ΠΜΣ ΒΙΟΣΤΑΤΙΣΤΙΚΗ

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΙΑΤΡΙΚΗ ΣΧΟΛΗ

ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΙΩΑΝΝΗΣ ΞΑΝΘΑΚΗΣ

«Αξιολόγηση της προγνωστικής ακρίβειας της κλίμακας εκτίμησης κινδύνου SAPS II για την πρόβλεψη της θνησιμότητας ασθενών μονάδων εντατικής θεραπείας: Συστηματική ανασκόπηση και μετα-ανάλυση»

ΑΘΗΝΑ,2023

Η παρούσα διπλωματική εργασία εκπονήθηκε στο πλαίσιο των σπουδών για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στη

ΒΙΟΣΤΑΤΙΣΤΙΚΗ

που απονέμει η Ιατρική Σχολή και το τμήμα Μαθηματικών του Εθνικού & Καποδιστριακού Πανεπιστημίου Αθηνών.

Εγκρίθηκε την ***6/2/2023*** από την εξεταστική επιτροπή:

ΟΝΟΜΑΤΕΠΩΝΥΜΟ	***ΒΑΘΜΙΔΑ***	***ΥΠΟΓΡΑΦΗ***
1. Ε. ΚΡΙΤΣΩΤΑΚΗΣ	**ΑΝΑΠΛ. ΚΑΘΗΓΗΤΗΣ**	
2. Ε. ΣΑΜΟΛΗ	**ΑΝΑΠΛ. ΚΑΘΗΓΗΤΡΙΑ**	EVANGELIA SAMOLI 24.02.2023 14:00
3. **Μ. ΚΑΣΔΑΓΛΗ**	**ΕΠΙΣΤΗΜ. ΣΥΝΕΡΓΑΤΙΔΑ**	 …….

Ευχαριστίες

Θα ήθελα να πω ένα μεγάλο ευχαριστώ στον επιβλέποντα καθηγητή μου, κύριο Ευάγγελο Κριτσωτάκη, για την άριστη επικοινωνία, την εξαιρετική συνεργασία και τις πολύτιμες συμβουλές του καθ' όλη τη διάρκεια εκπόνησης της διπλωματικής μου εργασίας. Θα ήθελα επιπλέον να ευχαριστήσω και τα υπόλοιπα μέλη της Επιτροπής, την αναπληρώτρια καθηγήτρια κυρία Ευαγγελία Σαμόλη και την επιστημονική συνεργάτιδα κυρία Μαρία Κάσδαγλη για τα χρήσιμα σχόλια τους και τη συμβολή τους στην ολοκλήρωση της εργασίας. Τέλος, ένα μεγάλο ευχαριστώ στην οικογένεια και τους φίλους μου που είναι πάντοτε δίπλα μου και με στηρίζουν.

Περιεχόμενα

Κατάλογος Πινάκων

Κατάλογος Γραφημάτων

ΚΕΦΑΛΑΙΟ 1
ΕΙΣΑΓΩΓΗ

1.1 Θεωρητικό πλαίσιο

Η συνεχόμενη εξέλιξη της ιατρικής και η πολυπλοκότητα των χειρουργικών επεμβάσεων έχουν οδηγήσει στη χρήση των Μονάδων Εντατικής Θεραπείας (ICU) ως ένα βασικό μέσο μετεγχειρητικής φροντίδας των ασθενών, με στόχο τη μείωση των μετεγχειρητικών επιπλοκών και του χρόνου νοσηλείας (Ghaffar, Pearse and Gillies, 2017). Παρ' όλα αυτά, η εισαγωγή σε ICU δεν ενδείκνυται για κάθε ασθενή και συχνά εγκυμονεί κινδύνους και ενδέχεται ορισμένες φορές να είναι περισσότερο βλαπτική παρά ωφέλιμη (Niederman and Berger, 2010; Ghaffar, Pearse and Gillies, 2017). Για το λόγο αυτό , κλινικά μοντέλα πρόβλεψης και σκορ σοβαρότητας της νόσου γίνονται όλο και περισσότερο διαδεδομένα στην κλινική πράξη. Τα προβλεπτικά αυτά μοντέλα χρησιμοποιούν ατομικά δεδομένα, όπως η ηλικία, το ιατρικό ιστορικό και διάφορες κλινικές και βιοχημικές μετρήσεις για να υπολογίσουν τον εξατομικευμένο κίνδυνο εκδήλωσης μιας συγκεκριμένης έκβασης υγείας. Ο εκτιμώμενος κίνδυνος, δηλαδή η πιθανότητα να συμβεί η συγκεκριμένη έκβαση εντός ορισμένου χρονικού διαστήματος, μπορεί να χρησιμοποιηθεί ως μέτρο για τη σοβαρότητα της νόσου ,ως κριτήριο για τη λήψη αποφάσεων θεραπείας και διαχείρισης των ασθενών αλλά και για την ενημέρωση των ασθενών των οικογενειών τους για τις πιθανές εκβάσεις (Damen *et al.*, 2022).

Πολλές κλίμακες εκτίμησης κινδύνου (κλίμακες και βαθμολογίες βαρύτητας) έχουν προταθεί για τη γρήγορη και ακριβή ταυτοποίηση των βαρέως πασχόντων ασθενών – μερικά από τα πιο γνωστά συστήματα περιλαμβάνουν τα Acute Physiology and Chronic Health Disease Classification System (APACHE) I-IV, Sequential Organ Failure Assessment (SOFA), Simplified Acute Physiology Score (SAPS) I-III, Mortality Prediction Model (MPM) και Multiple Organ Dysfunction Syndrome (MODS) (Keuning *et al.*, 2020). Τα προγνωστικά αυτά μοντέλα χρησιμοποιούν κλινικές παρατηρήσεις και εργαστηριακές μετρήσεις κατά τις πρώτες 24 ώρες νοσηλείας στη ΜΕΘ για να προβλέψουν την ενδονοσοκομειακή θνησιμότητα και το αποτέλεσμα τους είναι μια εκτιμώμενη πιθανότητα του θανάτου εντός του νοσοκομείου, η οποία θεωρείται ως μία εκτίμηση της σοβαρότητας της νόσου. Ωστόσο, αυτή η πληθώρα μοντέλων έχει

δημιουργήσει δυσκολίες στους κλινικούς, καθώς δημιουργεί το ερώτημα για το ποιο είναι το πιο κατάλληλο μοντέλο να χρησιμοποιήσουν κάθε φορά (Debray *et al.*, 2019).

Η αξιολόγηση της προβλεπτικής ικανότητας του μοντέλου στα ίδια δεδομένα από τα οποία αυτό κατασκευάστηκε θα οδηγήσει σε μία πολύ αισιόδοξη εκτίμηση, λόγω της υπερπροσαρμογής του μοντέλου (overfitting) και της χρήσης αλγορίθμων επιλογής μεταβλητών, ιδιαίτερα όταν είναι λίγες οι εκβάσεις συγκριτικά με τον αριθμό των μεταβλητών πρόβλεψης. Για αυτό το λόγο, οι μελέτες που αναπτύσσουν νέα προβλεπτικά μοντέλα θα πρέπει να συμπεριλαμβάνουν κάποια μορφή εσωτερικής επικύρωσης (internal validation) του μοντέλου και να κάνουν διόρθωση για overfitting. Η εσωτερική επικύρωση περιλαμβάνει μόνο το αρχικό δείγμα και μερικές μέθοδοι εφαρμογής της είναι ο διαχωρισμός του δείγματος (split sample), η χρήση επαναληπτικών δειγμάτων bootstrap και η διεπικύρωση (cross-validation) (Moons *et al.*, 2015).

Μετά την ανάπτυξη ενός μοντέλου, προτείνεται η εκτίμηση της επίδοσής του σε διαφορετικά δεδομένα από αυτά που χρησιμοποιήθηκαν αρχικά. Συνήθως τα μοντέλα δείχνουν να έχουν καλύτερες επιδόσεις στα αρχικά δεδομένα που χρησιμοποιήθηκαν για την ανάπτυξή του συγκριτικά με τα νέα δεδομένα(Ramspek *et al.*, 2021). Γι’ αυτό έχει πολλή μεγάλη σημασία να γίνεται εξωτερική επικύρωση(external validation) των μοντέλων σε νέα σύνολα δεδομένων από παρεμφερείς νέους πληθυσμούς, σε διαφορετικά νοσοκομεία, διαφορετικές χώρες ή ακόμα και σε διαφορετικούς τύπους ασθενών, ώστε να εκτιμάται αν η αναπαραγωγιμότητα και η δυνατότητα γενίκευσής τους είναι εφικτή (Collins *et al.*, 2015; Debray *et al.*, 2019). Ουσιαστικά στην εξωτερική αξιολόγηση εκτιμάται η πιθανότητα έκβασης για κάθε νέο ασθενή χρησιμοποιώντας το αρχικό μοντέλο και οι εκτιμήσεις αυτές συγκρίνονται με τις παρατηρήθηκες τιμές. Σε περιπτώσεις μεγάλης ασυμφωνίας υπάρχει η δυνατότητα το μοντέλο να ανανεωθεί ή να προσαρμοστεί με βάση το νέο σύνολο δεδομένων, για παράδειγμα με τη χρήση αναβαθμονόμησης (recalibration) ή με την προσθήκη νέων μεταβλητών πρόβλεψης (Moons *et al.*, 2015).

Τα μοντέλα λοιπόν στα οποία γίνεται εξωτερική επικύρωση εφαρμόζονται σε νέα ομάδα ασθενών των οποίων τα δεδομένα δεν χρησιμοποιήθηκαν για να κατασκευαστεί το μοντέλο, συνήθως χρησιμοποιώντας τις ίδιες μεταβλητές πρόβλεψης και την ίδια μεταβλητή έκβασης(Collins *et al.*, 2015). Η επίδοση τους έπειτα αξιολογείται συγκρίνοντας τα προβλεπόμενα και τα παρατηρούμενα αποτελέσματα σε όλους τους

ασθενείς και υπολογίζοντας μέτρα προβλεπτικής ακρίβειας, που εξετάζουν τη διακριτική ικανότητα (discrimination) και τη βαθμονόμηση (calibration) (Debray *et al.*, 2019).

1.2 SAPS II

Το Simplified Acute Physiology Score (SAPS),το οποίο αναπτύχθηκε πρώτη φορά και επικυρώθηκε το 1984 στη Γαλλία, χρησιμοποιούσε 13 σταθμισμένες φυσιολογικές μεταβλητές και την ηλικία για να εκτιμήσει τον κίνδυνο θανάτου των ενηλίκων ασθενών που βρίσκονταν σε μονάδες εντατικής θεραπείας. Ο υπολογισμός του γινόταν με τις χειρότερες τιμές των μεταβλητών τις πρώτες 24 ώρες από την εισαγωγή των ασθενών στις ΜΕΘ και δεν είχε σχεδιαστεί για ατομική πρόγνωση (Vincent, 2009). Το 1993 αναπτύχθηκε το SAPS II μέσω ανάλυσης λογαριθμιστικής παλινδρόμησης για να επιλεχθούν και να σταθμιστούν οι μεταβλητές. Το SAPS II αποτελείται από 17 μεταβλητές: 12 μεταβλητές σχετικές με τις οργανικές διαδικασίες του ασθενή (καρδιακός παλμός, συστολική αρτηριακή πίεση, θερμοκρασία σώματος, αεριζόμενη ή συνεχής πνευμονική αρτηριακή πίεση, παραγωγή ούρων, επίπεδα ουρίας ορού, αριθμός λευκών αιμοσφαιρίων, κάλιο ορού, επίπεδο νατρίου στον ορό, επίπεδο διττανθρακικών ορού, επίπεδο χολερυθρίνης, κλίμακα κώματος Γλασκώβης), την ηλικία του ασθενή, το είδος εισαγωγής στις ΜΕΘ (προγραμματισμένη χειρουργική, απρογραμμάτιστη χειρουργική ή ιατρική) και τρεις μεταβλητές σχετιζόμενες με την υποκείμενη νόσο (σύνδρομο επίκτητης ανοσοανεπάρκειας, μεταστατικός καρκίνος και αιματολογική κακοήθεια). Για τις φυσιολογικές μεταβλητές, όπως και στο SAPS, χρησιμοποιούνται οι χειρότερες τιμές τις πρώτες 24 ώρες από την εισαγωγή των ασθενών στις ΜΕΘ για τον υπολογισμό του score (Vincent, 2009). Οι τιμές που παίρνει είναι ακέραιες και κυμαίνονται σε κλίμακα αυξανόμενης σοβαρότητας από 0 μέχρι 163 και η προβλεπόμενη θνησιμότητα κυμαίνεται από 0% μέχρι 100%. Ο υπολογισμός του score γίνεται μόνο μία φορά και εάν ο ασθενής βγει και ξαναμπεί στη ΜΕΘ το SAPS II θα πρέπει να υπολογιστεί εκ νέου. Ο τύπος για τον υπολογισμό της θνησιμότητας είναι ο εξής (Le Gall *et al.*, 2005):

$$logit = -7.7631 + 0.0737 * Score + 0.9971 * \ln(Score + 1) \qquad (1.1)$$

$$Mortality = \frac{e^{logit}}{1 + e^{logit}} \qquad (1.2)$$

Όπου *Score* το SAPS II

1.3 Στόχος της εργασίας

Η παρούσα μελέτη έχει στόχο να συγκεντρώσει συστηματικά τις μελέτες εξωτερικής αξιολόγησης του προγνωστικού μοντέλου θνησιμότητας SAPS II σε ασθενείς ΜΕΘ, να αξιολογήσει τη μεθοδολογική ποιότητά τους και να μετα-αναλύσει τα μέτρα της προβλεπτικής ακρίβειας τους, όπως το κριτήριο c, την ευρεία βαθμονόμηση και την κλίση βαθμονόμησης κ.α.. Απώτερος σκοπός της εργασίας είναι να αναδείξει την προγνωστική αξία του μοντέλου αυτού και να ενθαρρύνει τη χρήση του στην καθημερινή κλινική πράξη ώστε να ενισχυθεί η φροντίδα και η θεραπεία των ασθενών.

Η διπλωματική αυτή χωρίζεται σε 5 κεφάλαια. Στο κεφάλαιο 2 γίνεται μία εισαγωγή στο θεωρητικό υπόβαθρο της αξιολόγησης των προβλεπτικών μοντέλων και περιγράφονται τα βασικά μέτρα συνολικής αξιολόγησης, διακριτικής ικανότητας, βαθμονόμησης και κλινικής χρησιμότητας. Στο κεφάλαιο 3 γίνεται παρουσίαση του γενικού πλαισίου εργασίας μετα-αναλύσεων των μέτρων αξιολόγησης προβλεπτικών μοντέλων. Στο κεφάλαιο 4 περιγράφονται η μέθοδος της αναζήτησης και εξαγωγής των δεδομένων και τα βήματα που ακολουθήθηκαν για τη στατιστική τους ανάλυση και έπειτα γίνεται η παρουσίαση των αποτελεσμάτων. Στο κεφάλαιο 5 γίνεται συζήτηση των βασικών ευρημάτων από την ανάλυση των δεδομένων, σύγκριση τους με αυτά από άλλες συναφείς μελέτες και παρουσιάζονται προβληματισμοί με σκοπό την πιθανή βελτίωση της ανάλυσης και την περαιτέρω διερεύνηση του θέματος.

ΚΕΦΑΛΑΙΟ 2
ΑΞΙΟΛΟΓΗΣΗ ΠΡΟΒΛΕΠΤΙΚΩΝ ΜΟΝΤΕΛΩΝ

2.1 Τύποι προβλεπτικών μοντέλων

Τα πολυμεταβλητά προβλεπτικά μοντέλα χωρίζονται σε δύο μεγάλες κατηγορίες: τα διαγνωστικά και τα προγνωστικά προβλεπτικά μοντέλα. Στα διαγνωστικά μοντέλα δύο ή περισσότερες μεταβλητές συνδυάζονται για να εκτιμήσουν την πιθανότητα μία συγκεκριμένη κατάσταση ή ασθένεια να υφίσταται(ή να απουσιάζει) τη στιγμή της πρόβλεψης. Το σύνολο δεδομένων από το οποίο κατασκευάζονται αυτά τα μοντέλα περιέχει άτομα που είναι υποψήφια να πάσχουν από αυτήν την ασθένεια και για αυτά τα άτομα προορίζονται τα διαγνωστικά μοντέλα. Από την άλλη πλευρά τα προγνωστικά προβλεπτικά μοντέλα συνδυάζουν πολλές μεταβλητές για να εκτιμήσουν την πιθανότητα μία συγκεκριμένη έκβαση (π.χ. θάνατος) να συμβεί σε μία συγκεκριμένη χρονική περίοδο στο μέλλον. Τα προγνωστικά μοντέλα κατασκευάζονται και χρησιμοποιούνται σε άτομα που βρίσκονται σε κίνδυνο να εκδηλώσουν την έκβαση, είτε αυτά έχουν διαγνωστεί με κάποια ασθένεια είτε πρόκειται για τελείως υγιή άτομα. Η βασική διαφορά ανάμεσα σε διαγνωστικά και προγνωστικά προβλεπτικά μοντέλα είναι το χρονικό πλαίσιο. Οι μελέτες των διαγνωστικών μοντέλων είναι συγχρονικές (cross-sectional), ενώ των προγνωστικών είναι διαχρονικές προοπτικές (longitudinal) (Moons *et al.*, 2015). Τέλος τα μοντέλα είναι συνήθως μοντέλα λογαριθμιστικής παλινδρόμησης ή μοντέλα αναλογικών κινδύνων Cox, αφού μελετούν την πιθανότητα ή τον κίνδυνο μιας έκβασης, ενώ πρόσφατα έχουν αρχίσει να αναπτύσσονται και προβλεπτικά μοντέλα μηχανικής μάθησης. Συνήθως η λογαριθμιστική παλινδρόμηση χρησιμοποιείται για συγχρονικές (διαγνωστικές) και βραχυπρόθεσμες προγνωστικές εκβάσεις (πχ θνησιμότητα εντός 30 ημερών) και η παλινδρόμηση Cox για μακροπρόθεσμες προγνωστικές εκβάσεις (πχ δεκαετή κίνδυνο θανάτου) (Moons *et al.*, 2015).

Βασική προϋπόθεση για να γίνει η εξωτερική επικύρωση, για όλους τους τύπους προβλεπτικών μοντέλων, είναι ο υπολογισμός των προβλεπόμενων κινδύνων σε νέο σύνολο δεδομένων ή κοορτή ατόμων (validation cohort ή testing sample) διαφορετικό από το σύνολο δεδομένων που χρησιμοποιήθηκε για την κατασκευή του μοντέλου

(development ή training sample). Για τους υπολογισμούς αυτούς απαραίτητο είναι να γνωρίζουμε την εξίσωση του μοντέλου και τις τιμές των μεταβλητών πρόβλεψης για το κάθε άτομο. Βασικό στοιχείο του προβλεπτικού μοντέλου είναι ο προβλεπτικός δείκτης (prediction index -PI) ή αλλιώς γραμμικός προγνωστικός παράγοντας (linear predictor). Ο προβλεπτικός δείκτης υπολογίζεται ως ο γραμμικός συνδυασμός (το άθροισμα των γινομένων) των μεταβλητών πρόβλεψης με τους αντίστοιχους συντελεστές παλινδρόμησης (β), και έπειτα μετασχηματίζεται σε ποσοστό/κίνδυνο (μεταξύ 0 και 1) . Ο μετασχηματισμός αυτός διαφέρει ανάλογα με τον τύπο παλινδρόμησης. Στη λογαριθμιστική παλινδρόμηση απαιτείται η σταθερά του μοντέλου (συντελεστής τομής,intercept) για τον υπολογισμό του ατομικού κινδύνου , ενώ στην παλινδρόμηση Cox απαιτείται ο βασικός (baseline) κίνδυνος σε μία συγκεκριμένη χρονική στιγμή. Οι συντελεστές τομής και βασικού κινδύνου παραμένουν σταθερά για όλα τα άτομα, ενώ ο κίνδυνος διαφέρει (Ramspek *et al.*, 2021). Ο υπολογισμός του προβλεπτικού δείκτη και των ατομικών κινδύνων γίνεται με τους εξής τύπους:

$$Prognostic\ Index(PI) = \sum_{i=1}^{n} \beta_i x_i \qquad (2.1)$$

Logistic regression

$$\beta = \ln(odds\ ratio)$$

$$P(probability) = \frac{1}{1 + e^{-(\beta_0 + PI)}} \qquad (2.2)$$

$$\beta_0 = intercept$$

Cox proportional hazards regression

$$\beta = \ln(hazard\ ratio)$$

$$P(probability) = 1 - \left(S_0(t)\right)^{e^{PI}} \quad (2.3)$$

$$S_0(t) = baseline\ hazard\ at\ t$$

2.2 Ο συντελεστής προσδιορισμού R^2

Ο δείκτης R^2 είναι ένα συνολικό μέτρο ποσοτικοποίησης της πληροφορίας ενός μοντέλου για ένα συγκεκριμένο σύνολο δεδομένων. Ουσιαστικά εκφράζει το ποσοστό της μεταβλητότητας της έκβασης , η οποία εξηγείται από το προβλεπτικό μοντέλο . Με τον δείκτη R^2 καθίσταται δυνατή η σύγκριση της επίδρασης της αλλαγής κωδικοποίησης των μεταβλητών, ο τρόπος και η μορφή εισαγωγής των συνεχών μεταβλητών στο μοντέλο , ο διαφορετικός συνδυασμός μεταβλητών και η παρουσία όρων αλληλεπίδρασης. Οι τιμές του κυμαίνονται από 0% έως 100% , με R^2=100% να σημαίνει ότι το μοντέλο εξηγεί πλήρως τη μεταβλητότητα της έκβασης (Steyerberg, 2019).

Το R^2 αποτελεί το συνηθέστερο μέτρο αξιολόγησης για εκβάσεις που είναι συνεχείς. Στην κλασική γραμμική παλινδρόμηση ο τύπος του R^2 είναι ο εξής:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \quad (2.4)$$

όπου SS_{res} το άθροισμα των τετραγώνων των καταλοίπων (residual sum of squares) και SS_{tot} το συνολικό άθροισμα τετραγώνων (total sum of squares). Σε αυτήν την περίπτωση για SS_{res} ίσο με το 0 το R^2 μεγιστοποιείται , ενώ για SS_{res} = SS_{tot} ελαχιστοποιείται.

Αν και όχι τόσο συχνά όσο στη γραμμική παλινδρόμηση, γίνεται χρήση του R^2 και στα γενικευμένα γραμμικά μοντέλα. Επειδή στα γενικευμένα γραμμικά μοντέλα χρησιμοποιείται εκτίμηση μέσω μέγιστης πιθανοφάνειας και όχι μέσω ελαχίστων τετραγώνων, οι Cox & Snell πρότειναν τον εξής τύπο:

$$R^2 = 1 - exp\left[-\frac{2}{n}\{l(\hat{\beta}) - l(0)\}\right] = 1 - \left\{\frac{L(0)}{L(\hat{\beta})}\right\}^{\frac{2}{n}} \quad (2.5)$$

που $l(\hat{\beta}) = logL(\hat{\beta})$ και $l(0) = logL(0)$ δηλώνουν τη λογαριθμησμένη πιθανοφάνεια του fitted και του 'κενού' (Null) μοντέλου αντίστοιχα (Nagelkerke, 1991).

Ο Nagelkerke παρατήρησε ότι αυτός ο ορισμός του R^2 έχει τις εξής ιδιότητες:

i. Είναι σε συμφωνία με το κλασικό R^2 της γραμμικής παλινδρόμησης εφόσον αυτό μπορεί να υπολογιστεί,

ii. είναι συνεπής με τη μέγιστη πιθανοφάνεια ως μέθοδο εκτίμησης, εφόσον μεγιστοποιείται από την εκτίμηση μεγίστης πιθανοφάνειας του μοντέλου,

iii. είναι ασυμπτωτικά ανεξάρτητος από το μέγεθος δείγματος n,

iv. ερμηνεύεται ως το ποσοστό της μεταβλητότητας που εξηγείται από το μοντέλο,

v. δεν έχει μονάδες μέτρησης και

vi. αντικαθιστώντας τον παράγοντα 2/n στον τύπο του Cox-Snell με k/n προκύπτει μια γενίκευση του ποσοστού της k-οστής κεντρικής στιγμής που εξηγείται από το μοντέλο.

Παρ' όλα αυτά στη λογαριθμιστική παλινδρόμηση ,επειδή η λογαριθμησμένη πιθανοφάνεια μεγιστοποιείται στη μονάδα , το R^2 θα μεγιστοποιείται κάτω από τη μονάδα και συγκεκριμένα:

$$R^2_{max} = 1 - \big(L(0)\big)^{\frac{2}{n}} \tag{2.6}$$

Για αυτό το λόγο ο Nagelkerke πρότεινε μία άλλη εκδοχή (Nagelkerke, 1991) :

$$\bar{R}^2 = \frac{R^2}{R^2_{max}} \tag{2.7}$$

Το R^2 του Nagelkerke μπορεί επίσης να υπολογιστεί και για μοντέλα επιβίωσης , βάσει της διαφοράς -2 log likelihood ενός μοντέλου χωρίς μεταβλητές πρόβλεψης (null model) και ενός με μία ή περισσότερες μεταβλητές πρόβλεψης (Steyerberg *et al.*, 2010).

2.3 Το Brier Score

Το Brier Score είναι ένα σύνθετο μέτρο της διακριτικής ικανότητας και της βαθμονόμησης, που συνοψίζει την απόδοση ενός προβλεπτικού μοντέλου. Χρησιμοποιείται για δυαδικές (0/1) μεταβλητές έκβασης και ορίζεται ως η μέση τιμή των τετραγωνικών αποστάσεων μεταξύ των παρατηρούμενων εκβάσεων και των προβλέψεων(πιθανότητα να συμβεί η έκβαση) (*Ch 15: Evaluation of performance*, 2022). Οι τύποι του Brier score και της διακύμανσής του είναι οι εξής:

$$Brier = \frac{1}{N}\sum(y_i - \hat{y}_i)^2 \tag{2.7}$$

$$Var(Brier) = \frac{1}{N^2}\sum_{i=1}^{N}\hat{y}_i(1-\hat{y}_i)(2-\hat{y}_i)^2 \tag{2.8}$$

Όπου N ο αριθμός των παρατηρήσεων, y_i η παρατηρούμενη έκβαση και $\hat{y}_i$η προβλεπόμενη πιθανότητα της έκβασης για την i-οστή παρατήρηση. Οι τιμές του κυμαίνονται από 0% για το τέλειο μοντέλο μέχρι 25% για ένα μοντέλο μη πληροφοριακό εάν έχουμε 50% επίπτωση της έκβασης. Το μειονέκτημα αυτού του σκορ είναι ότι οι τιμές του επηρεάζονται από την επίπτωση της έκβασης(την μέση πιθανότητα p της έκβασης) , με αποτέλεσμα να μην είναι εύκολα ερμηνεύσιμο.

Για αυτό το λόγο προτιμάται να χρησιμοποιείται το $Brier_{scaled}$ που κυμαίνεται μεταξύ 0% και 100% και έχει ευκολότερη ερμηνεία (όσο μεγαλύτερο, τόσο καλύτερα). Αυτό υπολογίζεται ως εξής (Steyerberg, 2019):

$$Brier_{scaled} = 1 - \frac{Brier}{Brier_{max}} \quad (2.9)$$

Όπου $Brier_{max} = \frac{1}{N}\sum(y_i - \bar{y}_i)^2$ και $\bar{y}_i$ ο ρυθμός εκβάσεων μεταξύ των παρατηρούμενων εκβάσεων. Από τους τύπους φαίνεται ότι το $Brier_{scaled}$ ερμηνεύεται ως η ποσοστιαία μείωση του μέσου τετραγωνισμένου σφάλματος ενός μοντέλου που δεν παρέχει πληροφορία (null, χωρίς predictors) (Wong *et al.*, 2018).

Ένας άλλος τρόπος έκφρασης του $Brier_{max}$ είναι ο εξής :

$$Brier_{max} = mean(p) * \big(1 - mean(p)\big)^2 + \big(1 - mean(p)\big) * mean(p)^2$$

ή

$$Brier_{max} = mean(p) * \big(1 - mean(p)\big) \quad (2.10)$$

Όπου $mean(p)$ η μέση πιθανότητα να συμβεί η έκβαση και $mean(p) * (1 - mean(p))$ η διακύμανση της δυαδικής μεταβλητής Y με κατανομή Bernoulli(p).

2.4 Διακριτική ικανότητα (discrimination)

Διακριτική ή ταξινομική (discriminatory) ονομάζεται η ικανότητα ενός μοντέλου ,στο οποίο η έκβαση είναι μία δίτιμη μεταβλητή, να κάνει διάκριση ανάμεσα στα άτομα που έχουν την έκβαση και αυτά που δεν την έχουν. Το πιο ευρέως διαδεδομένο μέτρο διακριτικής ικανότητας στα γενικευμένα γραμμικά μοντέλα είναι η στατιστική ή δείκτης συμφωνίας (concordance c statistic, c index). Για δίτιμες εκβάσεις το κριτήριο c ταυτίζεται με το εμβαδόν AUC της χαρακτηριστικής καμπύλης ROC (area under the receiver operating characteristic curve). Η καμπύλη ROC είναι ένα γράφημα της ευαισθησίας (true positive rate) έναντι του $1 - ειδικότητα$ (false positive rate) για διαδοχικά όρια της πιθανότητας της έκβασης (Debray *et al.*, 2017; Steyerberg, 2019).

Η ευαισθησία ορίζεται ως το κλάσμα των αληθώς θετικών ταξινομήσεων προς το σύνολο των ατόμων που έχουν τη νόσο/έκβαση και ειδικότητα ορίζεται ως το κλάσμα

των ψευδώς αρνητικών ταξινομήσεων προς το σύνολο των ατόμων χωρίς τη νόσο/έκβαση. Για την ταξινόμηση του δείγματος σε θετικούς ή αρνητικούς είναι αναγκαίο να ορίσουμε ένα κατώφλι στην προβλεπόμενη πιθανότητα. Εάν η πρόβλεψη για ένα συγκεκριμένο άτομο είναι πάνω από το κατώφλι τότε θα ταξινομείται ως θετικός , διαφορετικά ως αρνητικός. (Steyerberg, 2019)

Για να κατασκευάσουμε την καμπύλη ROC θα πρέπει να κάνουμε αυτήν την ταξινόμηση για όλα τα πιθανά κατώφλια από 0 έως 100% . Ο δείκτης AUC ή c-statistic μπορεί να ερμηνευτεί ως η δεσμευμένη πιθανότητα, στην οποία για οποιοδήποτε ζεύγος ατόμων που το ένα έχει την έκβαση και το άλλο όχι, ο προβλεπόμενος κίνδυνος για την έκβαση είναι υψηλότερος για το άτομο με την έκβαση (Riley *et al.*, 2019)(Debray *et al.*, 2019). Γενικά, το c-statistic ορίζεται έτσι ώστε να παίρνει τιμές μεταξύ 0.5 και 1.Όταν η τιμή του είναι κοντά στο 0.5 αυτό σημαίνει ότι το ποσοστό των ζευγών στα οποία το άτομο με το μεγαλύτερο κίνδυνο είναι εκείνο που έχει την έκβαση είναι περίπου 50%. Αυτό υπονοεί ότι το προβλεπτικό μοντέλο δεν συμπεριφέρεται καλύτερα από μία τυχαία διάκριση των ατόμων (όπου η πιθανότητα στον πληθυσμό είναι 50%). Αντίστοιχα, τιμή c-statistic κοντά στη μονάδα υποδηλώνει πως άτομα με την έκβαση έχουν τον υψηλότερο κίνδυνο στο ζεύγος σε ποσοστό περίπου 100% και άρα το μοντέλο διακρίνει πάντα τα άτομα με και χωρίς την έκβαση. Συμπερασματικά, ένα μη πληροφοριακό μοντέλο θα έχει τιμή AUC ίση με 0.5 , ενώ το ιδανικό μοντέλο θα έχει AUC ίσο με 1 (Steyerberg, 2019).

Ένα άλλο μέτρο διακριτικής ικανότητας , που μετράει πόσο καλά διαχωρίζονται τα άτομα με και χωρίς την έκβαση , είναι η διακριτική κλίση (discrimination slope). Υπολογίζεται πολύ εύκολα, ως η απόλυτη διαφορά των μέσων των προβλέψεων στα άτομα με και χωρίς την έκβαση και μπορεί να απεικονιστεί μέσω των boxplot (Steyerberg, 2019). Γενικά προτιμάται η στατιστική c έναντι της διακριτικής κλίσης, διότι η πρώτη είναι στατιστική διατεταγμένων βαθμών (rank order) και δεν επηρεάζεται από την ποιότητα της βαθμονόμησης (calibration) σε αντίθεση με τη διακριτική κλίση.

2.5 Βαθμονόμηση (calibration)

Η βαθμονόμηση αναφέρεται στην ακρίβεια του μοντέλου να εκτιμά τις προβλεπόμενες πιθανότητες κινδύνου και είναι μέτρο του κατά πόσο υπάρχει συμφωνία μεταξύ των

αναμενόμενων (βάσει του μοντέλου) και των παρατηρούμενων συχνοτήτων των εκβάσεων (Debray *et al.*, 2017). Υπάρχουν τέσσερα επίπεδα βαθμονόμησης στα οποία μπορεί να βαθμονομηθεί ένα μοντέλο: μέση (mean), ασθενής (weak), μέτρια (moderate) και ισχυρή (strong) βαθμονόμηση (Van Calster *et al.*, 2016).

Για τη μέση ή συνολική βαθμονόμηση (mean calibration ή calibration-in-the-large) ενός προβλεπτικού μοντέλου ο μέσος προβλεπόμενος κίνδυνος συγκρίνεται με τον παρατηρούμενο συνολικό ρυθμό των υπό μελέτη εκβάσεων (events). Ένας τρόπος να διερευνηθεί η μέση βαθμονόμηση είναι μέσω λογαριθμιστικής αναβαθμονόμησης (logistic recalibration) , υπολογίζοντας τη σταθερά βαθμονόμησης (calibration intercept) $\alpha|b_L = 1$ ή και ελέγχοντας τη μηδενική υπόθεση ότι η σταθερά αυτή ισούται με το μηδέν μέσω του ελέγχου αναλογιών πιθανοφάνειας (likelihood ratio test) με ένα βαθμό ελευθερίας. Το b_L που είναι η κλίση της βαθμονόμησης είναι σταθερή κι ίση με μονάδα σε αυτήν την περίπτωση για να διατηρηθούν σταθεροί οι σχετικοί κίνδυνοι. Εάν η σταθερά που υπολογίσουμε είναι θετική , αυτό είναι ένδειξη υποεκτίμησης του προβλεπόμενου κινδύνου έκβασης , ενώ αντίθετα αρνητική σταθερά υποδηλώνει υπερεκτίμηση του προβλεπόμενου κινδύνου. Ένας άλλος τρόπος εκτίμησης της μέσης βαθμονόμησης είναι με τον υπολογισμό της συνολικής αναλογίας παρατηρούμενων προς αναμενόμενων αριθμών εκβάσεων (total O:E ratio) (Debray *et al.*, 2017). Η μέση βαθμονόμηση σα μέτρο βαθμονόμησης είναι ανεπαρκής ως μοναδικό κριτήριο , αφού για παράδειγμα ικανοποιείται όταν ο προβλεπόμενος κίνδυνος για κάθε άτομο ισούται με το πραγματικό ρυθμό εκβάσεων(Van Calster *et al.*, 2016).

Η ασθενής βαθμονόμηση (weak calibration) καθορίζεται από τη σταθερά βαθμονόμησης ($\alpha|b_L = 1$) και την κλίση βαθμονόμησης. Εάν η σταθερά είναι ίση με μηδέν και η κλίση ίση με ένα τότε αυτό είναι ένδειξη απουσίας υπερεφαρμογής (overfitting) ή υποεφαρμογής (underfitting) και υπερεκτίμησης ή υποεκτίμησης των προβλεπόμενων κινδύνων εκδήλωσης της έκβασης. Αποκλίσεις από αυτές τις τιμές μπορούν να εκτιμηθούν με την κατασκευή διαστημάτων εμπιστοσύνης ή να ελεγχθούν με ένα έλεγχο αναβαθμονόμησης στους δύο βαθμούς ελευθερίας, με μηδενική υπόθεση $H0$: $\alpha|b_L = 1 = 0$ & $b_L = 1$. Στην περίπτωση αυτή η βαθμονόμηση θεωρείται ασθενής γιατί συνοψίζεται από μόνον δύο παραμέτρους και δεν διασφαλίζει κατ' ανάγκη την ύπαρξη καλής βαθμονόμησης σε ολόκληρο το εύρος των προβλεπόμενων πιθανοτήτων. Επιπλέον, η ασθενής βαθμονόμηση δεν είναι αξιόπιστο μέτρο σε επίπεδο εσωτερικής

επικύρωσης διότι ικανοποιείται για τα δεδομένα με τα οποία αναπτύχθηκε το μοντέλο εφόσον χρησιμοποιηθεί η τυπική μέθοδος εκτίμησης , όπως για παράδειγμα η μέγιστη πιθανοφάνεια για τα μοντέλα λογαριθμιστικής παλινδρόμησης χωρίς να έχει μεγάλη σημασία ο τρόπος διαχείρισης των συνεχών μεταβλητών ή η διαχείριση των μεταβλητών αλληλεπίδρασης. Από την άλλη πλευρά, αποτελεί πιο χρήσιμο μέτρο σε επίπεδο εξωτερικής επικύρωσης ,αφού η σταθερά και η κλίση βαθμονόμησης δίνουν μια γενική και συνοπτική εικόνα για τα πιθανά προβλήματα της βαθμονόμησης του κινδύνου , και για δείγματα επικύρωσης που είναι σχετικά μικρά. (Van Calster *et al.*, 2016)

Η μέτρια (moderate) βαθμονόμηση μελετά αν μεταξύ ασθενών που έχουν τον ίδιο προβλεπόμενο κίνδυνο για μία έκβαση, ο παρατηρούμενος ρυθμός έκβασης ισούται με τον προβλεπόμενο κίνδυνο. Για παράδειγμα, σε άτομα με εκτιμώμενο κίνδυνο 10%, 10 στους 100 έχουν ή αναπτύσσουν την έκβαση. Αυτό εκτιμάται με γραφήματα (calibration plots ή calibration curves) που δείχνουν τη σχέση ανάμεσα στον αναμενόμενο κίνδυνο (οριζόντιος άξονας) και τον παρατηρούμενο κίνδυνο (κατακόρυφος άξονας), με τη χρήση συναρτήσεων loess ή splines (Calster *et al.*, 2019). Τα γραφήματα βαθμονόμησης μπορούν επίσης να κατασκευαστούν με καμπύλες εξομάλυνσης loess, κάνοντας απευθείας παλινδρόμηση σε μετασχηματισμούς των αναμενόμενων έναντι των παρατηρούμενων εκβάσεων. Ιδανικά, σε ένα διάγραμμα διασποράς των αναμενόμενων πιθανοτήτων και των παρατηρούμενων αναλογιών, τα σημεία της καμπύλης θα πρέπει να είναι πάνω στη διαγώνιο (Debray *et al.*, 2019). Αυτή η προσέγγιση είναι περισσότερο ευέλικτη από τη λογαριθμιστική αναβαθμονόμηση, καθώς μπορεί να αποκαλύψει δυσβαθμονόμηση που οφείλεται σε ισχυρές αλληλεπιδράσεις και περιπτώσεις μη γραμμικότητας. Ένα σχετικά πρόσφατο μέτρο που έχει αναπτυχθεί και έχει ομοιότητες με το Brier score είναι ο εκτιμώμενος δείκτης βαθμονόμησης (estimated calibration index,ECI). Ο δείκτης αυτός βασίζεται σε μία ευέλικτη ανάλυση βαθμονόμησης με τον υπολογισμό της μέσης τετραγωνικής διαφοράς ανάμεσα στον προβλεπόμενο και παρατηρούμενο κίνδυνο και το μετασχηματισμό της ώστε να παίρνει τιμές μεταξύ 0 και 1. Εάν η ευέλικτη καμπύλη βαθμονόμησης είναι τέλεια, το ECI ισούται με 0. Επειδή το μέτρο αυτό συνοψίζει μια ευέλικτη καμπύλη σε ένα απλό αριθμό , η χρήση του έχει προταθεί για συγκρίσεις βαθμονόμησης μεταξύ μοντέλων (Van Calster *et al.*, 2016). Πέρα από τις γραφικές αναπαραστάσεις, μία άλλη μέθοδος εκτίμησης της μέτριας βαθμονόμησης είναι με την ανάλυση ανά ομάδες προβλεπόμενων κινδύνων, για

παράδειγμα με τον έλεγχο του Hosmer-Lemeshow που δεν προτείνεται όμως πολύ σα μέθοδος καθώς έχει χαμηλή απόδοση (Van Calster *et al.*, 2016).

Τέλος η ισχυρή βαθμονόμηση, η πιο αυστηρή έκφανση της βαθμονόμησης, απαιτεί οι προβλεπόμενοι κίνδυνοι να αντιστοιχούν με τους παρατηρούμενους ρυθμούς έκβασης για κάθε μοτίβο συμμεταβλητών (covariate pattern). Αυτός ο ορισμός της ισχυρής βαθμονόμησης διαχωρίζει διαφορετικά μοτίβα συμμεταβλητών που σχετίζονται με τον ίδιο προβλεπόμενο κίνδυνο. Από κλινικής άποψης είναι σημαντικό να υπάρχει ισχυρή βαθμονόμηση, αφού με μέτρια μόνο βαθμονόμηση μπορεί οι προβλέψεις κινδύνου για κάποια άτομα να είναι μεροληπτικές. Παρ' όλα αυτά , η εκτίμηση της ισχυρής βαθμονόμησης είναι εξαιρετικά δύσκολη, γιατί απαιτεί μεγάλο δείγμα ώστε να υπάρχουν αρκετά άτομα ανά covariate pattern. Επιπλέον η βαθμονόμηση πάντα εκτιμάται σε σχέση με τις μεταβλητές πρόβλεψης του μοντέλου. Επομένως σε μία στρωματοποίηση ατόμων ως προς μία μεταβλητή που δε συμπεριλαμβάνεται ως προβλεπτικός παράγοντας στο μοντέλο, μπορεί να προκύψουν άτομα στο ίδιο covariate pattern αλλά με διαφορετικούς προβλεπόμενους κινδύνους χωρίς αυτό να ακυρώνει την ισχυρή βαθμονόμηση του μοντέλου (Van Calster *et al.*, 2016).

2.6 Net Reclassification Improvement (NRI)

Πολλές φορές η προσθήκη μιας μεταβλητής στο μοντέλο έχει μικρή επίδραση στο c statistic. Παρ' όλα αυτά μπορεί να προκαλέσει αλλαγή στην ταξινόμηση των ομάδων κινδύνου μεταξύ των ατόμων. Η αλλαγή αυτή από μόνη της δεν είναι ικανή να δώσει πληροφορίες για τη βελτίωση της στρωματοποίησης του κινδύνου (Steyerberg *et al.*, 2010). Για αυτό , μελετάμε την αλλαγή στην ομάδα κινδύνου ξεχωριστά για αυτούς που έχουν την έκβαση και αυτούς που δεν την έχουν. Για όσους την έχουν εκδηλώσει, οποιαδήποτε μετακίνηση σε υψηλότερη ομάδα κινδύνου σημαίνει βελτίωση της ταξινόμησης, ενώ μετακίνηση προς τα κάτω δείχνει χειρότερη επαναταξινόμηση. Ακριβώς τα αντίστροφα ισχύουν για τα άτομα δίχως την έκβαση. Η βελτίωση της επαναταξινόμησης ποσοτικοποιείται με το Net Reclassification Improvement (NRI) . Το NRI είναι το άθροισμα των διαφορών στα ποσοστά αυτών που ανέβηκαν τάξη κινδύνου μείον αυτών που κατέβηκαν τάξη , μεταξύ των ατόμων με την έκβαση , και αυτών που

κατέβηκαν τάξη μείον όσων ανέβηκαν τάξη , μεταξύ όσων δεν εκδήλωσαν την έκβαση (Steyerberg *et al.*, 2010).

2.7 Κλινική χρησιμότητα

Ένα κλινικά χρήσιμο μοντέλο είναι αυτό που μπορεί να χρησιμοποιηθεί από τους εργαζομένους στο χώρο της υγείας για τη βελτίωση της θεραπείας και της φροντίδας του ασθενή. Για να είναι χρήσιμο κλινικά ένα μοντέλο είναι απαραίτητο το μοντέλο αυτό να έχει καλή προβλεπτική ικανότητα , για παράδειγμα υψηλή διακριτική ικανότητα και ακριβή βαθμονόμηση. Ωστόσο, αυτό από μόνο του δεν αρκεί, γιατί μοντέλο με τέλεια προβλεπτική ικανότητα πρακτικά δεν υπάρχει. Οπότε έχουν δημιουργηθεί κάποια μέτρα για να εκτιμηθεί και ποσοτικά η κλινική χρησιμότητα (Baker and Gerdin, 2017).

Καταρχάς , είναι απαραίτητο να οριστεί ένα όριο αποκοπής ή κατώφλι (threshold) στην προβλεπόμενη πιθανότητα . Οι ασθενείς με πρόβλεψη μεγαλύτερη από το κατώφλι θα κατατάσσονται ως θετικοί , ενώ όσοι έχουν πρόβλεψη κάτω από το κατώφλι θα κατατάσσονται ως αρνητικοί. Ο όρος κλινική χρησιμότητα (clinical usefulness) αναφέρεται στην ικανότητα του μοντέλου να κάνει σωστά αυτήν την ταξινόμηση. Το ζήτημα τώρα είναι η επιλογή του κατωφλιού πρόβλεψης. Κατώφλι 50% θα σήμαινε ότι τα ψευδώς θετικά και ψευδώς αρνητικά άτομα έχουν την ίδια σημασία , πράγμα μάλλον σπάνιο για προβλήματα πρόβλεψης της υγείας (Steyerberg, 2019).

Για μια ανάλυση λήψης απόφασης θα έπρεπε να υπολογιστούν το όφελος και η ζημιά που θα οδηγούσαν στο ιδανικό κατώφλι. Αυτό όμως δεν είναι πάντα εύκολο να υπολογιστεί. Είτε λόγω έλλειψης επαρκούς πληροφορίας για τα οφέλη και τη ζημιά , σε επίπεδο πληθυσμού , είτε επειδή τα σχετικά βάρη οφέλους και ζημιάς μπορεί να διαφέρουν από ασθενή σε ασθενή οπότε να είναι απαραίτητα ατομικά κατώφλια για κάθε ασθενή. Το βασικό στην «ανάλυση καμπύλης απόφασης (decision curve analysis) είναι ότι ένα κατώφλι πιθανότητας μπορεί να χρησιμοποιηθεί και για την κατάταξη των ασθενών σε θετικούς ή αρνητικούς αλλά και για τον υπολογισμό των σχετικών βαρών των ψευδώς θετικών και ψευδώς αρνητικών ταξινομήσεων. Εάν η ζημιά από την μη απαραίτητη θεραπεία είναι περιορισμένη (ψευδώς θετικό) ,τότε το κατώφλι θα πρέπει να

είναι μικρό. Αντίθετα εάν η υπερθεραπεία είναι βλαβερή για τους ασθενείς τότε το κατώφλι θα πρέπει να είναι υψηλό πριν πάρουμε την απόφαση για θεραπεία. Η επίδοση του προγνωστικού μοντέλου αξιολογείται με το Net Benefit , που δίνεται από τον εξής τύπο :

$$Net\ Benefit = \frac{TP - w * FP}{N} \tag{2.11}$$

$$w = \frac{p_t}{1 - p_t} \tag{2.12}$$

Το TP είναι οι αληθώς θετικές ταξινομήσεις , το FP οι ψευδώς θετικές ταξινομήσεις , το N ο συνολικός αριθμός των ασθενών , p_t το κατώφλι πιθανότητας και w η στάθμη που ισούται με τα odds του κατωφλιού δηλαδή την αναλογία ζημιάς προς όφελος (Steyerberg *et al.*, 2010; Steyerberg, 2019)

ΚΕΦΑΛΑΙΟ 3
ΜΕΘΟΔΟΛΟΓΙΑ ΣΥΣΤΗΜΑΤΙΚΩΝ ΑΝΑΣΚΟΠΗΣΕΩΝ ΚΑΙ ΜΕΤΑ ΑΝΑΛΥΣΕΩΝ ΠΡΟΒΛΕΠΤΙΚΩΝ ΜΟΝΤΕΛΩΝ

Η μετα-ανάλυση είναι μια ερευνητική διαδικασία που χρησιμοποιείται για τη συστηματική σύνθεση ή τη συγχώνευση των ευρημάτων πολλών ανεξάρτητων μελετών, χρησιμοποιώντας στατιστικές μεθόδους για τον υπολογισμό μιας συνολικής επίδρασης. Οι μετα-αναλύσεις δεν αντλούν απλώς δεδομένα από μικρότερες μελέτες για να επιτύχουν ένα μεγαλύτερο μέγεθος δείγματος. Οι αναλυτές χρησιμοποιούν καλά αναγνωρισμένες , συστηματικές μεθόδους για να λάβουν υπόψη τους τις διαφορές στα μεγέθη δείγματος , την ετερογένεια στη μεθοδολογική προσέγγιση και τα ευρήματα και να ελέγξουν πόσο ευαίσθητα είναι τα αποτελέσματα τους στο δικό τους πρωτόκολλο συστηματικής ανασκόπησης (επιλογή μελετών και στατιστική ανάλυση)(Shorten and Shorten, 2013).

Γενικά τα μοντέλα των μετα-αναλύσεων μπορούν να διακριθούν σε μοντέλα σταθερών (ή κοινών) επιδράσεων (fixed effects) και τυχαίων επιδράσεων (random effects). Σε μοντέλο σταθερών επιδράσεων, όλες οι μελέτες θεωρούνται ισοδύναμες και οι διαφορές στα μέτρα προβλεπτικής ικανότητας μεταξύ των μελετών αποδίδονται στην τύχη. Επομένως, οι εκτιμητές ακρίβειας των μέτρων διακριτικής ικανότητας και βαθμονόμησης χρησιμοποιούνται για να δώσουν στάθμη στην κάθε μελέτη, για τον υπολογισμό των αντίστοιχων συνολικών μέτρων απόδοσης του μοντέλου. Σε μοντέλο τυχαίων επιδράσεων, θεωρείται ότι οι διαφορές στα μέτρα προβλεπτικής απόδοσης μεταξύ των μελετών δεν οφείλονται μόνο στην τύχη, αλλά αποδίδονται και σε πιθανή παρουσία ετερογένειας μεταξύ των μελετών. Κατά συνέπεια, τα μοντέλα τυχαίων επιδράσεων έχουν κατά κανόνα ευρύτερα διαστήματα εμπιστοσύνης και οι στάθμες των επιμέρους μελετών είναι περισσότερο παρόμοιες μεταξύ τους συγκριτικά με τα μοντέλα σταθερών επιδράσεων (Debray *et al.,* 2019). Παρόλο που κι οι δύο τύποι μοντέλων έχουν τα πλεονεκτήματα και τα μειονεκτήματα τους, για τη μετα-ανάλυση των προβλεπτικών μέτρων απόδοσης των προγνωστικών μοντέλων συστήνεται η χρήση μοντέλων τυχαίων επιδράσεων. Αυτό γιατί τόσο η διακριτική ικανότητα όσο και η βαθμονόμηση εξαρτώνται από τα χαρακτηριστικά των ασθενών (case mix variation) και δεν έχει νόημα να υποτεθεί ότι διαφέρουν μεταξύ των μελετών εξαιτίας μόνο τυχαίας δειγματοληπτικής

διακύμανσης. Για παράδειγμα η ετερογένεια του c-statistic μπορεί να παρατηρείται όταν διαφέρουν οι επιδράσεις των μεταβλητών πρόβλεψης μεταξύ των μελετών, λόγω διαφορετικής μεθόδου εκτίμησης τους, ή λόγω χρήσης διαφορετικής έκφανσης του c-statistic σε κάθε μελέτη (Debray *et al.*, 2017).

Στη συνέχεια λοιπόν γίνεται μία γενική περιγραφή του πλαισίου εργασίας για την διεξαγωγή συστηματικών ανασκοπήσεων και μετα-αναλύσεων της απόδοσης προβλεπτικών μοντέλων

3.1 Ερευνητικό ερώτημα

Το πρώτο βήμα όταν σχεδιάζουμε μια συστηματική ανασκόπηση ή μετα-ανάλυση είναι να διατυπώσουμε με σαφήνεια το ερευνητικό ερώτημα. Αυτό είναι πολύ σημαντικό γιατί όλα τα μεταγενέστερα βήματα καθορίζονται από το ερευνητικό ερώτημα, συμπεριλαμβανομένων της στρατηγικής αναζήτησης και των κριτηρίων επιλογής των μελετών, της εξαγωγής των δεδομένων από τις μελέτες , των μεθόδων μετα-ανάλυσης και της ερμηνείας των αποτελεσμάτων (Damen *et al.*, 2022).

Σύμφωνα με αναγνωρισμένους οδηγούς σύνταξης συστηματικών ανασκοπήσεων, όπως το Critical Appraisal and Data Extraction for Systematic Reviews of Prediction Modelling Studies (CHARMS) checklist και το Cochrane Prognosis Methods Group , προτείνεται η χρήση του συστήματος PICOTS για την διαμόρφωση του ερευνητικού ερωτήματος σε μία μετα-ανάλυση προβλεπτικών μοντέλων. (Debray *et al.*, 2017; Damen *et al.*, 2022) Το ακρωνύμιο PICOTS προέρχεται από τα αρχικά των λέξεων Population, Intervention, Comparator, Outcome(s), Timing και Setting σύμφωνα με τον Πίνακα 3.1.

Πίνακας 3.1: Το σύστημα PICOTS

PICOTS	
Population	Πληθυσμός στόχος στον οποίο θα εφαρμοστεί το μοντέλο πρόγνωσης
Intervention	Το υπό μελέτη προγνωστικό μοντέλο
Comparator	Εάν υφίσταται, το συγκρίσιμο μοντέλο με το υπό μελέτη μοντέλο
Outcome(s)	Εκβάσεις ενδιαφέροντος που προβλέπονται από το μοντέλο
Timing	1.Χρονική στιγμή που χρησιμοποιείται το προγνωστικό μοντέλο 2.Χρονική περίοδος στην οποία η εμφάνιση της έκβασης προβλέπεται από το μοντέλο
Setting	Γενικό πλαίσιο εφαρμογής του μοντέλου και ο επιθυμητός ρόλος του.

3.2 Αναζήτηση και επιλογή των άρθρων

Όταν κάνουμε ανασκόπηση των μελετών που αξιολογούν την προβλεπτική ικανότητα ενός συγκεκριμένου προγνωστικού μοντέλου, είναι σημαντικό να εξασφαλίσουμε ότι η στρατηγική αναζήτησης σε βιβλιογραφικές βάσεις θα εντοπίσει όλες εκείνες τις δημοσιεύσεις που αξιολόγησαν το μοντέλο για τον πληθυσμό στόχο , το πλαίσιο εφαρμογής και τις εκβάσεις που μας ενδιαφέρουν. Επομένως, η στρατηγική αναζήτησης θα πρέπει να κατασκευάζεται με γνώμονα το ερευνητικό ερώτημα που τέθηκε σύμφωνα με το σύστημα PICOTS (Debray *et al.*, 2017). Αυτό δεν είναι πάντοτε εύκολο, καθώς οι δημοσιεύσεις που μας ενδιαφέρουν μπορεί να μην έχουν καταταγεί στις βιβλιογραφικές βάσεις ως προγνωστικές μελέτες ή να μην περιορίζονται σε έναν μοναδικό σχεδιασμό μελέτης. Συχνά δεν είναι εύκολο να καταλάβουμε από τον τίτλο μιας μελέτης εάν αυτή αφορά προγνωστικό μοντέλο με αποτέλεσμα να αυξάνεται κατά πολύ ο όγκος των άρθρων που πρέπει να μελετηθούν διεξοδικά σε μια βιβλιογραφική αναζήτηση.

Μπορούμε να περιορίσουμε τα αποτελέσματα της αναζήτησης προσθέτοντας το όνομα ή το ακρωνύμιο του προγνωστικού μοντέλου (αν υπάρχει) ή ακόμη να χρησιμοποιήσουμε υπάρχοντα φίλτρα αναζήτησης, όπως για παράδειγμα είναι το φίλτρο των Ingui et al. για την εύρεση πολυπαραγοντικών κλινικών προβλεπτικών μοντέλων (Geersing *et al.*, 2012; Debray *et al.*, 2017). Ακόμη, είναι σημαντικό να ελέγχουμε τη βιβλιογραφία της αρχικής μελέτης που ανέπτυξε το προγνωστικό μοντέλο και να αναζητούμε αναφορές (citations) προς τη μελέτη αυτή για να εντοπίσουμε τα σχετικά άρθρα. Παραδείγματα βάσεων βιβλιογραφικών δεδομένων όπου γίνεται η αναζήτηση είναι το PubMed/MEDLINE και το Google Scholar (Debray *et al.*, 2017; Damen *et al.*, 2022).

Αφού ολοκληρωθεί η αναζήτηση, επιλέγονται οι βιβλιογραφικές αναφορές που είναι συναφείς με το ερευνητικό ερώτημα και τα προαποφασισμένα κριτήρια επιλογής, ενώ οι υπόλοιπες απορρίπτονται. Ιδανικά κάθε άρθρο αξιολογείται από τουλάχιστον δύο άτομα , πρώτα βάσει τίτλου και περίληψης και έπειτα βάσει ολόκληρου του κειμένου. Τα κριτήρια επιλογής των μελετών θα πρέπει να έχουν καθοριστεί εκ των προτέρων, έτσι ώστε να είναι σε συμφωνία με όλα τα επίπεδα του συστήματος PICOTS (Population, Intervention, Comparator, Outcome, Timing και Setting) και να περιλαμβάνουν γενικότερα στοιχεία όπως η γλώσσα, και να γίνει πιλοτική δοκιμή τους σε ένα μέρος των επιλεγμένων άρθρων (Damen *et al.*, 2022).

3.3 Εξαγωγή των δεδομένων

Μετά τη διαλογή των μελετών το επόμενο βήμα είναι η εξαγωγή των απαραίτητων δεδομένων από τις αναφορές των μελετών που θα συμπεριληφθούν στη μετα-ανάλυση. Η εξαγωγή αυτή μας παρέχει τις απαιτούμενες πληροφορίες για να κατασκευάσουμε ένα πίνακα με τα περιγραφικά στοιχεία των μελετών και επιτρέπει την ποιοτική και, εάν επιθυμούμε, την ποσοτική σύνοψη των ευρημάτων του έτους (Damen *et al.*, 2022).

Ένας οδηγός για την εξαγωγή των δεδομένων σε περιπτώσεις αξιολόγησης μελετών προγνωστικών μοντέλων αποτελεί το CHARMS checklist. Σύμφωνα με αυτό και με γνώμονα πάντα το σύστημα PICOTS, σημαντικές πληροφορίες που πρέπει να συλλέξουμε είναι οι ακόλουθες (Palazón-Bru *et al.*, 2020):

- Πηγή των δεδομένων : ο σχεδιασμός της μελέτης
- Συμμετέχοντες : ο πληθυσμός εφαρμογής ή αξιολόγησης του μοντέλου
- Προβλεπόμενη έκβαση: ποιο είναι το γεγονός που μας ενδιαφέρει να προβλέψουμε και ποιο το χρονικό διάστημα της πρόβλεψης
- Υποψήφιοι προγνωστικοί παράγοντες: οι μεταβλητές που θα μπορούσαν να είναι παράγοντες συσχέτισης με την έκβαση
- Μέγεθος δείγματος: τεκμηρίωση ή ενδείξεις ότι η μελέτη έχει αρκετούς συμμετέχοντες για να αξιολογήσει το μελετώμενο μοντέλο
- Ελλείποντα δεδομένα: πως διαχειρίστηκαν τα ελλείποντα δεδομένα
- Κατασκευή του μοντέλου: επεξήγηση για τον τρόπο επιλογής των προγνωστικών παραγόντων που συμπεριλήφθηκαν στο τελικό μοντέλο
- Απόδοση του μοντέλου: μέτρα διακριτικής ικανότητας και βαθμονόμησης με τις τυπικές αποκλίσεις τους ή τα διαστήματα εμπιστοσύνης τους αν δίνονται
- Αξιολόγηση του μοντέλου: τρόπος επικύρωσης του μοντέλου
- Μέτρα κλινικής χρησιμότητας (αν δίνονται): Αποτελέσματα από αναλύσεις καμπυλών απόφασης και Net Benefit

Όσον αφορά τα μέτρα απόδοσης του μοντέλου , που είναι και τα μέτρα που θα μετα-αναλύσουμε στη συνέχεια , σε περίπτωση που δε δίνονται τα διαστήματα εμπιστοσύνης τους , υπάρχει τρόπος να τα κατασκευάσουμε υπολογίζοντας το τυπικό τους σφάλμα από άλλες πληροφορίες . Συγκεκριμένα πρώτα εξάγουμε από τις μελέτες τη σημειακή εκτίμηση του c-statistic, το συνολικό μέγεθος δείγματος και το συνολικό αριθμό εκβάσεων. Το τυπικό σφάλμα υπολογίζεται ως εξής (Debray *et al.*, 2019):

$$SE(c) \approx \sqrt{\frac{c(1-c)\left[1 + n^{*}\frac{1-c}{2-c} + \frac{m^{*}c}{1+c}\right]}{mn}} \qquad (3.1)$$

όπου c η εκτίμηση του c-statistic, n ο αριθμός των συνολικών εκβάσεων, m ο αριθμός των συνολικών μη εκβάσεων και $n^* = m^* = \frac{1}{2}(m+n) - 1$

Πολλές φορές όμως εξυπηρετεί να μετατρέψουμε τις εκτιμήσεις το c statistic και του τυπικού του σφάλματος σε logit κλίμακα πριν τη μετα-ανάλυση, ώστε να πληρούνται οι υποθέσεις του μοντέλου τυχαίων επιδράσεων όπως εξηγούμε παρακάτω. Σε αυτήν την περίπτωση το τυπικό σφάλμα υπολογίζεται εφαρμόζοντας τη μέθοδο Δέλτα και ο τύπος που προκύπτει είναι (Debray *et al.*, 2019) :

$$SE\big(logit(c)\big) \approx \frac{SE(c)}{c(1-c)} \cong \sqrt{\frac{1 + n^* \frac{1-c}{2-c} + \frac{m^* c}{1+c}}{mnc(1-c)}} \qquad (3.2)$$

Εναλλακτικά , εάν δίνονται τα όρια του διαστήματος εμπιστοσύνης μπορούμε να υπολογίσουμε το τυπικό σφάλμα του c statistic σε logit κλίμακα ως εξής:

$$SE\big(logit(c)\big) \cong \frac{logit(c_{ub}) - logit(c_{lb})}{2z^*} \qquad (3.3)$$

όπου z^* είναι το 100(1-α/2) εκατοστημόριο της κανονικής κατανομής. Για τα 95% διαστήματα εμπιστοσύνης, $z^* = 1{,}96$.

Για τα μέτρα βαθμονόμησης, ο αντίστοιχος τύπος για το τυπικό σφάλμα της μετασχηματισμένης αναλογίας των συνολικών παρατηρούμενων προς αναμενόμενων εκβάσεων δίνεται παρακάτω. Αυτή τη φορά ο μετασχηματισμός γίνεται σε κλίμακα (φυσικού) λογαρίθμου πριν τη μετα-ανάλυση

$$SE(\ln(O:E)) \approx \frac{SE(O)}{O} \approx \sqrt{\frac{1-P_O}{O}} \qquad (3.4)$$

όπου O ο αριθμός των παρατηρούμενων εκβάσεων και P_o η παρατηρούμενη πιθανότητα της έκβασης (Debray *et al.*, 2019).

3.4 Εκτίμηση της ποιότητας των μελετών και του κινδύνου συστηματικού σφάλματος

Ο κίνδυνος για συστηματικό σφάλμα (risk of bias) υφίσταται όταν υπάρχουν ελλείψεις ή ελαττώματα στο σχεδιασμό ή την ανάλυση που είναι πιθανό να οδηγήσουν σε λανθασμένα ή τροποποιημένα αποτελέσματα. Η αξιολόγηση αυτή γίνεται ιδανικά από δύο άτομα και σε περίπτωση ασυμφωνίας επεμβαίνει και τρίτος μελετητής. Για μετα-αναλύσεις μελετών προγνωστικών μοντέλων η εκτίμηση του κινδύνου συστηματικού σφάλματος και της καταλληλόλητας των συμπεριλαμβανομένων άρθρων μπορεί να πραγματοποιηθεί με το Prediction model Risk of Bias ASsessment Tool (PROBAST) (Damen *et al.*, 2022), το οποίο συνίσταται σε τέσσερις τομείς αξιολόγησης:

- Συμμετέχοντες: πιθανές πηγές συστηματικού σφάλματος που αφορούν τις μεθόδους επιλογής συμμετεχόντων και τις πηγές δεδομένων
- Προγνωστικοί παράγοντες: πιθανές πηγές συστηματικού σφάλματος που σχετίζονται με τον ορισμό και τη μέτρηση των παραγόντων που θα συμπεριληφθούν στο μοντέλο

- Έκβαση: πιθανές πηγές συστηματικού σφάλματος που σχετίζονται με τον ορισμό και τη μέτρηση της έκβασης που προβλέπεται από το μοντέλο

- Ανάλυση: πιθανές πηγές συστηματικού σφάλματος στις μεθόδους στατιστικής ανάλυσης , όπως η επιλογή των μεθόδων ανάλυσης και η διαχείριση των ελλειπουσών τιμών

Ο κάθε τομέας PROBAST περιέχει συγκεκριμένες ερωτήσεις που απαντώνται με "Ναι", "Πιθανώς ναι", "Πιθανώς όχι", "Όχι" ή "Καμία πληροφορία". Οι ερωτήσεις είναι δομημένες με τέτοιο τρόπο ώστε η απάντηση "Ναι" να υποδηλώνει απουσία συστηματικού σφάλματος. Έπειτα είναι στην ευχέρεια του ερευνητή και βάσει των ερωτήσεων να βαθμολογήσει τον κάθε τομέα του PROBAST όσον αφορά τον κίνδυνο για συστηματικό σφάλμα ως "χαμηλό", "υψηλό" ή "ασαφές". Από τις επιμέρους βαθμολογίες προκύπτει και η συνολική βαθμολογία της μελέτης για τον κίνδυνο συστηματικού σφάλματος. Χαμηλή βαθμολογία λαμβάνουν οι μελέτες με χαμηλό κίνδυνο σε όλους τους τομείς , ενώ υψηλή βαθμολογία λαμβάνουν όσες έχουν έναν τουλάχιστον τομέα με υψηλό κίνδυνο. Ασαφή βαθμολογία λαμβάνουν οι μελέτες με έναν τουλάχιστον τομέα να κρίνεται ασαφής και με τους υπόλοιπους τομείς να αξιολογούνται με χαμηλό κίνδυνο (Wolff *et al.*, 2019). Αντίθετα, η καταλληλότητα των μελετών κρίνεται μόνο από τους τρεις πρώτους τομείς (συμμετέχοντες, προγνωστικοί παράγοντες και έκβαση) και αξιολογείται ανάλογα με το βαθμό που ο κάθε τομέας ανταποκρίνεται στο ερευνητικό ερώτημα.

3.5 Ανάλυση των δεδομένων

Αφού βρεθούν όλες οι σχετικές μελέτες και γίνει η εξαγωγή των απαραίτητων δεδομένων από αυτές , οι εκτιμήσεις της διακριτικής ικανότητας και της βαθμονόμησης μπορούν να συνοψιστούν σε ένα σταθμισμένο μέσο όρο. Επειδή συνήθως οι μελέτες εξωτερικής αξιολόγησης ποικίλουν όσον αφορά το σχεδιασμό , την εκτέλεση και το

φάσμα των ασθενών , οι διαφορές στα αποτελέσματα τους δε μπορούν να αποδοθούν σε τυχαίο δειγματοληπτικό σφάλμα. Επομένως, η μετα-ανάλυση πρέπει να επιτρέπει την παρουσία ετερογένειας και να στοχεύει στον υπολογισμό ενός συνοπτικού αποτελέσματος που να ποσοτικοποιεί τη μέση απόδοση όλων των μελετών. Αυτό επιτυγχάνεται εφαρμόζοντας μοντέλο τυχαίων επιδράσεων, το οποίο συνυπολογίζει την ετερογένειας. Οι στάθμες που χρησιμοποιούνται για τον υπολογισμό της σταθμισμένης εκτίμησης , προκύπτουν από το τυπικό σφάλμα των μελετών και την εκτίμηση του μεγέθους τα ετερογένειας.

Ένα σημείο που θέλει προσοχή είναι η μορφή με την οποία θα χρησιμοποιηθούν τα μέτρα της διακριτικής ικανότητας και της βαθμονόμησης κατά την ανάλυση τους. Το μοντέλο τυχαίων επιδράσεων θεωρεί ότι:

$$Y_i \sim N\left(\mu_i, S_i^2\right), \quad \mu_i \sim N(\mu, \tau^2)$$

όπου Y_i είναι το στατιστικό μέτρο που μελετάμε για κάθε μελέτη i=1,2,..,k. Η μετα ανάλυση λοιπόν υποθέτει κανονικότητα του στατιστικού μέτρου, τόσο εντός κάθε μελέτης όσο και μεταξύ των διαφόρων μελετών. Σε επίπεδο εντός των μελετών, μπορούμε να περιμένουμε ότι η κανονικότητα πληρείται εάν το μέγεθος δείγματος είναι αρκετά μεγάλο βάσει του Κεντρικού Οριακού Θεωρήματος. Για την κανονικότητα όμως των τυχαίων επιδράσεων (επίπεδο μεταξύ των μελετών), μελέτη προσομοίωσης έχει ανέδειξε ότι δουλεύοντας σε logit κλίμακα πληρείται πιο συχνά η υπόθεση κανονικότητας για τις τυχαίες επιδράσεις (Snell *et al.*, 2018). Με το μετασχηματισμό logit, το μοντέλο τυχαίων επιδράσεων γράφεται ως εξής:

$$logit(c_i) \sim N\left(\mu_{discr}, Var\left(logit(c_i)\right) + \tau_{discr}^2\right) \qquad (3.5)$$

όπου logit(c_i) ο logit μετασχηματισμός του c statistic για την i-οστή μελέτη και τ^2 η ετερογένεια μεταξύ των μελετών. Έπειτα από την εκτίμηση του μοντέλου, μπορούμε να

υπολογίσουμε το c-statistic στην αρχική του κλίμακα θα χρησιμοποιήσουμε το $logit^{-1}(\hat{\mu}_{discr})$, ή ισοδύναμα $1/(1+\exp(-\hat{\mu}_{discr}))$.

Όσον αφορά τα μέτρα βαθμονόμησης η κλίση βαθμονόμησης και η μέση βαθμονόμηση δεν απαιτούν κάποιον μετασχηματισμό , αφού ικανοποιείται η κανονικότητα μεταξύ των μελετών. Για την αναλογία όπως των παρατηρούμενων προς τις αναμενόμενες εκβάσεις απαιτείται μετασχηματισμός εάν διαφέρουν αρκετά οι επιδράσεις των προβλεπτικών παραγόντων. Στη συγκεκριμένη περίπτωση συνίσταται ο λογαριθμικός μετασχηματισμός έτσι ώστε(Snell *et al.*, 2018; Debray *et al.*, 2019):

$$\ln(O:E)_i \sim N(\mu_{cal.OE}, Var(\ln(O:E)_i) + \tau^2_{cal.OE}) \tag{3.6}$$

Για τον υπολογισμό των συγκεντρωτικών εκτιμήσεων των δύο μέτρων αξιολόγησης των μοντέλων, αυτό μπορεί να γίνει με τη βελτιστοποίηση της (λογαριθμησμένης) πιθανοφάνειας των μοντέλων (3.5) και (3.6) για να προκύψουν οι εξής εκτιμήτριες:

$$\hat{\mu}_{pooled} = \frac{\Sigma_{i=1}^{K}\left(\frac{\hat{\theta}_i}{\tau^2 + Var(\hat{\theta}_i)}\right)}{\Sigma_{i=1}^{K}\left(\frac{1}{\tau^2 + Var(\hat{\theta}_i)}\right)} \tag{3.7}$$

$$SE(\hat{\mu}_{pooled}) = \sqrt{\frac{1}{\Sigma_{i=1}^{K}(1/(\tau^2 + Var(\hat{\theta}_i)))}} \tag{3.8}$$

όπου το $\hat{\theta}_\iota$ αντιπροσωπεύει την εκτίμηση της παραμέτρου σε κάθε μελέτη i ($logit$(c), log (O:E), calibration slope, ή calibration-in-the-large) και το $1/(\tau^2 + Var(\hat{\theta}_i))$ είναι η τιμή που παίρνει η στάθμη που δίνεται στις εκτιμήσεις των επιμέρους μελετών με το μοντέλο

των τυχαίων επιδράσεων. Στην κλασική προσέγγιση των DerSimonian και Laird το τ^2 υπολογίζεται ξεχωριστά και έπειτα προστίθεται στις εξισώσεις παραπάνω για τον υπολογισμό των εκτιμητριών. Η βιβλιογραφία όμως προτείνει τη χρήση restricted maximum likelihood estimation (REML) για τον υπολογισμό του τ^2, που θεωρείται βελτίωση σε σχέση με τον αρχικό εκτιμητή μεγίστης πιθανοφάνειας, αφού δε θεωρεί δεδομένη τη μέση τιμή των εκτιμήσεων $\hat{\theta}_\iota$ (Debray *et al.*, 2019; Langan *et al.*, 2019).

Μείζονος ενδιαφέροντος είναι και η κατασκευή διαστημάτων εμπιστοσύνης και διαστημάτων πρόβλεψης των εκτιμήσεων των συνολικών μέτρων διακριτικής ικανότητας και βαθμονόμησης. Τα διαστήματα πρόβλεψης έχουν ιδιαίτερη σημασία καθώς θα μας δώσουν μια ένδειξη του εύρους της ετερογένειας μεταξύ των μελετών και του πόσο καλά θα συμπεριφέρεται το μοντέλο που μελετάμε σε μία νέα μελέτη. Τα διαστήματα αυτά περιλαμβάνουν την αβεβαιότητα γύρω από την εκτίμηση αλλά και την ετερογένεια μεταξύ των μελετών. Για το λόγο αυτό συνήθως είναι πιο ευρεία από τα διαστήματα εμπιστοσύνης (Damen *et al.*, 2022).

Για τον υπολογισμό των τυπικών σφαλμάτων και την αποφυγή πιθανού συστηματικού σφάλματος όταν η μετα-ανάλυση δεν περιλαμβάνει πολλές μελέτες, η βιβλιογραφία προτείνει τη διόρθωση του τυπικού σφάλματος της εκτίμησης με τη μέθοδο των Hartung και Knapp (Debray *et al.*, 2019). Το διορθωμένο τυπικό σφάλμα υπολογίζεται ως εξής:

$$SE_{HK}(\hat{\mu}_{pooled}) = SE(\hat{\mu}_{pooled}) \sqrt{\frac{1}{K-1} \Sigma_{i=1}^{K} \frac{(\hat{\theta}_i - \hat{\mu})^2}{Var(\theta_i) + \hat{\tau}^2}} \qquad (3.9)$$

Οπότε το διάστημα εμπιστοσύνης δίνεται από τον τύπο

$$\hat{\mu}_{pooled} \pm t_{K-1,0.975} \widehat{SE}_{HK}(\hat{\mu}_{pooled}) \qquad (3.10)$$

όπου t_{K-1} δηλώνει το 100(1-α/2) εκατοστημόριο της Student-t κατανομής με K-1 βαθμούς ελευθερίας , με α συνήθως ίσο με 0.05 για να δώσει επίπεδο στατιστικής σημαντικότητας 5% και άρα 95% όρια αξιοπιστίας. Επιλέγεται η κατανομή Student-t, αντί της κανονικής για να ληφθεί υπόψη και η αβεβαιότητα του τ^2.

Αντίστοιχα το διάστημα πρόβλεψης της εκτίμησης δίνεται από τον τύπο

$$\hat{\mu}_{pooled} \pm t_{K-2,0.975}\sqrt{\hat{\tau}^2 + (\widehat{SE}(\hat{\mu}_{pooled}))^2} \qquad (3.11)$$

όπου γίνεται χρήση της Student-t κατανομής , αντί της κανονικής, για να ληφθεί υπόψη η αβεβαιότητα του $\hat{\tau}$ (Debray *et al.*, 2019).

Μετά την ολοκλήρωση της μετα-ανάλυσης και τον υπολογισμό του συνθετικού δείκτη πραγματοποιείται και ανάλυση ευαισθησίας, η οποία χρησιμοποιείται για τη διερεύνηση της επίδρασης μεμονωμένων μελετών στο συνθετικό δείκτη. Έτσι για παράδειγμα για να διερευνηθεί η επίδραση των μελετών χαμηλής ποιότητας, επαναλαμβάνεται η μετα-ανάλυση εξαιρώντας τις μελέτες με υψηλό κίνδυνο για συστηματικά σφάλματα.

3.6 Διερεύνηση της ετερογένειας μεταξύ των μελετών

Όταν παρατηρούμε ετερογένεια μεταξύ των μελετών στα μέτρα της διακριτικής ικανότητας και βαθμονόμησης ενός προβλεπτικού μοντέλου, είναι πολύ σημαντικό να κάνουμε διερεύνηση των πιθανών πηγών της ετερογένειας αυτής. Αυτό είναι ένα σημαντικό βήμα για να κατανοήσουμε σε ποιες περιπτώσεις η επίδοση του μοντέλου παραμένει επαρκής και πότε το μοντέλο ενδέχεται να χρειάζεται βελτίωση. Βασικές πηγές ετερογένειας είναι οι αποτελούν διαφορές στο σχεδιασμό και στους πληθυσμούς των μελετών.

Τα πιο γνωστά και ευρέως χρησιμοποιούμενα μέτρα ετερογένειας στις μετα αναλύσεις τυχαίων επιδράσεων είναι η στατιστική συνάρτηση Q του Cochran, ο δείκτης I^2των Higgins και Thompson και η διακύμανση (ετερογένεια) τ^2μεταξύ των μελετών. Το στατιστικό κριτήριο Q του Cochran χρησιμοποιείται για να διαχωρίσει το δειγματικό σφάλμα των μελετών από την πραγματική ετερογένεια μεταξύ των μελετών. Ορίζεται ως ένα σταθμισμένο άθροισμα τετραγώνων και χρησιμοποιεί ουσιαστικά την απόκλιση της παρατηρούμενης επίδρασης $\hat{\theta}_\kappa$ κάθε μελέτης από τη συνθετική επίδραση $\hat{\theta}_{pooled}$ σταθμισμένη με το αντίστροφο της διακύμανσης της κάθε μελέτης, w_κ. Ο τύπος είναι (Harrer *et al.*, 2022):

$$Q = \sum_{k=1}^{K} w_\kappa \left(\hat{\theta}_\kappa - \hat{\theta}_{pooled}\right)^2 \tag{2.12}$$

Ο έλεγχος ετερογένειας ή αλλιώς Cochran's Q test, βασίζεται στην παραδοχή ότι το Q ακολουθεί χ^2 κατανομή με k-1 βαθμούς ελευθερίας και είναι ένας έλεγχος της υπόθεσης ότι δεν υπάρχει ετερογένεια μεταξύ των μελετών.

Ο δείκτης I^2 είναι ένας άλλος τρόπος ποσοτικοποίησης της ετερογένειας μεταξύ των μελετών και βασίζεται ευθέως στο Q. Ορίζεται ως το ποσοστό της μεταβλητότητας των δεικτών επίδρασης που δεν οφείλεται σε δειγματικό σφάλμα. Ο τύπος του είναι:

$$I^2 = \frac{Q - (k-1)}{Q} \tag{2.13}$$

όπου k ο συνολικός αριθμός μελετών. Το I^2 δεν μπορεί να πάρει αρνητικές τιμές κι έτσι εάν το Q είναι μικρότερο από το K-1, το I^2 παίρνει αυτόματα την τιμή μηδέν (Harrer *et al.*, 2022).

Μία μέθοδος διερεύνησης των πηγών ετερογένειας είναι η μετα-παλινδρόμηση όταν υπάρχει σχετικά μεγάλος αριθμός μελετών που συνεισφέρουν στη μετα-ανάλυση, όπου εξαρτημένη μεταβλητή είναι η (μετασχηματισμένη) εκτίμηση του μέτρου προβλεπτικής επίδοσης του μοντέλου. Ως ανεξάρτητες μεταβλητές χρησιμοποιούνται παράγοντες στους οποίους μπορεί να οφείλεται αυτή η διαφορά των μέτρων επίδοσης όπως η ποιότητα

(κίνδυνος συστηματικού σφάλματος), τα χαρακτηριστικά των μελετών (έτος διεξαγωγής, ερευνητικός σχεδιασμός, τρόπος συλλογής των δεδομένων κλπ), συνοπτικά χαρακτηριστικά των ασθενών (πχ μέση ηλικία) ή ακόμη και στοιχεία της στατιστικής ανάλυσης. Στη μετα-παλινδρόμηση, ο συντελεστής παλινδρόμησης β εκφράζει πως η εξαρτημένη μεταβλητή, στην προκειμένη περίπτωση το logit(c), το log(O:E), το calibration slope ή το calibration-in-the-large, μεταβάλλεται μεταξύ των υποομάδων των μελετών εάν η ανεξάρτητη μεταβλητή είναι κατηγορική ή ανά μία μονάδα αύξησης για μία συνεχή ανεξάρτητη μεταβλητή. Όμως η ισχύς των ελέγχων στατιστικής σημαντικότητας σε αυτές τις αναλύσεις είναι συνήθως χαμηλή, ειδικά εάν ο αριθμός των μελετών είναι μικρός. Τέλος οι μετα-παλινδρομήσεις και οι αναλύσεις υπο-ομάδων είναι επιρρεπείς σε οικολογικά συστηματικά σφάλματα (ecological fallacy) όταν αναλύουν συνοπτικά δεδομένα ασθενών (Debray *et al.*, 2017). Εναλλακτικά, μπορεί να εφαρμοστεί ανάλυση υπο-ομάδων η οποία συνοψίζει την επίδοση του μοντέλου σε διαφορετικές υποομάδες.

3.7 Παρουσίαση και ερμηνεία των αποτελεσμάτων

Πολύ βασικό κομμάτι των μετα-αναλύσεων είναι η παρουσίαση και ερμηνεία των ευρημάτων και η διατύπωση των συμπερασμάτων του ερευνητή. Πρέπει να παρουσιαστούν και να αξιολογηθούν τόσο τα συνθετικά μέτρα απόδοσης που εκτιμήθηκαν από τη στατιστική ανάλυση όσο και η αβεβαιότητα τους (Damen *et al.*, 2022). Ένα εργαλείο αξιολόγησης των ευρημάτων είναι το Grading of Recommendations Assessment, Development and Evaluation (GRADE). Οι πέντε τομείς που εξετάζει η κλίμακα GRADE είναι ο κίνδυνος συστηματικού σφάλματος , η ανακρίβεια, η ασυνέπεια, η έλλειψη αμεσότητας και το σφάλμα δημοσίευσης (Foroutan *et al.*, 2020). Πρόσφατα δημοσιεύτηκε μία μελέτη για τη χρήση του GRADE σε συστηματικές ανασκοπήσεις και μετα-αναλύσεις προβλεπτικών μοντέλων (Foroutan *et al.*, 2022). Το ζήτημα όμως χρήζει περαιτέρω διερεύνησης, αφού το άρθρο εστίαζε μόνο στα μέτρα βαθμονόμησης. Πέραν αυτών είναι χρήσιμο να παρουσιάζονται πληροφορίες για τους πληθυσμούς των μελετών εξωτερικής αξιολόγησης που χρησιμοποιήθηκαν για τη μετα-ανάλυση και όρια αξιοπιστίας ή και πρόβλεψης των εκτιμήσεων της μετα-ανάλυσης (Damen *et al.*, 2022).

Η δημοσίευση των αποτελεσμάτων μιας συστηματικής ανασκόπησης και μετα-ανάλυσης πρέπει να ακολουθεί τις οδηγίες του PRISMA (Preferred Reporting Itmes for Systematic reviews and Meta-Analysis) που υποδεικνύει τα βασικά σημεία αναφοράς για μια μετα-ανάλυση μελετών παρέμβασης, που είναι κάπως σχετική με μετα-αναλύσεις μελετών εξωτερικής αξιολόγησης. Επιπλέον μπορεί να χρησιμοποιηθεί και το TRIPOD (transparent reporting of a multivariate prediction model for individual prognosis or diagnosis) που παρέχει συστάσεις για αναφορά μελετών ανάπτυξης , επικύρωσης και αναβάθμισης προβλεπτικών μοντέλων. Ιδανικά για μετα-αναλύσεις προγνωστικών μοντέλων θα βοηθούσε ο συνδυασμός του TRIPOD με το PRISMA (Debray *et al.*, 2017; Damen *et al.*, 2022).

ΚΕΦΑΛΑΙΟ 4
ΣΥΣΤΗΜΑΤΙΚΗ ΑΝΑΣΚΟΠΗΣΗ ΚΑΙ ΜΕΤΑ-ΑΝΑΛΥΣΗ ΤΗΣ ΠΡΟΓΝΩΣΤΙΚΗΣ ΙΚΑΝΟΤΗΤΑΣ ΤΟΥ ΜΟΝΤΕΛΟΥ SAPS II

4.1 Υλικό και Μέθοδος

Πραγματοποιήθηκε συστηματική ανασκόπηση και μετα-ανάλυση της σχετικής βιβλιογραφίας. Οι μέθοδοι αναζήτησης και αναφοράς που χρησιμοποιήθηκαν βασίστηκαν στους οδηγούς του Debray (Debray *et al.*, 2017, 2019) και του Damen (Damen *et al.*, 2022)

Το ερευνητικό ερώτημα της μελέτης μας είναι

"Αξιολόγηση της προβλεπτικής εγκυρότητας του μοντέλου SAPS II για την πρόβλεψη της θνησιμότητας σε ασθενείς που βρίσκονται σε ΜΕΘ". Ο παρακάτω πίνακας δείχνει τη σύνδεση με το σύστημα PICOTS

Πίνακας 4.1: Το ερευνητικό ερώτημα βάσει του συστήματος PICOTS

PICOTS SYSTEM	
Population	Ασθενείς σε μονάδες εντατικής θεραπείας (ΜΕΘ)
Intervention	SAPS II
Comparator model	Μη εφαρμόσιμο
Outcome	Ενδονοσοκομειακή θνησιμότητα
Timing	Εντός του πρώτου εικοσιτετραώρου από την εισαγωγή του ασθενή στη ΜΕΘ
Setting	Η αξιολόγηση της προβλεπτικής ικανότητας του μοντέλου και η αναπαραγωγιμότητα του με στόχο την ένταξη του στην καθημερινή κλινική πράξη για τη λήψη αποφάσεων

4.1.1 Αναζήτηση και επιλογή άρθρων

Πραγματοποιήθηκε βιβλιογραφική ανασκόπηση στην ελεύθερη ηλεκτρονική πλατφόρμα αναζήτησης PubMed την περίοδο από 16 Οκτωβρίου 2022 μέχρι 20 Νοέμβρη 2022. Το PubMed επιλέχθηκε επειδή αποτελεί τη δημόσια επιφάνεια της βάσης MEDLINE όπως και άλλων βάσεων δεδομένων, καθιστώντας το ως την πρωταρχική πηγή βιοιατρικής βιβλιογραφίας και μία από τις πιο ευρέως προσβάσιμες πηγές στον κόσμο. Η αναζήτηση βασίστηκε στο σύστημα PICOTS με τη βοήθεια του οποίου διατυπώθηκε το ερευνητικό ερώτημα. Οι όροι αναζήτησης που χρησιμοποιήθηκαν για την εύρεση κατάλληλων μελετών ως προς τον τύπο του προγνωστικό μοντέλο ήταν οι “SAPS”,”SAPS II”,”SAPS2”,”SAPS 2”,”SAPS 3”,” Simplified Acute Physiology Score” και ταυτόχρονα εισήχθησαν οι όροι "prediction model" "predictive model", "risk score",“nomogram”, "clinical prediction model", "decision support models", "prediction rule", "prognostic model", "diagnostic model", “specificity and sensitivity” για την εύρεση μελετών αξιολόγησης των προβλεπτικών μοντέλων. Χρησιμοποιήθηκαν επίσης και δύο Mesh όροι που είναι ευρέως χρησιμοποιούμενο εργαλείο για την αναζήτηση στο PubMed και συγκεκριμένα οι "Simplified Acute Physiology Score"[Mesh] και “specificity and sensitivity” [Mesh].

Ακολούθησε η φάση διαλογής, η οποία πραγματοποιήθηκε σε δύο στάδια. Το πρώτο στάδιο έγινε με βάση τον τίτλο και την περίληψη και έπειτα σε όσα άρθρα επιλέχθηκαν στο πρώτο στάδιο έγινε περαιτέρω διαλογή με βάση το πλήρες κείμενο.

Κριτήρια επιλογής για τις μελέτες ήταν 1) να περιλαμβάνει ασθενείς που είχαν εισαχθεί σε κάποια ΜΕΘ, 2) να μελετά τη θνησιμότητά τους, 3) να κάνει εξωτερική επικύρωση του SAPS II είτε ως προς τη διακριτική του ικανότητα είτε ως προς τη βαθμονόμηση, ακόμα κι αν αυτό δεν ήταν ο πρωταρχικός στόχος της μελέτης, 4) να παρέχει πληροφορίες για τα μέτρα αξιολόγησης του μοντέλου και το τυπικό τους σφάλμα ή το 95% διάστημα εμπιστοσύνης τους, 5) να είναι προοπτικές ή αναδρομικές μελέτες και 6) να έχουν δημοσιευθεί μέσα στο 2022 λόγω χρονικών περιορισμών της παρούσας διπλωματικής εργασίας.

Τα κριτήρια αποκλεισμού των μελετών ήταν 1) να ήταν γραμμένες σε κάποια γλώσσα πέρα από την αγγλική, 2) να είχαν δημοσιευθεί πριν από το 2003 και 3) να μην ήταν διαθέσιμο το πλήρες κείμενό τους.

4.1.2 Εξαγωγή των δεδομένων

Η εξαγωγή των δεδομένων έγινε με οδηγό το CHARMS checklist που είναι ειδικά σχεδιασμένο για συστηματικές ανασκοπήσεις προγνωστικών προβλεπτικών μοντέλων (Palazón-Bru *et al.,* 2020). Τα δεδομένα που συλλέχθηκαν ήταν 1) η πηγή των δεδομένων, 2) το είδος της μελέτης, 3) ο ορισμός της έκβασης, 4) χαρακτηριστικά των ασθενών (ηλικία, τύπος της ΜΕΘ, τύπος ασθένειας), 5) μέγεθος δείγματος και αριθμός εκβάσεων, 6) διαχείριση ελλειπουσών τιμών και 7) τιμές των μέτρων διακριτικής ικανότητας και βαθμονόμησης είχαν αναφερθεί στις μελέτες. Τα μέτρα διακριτικής ικανότητας περιελάμβαναν το εμβαδό της καμπύλης ROC και τη διακριτική κλίση, ενώ τα πιθανά μέτρα βαθμονόμησης ήταν η σταθερά βαθμονόμησης, η κλίση βαθμονόμησης, η αναλογία των παρατηρούμενων και των αναμενόμενων εκβάσεων και το Hosmer-Lemeshow test. Επιπλέον καταγράφηκαν συνολικά μέτρα όπως το Brier σκορ και το R squared και πληροφορίες για όσες μελέτες είχαν κάνει ανάλυση της καμπύλης απόφασης (Decision curve analysis, DCA).

4.1.3 Εκτίμηση της ποιότητας των μελετών και του κινδύνου συστηματικού σφάλματος

Η ποιότητα των επιλεγμένων μελετών εκτιμήθηκε με το εργαλείο PROBAST που αξιολογεί τον κίνδυνο συστηματικού σφάλματος και την καταλληλότητα τους. Για τον κίνδυνο συστηματικού σφάλματος λήφθηκαν υπόψη τέσσερεις τομείς (συμμετέχοντες, προβλεπτικοί παράγοντες, έκβαση και ανάλυση), ενώ για την καταλληλότητα μόνο οι πρώτοι τρεις. Κάθε τομέας βαθμολογείται με χαμηλό, ασαφές ή υψηλό κίνδυνο και ανάλογα με τις επιμέρους βαθμολογίες προκύπτει και η συνολική βαθμολογία. Συγκεκριμένα εάν όλοι οι τομείς βαθμολογηθούν με χαμηλό κίνδυνο τότε ο συνολικός κίνδυνος θα είναι χαμηλός, ενώ εάν υπάρχει έστω κι ένας τομέας με ασαφή ή υψηλό κίνδυνο τότε ο συνολικός κίνδυνος θα χαρακτηριστεί ως ασαφής ή υψηλός αντίστοιχα. (Wolff *et al.,* 2019).

4.1.4 Μεθοδολογία μετα-ανάλυσης

Πριν την διεξαγωγή της μετα-ανάλυσης ήταν απαραίτητος ο υπολογισμός των τυπικών σφαλμάτων και των 95% διαστημάτων εμπιστοσύνης των AUROC εκτιμήσεων, σε όσες μελέτες δεν είχαν δοθεί απευθείας αυτά τα μέτρα. Πιο συγκεκριμένα, 2 μόνο μελέτες ανέφεραν και το τυπικό σφάλμα και το 95% διάστημα εμπιστοσύνης (Krasselt *et al.*, 2022; Wang *et al.*, 2022), 6 δεν ανέφεραν κανένα από τα δύο (Hai and Viet Hoa, 2022; Mirzakhani *et al.*, 2022; Qi *et al.*, 2022; Rong *et al.*, 2022; Álvarez-Avello *et al.*, 2022; Zou *et al.*, 2022), ενώ οι υπόλοιπες 17 έδιναν μόνο το 95% διάστημα εμπιστοσύνης. Το AUROC μετασχηματίστηκε σε logit κλίμακα, ώστε να εξασφαλιστεί η προϋπόθεση της κανονικής κατανομής μεταξύ των μελετών (Snell *et al.*, 2018; Debray *et al.*, 2019). Ο υπολογισμός του τυπικού σφάλματος έγινε με τα όρια του διαστήματος εμπιστοσύνης με τον τύπο (3.3) , αλλά σε περίπτωση που έλειπε αυτή η πληροφορία ο υπολογισμός έγινε με τη βοήθεια του μεγέθους δείγματος και τον αριθμό των εκβάσεων με τον τύπο (3.2)(Debray *et al.*, 2019).

Στη συνέχεια εφαρμόστηκε μοντέλο τυχαίων επιδράσεων των μετασχηματισμένων AUROC εκτιμήσεων και οι στάθμες τους υπολογίστηκαν με τη μέθοδο της αντίστροφης διακύμανσης. Επίσης, χρησιμοποιήθηκε ο restricted maximum likelihood (REML) εκτιμητής για τον υπολογισμό του τ^2 και πραγματοποιήθηκε διόρθωση του τυπικού σφάλματος της συνοπτικής εκτίμησης με τη μέθοδο των Hartung-Knapp. Μερικές μελέτες εξέταζαν πάνω από μία έκβαση θανάτου σε διαφορετικούς υποπληθυσμούς, οπότε αυτές συνείσφεραν περισσότερες από μία μετρήσεις για την ανάλυση. Η κατανομή των εκτιμήσεων της AUROC στις επιμέρους μελέτες και η συνοπτική εκτίμηση με τα αντίστοιχα 95% διαστήματα εμπιστοσύνης και πρόβλεψης και τα μέτρα ετερογένειας παρουσιάστηκαν με τη βοήθεια των διαγραμμάτων δάσους (forest plots).

Για τη διερεύνηση της ετερογένειας μεταξύ των μελετών έγιναν δύο αναλύσεις υπο-ομάδων, για να διερευνηθεί ενδεχόμενη διαφοροποίηση σε σχέση με το είδος της έκβασης και με το συνολικό κίνδυνο για συστηματικό σφάλμα, αντίστοιχα. Αρχικά υπολογίστηκε μια συνοπτική εκτίμηση του logit(AUC) ανά υποομάδα με τη μέθοδο των τυχαίων επιδράσεων και έπειτα υπολογίστηκε το στατιστικό κριτήριο Q για να ελεγχθεί εάν υπάρχει διαφορά μεταξύ των εκτιμήσεων των υποομάδων (συγκρίνοντας με την οριακή τιμή της χ^2 κατανομής με G-1 βαθμούς ελευθερίας, όπου G ο αριθμός των υποομάδων (εδώ G = 3 για τις υπο-ομάδες ανά έκβαση και G =3 ανά κατηγορία κινδύνου

σφάλματος) (Harrer *et al.*, 2022). Επειδή στην ανάλυση των ανά έκβαση υπήρχε μεγάλη ανομοιομορφία μεταξύ των ομάδων, υπολογίστηκε και χρησιμοποιήθηκε ένα συνοπτικό τ^2 που ήταν κοινό για όλες τις ομάδες. Αντίθετα για την ανάλυση των υποομάδων ανά κατηγορία κινδύνου για συστηματικό σφάλμα, οι υποομάδες δεν ήταν τόσο ανομοιόμορφες και υποτέθηκε ότι η ετερογένεια μεταξύ των μελετών θα είναι διαφορετική σε κάθε υποομάδα.

Πέρα από την ανάλυση υπο-ομάδων πραγματοποιήθηκαν και αναλύσεις ευαισθησίας αποκλείοντας κάθε φορά πιθανές μελέτες με υψηλό κίνδυνο για συστηματικό σφάλμα και μελέτες με πολύ μικρά ή πολύ μεγάλα μεγέθη δείγματος.

Ο κίνδυνος για συστηματικό σφάλμα δημοσίευσης εξετάστηκε με την κατασκευή funnel plot και με τον έλεγχο του Egger. Το funnel plot σε απουσία συστηματικού σφάλματος υποθέτει ότι μελέτες με υψηλή ακρίβεια θα συγκεντρώνονται κοντά στην τιμή του συνθετικού δείκτη και μελέτες με μικρότερη ακρίβεια θα κατανέμονται συμμετρικά γύρω από το συνθετικό δείκτη, δίνοντας έτσι στο γράφημα ένα περίπου τριγωνικό σχήμα. Απόκλιση από το σχήμα αυτό είναι ένδειξη παρουσίας συστηματικού σφάλματος δημοσίευσης. Όσον αφορά τον έλεγχο του Egger, αυτός αποτελεί ουσιαστικά μία γραμμική παλινδρόμηση των εκτιμήσεων της επίδρασης προς τα τυπικά σφάλματα τους σταθμισμένων με την αντίστροφη διακύμανσή τους. Εάν η γραμμή παλινδρόμησης διέρχεται από την αρχή των αξόνων αυτό σημαίνει απουσία συστηματικού σφάλματος δημοσίευσης.

Η ανάλυση των δεδομένων πραγματοποιήθηκε στην έκδοση 4.0.3 της R.

4.2 ΑΠΟΤΕΛΕΣΜΑΤΑ

4.2.1 Αποτελέσματα αναζήτησης

Η βιβλιογραφική αναζήτηση στη βάση PubMed έδωσε αρχικά 728 αποτελέσματα (Γράφημα 4.1). Από αυτά, μέσω της διαλογής βάσει τίτλου και περίληψης αποκλείστηκαν συνολικά 422 (58%) άρθρα. Ανάμεσα σε αυτά υπήρχαν 8 (1.9%) ανασκοπήσεις, 105 (24.9%) άρθρα οποία ήταν δημοσιευμένα πριν το 2003, ένα(0.2%) είχε αποσυρθεί και τα υπόλοιπα 308(73%) δεν ήταν σχετικά με το ερευνητικό ερώτημα. Εν συνεχεία, πραγματοποιήθηκε διαλογή βασισμένη στο πλήρες κείμενο σε 306 άρθρα,

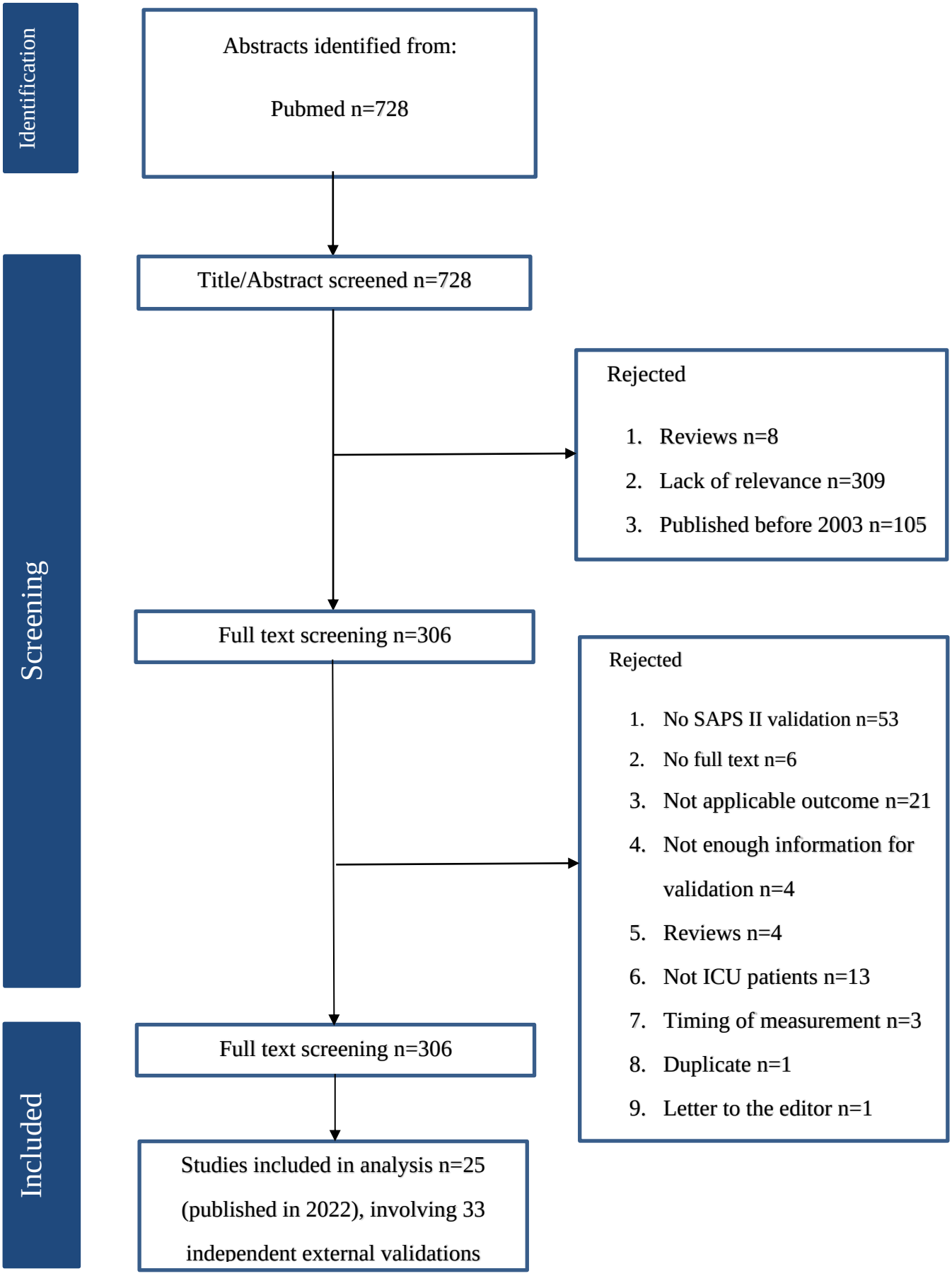
εκ των οποίων κατάλληλα κρίθηκαν τα 200(65.4%). Σχεδόν στο σύνολό τους, το μέτρο αξιολόγησης της διακριτικής ικανότητας των προβλεπτικών μοντέλων που χρησιμοποίησαν ήταν το AUROC, ενώ αντίθετα τα περισσότερα άρθρα δεν αξιολόγησαν καθόλου την βαθμονόμηση των μοντέλων με κάποιο από τα διαθέσιμα μέτρα. Στις περιπτώσεις που υπήρχε αξιολόγηση της βαθμονόμησης αυτό θα γινόταν κυρίως είτε με τον έλεγχο του Hosmer-Lemeshow είτε με την αναλογία παρατηρούμενων προς αναμενόμενων εκβάσεων, που δεν είναι και τα πιο ακριβή και αξιόπιστα μέτρα. Από τα 106 άρθρα που απορρίφθηκαν στο δεύτερο στάδιο της διαλογής, τα 53(50%) δεν έκαναν αξιολόγηση του SAPS II μοντέλου, 6(5.7%) δεν είχαν διαθέσιμο το πλήρες κείμενο, 4(3.8%) ήταν ανασκοπήσεις, 13(12.3%) δεν χρησιμοποίησαν δείγμα ασθενών που να είχαν εισαχθεί σε κάποια ΜΕΘ, 21(19.8%) είχαν έκβαση που δεν αντιστοιχούσε στο ερευνητικό ερώτημα, 3(2.8%) δεν είχαν τη ζητούμενη εκτίμηση του SAPS II σκορ στο ζητούμενο χρονικό πλαίσιο, 4(3.8%) είχαν ελλιπείς πληροφορίες για την αξιολόγηση του μοντέλου, 1(0.9%) ήταν γράμμα στον συντάκτη και 1(0.9%) ήταν διπλότυπο.

Από τις 200 επιλέξιμες μελέτες που προέκυψαν από την αναζήτηση και την ολοκλήρωση της διαλογής, στην παρούσα εργασία χρησιμοποιήθηκαν μόνο οι μελέτες του τελευταίου έτους ώστε να είναι εφικτή η διαχείριση του όγκου της εργασίας, δηλαδή 25 μελέτες. Μερικές μελέτες εξέταζαν την αξιολόγηση του μοντέλου για πάνω από μία πιθανές εκβάσεις και κάποιες άλλες αξιολογούσαν το μοντέλο σε δύο διαφορετικά δείγματα εξωτερικής επικύρωσης. Οι εκβάσεις στις οποίες αναφερόταν το SAPS II σκορ και οι οποίες συμπεριλήφθηκαν στη ανάλυση ήταν η θνησιμότητα εντός 28 ημερών, η θνησιμότητα εντός 30 ημερών, η νοσοκομειακή θνησιμότητα και η θνησιμότητα στις ΜΕΘ. Η θνησιμότητα εντός 28 ημερών και η θνησιμότητα εντός 30 ημερών εντάχθηκαν στην ίδια κατηγορία στο στάδιο της ανάλυσης υπο-ομάδων βάσει της έκβασης. Υπήρχε μία εκτίμηση για τη θνησιμότητα εντός 72 ωρών και μία για την ετήσια θνησιμότητα, αλλά αυτές αποκλείστηκαν από την ανάλυση. Τελικά, η μετα-ανάλυση πραγματοποιήθηκε σε 33 διαφορετικές και ανεξάρτητες παρατηρήσεις από 25 μελέτες.

4.2.2 Χαρακτηριστικά των μελετών

Τα κύρια χαρακτηριστικά των 25 μελετών που εξετάστηκαν συνοψίζονται στον πίνακα 4.2. Από τις 25 μελέτες οι 3 (12%) ήταν προοπτικές μελέτες κοορτής και οι υπόλοιπες

Γράφημα 4.1: Στρατηγική αναζήτησης και διαδικασία διαλογής



22(88%) ήταν αναδρομικές μελέτες κοορτής. Σε 13(52%) μελέτες τα δεδομένα συλλέχθηκαν από δημόσια προσβάσιμες βάσεις δεδομένων, σε 8(32%) χρησιμοποιήθηκαν αναδρομικά δεδομένα από εθνικά και πανεπιστημιακά νοσοκομεία και κέντρα υγείας, σε 3 (12%) τα δεδομένα προέρχονταν από κοορτές πανεπιστημιακών και στρατιωτικών νοσοκομείων και μία (4%) μελέτη πήρε δεδομένα και από δημόσια βάση δεδομένων και από νοσοκομείο. Όσον αφορά την χώρα προέλευσης των δεδομένων υπήρχε μεγάλη ετερογένεια., με τις μελέτες να έχουν συλλέξει δεδομένα από συνολικά 13 χώρες, με επικρατέστερες τις ΗΠΑ (n=11 μελέτες, 44%). Μία εκ των μελετών ήταν πολυεθνική σε 3 χώρες (Δανία, Σουηδία και Νορβηγία). Η πλειοψηφία των μελετών (n= 18, 72%) είχαν ως στόχο τη σύγκριση της απόδοσης μεταξύ διαφορετικών προβλεπτικών μοντέλων, συμπεριλαμβανομένων του SAPS II. Υπήρχαν όμως και 3(12%) μελέτες που διερευνούσαν παράγοντες κινδύνου της θνησιμότητας, 2(8%) ήταν μελέτες εξωτερικής επικύρωσης προβλεπτικών μοντέλων θνησιμότητας συμπεριλαμβανομένου και του SAPS II, μία(4%) διερευνούσε την προγνωστική αξία βιοδεικτών και μία(4%) έκανε σύγκριση της δυνατότητας κλινικής εφαρμογής μεταξύ διάφορων μοντέλων.

4.2.3 Συμμετέχοντες

Οι κοορτές των ασθενών στις διάφορες μελέτες παρουσίαζαν μεγάλη ετερογένεια ως προς το μέγεθος τους, αφού κυμαίνονταν από 39 έως 61589 ασθενείς με διάμεσο τους 616 ασθενείς. Αντίστοιχη διακύμανση τιμών παρατηρήθηκε και στον αριθμό των εκβάσεων που κυμάνθηκε από 18 έως 7506 θανάτους (διάμεσος 132.5), έχοντας όμως 5 ελλείπουσες τιμές για τον παρατηρούμενο αριθμό θανάτων. Στις 21(84%) μελέτες οι ασθενείς προέρχονταν από γενικές/μικτές ΜΕΘ (ICU), ενώ 2 (8%) ήταν από ιατρικές ΜΕΘ (Medical ICU, MICU) και 2 (8%) από μονάδες στεφανιαίας φροντίδας (Coronary Care Unit,CCU). Στις 5 (20%) μελέτες χρησιμοποιήθηκαν δεδομένα από ασθενείς που είχαν νοσηλευτεί για οποιαδήποτε αιτία σε ΜΕΘ, ενώ στις υπόλοιπες 20 (80%) τα δεδομένα προέρχονταν από ασθενείς με συγκεκριμένη παθολογία: ενδοκρανιακή αιμορραγία (Zou *et al.*, 2022), οξεία νεφρική βλάβη (Wang *et al.*, 2022; Wu *et al.*, 2022), βακτηριαιμία (Tokur *et al.*, 2022), καρδιογενές σοκ (Choi *et al.*, 2022; Rong *et al.*, 2022; Álvarez-Avello *et al.*, 2022), σήψη (Hai and Viet Hoa, 2022; J. Liu *et al.*, 2022; Moreno-Torres *et al.*, 2022), θρομβοκυτταροπενία (Lu, Zhang and Jiang, 2022), οξεία

παγκρεατίτιδα (Z. Liu *et al.*, 2022), γαστρεντερική εκτομή (Qi *et al.*, 2022), νεκρωτικές βλάβες (Katz *et al.*, 2022), ασθένειες του συνδετικού ιστού (Krasselt *et al.*, 2022), ηπατικές σηψαιμικές βλάβες (Y. Liu *et al.*, 2022), καρδιακή ανεπάρκεια ή χρόνιες νεφρικές ασθένειες (Chen *et al.*, 2022), αναπνευστικές παθήσεις (Han *et al.*, 2022; Ren *et al.*, 2022) και κρίσιμες παθολογικές καταστάσεις (Mirzakhani *et al.*, 2022). Οι μέσες τιμές της ηλικίας των ασθενών μαζί με τις αντίστοιχες τυπικές αποκλίσεις αναφέρθηκαν μόνο σε 6 μελέτες με ελάχιστη τιμή τα 52 έτη, μέγιστη τα 66,5 έτη και διάμεσο τα 61,4 έτη. Σε 11 μελέτες, οι ερευνητές ανέφεραν τις διάμεσες τιμές της ηλικίας και τα ενδοτεταρτημοριακά τους εύρη. Οι διάμεσοι κυμαίνονταν από 59 έως 77,7 έτη, ενώ 8 μελέτες δεν ανέφεραν κανένα περιγραφικό μέτρο της ηλικίας των ασθενών.

4.2.4 Δείκτης υγείας

Οι 23 στους 33 δείκτες υγείας (69.7%) αναφέρονταν στην ενδονοσοκομειακή θνησιμότητα, 5 (15.2%) στη θνησιμότητα εντός 30 ημερών, 3 (9.1%) στη θνησιμότητα εντός 28 ημερών και 2 (6%) στη θνησιμότητα εντός της ΜΕΘ. Ο υπολογισμός του σκορ σε 20 (60.6%) περιπτώσεις έγινε εντός του πρώτου εικοσιτετράωρου από την εισαγωγή των ασθενών στη ΜΕΘ, ενώ σε μία έγινε εντός του πρώτου εικοσιτετράωρου από το καρδιακό επεισόδιο των ασθενών. Στις υπόλοιπες 12 εκτιμήσεις δεν αναφέρθηκε σαφώς ο ακριβής χρόνος υπολογισμού του SAPS II σκορ.

Πίνακας 4.2Α: Χαρακτηριστικά των μελετών της μετα-ανάλυσης

Author	Aim type	Study Design	Data source	Country	Study date
Zou et al.	Clinical applicability comparison	Retrospective Cohort Study	MIMIC-III	USA	2001- 2012
Wu et al.	Predictive performance comparison	Retrospective Cohort Study	MIMIC-III	USA	2001- 2012
Wang et al.	Predictive performance comparison	Retrospective Cohort Study	Beijing Acute Kidney Injury Trial (BAKIT)	China	1/03/2012-31/08/2012
Tokur et al.	Investigate risk factors	Retrospective Cohort Study	A Training and Research Hospital	Turkey	01/2014-03/2020
Rong et al.	Predictive performance comparison	Retrospective Cohort Study	MIMIC-III,Wenzhou Medical University Hospital	USA,China	
Ren et al.	Predictive performance comparison	Retrospective Cohort Study	MIMIC-III	USA	
Rahmatinejad et al.	Predictive performance comparison	Retrospective Cohort Study	Cohorts from 5 hospitals	Iran	08/2018-08/2020
Qi et al.	Predictive performance comparison	Retrospective Cohort Study	MIMIC-III	USA	2002- 2012
Nistal-Nuño	Predictive performance comparison	Retrospective Cohort Study	MIMIC-III	USA	2001- 2012
Moser et al.	Predictive performance comparison	Retrospective Cohort Study	Finnish Intensive Care Consortium (FICC)	Finland	2015- 2017
Moreno-Torres et al.	Investigate risk factors	Retrospective Cohort Study	Spanish tertiary University Hospital	Spain	1/01/2019-31/12/2019
Mirzakhan et al.	Predictive performance comparison	Retrospective Cohort Study	Razi Educational and Medical Center of Ghaemshahr	Iran	03/2017-09/2019
Lu et al	Predictive performance comparison	Retrospective Cohort Study	MIMIC-IV	USA	2008- 2019
Zhanxiao Liu et al.	Predictive performance comparison	Retrospective Cohort Study	MIMIC-III	USA	2001- 2012
Yousheng Liu et al.	Predictive performance comparison	Retrospective Cohort Study	MIMIC-IV	USA	2008- 2019
Jie Liu et al.	Investigate risk factors	Retrospective Cohort Study	MIMIC-III	USA	2001- 2012
Krasselt et al.	Predictive performance comparison	Retrospective Cohort Study	University Hospital of Leipzig	Germany	2006- 2019
Katz et al.	Predictive performance comparison	Retrospective Cohort Study	INFECT dataset	Denmark,Sweden,Norway	02/2013-06/2017
Kahraman et al.	External validation	Retrospective Cohort Study*	Tertiary referral hospital	Turkey	1/10/2019-31/05/2020
Han et al.	Predictive performance comparison	Retrospective Cohort Study	MIMIC-III	USA	
Hai and Viet Hoa	Predictive performance comparison	Prospective Cohort Study*	108 Military Central Hospital	Vietnam	12/2016-12/2018
Csiszar et al.	Investigate biomarker's prognostic value	Prospective Cohort Study	Three ICU units at the University of Pécs	Hungary	01/2018-01/2019
Choi et al.	Predictive performance comparison	Retrospective Cohort Study	EHR of Severance Hospital in South Korea	South Korea	2006- 2020
Chen et al.	Predictive performance comparison	Retrospective Cohort Study	MIMIC-IV	USA	2008- 2019
Álvarez-Avello et al.	External validation	Prospective Cohort Study	A cohort of CS patients	Spain	09/2014-01/2019

*Οι συγγραφείς ανέφεραν τη μελέτη αυτή ως συγχρονική

Πίνακας 4.2Β: Χαρακτηριστικά των μελετών της μετα-ανάλυσης

Author	*Patients*	*Outcome*	*Timing of measurement*	*Sample size*	*Events*	*Mean age*	*SD*	*Median age*	*IQR lower*	*IQR upper*
Zou et al.	Intracerebral hemorrhage(ICH)	30-day mortality	Not recorded	623	199			71	58	81
Zou et al.	Intracerebral hemorrhage(ICH)	30-day mortality	Not recorded	267	99			70	59	80
Wu et al.	Acute kidney injury(AKI)	In-hospital mortality	Within 24h from ICU admission	24166	7805			68	56	78
Wang et al.	Acute kidney injury(AKI)	28-day mortality	Not recorded	1506	407			67	53	78
Tokur et al.	A. baumannii bacteremia(BD)	28-day mortality	Within 24h from ICU admission	39	25			72	58	84
Rong et al.	Cardiogenic shock(CS)	30-day mortality	Within 24h from ICU admission	804	304					
Rong et al.	Cardiogenic shock(CS)	30-day mortality	Within 24h from ICU admission	115	62					
Ren et al.	Sepsis & Lung infection	In-hospital mortality	Not recorded	1173						
Ren et al.	Sepsis & Lung infection	In-hospital mortality	Not recorded	503						
Rahmatinejad et al.	ICU	In-hospital mortality	Within 24h from ICU admission	3455	916	56,65	22			
Qi et al.	Post-gastrointestinal resection surgery	In-hospital mortality	Not recorded	596	82	63,05	16			
Qi et al.	Post-gastrointestinal resection surgery	In-hospital mortality	Not recorded	199	26	64,82	16			
Nistal-Nuño	ICU	ICU mortality	Within 24h from ICU admission	1979						
Moser et al.	ICU	In-hospital mortality	Not recorded	61224	6463			63		
Moreno-Torres et al.	Sepsis	In-hospital mortality	Within 24h from ICU admission	203	64	63,1	14			
Mirzakhan et al.	Burn/CAD/heart surgery/brain dead	In-hospital mortality	Within 24h from ICU admission	840	333	66,49	18			

Πίνακας 4.2Β(συν.): Χαρακτηριστικά των μελετών της μετα-ανάλυσης

Author	*Patients*	*Outcome*	*Timing of measurement*	*Sample size*	*Events*	*Mean age*	*SD*	*Median age*	*IQR lower*	*IQR upper*
Lu et al	Severe thrombocytopenia	In-hospital mortality	Within 24h from ICU admission	749	344					
Lu et al	Severe thrombocytopenia	In-hospital mortality	Within 24h from ICU admission	302	166					
Zhanxiao Liu et al.	Acute pancreatitis (AP)	In-hospital mortality	Within 24h from ICU admission	410	56			59	46,93	71,61
Zhanxiao Liu et al.	Acute pancreatitis (AP)	In-hospital mortality	Within 24h from ICU admission	221	32			61,2	47,41	74,3
Yousheng Liu et al.	Sepsis-associated liver injury (SALI)	In-hospital mortality	Not recorded	616						
Yousheng Liu et al.	Sepsis-associated liver injury (SALI)	In-hospital mortality	Not recorded	154						
Jie Liu et al.	Sepsis	28-day mortality	Within 24h from ICU admission	3367	960			66	55	77
Krasselt et al.	Connective tissue diseases(CTD)	In-hospital mortality	On ICU admission	44	18	59,8	16			
Katz et al.	Necrotizing soft-tissue infections (NSTI)	30-day mortality	Within 24h from ICU admission	405	56					
Kahraman et al.	CCU	In-hospital mortality	Within 24h from ICU admission	871	83			66	58	75
Han et al.	Ventilator-associated pneumonia(VAP)	In-hospital mortality	Not recorded	1984	353					
Hai and Viet Hoa	Sepsis	In-hospital mortality	Within 24h from ICU admission	194	73			69	59	80
Csiszar et al.	IHCA or OHCA	ICU mortality	Within 24h from cardiac arrest	54	33					
Choi et al.	ICU	In-hospital mortality	Within 24h from ICU admission	61589	7276			67	57	74
Choi et al.	ICU	In-hospital mortality	Within 24h from ICU admission	23557	4099			65	53	75
Chen et al.	CHF and CKD	In-hospital mortality	Not recorded	4638	707			77,7	68,7	85,2
Álvarez-Avello et al.	Cardiogenic shock (CS)	In-hospital mortality	Within 24h from ICU admission	130	56	52	15			

SD: τυπική απόκλιση της ηλικίας

IQR (lower/upper): ενδοτεταρτομοριακό εύρος (Q1/Q3)

4.2.5 Μέτρα αξιολόγησης του SAPS II

Τα μέτρα αξιολόγησης που μελετήθηκαν δίνονται στους πίνακες 4.3 και 4.4. Ως μέτρο διακριτικής ικανότητας, ο συντελεστής προσδιορισμού (R^2) δεν αναφέρθηκε σε καμία μελέτη, ενώ το Brier σκορ δόθηκε μόνον σε 3 (12%) μελέτες (Moser *et al.*, 2022; Nistal-Nuño, 2022; Rahmatinejad *et al.*, 2022) χωρίς όμως να συνοδεύεται με τυπικό σφάλμα ή 95% όρια εμπιστοσύνης. Αρκετές μελέτες έκαναν αναφορά στην ευαισθησία (sensitivity) και την ειδικότητα (specificity), που είναι μέτρα ταξινομικής ικανότητας του μοντέλου. Συνολικά, υπήρξαν 13 παρατηρήσεις για την ευαισθησία (διάμεσος 0,66, εύρος 0,50-0,84) και 13 παρατηρήσεις για την ειδικότητα (διάμεσος 0,748, εύρος 0,44-0,918). Παρόλο που οι τιμές της ευαισθησίας και της ειδικότητας έχουν νόημα μόνο σε σχέση με ένα συγκεκριμένο κατώφλι του SAPS II, δεν ανέφεραν όλες οι μελέτες ποιο ήταν το κατώφλι που χρησιμοποίησαν ή με ποιον τρόπο το υπολόγισαν. Μία μέθοδος που εφαρμόστηκε σε 4 παρατηρήσεις ήταν η μεγιστοποίηση του δείκτη του Youden, δηλαδή του αθροίσματος της ειδικότητας και της ευαισθησίας, θεωρώντας ότι οι δύο αυτοί δείκτες είναι εξίσου σημαντικοί. Σε 5 παρατηρήσεις δεν αναφερόταν η μέθοδος υπολογισμού, σε 1 έγινε βάσει του αντισταθμίσματος ευαισθησίας και ειδικότητας, ενώ 3 παρατηρήσεις δεν ανέφεραν ποιο ήταν το κατώφλι που χρησιμοποιήθηκε

Όλες οι μελέτες αξιολόγησαν το SAPS II ως προς τη διακριτική του ικανότητα μέσω του εμβαδού της καμπύλης ROC. Δύο (6%) παρατηρήσεις συνοδεύτηκαν από το τυπικό τους σφάλμα, ενώ οι 24 (72.7%) παρείχαν το 95% διάστημα εμπιστοσύνης του. Από την άλλη πλευρά καμία μελέτη δεν έκανε αναφορά στη κλίση διακριτικής ικανότητας.

Τα μέτρα αξιολόγησης της βαθμονόμησης απουσίαζαν σχεδόν ολοκληρωτικά από τα υπό μελέτη άρθρα. Μόνο μία μελέτη εκτίμησε τη σταθερά και την κλίση βαθμονόμησης και τα 95% όρια αξιοπιστίας τους (Moser *et al.*, 2022) και μία άλλη είχε τη μοναδική εκτίμηση της αναλογίας παρατηρούμενων προς αναμενόμενων εκβάσεων χωρίς όμως να κάνει λόγο για τυπικό σφάλμα ή 95% όρια εμπιστοσύνης (Kahraman *et al.*, 2022). Υπήρχαν και τέσσερις μελέτες που χρησιμοποίησαν τον έλεγχο των Hosmer-Lemeshow για μία αδρή αξιολόγηση της βαθμονόμησης (Alvarez *et al.*, 1998; Kahraman *et al.*, 2022; Moser *et al.*, 2022; Rahmatinejad *et al.*, 2022).

Πίνακας 4.3: Πληροφορίες για τα συνολικά μέτρα αξιολόγησης των προβλεπτικών μοντέλων και τα μέτρα διακριτικής τους ικανότητας

Author	*Brier score*	*Sensitivity*	*Specificity*	*Cutoff point*	*AUC*	*SE*	*95% CI lb*	*95% CI ub*	*logitAUC*	*SE*	*95% CI lb*	*95% CI ub*
Zou et al.					0,728	0,022	0,685	0,770	0,983	0,109	0,770	1,196
Zou et al.					0,733	0,032	0,671	0,795	1,011	0,162	0,694	1,328
Wu et al.		0,704	0,44	47	0,598	0,011	0,576	0,620	0,397	0,047	0,306	0,490
Wang et al.		0,6137	0,806	39	0,767	0,015	0,739	0,796	1,191	0,084	1,041	1,361
Tokur et al.		0,84	0,86	55,5	0,890	0,043	0,820	0,990	2,091	0,785	1,516	4,595
Rong et al.					0,677	0,019	0,639	0,715	0,740	0,089	0,566	0,914
Rong et al.					0,724	0,047	0,631	0,817	0,964	0,237	0,499	1,429
Ren et al.					0,707	0,019	0,668	0,741	0,881	0,090	0,699	1,051
Ren et al.					0,664	0,026	0,613	0,715	0,681	0,117	0,460	0,920
Rahmatinejad et al.	0,17	0,662	0,73	44,5	0,767	0,010	0,750	0,790	1,191	0,058	1,099	1,325
Qi et al.					0,784	0,027	0,731	0,838	1,290	0,161	0,975	1,605
Qi et al.					0,732	0,053	0,629	0,835	1,006	0,268	0,481	1,532
Nistal-Nuño	0,143	0,771	0,679	44	0,793	0,014	0,766	0,820	1,343	0,084	1,186	1,516
Moser et al.	0,1				0,864	0,002	0,860	0,869	1,849	0,020	1,815	1,892
Moreno-Torres et al.					0,734	0,038	0,660	0,809	1,015	0,199	0,663	1,444
Mirzakhan et al.		0,6726	0,7337		0,771	0,017	0,739	0,803	1,214	0,094	1,030	1,398
Lu et al					0,766	0,025	0,718	0,815	1,186	0,140	0,935	1,483
Lu et al					0,789	0,021	0,747	0,831	1,319	0,130	1,083	1,593

Πίνακας 4.3(συν.): Πληροφορίες για τα συνολικά μέτρα αξιολόγησης των προβλεπτικών μοντέλων και τα μέτρα διακριτικής τους ικανότητας

Author	Brier score	Sensitivity	Specificity	Cutoff point	AUC	SE	95% CI lb	95% CI ub	logitAUC	SE	95% CI lb	95% CI ub
Zhanxiao Liu et al.		0,607	0,749		0,725	0,035	0,656	0,794	0,969	0,180	0,646	1,349
Zhanxiao Liu et al.		0,594	0,831		0,792	0,044	0,706	0,879	1,337	0,282	0,876	1,983
Yousheng Liu et al.					0,612	0,023	0,567	0,657	0,456	0,097	0,270	0,650
Yousheng Liu et al.					0,629	0,046	0,537	0,718	0,528	0,201	0,148	0,935
Jie Liu et al.		0,503	0,679	49	0,610	0,011	0,589	0,631	0,447	0,045	0,360	0,537
Krasselt et al.		0,73	0,78	47	0,772	0,084	0,608	0,937	1,220	0,477	0,439	2,700
Katz et al.					0,88	0,023	0,830	0,920	1,992	0,219	1,586	2,442
Kahraman et al.		0,8	0,918	5,55	0,908	0,020	0,869	0,947	2,289	0,253	1,892	2,883
Han et al.		0,652	0,722	0,173	0,76	0,014	0,733	0,787	1,153	0,076	1,010	1,307
Hai and Viet Hoa		0,616	0,777	48	0,73	0,037	0,657	0,803	0,995	0,189	0,625	1,364
Csiszar et al.					0,747	0,074	0,602	0,891	1,083	0,430	0,414	2,101
Choi et al.				45	0,766	0,003	0,759	0,772	1,186	0,018	1,147	1,220
Choi et al.				44	0,792	0,004	0,784	0,799	1,337	0,023	1,289	1,380
Chen et al.					0,747	0,010	0,726	0,767	1,083	0,055	0,974	1,191
Álvarez-Avello et al.					0,7516	0,043	0,667	0,836	1,107	0,231	0,655	1,559

AUC: εμβαδό της καμπύλης ROC

SE: τυπικό σφάλμα

95% CI lb/ub: 95% όρια αξιοπιστίας (κατώτατο/ανώτατο όριο)

Πίνακας 4.4: Πληροφορίες για τα μέτρα βαθμονόμησης των μοντέλων και για την ανάλυση καμπύλης απόφασης

Author	*Calibration intercept*	*95% CI Lower Bound*	*95% CI Upper Bound*	*Calibration Slope*	*95% CI Lower Bound*	*95% CI Upper Bound*	*Hosmer-Lemeshow Test*	*p-value*	*Total O:E ratio*	*Decision Curve Analysis*	*Positive Net Benefit Range*
Zou et al.										Yes	(0.2,0.7)
Zou et al.										Yes	(0.2,0.9)
Wu et al.										No	
Wang et al.										No	
Tokur et al.										No	
Rong et al.										No	
Rong et al.										No	
Ren et al.										Yes	(0.2,0.6)
Ren et al.										Yes	(0.2,0.4)
Rahmatinejad et al.								0.073		No	
Qi et al.										No	
Qi et al.										No	
Nistal-Nuño										No	
Moser et al.	-1,419	-1,45	-1,388	0,872	0,853	0,891		<0.001		No	
Moreno-Torres et al.										No	

Πίνακας 4.4(συν): Πληροφορίες για τα μέτρα βαθμονόμησης των μοντέλων και για την ανάλυση καμπύλης απόφασης

Author	*Calibration intercept*	*95% CI Lower Bound*	*95% CI Upper Bound*	*Calibration Slope*	*95% CI Lower Bound*	*95% CI Upper Bound*	*Hosmer-Lemeshow Test*	*p-value*	*Total O:E ratio*	*Decision Curve Analysis*	*Positive Net Benefit Range*
Mirzakhan et al.										No	
Lu et al										Yes	(0.2,0.8)
Lu et al										Yes	(0.2,0.7)
Zhanxiao Liu et al.										Yes	
Zhanxiao Liu et al.										Yes	
Yousheng Liu et al.										Yes	(0.35,0.55)
Yousheng Liu et al.										Yes	(0.35,0.63)
Jie Liu et al.										No	
Krasselt et al.										No	
Katz et al.										No	
Kahraman et al.							7,668	0.746	1,88	No	
Han et al.										No	
Hai and Viet Hoa										No	
Csiszar et al.										No	
Choi et al.										No	
Choi et al.										No	
Chen et al.										Yes	(0.1,0.7)
Álvarez-Avello et al.								0.0783		No	

Σε 9 (2.3%) παρατηρήσεις είχε γίνει ανάλυση καμπύλης απόφασης ώστε να εκτιμηθεί η κλινική αξία του SAPS II και για το εύρος των κατωφλιών για τα οποία η χρήση του μοντέλου είναι πιο αποδοτική από τις ακραίες υποθέσεις ότι όλοι οι ασθενείς έχουν την έκβαση ή ότι δεν την έχει κανένας.

Όσον αφορά τις ελλείπουσες τιμές, παρατηρήθηκαν πολλές εναλλακτικές ως προς τον τρόπο διαχείρισης τους. Η μέθοδος του πολλαπλού καταλογισμού (multiple imputation) εφαρμόστηκε σε 10 (30.3%) παρατηρήσεις, ενώ του διάμεσου καταλογισμού για δύο (6.1%) παρατηρήσεις. Σε δύο (6.1%) περιπτώσεις διεγράφησαν οι παρατηρήσεις των ασθενών με ελλείπουσες τιμές (listwise deletion) και σε 6 (18.2%) χρησιμοποιήθηκε συνδυασμός του listwise deletion και του imputation. Επιπλέον σημειώθηκε η μέθοδος της τελευταίας παρατήρησης (last observation carried forward) σε 1 μελέτη (3%), η τεχνική του predictive mean matching για 1 (3%) άλλη μελέτη και τα normal domain values για 1 ακόμη (3%) μελέτη. Οι υπόλοιπες 10(30.3%) παρατηρήσεις δεν ανέφεραν τον τρόπο διαχείρισης των ελλειπουσών τιμών.

4.2.6 Αξιολόγηση του κίνδυνου συστηματικού σφάλματος

Τα αποτελέσματα της ανάλυσης του κινδύνου για συστηματικό σφάλμα στην επιλογή των συμμετεχόντων, των προβλεπτικών παραγόντων, τον ορισμό της έκβασης και την εκτέλεση της ανάλυσης παρουσιάζονται στον Πίνακα 4.5. Επίσης παρουσιάζονται και τα αποτελέσματα της ανάλυσης της καταλληλόλητας των μελετών όσον αφορά τους συμμετέχοντες, τους προβλεπτικούς παράγοντες και την έκβαση. Όσον αφορά τον κίνδυνο συστηματικού σφάλματος 3 (12%) μελέτες κρίθηκαν ότι είχαν υψηλό κίνδυνο, 11 (44%) βαθμολογήθηκαν με ασαφή κίνδυνο και οι υπόλοιπες 11 (44%) με χαμηλό κίνδυνο. Ο υψηλός και ο ασαφής κίνδυνος οφειλόταν κυρίως στην απουσία πληροφορίας για τον αριθμό των εκβάσεων και τη διαχείριση των ελλειπουσών τιμών. Για την καταλληλότητα των μελετών σε 18 (72%) δόθηκε χαμηλή βαθμολογία για κίνδυνο, που συνεπάγεται υψηλή καταλληλότητα, και σε 7 (28%) ασαφής βαθμολογία.

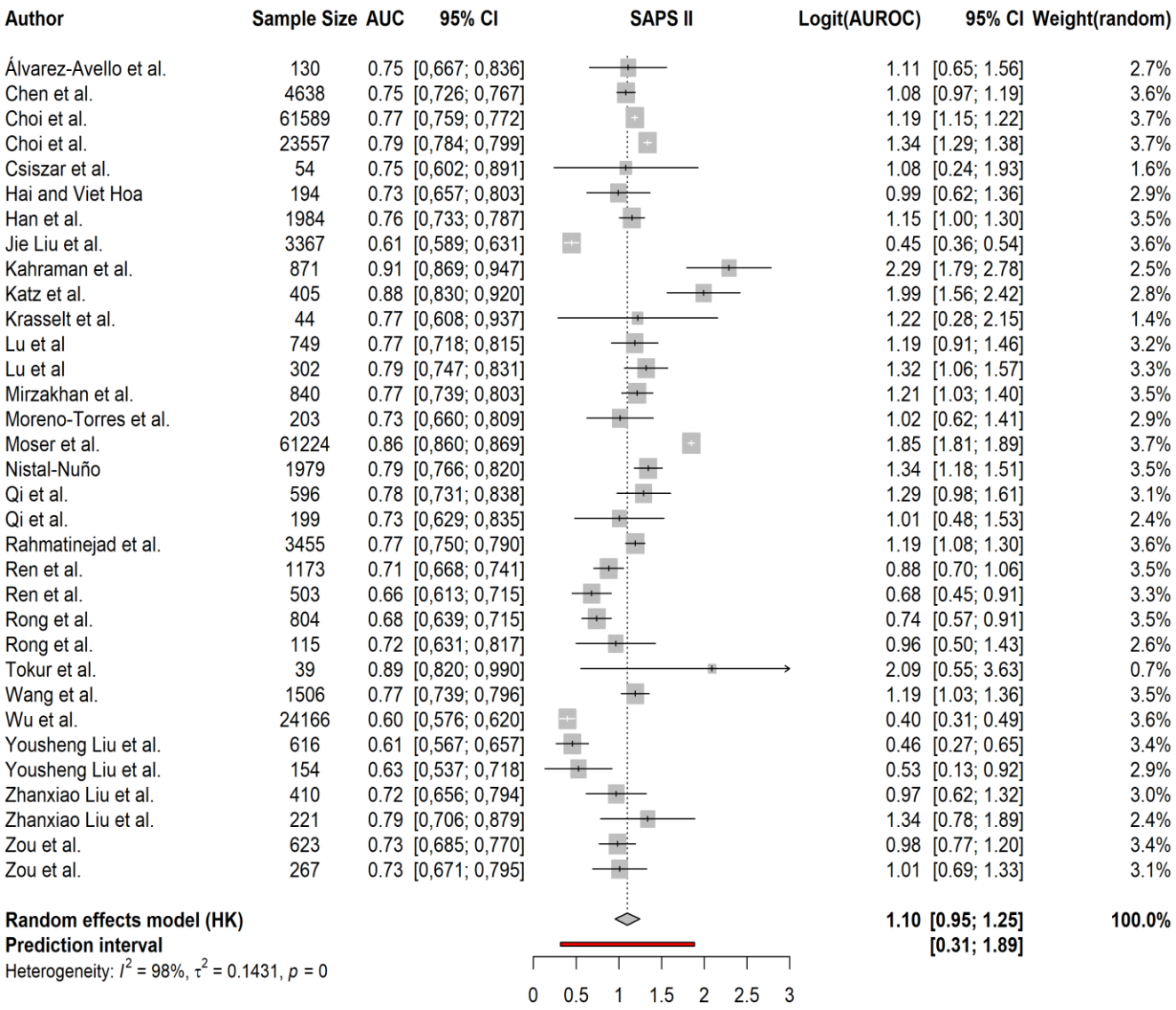
Πίνακας 4.5 Εκτίμηση του κινδύνου για συστηματικό σφάλμα και της καταλληλότητας των μελετών βάσει του εργαλείου PROBAST

	Risk of Bias					Applicability			
Author	**Participants**	**Predictors**	**Outcome**	**Analysis**	**Overall**	**Participants**	**Predictors**	**Outcome**	**Overall**
Zou et al.	Low	Low	Low	Low	Low	Low	Unclear	Low	Unclear
Wu et al.	Low	Low	Low	Unclear	Unclear	Low	Low	Low	Low
Wang et al.	Low	Low	Low	High	High	Low	Unclear	Low	Unclear
Tokur et al.	Low	Low	Low	High	High	Low	Low	Low	Low
Rong et al.	Low	Low	Low	Low	Low	Low	Low	Low	Low
Ren et al.	Low	Low	Low	Unclear	Unclear	Low	Low	Unclear	Unclear
Rahmatinejad et al.	Low	Low	Low	Low	Low	Low	Low	Low	Low
Qi et al.	Low	Low	Low	Low	Low	Low	Low	Low	Low
Nistal-Nuño	Low	Low	Low	Unclear	Unclear	Low	Low	Unclear	Unclear
Moser et al.	Low	Low	Low	Low	Low	Low	Unclear	Low	Unclear
Moreno-Torres et al.	Unclear	Low	Low	Low	Unclear	Low	Low	Low	Low
Mirzakhan et al.	Low	Low	Low	Unclear	Unclear	Low	Low	Low	Low
Lu et al	Low	Low	Low	Low	Low	Low	Low	Low	Low
Zhanxiao Liu et al.	Low	Low	Low	Low	Low	Low	Low	Low	Low
Yousheng Liu et al.	Low	Low	Low	Unclear	Unclear	Low	Low	Unclear	Unclear
Jie Liu et al.	Low	Low	Low	Unclear	Unclear	Low	Low	Low	Low
Krasselt et al.	Low	Low	Low	Unclear	Unclear	Low	Low	Low	Low
Katz et al.	Low	Low	Low	Low	Low	Low	Low	Low	Low
Kahraman et al.	Low	Low	Low	Unclear	Unclear	Low	Low	Low	Low
Han et al.	Low	Low	Low	Low	Low	Low	Low	Low	Low
Hai and Viet Hoa	Low	Low	Low	Unclear	Unclear	Low	Low	Low	Low
Csiszar et al.	Low	Unclear	Low	Unclear	Unclear	Low	Low	Unclear	Unclear
Choi et al.	Low	Low	Low	Low	Low	Low	Low	Low	Low
Chen et al.	Low	Low	Low	Low	Low	Low	Low	Low	Low
Álvarez-Avello et al.	Low	Low	Low	High	High	Low	Low	Low	Low

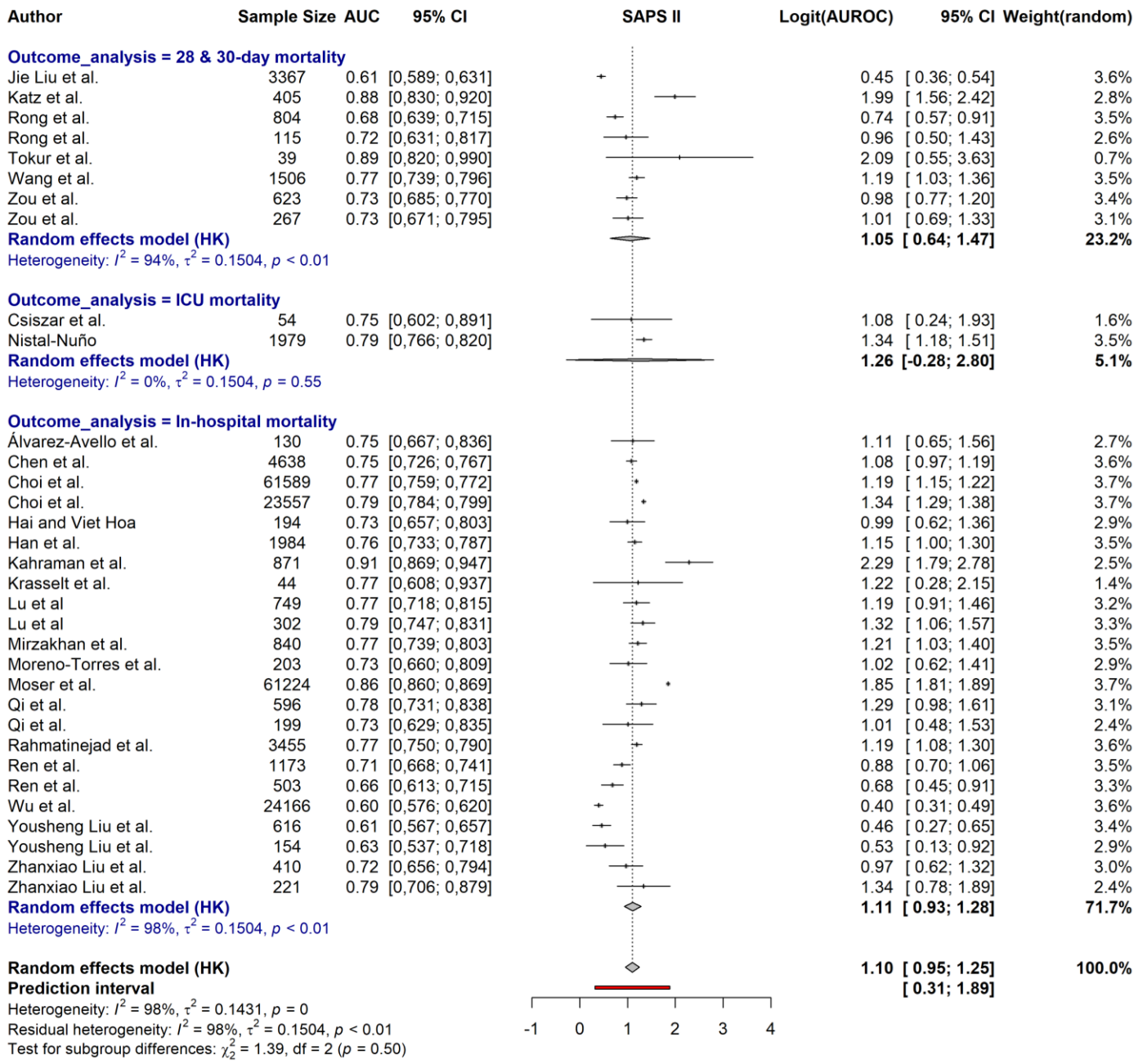
4.2.7 Μετα-ανάλυση

Τα αποτελέσματα της μετα-ανάλυσης του logit-μετασχηματισμένου εμβαδού της καμπύλης ROC παρουσιάζονται στο Γράφημα 4.2. Η συνοπτική εκτίμηση του logit(AUC) που προέκυψε από την ανάλυση ήταν 1,10 (95% CI: 0,95-1,25), ενώ τα μέτρα ετερογένειας ήταν το I^2 = 98,2% (95% CI: 97,9%-98,5%), το τ^2 = 0,1431 (95% CI: 0,0826-0,2725), το p-value του Cochrane's Q που ήταν σχεδόν ίσο με το μηδέν και το 95% διάστημα πρόβλεψης που ήταν (0,31-1,89).

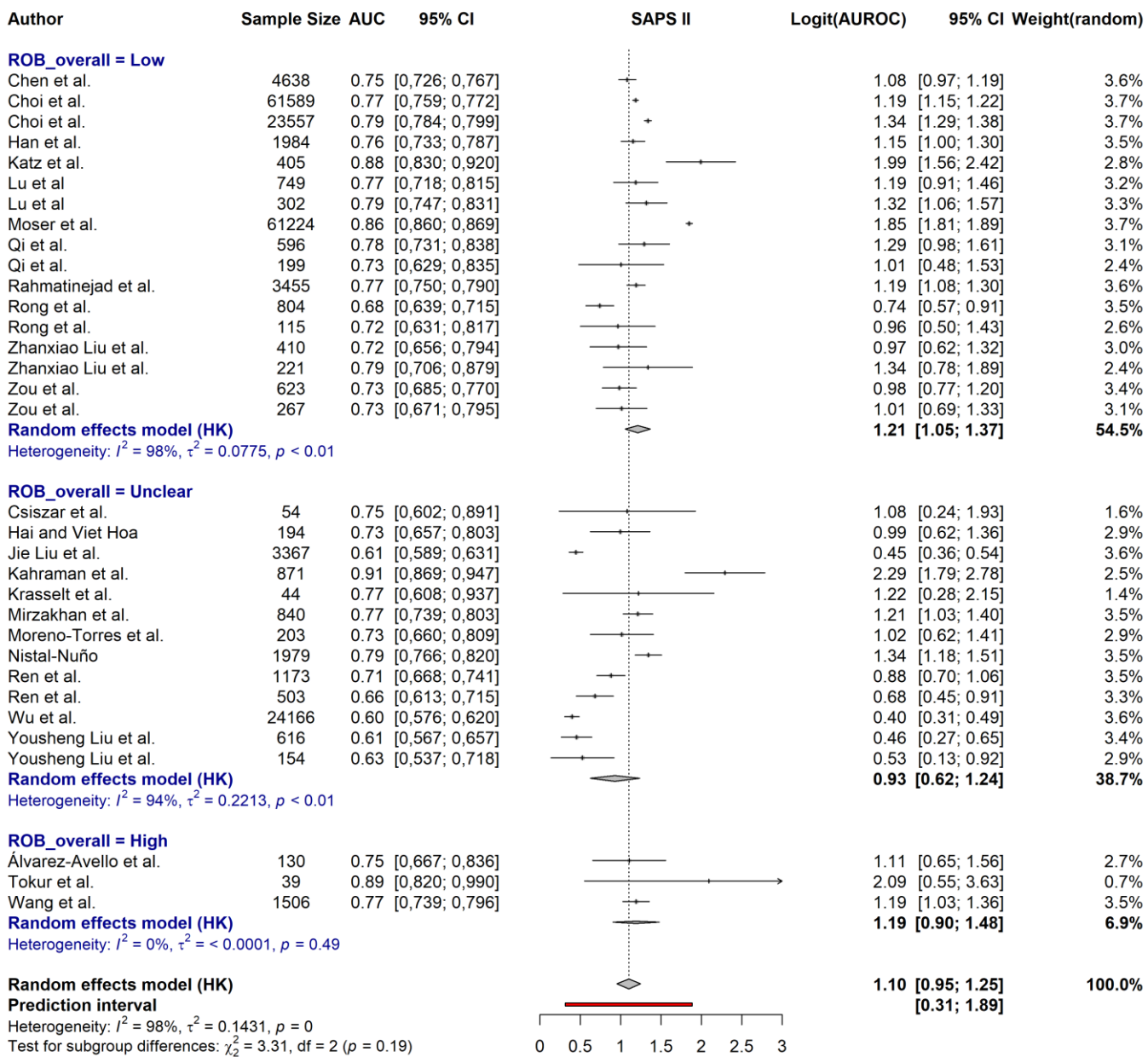
Τα αποτελέσματα των μετα-αναλύσεων σε υποομάδες παρουσιάζονται στα Γραφήματα 4.3,4.4. Στην ανάλυση των υποομάδων βάσει της έκβασης, λόγω του μικρού αριθμού παρατηρήσεων ανά ομάδα υπολογίστηκε μια εκτίμηση του τ^2 που θωρήθηκε κοινή μεταξύ των ομάδων και ήταν ίση με 0,1504. Για τη θνησιμότητα εντός 28 και 30 ημερών η συνοπτική εκτίμηση ήταν 1,05 (95% CI: 0,64-1,47) με I^2 = 93,9% και Cochrane's Q p-value<0,01. Για τη θνησιμότητα εντός του νοσοκομείου, που ήταν και η πλειοψηφία των παρατηρήσεων, η συνοπτική εκτίμηση ήταν 1,11 (95%CI: 0,93-1,28) με I^2 =98% και Cochrane's Q p-value<0,01. Τέλος η θνησιμότητα εντός της ΜΕΘ είχε 2 παρατηρήσεις, με συνοπτική εκτίμηση 1,26 (95% CI: -0,28-2,8), I^2 =0% και Cochrane's Q p-value=0,55. Το χ^2 (chi-square) στατιστικό στους 2 βαθμούς ελευθερίας είχε τιμή ίση με 1,39 και p-value=0,5 που σημαίνει ότι οι διαφορές των εκτιμήσεων μεταξύ των υποομάδων δεν είναι στατιστικά σημαντικές. Από την άλλη πλευρά, στη μετα-ανάλυση υποομάδων με βάσει τον κίνδυνο συστηματικού σφάλματος απουσίαζε το πρόβλημα των μικρών ομάδων, οπότε επιτράπηκε στο τ^2 να παίρνει διαφορετικές τιμές μεταξύ των υποομάδων. Η συνοπτική εκτίμηση των παρατηρήσεων με χαμηλό κίνδυνο συστηματικού σφάλματος ισούταν με 1,21 (95% CI: 1,05-1,37), ενώ η ετερογένεια περιγραφόταν με I^2=98%, τ^2=0,0775 και Cochrane's Q p-value<0,01. Για τις παρατηρήσεις με ασαφή κίνδυνο για συστηματικό σφάλμα η εκτίμηση του logit(AUC) ισούταν με 0,93 (95% CI: 0,62-1,24), το I^2=94%, το τ^2=0,2213 και το Cochrane's Q p-value<0,01. Τέλος για τις 3 παρατηρήσεις με υψηλό κίνδυνο συστηματικού σφάλματος εκτιμήθηκε: logit(AUROC)=1,19 (95% CI: 0,90-1,48), I^2=0%, τ^2<0,0001 και Cochrane's Q p-value=0,49. Σε αυτήν την περίπτωση το στατιστικό χ^2 στους 2 βαθμούς ελευθερίας είχε p-value= 0,19, που δεν είναι στατιστικά σημαντικό στο επίπεδο του 5% .



Γράφημα 4.2: Forest plot της μετα-ανάλυσης του logit-μετασχηματισμένου εμβαδού της καμπύλης ROC για την αξιολόγηση του SAPS II μοντέλου



Γράφημα 4.3: Μετα-ανάλυση σε υποομάδες ανάλογα με την κατηγορία έκβασης των ασθενών που μελετήθηκε (28 & 30-day mortality, in-hospital mortality, ICU mortality)

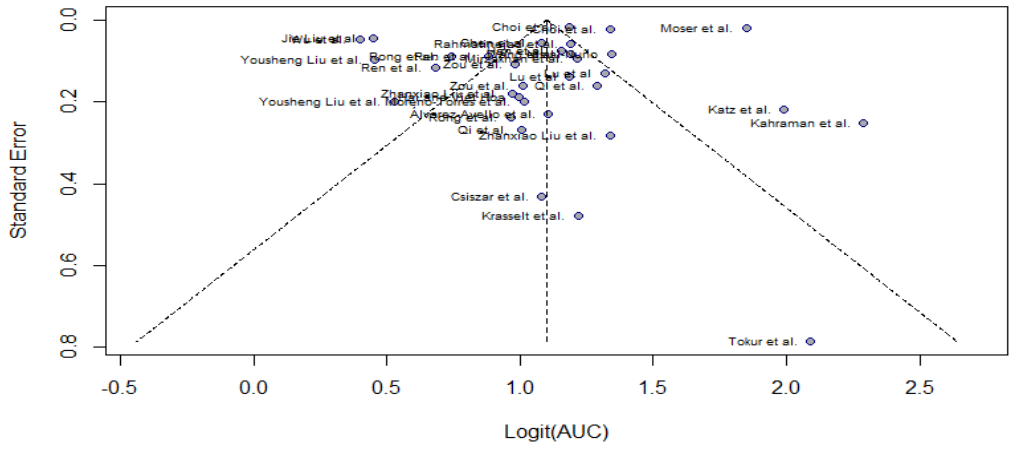


Γράφημα 4.4: Μετα-ανάλυση σε υποομάδες ανάλογα με την συνολική αξιολόγηση των παρατηρήσεων ως προς το βαθμό κινδύνου για συστηματικό σφάλμα στη μελέτη (Low/Unclear/High)

Πέρα από ανάλυση σε υποομάδες έγιναν και αναλύσεις ευαισθησίας αποκλείοντας κάποιες μελέτες κάθε φορά για να διερευνηθεί εάν είχαν μεγάλη επίδραση στη διαμόρφωση της τελικής εκτίμησης. Αρχικά εξαιρέθηκαν οι μελέτες στις οποίες παρατηρήθηκε υψηλός κίνδυνος συστηματικού σφάλματος (Tokur *et al.*, 2022; Álvarez-Avello *et al.*, 2022; Wang *et al.*, 2022). Τα αποτελέσματα της ανάλυσης μεταβλήθηκαν ελάχιστα συγκριτικά με τη συνολική ανάλυση, επιβεβαιώνοντας εν μέρη έτσι και το συμπέρασμα της ανάλυσης υποομάδων ότι ο κίνδυνος συστηματικού σφάλματος δεν είναι σημαντικός παράγοντας επίδρασης της διακριτικής ικανότητας του SAPS II. Επιπλέον πραγματοποιήθηκαν αναλύσεις εξαιρώντας μελέτες οι οποίες είχαν πολύ μικρό μέγεθος δείγματος (Csiszar *et al.*, 2022; Krasselt *et al.*, 2022; Tokur *et al.*, 2022) και πολύ μεγάλο μέγεθος δείγματος (Choi *et al.*, 2022; Moser *et al.*, 2022). Σε καμία από τις δύο περιπτώσεις δεν παρατηρήθηκαν αισθητές διαφορές στη συνολική εκτίμηση. Αυτό σημαίνει ότι η παρουσία των μικρών μελετών δεν είχε ανεπιθύμητες επιδράσεις στην ανάλυση, αλλά και ότι η παρουσία των μεγάλων μελετών δεν είναι αυτή που καθόρισε αποκλειστικά την τιμή της συνθετικής εκτίμησης της ανάλυσης. Το γεγονός αυτό οφείλεται στη χρήση μοντέλου τυχαίων επιδράσεων για την ανάλυση έτσι ώστε όλες οι μελέτες να λάβουν παρόμοιες στάθμες ανεξαρτήτως μεγέθους δείγματος.

Όσον αφορά τη διερεύνηση της επίδρασης του small-study effect και του σφάλματος δημοσίευσης παρουσιάζεται στο Γράφημα 4.5 το funnel plot. Από το funnel plot φαίνεται να μην υπάρχει σφάλμα δημοσίευσης, αφού η συμμετρία του είναι αρκετά ικανοποιητική. Επίσης ο έλεγχος του Egger έδωσε P-value=0,0887, δηλαδή ήταν μη στατιστικά σημαντικός στο επίπεδο στατιστικής σημαντικότητας του 5%.

Γράφημα 4.5: Funnel plot , με το logit-μετασχηματισμένο AUROC στον άξονα x και το τυπικό σφάλμα στον άξονα y.



Στον Πίνακα 4.6 παρουσιάζονται τα βασικά ευρήματα της μετα-ανάλυσης μετά τον αντίστροφο μετασχηματισμό των εκτιμήσεων με τη συνάρτηση

$$logit^{-1}(x) = \frac{\exp(x)}{1 + \exp(x)}$$

Πίνακας 4.6: Αποτελέσματα της μετα-ανάλυσης μετά τον αντίστροφο μετασχηματισμό των εκτιμήσεων

Analysis	*Pooled AUROC*	*95% Confidence Interval*	*95% Prediction Interval*
Overall	0.75	0.72-0.78	0.58- 0.87
28 & 30-day mortality	0.74	0.65-0.81	
In-hospital mortality	0.75	0.72-0.78	
ICU mortality	0.78	0.43- 0.94	
Low risk of bias	0.77	0.74-0.80	
Unclear risk of bias	0.72	0.65-0.78	
High risk of bias	0.77	0.71-0.81	

ΚΕΦΑΛΑΙΟ 5
ΣΥΖΗΤΗΣΗ

5.1 Κύρια ευρήματα

Στην παρούσα εργασία σχεδιάσαμε την πρώτη συστηματική αξιολόγηση της διεθνούς βιβλιογραφίας για την προγνωστική ικανότητα του μοντέλου SAPS II σε ασθενείς ΜΕΘ και παρουσιάσαμε προκαταρκτικά αποτελέσματα από τα δεδομένα των 25 πιο πρόσφατα δημοσιευμένων μελετών. Είναι αξιοσημείωτο ότι από τις 25 μελέτες όλες αξιολογούσαν τη διακριτική ικανότητα του μοντέλου αποκλειστικά μέσω του εμβαδού της καμπύλης ROC, ενώ ως προς τη βαθμονόμηση μόλις τέσσερις αξιολόγησαν το μοντέλο με κάποιο τρόπο, εκ των οποίων οι τρεις μάλιστα κάπως αδρά, και τρεις που έκαναν αναφορά στο Brier σκορ. Οπότε η μετα-ανάλυση πραγματοποιήθηκε μόνο στα μέτρα διακριτικής ικανότητας λόγω έλλειψης αρκετών παρατηρήσεων για τα υπόλοιπα μέτρα. Γενικώς, η πλειοψηφία των μελετών παρουσίαζε αποδεκτή τιμή διακριτικής ικανότητας (μεταξύ 0,7 και 0,8) του μοντέλου, ενώ μόλις τέσσερις υψηλή τιμή (>0,8) και πέντε χαμηλή τιμή (<0,7). Η συνοπτική εκτίμηση που προέκυψε από την ανάλυση μετά τον αντίστροφο μετασχηματισμό ήταν 0,75, που θεωρείται αποδεκτή τιμή, με τα 95% όρια αξιοπιστίας της να κυμαίνονται από 0,72 έως 0,78.

Η ερμηνεία όμως αυτού του αποτελέσματος θέλει περαιτέρω διερεύνηση. Η μετα-ανάλυση αποκάλυψε ότι υπήρχε μεγάλη ετερογένεια τόσο σε επίπεδο μεταξύ όσο και εντός των μελετών. Όσον αφορά την ετερογένεια εντός των μελετών, μόνο οκτώ από τις είκοσι πέντε μελέτες συνείσφεραν πάνω από μία μέτρηση στην ανάλυση. Από αυτές όμως μόνο στη μία οι εκτιμήσεις του AUROC φαίνεται να διαφέρουν σε σημαντικό βαθμό και να μην αλληλεπικαλύπτονται τα 95% διαστήματα εμπιστοσύνης τους. Άρα η ετερογένεια αφορά κατά κύριο λόγο το επίπεδο μεταξύ των μελετών. Πράγματι, το I^2 ισούταν με 98,2%, το τ^2 με 0.1431 και το p-value του χ^2 ελέγχου ήταν μικρότερο από 0.001, που είναι ενδείξεις πολύ μεγάλης ετερογένειας. Το 95% διάστημα πρόβλεψης που προέκυψε μετά τον αντίστροφο μετασχηματισμό είναι το διάστημα από 0,58 έως 0,87 και περιέχει τιμές του AUROC που χαρακτηρίζουν τη διακριτική ικανότητα του μοντέλου είτε υψηλή, είτε αποδεκτή αλλά είτε και χαμηλή. Το διάστημα αυτό λοιπόν είναι αρκετά

ευρύ και ενισχύει την υπόθεση παρουσίας ισχυρής ετερογένειας. Πιθανές πηγές αυτής της ετερογένειας αποτελούν η ποικιλία στα μεγέθη δείγματος των μελετών, η ποικιλία στη χώρα προέλευσης των δεδομένων, η παθολογία των ασθενών ανά μελέτη και η χρονική στιγμή πρόβλεψης του κινδύνου.

Δύο πιθανές πηγές ετερογένειας που διερευνήθηκαν μέσω της ανάλυσης σε υποομάδες ήταν η έκβαση της κάθε μελέτης και ο συνολικός κίνδυνος για συστηματικό σφάλμα. Οι πιθανές εκβάσεις που μελετήθηκαν ήταν η θνησιμότητα εντός 28 ημερών, η θνησιμότητα εντός 30 ημερών, η ενδονοσοκομειακή θνησιμότητα και η θνησιμότητα στις μονάδες εντατικής θεραπείας. Για αυτήν την ανάλυση επειδή ο αριθμός των παρατηρήσεων δεν ήταν αρκετά μεγάλος για όλα τα γκρουπ θεωρήθηκε ότι το τ^2 δε θα παίρνει διαφορετικές τιμές ανά ομάδα αλλά εκτιμήθηκε μόνο μία τιμή η οποία παρέμενε σταθερή για όλες τις εκβάσεις. Οι επιμέρους εκτιμήσεις για το κάθε γκρουπ δε διέφεραν αισθητά μεταξύ τους και προσέγγιζαν ικανοποιητικά τη συνολική συνοπτική εκτίμηση της μετα-ανάλυσης. Επιπλέον ο χ^2 έλεγχος για τη διαφορά των εκτιμήσεων ανάμεσα στα τέσσερα αυτά γκρουπ ήταν εμφανώς μη στατιστικά σημαντικός. Οπότε ο τύπος της έκβασης δε φαίνεται να επηρεάζει σημαντικά την τελική τιμή της εκτίμησης και συνεπώς δε θεωρείται σημαντικός παράγοντας ετερογένειας. Ομοίως, μικρή επίδραση στη συνολική εκτίμηση φαίνεται να έπαιξε και η βαρύτητα του συστηματικού σφάλματος των μελετών. Οι ομάδες στη δεύτερη ανάλυση υποομάδων ήταν τρεις, χαμηλός ή ασαφής ή υψηλός κίνδυνος συστηματικού σφάλματος. Οι κυριότεροι λόγοι για να βαθμολογηθεί μία μελέτη με υψηλό κίνδυνο συστηματικού σφάλματος ήταν να μην αναφέρει τη χρονική στιγμή υπολογισμού του SAPS II σκορ, να μην εξηγεί τον τρόπο διαχείρισης των ελλειπουσών τιμών και το πολύ μικρό μέγεθος δείγματος. Σε αντίθεση με την προηγούμενη ανάλυση υποομάδων, αν και η ομάδα του υψηλού κινδύνου είχε μόνο τρεις παρατηρήσεις, επιτράπηκε στο τ^2 να παίρνει διαφορετικές τιμές ανά ομάδα. Οι επιμέρους εκτιμήσεις ανά ομάδα κινδύνου ήταν και πάλι πολύ κοντά στη συνολική εκτίμηση και ο χ^2 έλεγχος για τη διαφορά των υποομάδων ήταν ξανά μη στατιστικά σημαντικός, αν και λιγότερο έντονα από την πρώτη ανάλυση. Μία άλλη μεταβλητή που είναι συχνά παράγοντας ετερογένειας στις μετα-αναλύσεις είναι η ηλικία των ασθενών. Δυστυχώς όμως δεν υπήρχε ένα κοινό μέτρο της ηλικίας μεταξύ των μελετών για να γίνει μία μετα-παλινδρόμηση, αφού μερικές ανέφεραν τη μέση τιμή της, μερικές τη διάμεσο και ορισμένες δεν περιέγραφαν καθόλου την ηλικία των ασθενών στα δείγματα που χρησιμοποίησαν. Κατά πάσα πιθανότητα όμως η ηλικία δε θα αποτελούσε σημαντικό

παράγοντα ετερογένειας σε αυτήν τη μετα-ανάλυση, καθώς δε φαίνεται να υπήρχαν μεγάλες διακυμάνσεις της ηλικίας μεταξύ των μελετών. Είναι βέβαια άξιο αναφοράς ότι οι ηλικίες των ασθενών είναι μάλλον μικρότερες από αυτές που θα περίμενε κανείς, δεδομένου ότι οι ασθενείς πρέπει αποκλειστικά να έχουν ενταχθεί σε κάποια μονάδα εντατικής θεραπείας. Συγκεκριμένα οι περισσότερες μέσες τιμές και οι διάμεσοι κυμαίνονται από 60 έως 70 έτη.

5.2 Ερμηνεία των ευρημάτων σε σχέση με παλαιότερες δημοσιεύσεις

Οι Nassar και οι συνεργάτες του (Nassar, Malbouisson and Moreno, 2014) στη συστηματική τους ανασκόπηση μελέτησαν τις επιδόσεις του SAPS III μοντέλου ως προγνωστικό μοντέλο εντατικής φροντίδας. Συγκέντρωσαν 28 μελέτες εξωτερικής επικύρωσης του SAPS III και εξήγαγαν τα μέτρα διακριτικής ικανότητας και βαθμονόμησης. Το μέτρο διακριτικής ικανότητας που αξιολόγησαν ήταν το AUROC, όπως και στη δική μας μελέτη, ενώ σα μέτρο βαθμονόμησης χρησιμοποίησαν το στατιστικό του ελέγχου Hosmer-Lemeshow. Το SAPS III μοντέλο, όπως και το SAPS II, φαίνεται να επιτυγχάνει ικανοποιητική διακριτική ικανότητα. Όσον αφορά τη βαθμονόμηση, μόνο 11 μελέτες είχαν μη στατιστικά σημαντική δυσβαθμονόμηση και αυτό φάνηκε να οφείλεται στο γεγονός ότι ο έλεγχος του Hosmer-Lemeshow είναι ευαίσθητος σε μεγάλα μεγέθη δείγματος.

Επιπλέον, μία πρόσφατη συστηματική ανασκόπηση και μετα-ανάλυση συνέκρινε την απόδοση μοντέλων μηχανικής μάθησης δίτιμης ταξινόμησης και μοντέλων σοβαρότητας της ασθένειας, συμπεριλαμβανομένου και του SAPS II, στην πρόβλεψη της θνησιμότητας των ασθενών εντός της ΜΕΘ (Barboi, Tzavelis and Muhammad, 2022). Πραγματοποιήθηκε μετα-ανάλυση τυχαίων επιδράσεων, καθώς κι εκείνοι παρατήρησαν μεγάλη ετερογένεια μεταξύ των μελετών. Εν αντιθέσει με μας, οι μελέτες που αφορούσαν το SAPS II και συμπεριλήφθηκαν στην ανάλυση ήταν μόνο 5, αλλά η συνοπτική εκτίμηση του AUROC ήταν κοντά στη δική μας (0,77 , 95% CI: 0,739-0,801). Όπως και στη δική μας έρευνα, η αναφορά των μέτρων βαθμονόμησης στις μελέτες ήταν πολύ περιορισμένη, με αποτέλεσμα να αδυνατούν να αξιολογήσουν τα μοντέλα ως προς την ικανότητα τους να επιτυγχάνουν ικανοποιητική βαθμονόμηση.

5.3 Δυνατά σημεία και περιορισμοί

Αυτή είναι η πρώτη συστηματική αξιολόγηση μελετών εξωτερικής επικύρωσης των μέτρων αξιολόγησης της προβλεπτικής ικανότητας του SAPS II μοντέλου πρόβλεψης θνησιμότητας ασθενών σε μονάδες εντατικής θεραπείας. Οι μελέτες που συμπεριλήφθηκαν δεν ήταν αποκλειστικά μελέτες εξωτερικής επικύρωσης, αλλά οποιαδήποτε έρευνα αξιολογούσε τη διακριτική ικανότητα ή τη βαθμονόμηση του SAPS II σε δείγμα διαφορετικό από αυτό που χρησιμοποιήθηκε για την κατασκευή του μοντέλου. Συγκεντρώθηκαν όλα τα διαθέσιμα μέτρα διακριτικής ικανότητας, βαθμονόμησης και συνολικής αξιολόγησης του μοντέλου ακόμα κι αν δεν χρησίμευαν για την ανάλυση.

Δυστυχώς αυτή η μελέτη κρύβει και κάποιους περιορισμούς. Καταρχάς η αναζήτηση πραγματοποιήθηκε μόνο στην πλατφόρμα PUBMED και από τις μελέτες που προέκυψαν για τη διπλωματική εργασία χρησιμοποιήθηκαν λόγω όγκου εργασίας μόνο τα άρθρα του τελευταίου έτους. Τα υπόλοιπα άρθρα που έγιναν δεκτά μέχρι και τη φάση της διαλογής βάσει του πλήρους κειμένου δεν απορρίφθηκαν τελικά, απλά δεν συμπεριλήφθηκαν στο κομμάτι της εξαγωγής των δεδομένων και συνεπώς ούτε και στο κομμάτι της ανάλυσης. Ένα άλλο σημείο πιθανής πρόκλησης συστηματικού σφάλματος είναι ότι κρατήθηκαν μόνο όσα άρθρα ήταν γραμμένα στην αγγλική γλώσσα, γεγονός που μπορεί να οδηγήσει σε συστηματικό σφάλμα επιλογής. Επίσης, μόνο 17 από τις 33 παρατηρήσεις προέρχονταν από μελέτες με χαμηλό κίνδυνο για συστηματικό σφάλμα σύμφωνα με το PROBAST εργαλείο. Ένας ακόμα πολύ σημαντικός περιορισμός είναι ότι η μετα-ανάλυση βασίστηκε μόνο στο εμβαδό της καμπύλης ROC για την αξιολόγηση της διακριτικής ικανότητας του μοντέλου. Μέτρα αξιολόγησης της βαθμονόμησης, συνολικά μέτρα αξιολόγησης και μέτρα κλινικής χρησιμότητας ελάχιστα αναφέρονταν στις μελέτες και ήταν αδύνατο να γίνει ανάλυση πάνω σε αυτά.

5.4 Προβληματισμοί

Η μαζική αυτή απουσία των συνολικών μέτρων αξιολόγησης, όπως το Brier score, και των μέτρων βαθμονόμησης των μοντέλων είναι ένα θέμα που πρέπει να προκαλέσει ανησυχία για την ποιότητα των αξιολογήσεων των μοντέλων πρόβλεψης σοβαρότητας μιας ασθένειας. Δεν αποτελεί πρόβλημα μόνο για τις μελέτες που αξιολογούν το SAPS II, αφού σχεδόν το σύνολο των μελετών που προέκυψαν από την αναζήτηση στο PUBMED αξιολογούσαν κι άλλα μοντέλα ταυτόχρονα, όπως το APACHE και το SOFA. Ακόμα και οι μελέτες που ήταν αποκλειστικά μελέτες εξωτερικής επικύρωσης των μοντέλων, δεν διερευνούσαν σε βάθος τη βαθμονόμηση των μοντέλων, παρά χρησιμοποιούσαν τον έλεγχο των Hosmer-Lemeshow που είναι μια αδρή και όχι πάντα αξιόπιστη ένδειξη της ποιότητας βαθμονόμησης. Μόνο μία από τις 25 μελέτες είχε υπολογίσει σταθερά και κλίση βαθμονόμησης μαζί με τα 95% όρια εμπιστοσύνης τους και μόλις τρεις το Brier score. Για να είναι πλήρης η αξιολόγηση ενός μοντέλου δεν αρκεί μόνο να έχει ικανοποιητική διακριτική ικανότητα, διότι είναι αδύνατο να βγουν σαφή συμπεράσματα για τη χρησιμότητα ενός προβλεπτικού μοντέλου μόνο από το εμβαδό της καμπύλης ROC. Το ερώτημα είναι γιατί οι μελέτες αρκούνται στην παρουσίαση μόνο αυτού του μέτρου. Ειδικά οι μελέτες εξωτερικής επικύρωσης που μοναδικό σκοπό έχουν τον έλεγχο των μοντέλων με απώτερο στόχο τη γενίκευση τους σε ευρύτερους πληθυσμούς, θα έπρεπε να είναι πιο ενδελεχείς και αυστηρές με τα κριτήρια τους. Η πλειοψηφία των μελετών αυτής της μετα-ανάλυσης είχαν πρωταρχικό στόχο την ανάπτυξη ενός νέου μοντέλου και τη σύγκριση του με παλιότερα εδραιωμένα μοντέλα πρόβλεψης θνησιμότητας. Η σύγκριση αυτή σχεδόν αποκλειστικά περιοριζόταν στα μέτρα διακριτικής ικανότητας, που σαφώς είναι πιο εύκολο να επιτύχουν ικανοποιητικά επίπεδα. Αντίθετα καλή βαθμονόμηση είναι πιο δύσκολο να επιτευχθεί, αφού είναι ο βαθμός που συμβαδίζουν οι προβλεπόμενες τιμές του μοντέλου με τις παρατηρούμενες τιμές κι ίσως αυτός να είναι κι ένας λόγος που πάρα πολλές μελέτες αποφεύγουν να την αξιολογήσουν. Το πρόβλημα αυτό δεν οφειλόταν στην επιλογή μόνο των μελετών του τελευταίου έτους για τη μετα-ανάλυση, διότι η απουσία της αξιολόγησης των μέτρων αυτών ήταν αισθητή στο σύνολο των μελετών που προέκυψαν από την αναζήτηση.

Πέρα από το εμβαδό της καμπύλης ROC, αρκετές μελέτες ανέφεραν την ευαισθησία και την ειδικότητα που είναι μέτρα της σωστής ταξινόμησης των ατόμων ως ασθενών ή υγιών. Μία μετα-ανάλυση σε αυτά τα μέτρα θα ήταν δύσκολη διότι η κάθε μελέτη χρησιμοποιούσε διαφορετικό κατώφλι για τον υπολογισμό τους. Δυστυχώς, δεν εφαρμόστηκε μία κοινή μέθοδος για την εκτίμηση του βέλτιστου κατωφλιού, εάν και αρκετές μελέτες χρησιμοποίησαν το δείκτη του Youden για αυτό το σκοπό. Επιπλέον, είναι απαραίτητη η διεύρυνση της χρήσης της ανάλυσης της καμπύλης απόφασης. Μέσω αυτής μπορεί να γίνει ορατή η κλινική χρησιμότητα των μοντέλων πρόβλεψης και πιο συγκεκριμένα να γίνουν διακριτά τα όρια της πιθανότητας εκδήλωσης της έκβασης μέσα στα οποία η χρήση των μοντέλων υπερτερεί έναντι άλλων μεθόδων.

5.5 Συμπεράσματα

Τα αποτελέσματα της ανάλυσης έδειξαν ότι το SAPS II ως μοντέλο πρόβλεψης της θνησιμότητας σε ασθενείς που βρίσκονται σε μονάδες εντατικής θεραπείας παρουσιάζει ικανοποιητική διακριτική ικανότητα έχοντας όμως μία αβεβαιότητα λόγω της μεγάλης ετερογένειας των μελετών. Παρ' όλα αυτά δε μπορεί να χρησιμοποιηθεί ως χρήσιμο εργαλείο πρόβλεψης για τους κλινικούς στην καθημερινή κλινική πράξη γιατί πρέπει να αξιολογηθούν τόσο τα συνολικά μέτρα απόδοσης όσο και τα μέτρα βαθμονόμησης για να παρθεί ένα τέτοιο συμπέρασμα, αλλιώς η αξιολόγηση θα είναι ελλιπής και ασυνεπής. Είναι προφανές ότι είναι επιτακτική η ανάγκη βελτίωσης της ποιότητας των μελετών εξωτερικής επικύρωσης και μεγαλύτερης αυστηρότητας στα κριτήρια δημοσίευσης τους. Η βελτίωση αυτή θα συνεισφέρει στη χωρική και χρονική γενίκευση αυτών των μοντέλων μειώνοντας έτσι την ετερογένεια που παρατηρείται μεταξύ διαφορετικών πληθυσμών. Επιπλέον, η εγκαθίδρυση των μοντέλων αυτών ως διαγνωστικών εργαλείων θα ενισχύσει την πρόωρη αναγνώριση και διάγνωση των ασθενών που υποτροπιάζουν και θα παρέχει στους κλινικούς πολύτιμο επιπλέον χρόνο να παρέμβουν ουσιαστικά.

ΠΕΡΙΛΗΨΗ

Εισαγωγή: Η σωστή διαχείριση των κρίσιμων νοσημάτων περιλαμβάνει τη γρήγορη και ακριβή ταυτοποίηση των βαρέως πασχόντων ασθενών. Ένα χρήσιμο εργαλείο που μπορεί να το επιτύχει αυτό είναι το ευρέως διαδομένο μοντέλο πρόβλεψης ενδονοσοκομειακής θνησιμότητας, SAPS II. Σκοπός της εργασίας αυτής είναι να συγκεντρώσει μελέτες εξωτερικής επικύρωσης του SAPS II μοντέλου από τη διεθνή βιβλιογραφία και να μετα-αναλύσει τα μέτρα προγνωστικής ακρίβειάς του**.**

Μέθοδοι: Πραγματοποιήθηκε βιβλιογραφική αναζήτηση στο PUBMED την περίοδο από 16 Οκτωβρίου 2022 μέχρι 20 Νοέμβρη 2022. Επιλέχθηκαν οι μελέτες που είχαν δημοσιευθεί από το 2003 έως το 2022 και αξιολογούσαν το SAPS II για την πρόβλεψη της θνησιμότητας σε ασθενείς ΜΕΘ. Οι πληροφορίες που εξήχθησαν ήταν τα χαρακτηριστικά των μελετών, τα χαρακτηριστικά των ασθενών, τα μέτρα προγνωστικής ακρίβειας και αξιολογήθηκε ο κίνδυνος συστηματικού σφάλματος για την κάθε μελέτη. Για τη μετα-ανάλυση χρησιμοποιήθηκε το εμβαδό της καμπύλής ROC (AUROC) που είναι μέτρο διακριτικής ικανότητας. Πραγματοποιήθηκε σύνθεση των αποτελεσμάτων, με την εφαρμογή μοντέλου τυχαίων επιδράσεων και διερεύνηση των πηγών ετερογένειας με αναλύσεις σε υπο-ομάδες.

Αποτελέσματα: Έγινε διαλογή 728 άρθρων, εκ των οποίων στα 306(42%) εκτιμήθηκε το πλήρες κείμενο. Κρίθηκαν σχετικά με τη μελέτη τα 200, αλλά συμπεριλήφθηκαν μόνο όσα είχαν δημοσιευθεί το 2022 (25). Από τις 25 μελέτες, 23(69,7%) μελετούσαν την ενδονοσοκομειακή θνησιμότητα, 2(6%) τη θνησιμότητα εντός της ΜΕΘ και 8(23,3%) τη θνησιμότητα εντός 1 μήνα, ενώ 11 (44%) κρίθηκαν με χαμηλό κίνδυνο συστηματικού σφάλματος, 11(44%) με ασαφή κίνδυνο και 3 (12%) με υψηλό κίνδυνο. Η συνθετική εκτίμηση του AUROC ήταν 0,75 (95% CI: 0,72-0,78). Παρατηρήθηκε σημαντικός βαθμός ετερογένειας (I^2= 98%) ιδιαίτερα σε επίπεδο μεταξύ των μελετών, με το 95% διαστημα πρόβλεψης του AUROC να διαμορφώνεται μεταξύ 0,58 και 0,87. Οι αναλύσεις σε υπο-ομάδες ανάλογα την έκβαση και τον κίνδυνο συστηματικού σφάλματος έδειξαν ότι δεν υπάρχει στατιστικά σημαντική διαφορά μεταξύ των υπο-ομάδων.

Συμπεράσματα: Το μοντέλο SAPS II ως εργαλείο πρόβλεψης θνησιμότητας σε ασθενείς ΜΕΘ επιτυγχάνει ικανοποιητική διακριτική ικανότητα. Η απουσία όμως των μέτρων

βαθμονόμησης και η περιορισμένη ανάλυση καμπύλης απόφασης καθιστούν αδύνατη την πλήρη και συνεπή αξιολόγηση του μοντέλου και την εδραίωση του ως ένα χρήσιμο εργαλείο για την καθημερινή κλινική πράξη.

ABSTRACT

Introduction: The ideal management of severe diseases includes the fast and precise identification of critical care patients. A useful tool that can achieve that is the widely used prognostic model of in-hospital mortality, SAPS II. The objective of this study is to gather external validation studies of SAPS II from the international literature and to meta-analyze its predictive performace measurements.

Methods: A systematic search was conducted between October 16 and November 20, 2022. Studies were chosen if they were published between 2003 and 2022, and reviewed SAPS II on predicting ICU patients' mortality. Data that were extracted included study and patients' characteristics, predictive performance measurements and the risk of bias was assessed for each study. The discrimination measure that was used for the meta-analysis was the Area Under the ROC curve (AUROC). We synthesized the results with the use of a random effects model and researched the source of heterogeneity by subgroup analysis.

Results: A total of 728 articles were screened, 306(42%) of which had their full text assessed. Of those 200 were judged as suitable, but only those that were published in 2022 (25) were included in the study. Of the 25 studies, 23 (69,7%) researched in-hospital mortality, 2 (6%) ICU mortality, 8 (23,3%) 1-month mortality, while 11 (44%) were found to have low risk of bias, 11 (44%) unclear risk of bias and 3 (12%) high risk of bias. The pooled estimate of AUROC was 0,75 (95% CI: 0,72-0,78). A high degree of between study heterogeneity was noted (I^2 = 98%), with a 95% prediction interval of AUROC between 0,58 and 0,87. Subgroup analysis depending on type of event and overall risk of bias showed no statistical difference between the groups.

Conclusions: The SAPS II model can achieve a decent discrimination as a tool of predicting ICU mortality. However, the absence of calibration measurements and the limited use of decision curve analysis deter the absolute and consistent review of the model and its establishment as a useful tool for daily clinical practice.

ΠΑΡΑΡΤΗΜΑ

PRISMA Checklist

Section and Topic	Item #	Checklist item	Location where item is reported
TITLE			
Title	1	Identify the report as a systematic review.	Σελ. 1
ABSTRACT			
Abstract	2	See the PRISMA 2020 for Abstracts checklist.	Σελ. 65
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of existing knowledge.	Σελ. 8-9
Objectives	4	Provide an explicit statement of the objective(s) or question(s) the review addresses.	Σελ. 10
METHODS			
Eligibility criteria	5	Specify the inclusion and exclusion criteria for the review and how studies were grouped for the syntheses.	Σελ. 35,36
Information sources	6	Specify all databases, registers, websites, organisations, reference lists and other sources searched or consulted to identify studies. Specify the date when each source was last searched or consulted.	Σελ. 35
Search strategy	7	Present the full search strategies for all databases, registers and websites, including any filters and limits used.	Σελ. 35,36
Selection process	8	Specify the methods used to decide whether a study met the inclusion criteria of the review, including how many reviewers screened each record and each report retrieved, whether they worked independently, and if applicable, details of automation tools used in the process.	Σελ. 35
Data collection process	9	Specify the methods used to collect data from reports, including how many reviewers collected data from each report, whether they worked independently, any processes for obtaining or confirming data from study investigators, and if applicable, details of automation tools used in the process.	Σελ. 36
Data items	10a	List and define all outcomes for which data were sought. Specify whether all results that were compatible with each outcome domain in each study were sought (e.g. for all measures, time points, analyses), and if not, the methods used to decide which results to collect.	Σελ. 36
	10b	List and define all other variables for which data were sought (e.g. participant and intervention characteristics, funding sources). Describe any assumptions made about any missing or unclear information.	Σελ. 35,36
Study risk of bias assessment	11	Specify the methods used to assess risk of bias in the included studies, including details of the tool(s) used, how many reviewers assessed each study and whether they worked independently, and if applicable, details of automation tools used in the process.	Σελ. 37
Effect measures	12	Specify for each outcome the effect measure(s) (e.g. risk ratio, mean difference) used in the synthesis or presentation of results.	Σελ. 38
Synthesis methods	13a	Describe the processes used to decide which studies were eligible for each synthesis (e.g. tabulating the study intervention characteristics and comparing against the planned groups for each synthesis (item #5)).	
	13b	Describe any methods required to prepare the data for presentation or synthesis, such as handling of missing summary statistics, or data conversions.	Σελ. 38
	13c	Describe any methods used to tabulate or visually display results of individual studies and syntheses.	

Section and Topic	Item #	Checklist item	Location where item is reported
	13d	Describe any methods used to synthesize results and provide a rationale for the choice(s). If meta-analysis was performed, describe the model(s), method(s) to identify the presence and extent of statistical heterogeneity, and software package(s) used.	Σελ. 37, 38
	13e	Describe any methods used to explore possible causes of heterogeneity among study results (e.g. subgroup analysis, meta-regression).	Σελ. 38
	13f	Describe any sensitivity analyses conducted to assess robustness of the synthesized results.	Σελ. 38
Reporting bias assessment	14	Describe any methods used to assess risk of bias due to missing results in a synthesis (arising from reporting biases).	
Certainty assessment	15	Describe any methods used to assess certainty (or confidence) in the body of evidence for an outcome.	
RESULTS			
Study selection	16a	Describe the results of the search and selection process, from the number of records identified in the search to the number of studies included in the review, ideally using a flow diagram.	Σελ. 39, 40
	16b	Cite studies that might appear to meet the inclusion criteria, but which were excluded, and explain why they were excluded.	Σελ. 40
Study characteristics	17	Cite each included study and present its characteristics.	Σελ. 39
Risk of bias in studies	18	Present assessments of risk of bias for each included study.	Σελ. 51
Results of individual studies	19	For all outcomes, present, for each study: (a) summary statistics for each group (where appropriate) and (b) an effect estimate and its precision (e.g. confidence/credible interval), ideally using structured tables or plots.	Σελ. 46-49
Results of syntheses	20a	For each synthesis, briefly summarise the characteristics and risk of bias among contributing studies.	Σελ. 52
	20b	Present results of all statistical syntheses conducted. If meta-analysis was done, present for each the summary estimate and its precision (e.g. confidence/credible interval) and measures of statistical heterogeneity. If comparing groups, describe the direction of the effect.	Σελ. 52
	20c	Present results of all investigations of possible causes of heterogeneity among study results.	Σελ. 52, 54, 55
	20d	Present results of all sensitivity analyses conducted to assess the robustness of the synthesized results.	
Reporting biases	21	Present assessments of risk of bias due to missing results (arising from reporting biases) for each synthesis assessed.	
Certainty of evidence	22	Present assessments of certainty (or confidence) in the body of evidence for each outcome assessed.	
DISCUSSION			
Discussion	23a	Provide a general interpretation of the results in the context of other evidence.	Σελ. 58-60
	23b	Discuss any limitations of the evidence included in the review.	Σελ. 59
	23c	Discuss any limitations of the review processes used.	Σελ. 61
	23d	Discuss implications of the results for practice, policy, and future research.	Σελ. 62
OTHER INFORMATION			
	24a	Provide registration information for the review, including register name and registration number, or state that the review was not registered.	

Section and Topic	Item #	Checklist item	Location where item is reported
Registration and protocol	24b	Indicate where the review protocol can be accessed, or state that a protocol was not prepared.	
	24c	Describe and explain any amendments to information provided at registration or in the protocol.	
Support	25	Describe sources of financial or non-financial support for the review, and the role of the funders or sponsors in the review.	
Competing interests	26	Declare any competing interests of review authors.	
Availability of data, code and other materials	27	Report which of the following are publicly available and where they can be found: template data collection forms; data extracted from included studies; data used for all analyses; analytic code; any other materials used in the review.	

PROBAST RISK OF BIAS AND APPLICABILITY ASSESSMENT

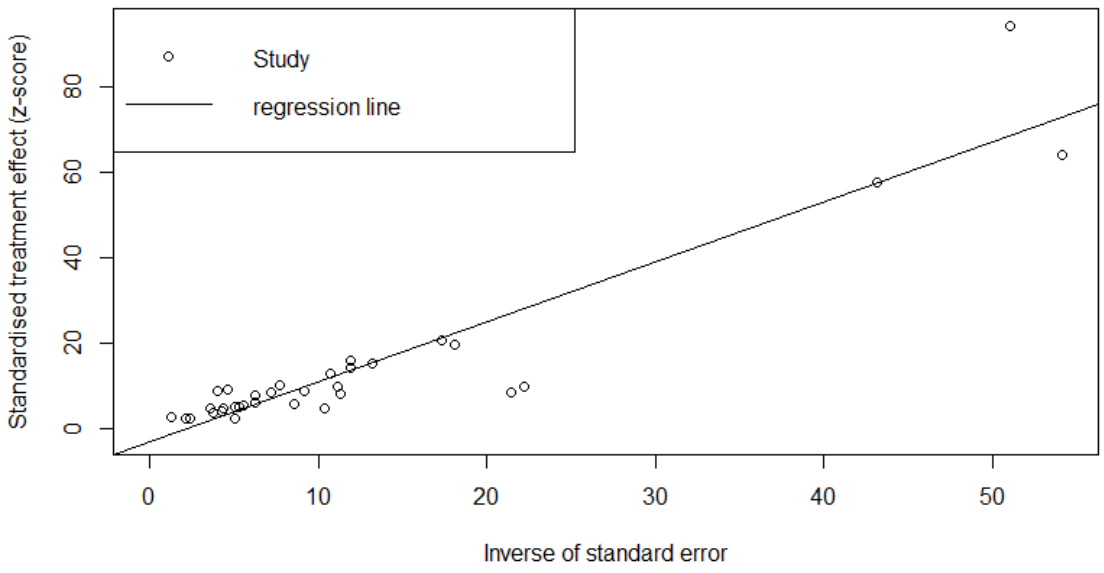
DOMAIN 1: Participants			
A. Risk of Bias			
Describe the sources of data and criteria for participant selection:			
		Dev	*Val*
1.1 Were appropriate data sources used, e.g. cohort, RCT or nested case-control study data?			
1.2 Were all inclusions and exclusions of participants appropriate?			
Risk of bias introduced by selection of participants	**RISK:** *(low/ high/ unclear)*		
Rationale of bias rating:			
B. Applicability			
Describe included participants, setting and dates:			
Concern that the included participants and setting do not match the review question	**CONCERN:** *(low/ high/ unclear)*		
Rationale of applicability rating:			

DOMAIN 2: Predictors			
A. Risk of Bias			
List and describe predictors included in the final model, e.g. definition and timing of assessment:			
		Dev	*Val*
2.1 Were predictors defined and assessed in a similar way for all participants?			
2.2 Were predictor assessments made without knowledge of outcome data?			
2.3 Are all predictors available at the time the model is intended to be used?			
Risk of bias introduced by predictors or their assessment	**RISK:** *(low/ high/ unclear)*		
Rationale of bias rating:			
B. Applicability			
Describe included participants, setting and dates:			
Concern that the definition, assessment or timing of predictors in the model do not match the review question	**CONCERN:** *(low/ high/ unclear)*		
Rationale of applicability rating:			

DOMAIN 3:Outcome			
A. Risk of Bias			
Describe the outcome, how it was defined and determined, and the time interval between predictor assessment and outcome determination:			
		Dev	*Val*
3.1 Was the outcome determined appropriately?			
3.2 Was a pre-specified or standard outcome definition used?			
3.3 Were predictors excluded from the outcome definition?			
3.4 Was the outcome defined and determined in a similar way for all participants?			
3.5 Was the outcome determined without knowledge of predictor information?			
3.6 Was the time interval between predictor assessment and outcome determination appropriate?			
Risk of bias introduced by the outcome or its determination	**RISK:** *(low/ high/ unclear)*		
Rationale of bias rating:			
B. Applicability			
At what time point was the outcome determined: *If a composite outcome was used, describe the relative frequency/distribution of each contributing outcome:*			
Concern that the outcome, its definition, timing or determination do not match the review question	**CONCERN:** *(low/ high/ unclear)*		
Rationale of applicability rating:			

DOMAIN 4:Analysis			
A. Risk of Bias			
Describe numbers of participants, number of candidate predictors, outcome events and events per candidate predictor:			
Describe how the model was developed (for example in regards to modelling technique (e.g. survival or logistic modelling), predictor selection, and risk group definition):			
Describe whether and how the model was validated, either internally (e.g. bootstrapping, cross validation, random split sample) or externally (e.g. temporal validation, geographical validation, different setting, different type of participants):			
Describe the performance measures of the model, e.g. (re)calibration, discrimination, (re)classification, net benefit, and whether they were adjusted for optimism:			
Describe any participants who were excluded from the analysis:			
Describe missing data on predictors and outcomes as well as methods used for missing data:			
		Dev	*Val*
4.1 Were there a reasonable number of participants with the outcome?			
4.2 Were continuous and categorical predictors handled appropriately?			
4.3 Were all enrolled participants included in the analysis?			
4.4 Were participants with missing data handled appropriately?			
4.5 Was selection of predictors based on univariable analysis avoided?			
4.6 Were complexities in the data (e.g. censoring, competing risks, sampling of controls) accounted for appropriately?			
4.7 Were relevant model performance measures evaluated appropriately?			
4.8 Were model overfitting and optimism in model performance accounted for?			
4.9 Do predictors and their assigned weights in the final model correspond to the results from multivariable analysis?			
Risk of bias introduced by the analysis	**RISK:** *(low/ high/ unclear)*		
Rationale of bias rating:			

Γράφημα. Έλεγχος του Egger, γραμμή παλινδρόμησης με τον αντίστροφο του τυπικού σφάλματος στον άξονα x



Ενδεικτικός κώδικας σε R (έκδοση 4.0.3)

```
library(readxl)

library(meta)

library(metamisc)

library(dplyr)


data<-read_excel("~/Data_Rc.xlsx", skip = 2)

data<-subset(data,data$Model=="SAPS II")

temp<-read_excel("~/newdata.xlsx", skip = 1)


##SE & 95% CI calculation

est1 <- ccalc(cstat = AUC, cstat.se = AUC_SE, cstat.cilb = AUC_CI_lb,

        cstat.ciub = AUC_CI_ub, N = Sample_size, O = Events, data = data, slab =Author)

##logit transformations

est2 <- ccalc(cstat = AUC, cstat.se = AUC_SE, cstat.cilb = AUC_CI_lb,

        cstat.ciub = AUC_CI_ub, N = Sample_size, O = Events, data = data, slab = Author,

        g = "log(cstat/(1-cstat))")

##merge with dataset

data$AUC_SE<-est1$theta.se

data$AUC_CI_lb<-est1$theta.cilb

data$AUC_CI_ub<-est1$theta.ciub

data$logitAUC<-est2$theta

data$logitAUC_SE<-est2$theta.se

data$logitAUC_CI_lb<-est2$theta.cilb

data$logitAUC_CI_ub<-est2$theta.ciub


##remove unwanted observations
```

```
newdata<-subset(data,data$Outcome!="1-year mortality"& data$Outcome!="72h mortality")

newdata=newdata[-c(5,6),]

newdata2=newdata[!duplicated(newdata$Study_ID),]


##unite 28-day & 30-day mortality into one category

newdata$Outcome=as.factor(newdata$Outcome)

newdata$Outcome_analysis=recode(newdata$Outcome, "28-day mortality"="28 & 30-day mortality",

                        "30-day mortality"="28 & 30-day mortality")

newdata$confint=temp$Conf_int


##random effect meta analysis

random_reml<-metagen(TE=newdata$logitAUC,seTE = newdata$logitAUC_SE,studlab =
newdata$Author,

                data=newdata,fixed=FALSE,random = TRUE,method.tau = "REML",hakn=TRUE,title =
"SAPS II validation")

random_reml


##subgroup analysis

sub_out<-update.meta(random_reml,subgroup = newdata$Outcome_analysis,tau.common = TRUE)

sub_rob<-update.meta(random_reml,subgroup = newdata$ROB_overall,tau.common = FALSE)


##inv logit transformation

invlogit=function(x){

  y=1/(1+(exp(-x)))

  return(y)

}
```

```
##forest plots

pdf(file = "forestplot.pdf", width = 12, height =14 )

forest.meta(random_reml, studlab = TRUE, sortvar = studlab, common = FALSE, print.tau2 = TRUE,
ref=NA, xlim = c(0,3), smlab = "SAPS II", leftcols=c("studlab","Sample_size", "AUC","confint"),
leftlabs=c("Author","Sample Size","AUC","95% CI"), rightlabs=c( "Logit(AUROC)","95%
CI","Weight(random)"))

dev.off()



png(file = "forestplot.png", width = 3500, height =2556,res = 300 )

forest.meta(random_reml, studlab = TRUE, sortvar = studlab, common = FALSE, prediction = TRUE,
print.tau2 = TRUE, ref=NA, xlim = c(0,3), smlab=”SAPS II”, leftcols=c("studlab","Sample_size",
"AUC","confint"), leftlabs=c("Author","Sample Size","AUC","95% CI"), rightlabs=c(
"Logit(AUROC)","95% CI","Weight(random)"))

dev.off()



pdf(file = "forestplot_out.pdf", width = 12, height =14 )



forest.meta(sub_out, studlab = TRUE, sortvar = studlab, common = FALSE, ref = NA, xlim = c(-1,4),
prediction = TRUE, print.tau2 = TRUE, smlab="SAPS II", colgap.forest.left = "1cm",
leftcols=c("studlab","Sample_size", "AUC","confint"), leftlabs=c("Author","Sample Size","AUC","95%
CI"), rightlabs=c( "Logit(AUROC)","95% CI","Weight(random)"), col.subgroup = "dark blue")

dev.off()



png(file = "forestplot_out.png", width = 3500, height =3500,res = 300 )

forest.meta(sub_out, studlab = TRUE, sortvar = studlab, common = FALSE, ref = NA, xlim = c(-1,4),
prediction = TRUE, print.tau2 = TRUE, smlab="SAPS II", colgap.forest.left = "1cm",
leftcols=c("studlab","Sample_size", "AUC","confint"), leftlabs=c("Author","Sample Size","AUC","95%
CI"), rightlabs=c( "Logit(AUROC)","95% CI","Weight(random)"), col.subgroup = "dark blue")

dev.off()



pdf(file = "forestplot_rob.pdf", width = 12, height =14 )
```

```
forest.meta(sub_rob, studlab = TRUE, sortvar = studlab, common = FALSE, ref = NA, xlim = c(-1,4),
prediction = TRUE, print.tau2 = TRUE, smlab="SAPS II", colgap.forest.left = "1cm",
leftcols=c("studlab","Sample_size", "AUC","confint"), leftlabs=c("Author","Sample Size","AUC","95%
CI"), rightlabs=c( "Logit(AUROC)","95% CI","Weight(random)"), col.subgroup = "dark blue")

dev.off()


png(file = "forestplot_rob.png", width = 3500, height =3500,res = 300 )


forest.meta(sub_rob, studlab = TRUE, sortvar = studlab, common = FALSE, ref = NA, xlim = c(-1,4),
prediction = TRUE, print.tau2 = TRUE, smlab="SAPS II", colgap.forest.left = "1cm",
leftcols=c("studlab","Sample_size", "AUC","confint"), leftlabs=c("Author","Sample Size","AUC","95%
CI"), rightlabs=c( "Logit(AUROC)","95% CI","Weight(random)"), col.subgroup = "dark blue")

dev.off()

##funnel plot

funnel(random_reml,xlab = "Logit(AUC)",random = TRUE,common = FALSE,lty.random = 2,lwd.random
= 1,col = "dark blue",studlab = TRUE,cex.studlab =0.6 )

title(main = "Funnel plot")

col.contour = c("gray75", "gray85", "gray95")

funnel.meta(random_reml, xlim = c(-0.5, 2), contour = c(0.9, 0.95, 0.99),col.contour = col.contour)

legend(x = 1.6, y = 0.01,  legend = c("p < 0.1", "p < 0.05", "p < 0.01"),fill = col.contour)


##small study bias/egger test

metabias(random_reml,method.bias="Egger",plotit=TRUE,)

legend(x="topleft",c("Study", "regression line"),lty = c(NA,1),pch = c(1,NA),col = c(1,1))

z<-metareg(random_reml,formula = ~newdata$ROB_overall)


###sensitivity analysis

newdata3=subset(newdata,subset = newdata$ROB_overall!="High")

random_reml3<-metagen(TE=newdata3$logitAUC,seTE = newdata3$logitAUC_SE,studlab =
newdata3$Author,data=newdata3,fixed=FALSE,random = TRUE,method.tau = "REML",hakn=TRUE,title
= "SAPS II validation"

newdata4=subset(newdata,subset = newdata$Study_ID!=10)
```

```
random_reml4<-metagen(TE=newdata4$logitAUC,seTE = newdata4$logitAUC_SE,studlab =
newdata4$Author, data=newdata4,fixed=FALSE,random = TRUE,method.tau = "REML",hakn=TRUE,title
= "SAPS II validation")

newdata5=subset(newdata,subset = newdata$Study_ID!=23)

random_reml5<-metagen(TE=newdata5$logitAUC,seTE = newdata5$logitAUC_SE,studlab =
newdata5$Author, data=newdata5,fixed=FALSE,random = TRUE,method.tau = "REML",hakn=TRUE,title
= "SAPS II validation")

newdata6=subset(newdata,subset = newdata$Study_ID!=4)

random_reml6<-metagen(TE=newdata6$logitAUC,seTE = newdata6$logitAUC_SE,studlab =
newdata6$Author, data=newdata6,fixed=FALSE,random = TRUE,method.tau = "REML",hakn=TRUE,title
= "SAPS II validation")

newdata7=subset(newdata,subset = newdata$Study_ID!=22)

random_reml7<-metagen(TE=newdata7$logitAUC,seTE = newdata7$logitAUC_SE,studlab =
newdata7$Author, data=newdata7,fixed=FALSE,random = TRUE,method.tau = "REML",hakn=TRUE,title
= "SAPS II validation")

newdata8=subset(newdata,subset = newdata$Study_ID!=17& newdata$Study_ID!=4
&newdata$Study_ID!=22)

random_reml8<-metagen(TE=newdata8$logitAUC,seTE = newdata8$logitAUC_SE,studlab =
newdata8$Author, data=newdata8,fixed=FALSE,random = TRUE,method.tau = "REML",hakn=TRUE,title
= "SAPS II validation")

###################

##alternative package for random-effects meta-analysis

library(metafor)

test<-rma.uni(yi=newdata$logitAUC,sei=newdata$logitAUC_SE,method = "REML",test = "knha")
```

ΒΙΒΛΙΟΓΡΑΦΙΑ

Alvarez, M. *et al.* (1998) ‘Mortality prediction in head trauma patients: performance of Glasgow Coma Score and general severity systems.’, *Critical care medicine*, 26(1), pp. 142–148. Available at: https://doi.org/10.1097/00003246-199801000-00030.

Álvarez-Avello, J.M. *et al.* (2022) ‘Usefulness of severity scales for cardiogenic shock in-hospital mortality. Proposal for a new prognostic model.’, *Revista espanola de anestesiologia y reanimacion*, 69(2), pp. 79–87. Available at: https://doi.org/10.1016/j.redare.2021.03.010.

Baker, T. and Gerdin, M. (2017) ‘European Journal of Internal Medicine The clinical usefulness of prognostic prediction models in critical illness’, *European Journal of Internal Medicine*, pp. 10–13. Available at: https://doi.org/10.1016/j.ejim.2017.09.012.

Barboi, C., Tzavelis, A. and Muhammad, L.N. (2022) ‘Comparison of Severity of Illness Scores and Artificial Intelligence Models That Are Predictive of Intensive Care Unit Mortality: Meta-analysis and Review of the Literature’, *JMIR Medical Informatics*, 10(5), p. e35293. Available at: https://doi.org/10.2196/35293.

Calster, B.V. *et al.* (2019) ‘Calibration : the Achilles heel of predictive analytics’, pp. 1–7.

Ch 15: Evaluation of performance (2022). Available at: https://www.clinicalpredictionmodels.org/extra-material/chapter-15.

Chen, J. *et al.* (2022) ‘A nomogram to predict the in-hospital mortality of patients with congestive heart failure and chronic kidney disease.’, *ESC heart failure* [Preprint]. Available at: https://doi.org/10.1002/ehf2.14042.

Choi, M.H. *et al.* (2022) ‘Mortality prediction of patients in intensive care units using machine learning algorithms based on electronic health records.’, *Scientific reports*, 12(1), p. 7180. Available at: https://doi.org/10.1038/s41598-022-11226-4.

Collins, G.S. *et al.* (2015) ‘Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement’, *Annals of Internal Medicine*, 162(1), pp. 55–63. Available at: https://doi.org/10.7326/M14-0697.

Csiszar, B. *et al.* (2022) ‘L-arginine, asymmetric and symmetric dimethylarginine for early outcome prediction in unselected cardiac arrest victims: a prospective cohort study.’, *Internal and emergency medicine*, 17(2), pp. 525–534. Available at: https://doi.org/10.1007/s11739-021-02767-z.

Damen, J.A.A. *et al.* (2022) ‘How to conduct a systematic review and meta-analysis of prognostic model studies’, *Clinical Microbiology and Infection* [Preprint], (xxxx). Available at: https://doi.org/10.1016/j.cmi.2022.07.019.

Debray, T.P.A. *et al.* (2017) ‘A guide to systematic review and meta-analysis of prediction model performance’, *BMJ (Online)*, 356. Available at: https://doi.org/10.1136/bmj.i6460.

Debray, T.P.A. *et al.* (2019) ‘A framework for meta-analysis of prediction model studies with binary and time-to-event outcomes’, *Statistical Methods in Medical Research*, 28(9), pp. 2768–2786. Available at: https://doi.org/10.1177/0962280218785504.

Foroutan, F. *et al.* (2020) ‘GRADE Guidelines 28: Use of GRADE for the assessment of evidence about prognostic factors: rating certainty in identification of groups of patients with different absolute risks’, *Journal of Clinical Epidemiology*, 121, pp. 62–70. Available at: https://doi.org/10.1016/j.jclinepi.2019.12.023.

Foroutan, F. *et al.* (2022) ‘GRADE concept paper 2: Concepts for judging certainty on the calibration of prognostic models in a body of validation studies’, *Journal of Clinical Epidemiology*, 143, pp. 202–211. Available at: https://doi.org/10.1016/j.jclinepi.2021.11.024.

Geersing, G.-J. *et al.* (2012) ‘Search filters for finding prognostic and diagnostic prediction studies in Medline to enhance systematic reviews’, *PloS One*, 7(2), p. e32844. Available at: https://doi.org/10.1371/journal.pone.0032844.

Ghaffar, S., Pearse, R.M. and Gillies, M.A. (2017) ‘ICU admission after surgery : who benefits ?’ Available at: https://doi.org/10.1097/MCC.0000000000000448.

Hai, P.D. and Viet Hoa, L.T. (2022) ‘The Prognostic Accuracy Evaluation of mNUTRIC, APACHE II, SOFA, and SAPS 2 Scores for Mortality Prediction in Patients with Sepsis.’, *Critical care research and practice*, 2022, p. 4666594. Available at: https://doi.org/10.1155/2022/4666594.

Han, X. *et al.* (2022) ‘Developing and validating a prediction model for in-hospital mortality in patients with ventilator-associated pneumonia in the ICU.’, *Annals of palliative medicine*, 11(5), pp. 1799–1810. Available at: https://doi.org/10.21037/apm-22-502.

Harrer, M. *et al.* (2022) *Doing meta-analysis with R: a hands-on guide*. First edition. Boca Raton: CRC Press.

Kahraman, F. *et al.* (2022) ‘Predictive outcomes of APACHE II and expanded SAPS II mortality scoring systems in coronary care unit.’, *International journal of cardiology*, pp. S0167-5273(22)01414–0. Available at: https://doi.org/10.1016/j.ijcard.2022.09.065.

Katz, S. *et al.* (2022) ‘Decision support system and outcome prediction in a cohort of patients with necrotizing soft-tissue infections.’, *International journal of medical informatics*, 167, p. 104878. Available at: https://doi.org/10.1016/j.ijmedinf.2022.104878.

Keuning, B.E. *et al.* (2020) ‘Mortality prediction models in the adult critically ill: A scoping review’, *Acta Anaesthesiologica Scandinavica*, 64(4), pp. 424–442. Available at: https://doi.org/10.1111/aas.13527.

Krasselt, M. *et al.* (2022) ‘Sepsis Mortality Is high in Patients With Connective Tissue Diseases Admitted to the Intensive Care Unit (ICU).’, *Journal of intensive care medicine*, 37(3), pp. 401–407. Available at: https://doi.org/10.1177/0885066621996257.

Langan, D. *et al.* (2019) ‘A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses’, *Research Synthesis Methods*, 10(1), pp. 83–98. Available at: https://doi.org/10.1002/jrsm.1316.

Le Gall, J.R. *et al.* (2005) ‘Mortality prediction using SAPS II: an update for French intensive care units.’, *Critical care (London, England)*, 9(6), pp. R645-652. Available at: https://doi.org/10.1186/cc3821.

Liu, J. *et al.* (2022) ‘Association of Red Cell Distribution Width-to-Platelet Ratio and Mortality in Patients with Sepsis.’, *Mediators of inflammation*, 2022, p. 4915887. Available at: https://doi.org/10.1155/2022/4915887.

Liu, Y. *et al.* (2022) ‘Development and validation of a predictive model for in-hospital mortality in patients with sepsis-associated liver injury.’, *Annals of translational medicine*, 10(18), p. 997. Available at: https://doi.org/10.21037/atm-22-4319.

Liu, Z. *et al.* (2022) ‘A prediction model with measured sentiment scores for the risk of in-hospital mortality in acute pancreatitis: a retrospective cohort study.’, *Annals of translational medicine*, 10(12), p. 676. Available at: https://doi.org/10.21037/atm-22-1613.

Lu, Y., Zhang, Q. and Jiang, J. (2022) ‘Development and validation of a prediction model for in-hospital mortality of patients with severe thrombocytopenia.’, *Scientific reports*, 12(1), p. 6316. Available at: https://doi.org/10.1038/s41598-022-10438-y.

Mirzakhani, F. *et al.* (2022) ‘Which model is superior in predicting ICU survival: artificial intelligence versus conventional approaches.’, *BMC medical informatics and decision making*, 22(1), p. 167. Available at: https://doi.org/10.1186/s12911-022-01903-9.

Moons, K.G.M. *et al.* (2015) ‘Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): Explanation and elaboration’, *Annals of Internal Medicine*, 162(1), pp. W1–W73. Available at: https://doi.org/10.7326/M14-0698.

Moreno-Torres, V. *et al.* (2022) ‘Red blood cell distribution width as prognostic factor in sepsis: A new use for a classical parameter.’, *Journal of critical care*, 71, p. 154069. Available at: https://doi.org/10.1016/j.jcrc.2022.154069.

Moser, A. *et al.* (2022) ‘Mortality prediction in intensive care units including premorbid functional status improved performance and internal validity.’, *Journal of clinical epidemiology*, 142, pp. 230–241. Available at: https://doi.org/10.1016/j.jclinepi.2021.11.028.

Nagelkerke, N.J.D. (1991) ‘Miscellanea A note on a general definition of the coefficient of determination’, p. 4.

Nassar, A.P., Malbouisson, L.M.S. and Moreno, R. (2014) ‘Evaluation of Simplified Acute Physiology Score 3 performance: a systematic review of external validation studies.’, *Critical care (London, England)*, 18(3), p. R117. Available at: https://doi.org/10.1186/cc13911.

Niederman, M.S. and Berger, J.T. (2010) ‘The delivery of futile care is harmful to other patients’, 38(10). Available at: https://doi.org/10.1097/CCM.0b013e3181f1cba5.

Nistal-Nuño, B. (2022) ‘Developing machine learning models for prediction of mortality in the medical intensive care unit.’, *Computer methods and programs in biomedicine*, 216, p. 106663. Available at: https://doi.org/10.1016/j.cmpb.2022.106663.

Palazón-Bru, A. *et al.* (2020) ‘A general presentation on how to carry out a CHARMS analysis for prognostic multivariate models’, *Statistics in Medicine*, 39(23), pp. 3207–3225. Available at: https://doi.org/10.1002/sim.8660.

Qi, J. *et al.* (2022) ‘A nomogram to predict in-hospital mortality for post-gastrointestinal resection surgery patients in intensive care units: A retrospective cohort study.’, *American journal of surgery*, 223(6), pp. 1162–1166. Available at: https://doi.org/10.1016/j.amjsurg.2021.11.031.

Rahmatinejad, Z. *et al.* (2022) ‘Internal Validation of the Predictive Performance of Models Based on Three ED and ICU Scoring Systems to Predict Inhospital Mortality for Intensive Care Patients Referred from the Emergency Department.’, *BioMed research international*, 2022, p. 3964063. Available at: https://doi.org/10.1155/2022/3964063.

Ramspek, C.L. *et al.* (2021) ‘External validation of prognostic models: What, why, how, when and where?’, *Clinical Kidney Journal*, 14(1), pp. 49–58. Available at: https://doi.org/10.1093/ckj/sfaa188.

Ren, Y. *et al.* (2022) ‘Risk factor analysis and nomogram for predicting in-hospital mortality in ICU patients with sepsis and lung infection.’, *BMC pulmonary medicine*, 22(1), p. 17. Available at: https://doi.org/10.1186/s12890-021-01809-8.

Riley, R.D. *et al.* (2019) ‘A guide to systematic review and meta-analysis of prognostic factor studies’, *BMJ (Online)*, 364. Available at: https://doi.org/10.1136/bmj.k4597.

Rong, F. *et al.* (2022) ‘Machine Learning for Prediction of Outcomes in Cardiogenic Shock.’, *Frontiers in cardiovascular medicine*, 9, p. 849688. Available at: https://doi.org/10.3389/fcvm.2022.849688.

Shorten, A. and Shorten, B. (2013) ‘What is meta-analysis ?’, 16(1), pp. 2012–2013.

Snell, K.I.E. *et al.* (2018) 'Meta-analysis of prediction model performance across multiple studies: Which scale helps ensure between-study normality for the C-statistic and calibration measures?', *Statistical Methods in Medical Research*, 27(11), pp. 3505–3522. Available at: https://doi.org/10.1177/0962280217705678.

Steyerberg, E.W. *et al.* (2010) 'Assessing the performance of prediction models: A framework for traditional and novel measures', *Epidemiology*, 21(1), pp. 128–138. Available at: https://doi.org/10.1097/EDE.0b013e3181c30fb2.

Steyerberg, E.W. (2019) *Statistics for Biology and Health Clinical Prediction Models A Practical Approach to Development, Validation, and Updating Second Edition*, pp. 301–330. Available at: http://www.springer.com/series/2848.

Tokur, M.E. *et al.* (2022) 'Mortality predictors on the day of healthcare-associated Acinetobacter baumanni bacteremia in intensive care unit.', *Journal of infection in developing countries*, 16(9), pp. 1473–1481. Available at: https://doi.org/10.3855/jidc.16902.

Van Calster, B. *et al.* (2016) 'A calibration hierarchy for risk models was defined: from utopia to empirical data', *Journal of Clinical Epidemiology*, 74, pp. 167–176. Available at: https://doi.org/10.1016/j.jclinepi.2015.12.005.

Vincent, J.L. (2009) 'CHAPTER 9 -General Illness Severity Scores', in *Critical Care Nephrology*. 2nd edition, pp. 55–60.

Wang, N. *et al.* (2022) 'The predictive value of the Oxford Acute Severity of Illness Score for clinical outcomes in patients with acute kidney injury.', *Renal failure*, 44(1), pp. 320–328. Available at: https://doi.org/10.1080/0886022X.2022.2027247.

Wolff, R.F. *et al.* (2019) 'PROBAST: A tool to assess the risk of bias and applicability of prediction model studies', *Annals of Internal Medicine*, 170(1), pp. 51–58. Available at: https://doi.org/10.7326/M18-1376.

Wong, J. *et al.* (2018) 'Derivation and validation of a multivariable model to predict when primary care physicians prescribe antidepressants for indications other than depression', *Clinical Epidemiology*, Volume 10, pp. 457–474. Available at: https://doi.org/10.2147/CLEP.S153000.

Wu, J. *et al.* (2022) ‘Red Cell Distribution Width to Platelet Ratio Is Associated with Increasing In-Hospital Mortality in Critically Ill Patients with Acute Kidney Injury.’, *Disease markers*, 2022, p. 4802702. Available at: https://doi.org/10.1155/2022/4802702.

Zou, J. *et al.* (2022) ‘Development and validation of a nomogram to predict the 30-day mortality risk of patients with intracerebral hemorrhage.’, *Frontiers in neuroscience*, 16, p. 942100. Available at: https://doi.org/10.3389/fnins.2022.942100.