=one_se_rule if model_name == "coxnet" else True,
                                      error_score=np.nan, n_jobs=1)
                search.fit(X_train, y_train)
                pipeline = search.best_estimator_
                record["params"] = json.dumps({k.replace("model__", ""): v for k, v in
    search.best_params_.items()})
```

```
                record["failed_configs"] =
    int(np.isnan(search.cv_results_["mean_test_score"]).sum())
            else:
                pipeline.fit(X_train, y_train)
                record["params"] = "{}"
        record["warnings"] = len(caught)
        risk = pipeline.predict(X_test)
        surv_h = survival_at(pipeline, X_test, [h])[:, 0]
        record.update(score(y_train, y_test, risk, surv_h, h))
        record["error"] = ""
    except Exception as e:  # noqa: BLE001  ένα fold που αποτυγχάνει καταγράφεται, δεν σταματά
    την ανάλυση
        record.update({m: np.nan for m in METRICS_H})
        record["error"] = f"{type(e).__name__}: {str(e)[:150]}"
    record["seconds"] = time.perf_counter() - started
    return record



def km_reference(y):
    """Brier score του Kaplan–Meier στον h (ίδια καμπύλη για όλους) ανά fold."""
    rows = []
    for h in HORIZONS:
        y_h = truncate_at(y, h)
        for repeat, fold, train, test in splits(y_h):
            t, s = kaplan_meier_estimator(y_h[train]["event"], y_h[train]["time"])
            s_h = s[np.searchsorted(t, h, side="right") - 1] if (t <= h).any() else 1.0
            rows.append({"horizon": h, "repeat": repeat, "fold": fold,
                         "brier": _safe(lambda: brier_score(y_h[train], y_h[test],
                                                            np.full((len(test), 1), s_h),
    [h])[1][0])})
    return pd.DataFrame(rows)



def run_horizons(feature_sets, y, models, n_jobs=-1, verbose=5, out=OUT):
    """Όλοι οι συνδυασμοί ορίζοντα × συνόλου × αλγορίθμου × fold. Αποθήκευση στο out/tables/.

    out: results/horizons/ για το notebook 04, results/horizons_panel/ για το notebook 05.
    """
    tasks = []
    for h in HORIZONS:
        y_h = truncate_at(y, h)
        for s in feature_sets:
            for m in models:
                tasks += [(feature_sets[s], y_h, h, s, m, *split) for split in splits(y_h)]
    started = time.perf_counter()
    records = Parallel(n_jobs=n_jobs, verbose=verbose)(delayed(run_fold)(*t) for t in tasks)
    tables = out / "tables"
    tables.mkdir(parents=True, exist_ok=True)
    metrics = pd.DataFrame(records)
    metrics.to_csv(tables / "metrics.csv", index=False)
    km_reference(y).to_csv(tables / "km_reference.csv", index=False)
    (out / "run_info.json").write_text(json.dumps({
        "finished_utc": datetime.now(timezone.utc).isoformat(timespec="seconds"),
        "minutes": round((time.perf_counter() - started) / 60, 1), "seed": SEED, "horizons":
    HORIZONS,
        "feature_sets": list(feature_sets), "models": models, "fits": len(tasks),
        "failed_fits": int(metrics.error.ne("").sum()),
    }, indent=2, ensure_ascii=False), encoding="utf-8")
    return load(out)



def load(out=OUT):
    tables = out / "tables"
    return {"metrics": pd.read_csv(tables / "metrics.csv",
    keep_default_na=True).fillna({"error": ""}),
            "km_reference": pd.read_csv(tables / "km_reference.csv")}



def truncated_outcome(cohort, h):
```

```
    """Χρόνος/συμβάν με αποκοπή στον h σε μορφή πίνακα, για τις αναλύσεις ανά γονίδιο
    (lifelines)."""
    late = cohort.time > h
    return pd.DataFrame({"time": cohort.time.where(~late, h + DELTA),
                         "event": cohort.event.where(~late, 0).astype(int)},
    index=cohort.index)



def summarize(metrics):
    """Ανά ορίζοντα × σύνολο × αλγόριθμο: μέσος όρος, διορθωμένο 95% ΔΕ και πόσα folds είχαν
    τιμή."""
    rows = []
    for (h, s, m), g in metrics.groupby(["horizon", "feature_set", "model"], sort=False):
        row = {"horizon": h, "feature_set": s, "model": m}
        for metric in METRICS_H:
            values = g[metric].dropna()
            row[f"{metric}_folds"] = len(values)
            if len(values) >= 2:
                row[metric], row[f"{metric}_low"], row[f"{metric}_high"], _ =
    corrected_ttest(values)
            else:
                row[metric] = row[f"{metric}_low"] = row[f"{metric}_high"] = np.nan
        rows.append(row)
    return pd.DataFrame(rows)



def paired(metrics, comparisons):
    """Διαφορά δύο συνόλων στον ίδιο ορίζοντα, αλγόριθμο και fold, με διορθωμένο 95% ΔΕ."""
    wide = metrics.set_index(["horizon", "feature_set", "model", "repeat",
    "fold"]).sort_index()
    rows = []
    for h in metrics.horizon.unique():
        for new, ref in comparisons:
            for model in metrics.model.unique():
                for metric in METRICS_H:
                    d = (wide.loc[(h, new, model), metric] - wide.loc[(h, ref, model),
    metric]).dropna()
                    if len(d) < 2:
                        continue
                    mean, low, high, p = corrected_ttest(d)
                    rows.append({"horizon": h, "new": new, "reference": ref, "model": model,
    "metric": metric,
                                 "difference": mean, "low": low, "high": high, "p": p,
    "folds": len(d)})
    return pd.DataFrame(rows)
```

Σημείωμα αναπαραγωγιμότητας

Ο κώδικας του Παραρτήματος Α ανέρχεται σε 676 γραμμές Python. Ο πλήρης κώδικας του έργου περιλαμβάνει επιπλέον τις ρυθμίσεις της ανάλυσης, την ανάλυση ανά αλλοίωση, τα γραφήματα, τα σενάρια εκτέλεσης και τα σημειωματάρια Jupyter 01, 02 και 05, από τα οποία προέρχονται τα αποτελέσματα του Κεφαλαίου 6.

Ο σπόρος τυχαιότητας (2026) είναι καθορισμένος σε όλα τα στάδια, ώστε οι εκτελέσεις να είναι επαναλήψιμες, και οι εκδόσεις των βιβλιοθηκών είναι καθορισμένες στο αρχείο requirements.txt. Τα σημειωματάρια φορτώνουν τα αποθηκευμένα αποτελέσματα. Με τα σενάρια run_analysis.py και run_horizons_panel.py η ανάλυση εκτελείται ξανά από την αρχή. Κάθε εκτέλεση καταγράφει ημερομηνία, διάρκεια, σπόρο και ρυθμίσεις στο αρχείο run_info.json. Τα σχήματα του Κεφαλαίου 6 σχεδιάστηκαν από τα αποθηκευμένα αποτελέσματα με τις συναρτήσεις του plots.py.

Τα δεδομένα είναι δημόσια διαθέσιμα στην πλατφόρμα cBioPortal (μελέτη prad_su2c_2019), και το Dataset S1 στο συμπληρωματικό υλικό της δημοσίευσης των Abida et al. (2019).