ORIGINAL ARTICLE DOI: 10.20515/otd. 1990468

Osmangazi Journal of Med, 2026;48(5):912-923

Osmangazi Journal of Medicine

e-ISSN: 2587-1579

Elektronik Sağlık Kayıtlarında Dengesiz Sınıf Probleminde Dengeleme Stratejilerinin Ayrım ve Kalibrasyon Üzerine Etkisi: Regülarize Lojistik Regresyon ile Gradient Boosting'in Karşılaştırmalı Bir Simülasyon Çalışması

The Effect of Class-Imbalance Correction Strategies on Discrimination and Calibration in Electronic Health Records: A Comparative Simulation Study of Regularized Logistic Regression and Gradient Boosting

Muzaffer Bilgin, Ertuğrul Çolak

Eskişehir Osmangazi Üniversitesi Tıp Fakültesi, Biyoistatistik Anabilim Dalı, Eskişehir, Türkiye

Correspondence / Sorumlu yazar
Muzaffer BİLGİN

Eskişehir Osmangazi Üniversitesi Tıp Fakültesi, Biyoistatistik Anabilim Dalı, Eskişehir, Türkiye

e-mail: muzaffer.bilgin@yahoo.com

Received : 09.07.2026

Accepted : 20.07. 2026

Ethics Committee Approval: This study did not use real patient data; all analyses were performed on synthetic data generated entirely by simulation. Therefore, Ethics Committee approval was not required.

Informed Consent: Informed consent was not required, as no real patient/participant data were used in this study.

Authorship Contributions: Konsept: MB, EÇ. Tasarım: MB, EÇ. Veri Toplama veya İşleme: MB. Analiz veya Yorum: MB, EÇ. Literatür Taraması: MB, EÇ. Yazma: MB, EÇ.

Copyright Transfer Form: Copyright Transfer Form was signed by author.

Peer-review: Externally peer-reviewed.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: The author declared that this study received no financial support

Abstract: Outcome events in EHR-based clinical risk prediction are often rare, causing class imbalance; resampling and cost-sensitive corrections have understudied effects on calibration and net benefit. This study compared regularized logistic regression (LR) and gradient boosting (GB) under four correction strategies, and evaluated the corrective effect of recalibration. A synthetic dataset (n=4000; event rate 12%, resembling acute kidney injury) with 15 correlated predictors was generated (Monte Carlo design, 200 replications). LR and GB (XGBoost) were evaluated under no balancing, SMOTE, class weighting, and undersampling via 5-fold nested cross-validation. AUROC, AUPRC, Brier score, calibration slope/intercept, and Integrated Calibration Index (ICI) were computed; clinical utility was assessed by decision curve analysis. The highest discrimination was achieved by unbalanced LR (AUROC 0.706; AUPRC 0.273). Balancing did not improve LR discrimination and worsened GB's (SMOTE: AUROC 0.681 to 0.632); the LR–GB difference favored LR throughout. Unbalanced models were well calibrated (intercept ≈0.00, slope 1.024, ICI 0.007); balancing severely distorted calibration (intercept ≈−2.0; ICI 0.32–0.34). Unbalanced LR gave the highest net benefit across all thresholds; SMOTE-applied GB the lowest. Intercept recalibration largely reversed the prevalence-driven harm (Brier 0.215→0.099) but did not repair GB's ranking damage. Balancing did not improve discrimination and impaired calibration and net benefit (the "balancing paradox"). Relying solely on AUROC is misleading; calibration and net benefit must always be reported, and balancing must be followed by recalibration. Regularized LR without balancing was a robust default choice.

Keywords: Class imbalance; calibration; logistic regression; gradient boosting; decision curve analysis; clinical prediction model

Özet: Elektronik sağlık kayıtlarına (ESK) dayalı klinik risk tahmininde sonuç olayları çoğunlukla nadirdir ve sınıf dengesizliği oluşur. Yaygın kullanılan yeniden örnekleme ve maliyet-duyarlı yöntemlerin kalibrasyon ve klinik net fayda üzerindeki etkisi yeterince incelenmemiştir. Bu çalışmada regülarize lojistik regresyon (LR) ile gradient boosting (GB), dört dengeleme stratejisi altında karşılaştırılmış; ayrıca yeniden kalibrasyonun düzeltici etkisi değerlendirilmiştir. Akut böbrek hasarına benzer, olay oranı %12 olan bir sonuç için 15 öngörücülü sentetik veri (n=4000; 200 replikasyonlu Monte Carlo tasarımı) üretildi. LR ve GB (XGBoost); dengeleme yok, SMOTE, sınıf ağırlıklandırma ve azaltarak örnekleme stratejileriyle 5-katmanlı iç içe çapraz doğrulamayla değerlendirildi. AUROC, AUPRC, Brier skoru, kalibrasyon eğimi/kesişimi ve Entegre Kalibrasyon İndeksi (ICI) hesaplandı; klinik fayda karar eğrisi analiziyle değerlendirildi. En yüksek ayrımı dengelemesiz LR sağladı (AUROC 0,706; AUPRC 0,273). Dengeleme LR ayrımını artırmadı, GB ayrımını kötüleştirdi (SMOTE: AUROC 0,681'den 0,632'ye); LR–GB farkı tüm stratejilerde LR lehineydi. Dengelemesiz modeller iyi kalibre idi (kesişim ≈0,00, eğim 1,024, ICI 0,007); dengeleme kalibrasyonu ciddi biçimde bozdu (kesişim ≈ −2,0; ICI 0,32–0,34). Dengelemesiz LR tüm eşiklerde en yüksek net faydayı sağladı; SMOTE uygulanmış GB en düşük faydayı verdi. Yeniden kalibrasyon, prevalans kaynaklı zararı büyük ölçüde geri döndürdü (Brier 0,215→0,099); ancak GB'nin sıralama hasarını gideremedi. ESK temelli dengesiz sınıf probleminde dengeleme, ayrımı iyileştirmemiş, kalibrasyonu ve net faydayı bozmuştur ("dengeleme paradoksu"). Yalnızca AUROC'a güvenmek yanıltıcıdır; kalibrasyon ve net fayda mutlaka raporlanmalı, dengeleme kullanılacaksa yeniden kalibrasyon izlemelidir. Regülarize LR, dengeleme uygulanmadan, hem kalibrasyon hem klinik fayda açısından güçlü bir varsayılan seçenek olmuştur.

Anahtar Kelimeler: Sınıf dengesizliği; kalibrasyon; lojistik regresyon; gradient boosting; karar eğrisi analizi; klinik tahmin modeli

How to cite/ Atıf için: Bilgin M, Colak E, The Effect of Class-Imbalance Correction Strategies on Discrimination and Calibration in Electronic Health Records: A Comparative Simulation Study of Regularized Logistic Regression and Gradient Boosting, Osmangazi Journal of Medicine, 2026;48(5):912-923

1. Giriş

Elektronik sağlık kayıtları (ESK), rutin klinik bakım sırasında biriken yüksek boyutlu ve heterojen verileriyle, klinik sonuçların tahminine yönelik istatistiksel ve makine öğrenmesi temelli risk modellerinin geliştirilmesinde başlıca veri kaynağı hâline gelmiştir. Hastane içi mortalite, yeniden yatış, sepsis ve akut böbrek hasarı (ABH) gibi sonuçların erken ve güvenilir biçimde öngörülmesi, klinik karar desteği ve kaynak planlaması açısından önemli bir potansiyel sunmaktadır. Ancak bu modellerin klinik kullanıma değer üretebilmesi, yalnızca olayları "ayırt edebilmelerine" değil, ürettikleri olasılıkların gerçek riskle tutarlı (iyi kalibre) ve klinik karar verme açısından net fayda sağlayacak biçimde olmasına bağlıdır (1,2]).

ESK temelli sonuçların önemli bir bölümü doğası gereği nadirdir; ilgilenilen olay, gözlemlerin yalnızca küçük bir kısmında görülür. Bu durum, sınıf dengesizliği olarak adlandırılan ve model eğitimini olumsuz etkileyen bir yapı ortaya çıkarır. Dengesiz veride model, çoğunluk (olaysız) sınıfa doğru yanlılık gösterebilir; tüm gözlemleri çoğunluk sınıfına atayan bir model bile yüksek "doğruluk (accuracy)" değeri üretebilir. Bu nedenle dengesiz veride doğruluk gibi ölçütler yanıltıcıdır ve azınlık (klinik açıdan kritik) sınıfın yakalanması özel bir metodolojik dikkat gerektirir (3).

Bu sorunla başa çıkmak için en yaygın iki yaklaşım yeniden örnekleme ve maliyet-duyarlı öğrenmedir. Yeniden örnekleme, azınlık sınıfını çoğaltmayı (ör. sentetik azınlık aşırı örnekleme tekniği, SMOTE) veya çoğunluk sınıfını azaltmayı (undersampling) içerir (4); maliyet-duyarlı yaklaşımlar ise yanlış sınıflandırma maliyetlerini sınıf ağırlıklarıyla yeniden tanımlar. Bu tekniklerin azınlık sınıfının duyarlılığını (recall) ve eşik-bağımlı ölçütleri iyileştirdiği sıklıkla rapor edilmiştir.

Klinik tahmin modeli ailesinde iki temel yaklaşım öne çıkar. Regülarize lojistik regresyon (LR), yüksek boyutlu ESK verisinde aşırı uyumu önlemek için L1/L2/elastik ağ cezaları kullanır; yorumlanabilirliği, olasılık çıktısının doğal kalibrasyonu ve değişken seçimi avantajlarıyla biyoistatistikte yerleşik bir araçtır. Buna karşın gradient boosting (GB) gibi ağaç-temelli yöntemler, doğrusal olmayan ilişkileri ve değişken etkileşimlerini esnek biçimde yakalayarak çoğu zaman yüksek ayrım gücü sergiler; ancak yorumlanabilirlikleri düşüktür, hiperparametre seçimine duyarlıdırlar ve olasılık kalibrasyonları bozulabilir.

Mevcut karşılaştırmalı çalışmaların önemli bir kısmı, model başarımını neredeyse yalnızca ayrım ölçütleriyle (özellikle AUROC) değerlendirmekte; dengeleme stratejilerinin kalibrasyon ve klinik net fayda üzerindeki etkisini göz ardı etmektedir (5,6). Oysa dengeleme, eğitim verisindeki sınıf dağılımını yapay biçimde değiştirdiği için, modelin ürettiği olasılıkların gerçek prevalansla ilişkisini bozma potansiyeline sahiptir. van den Goorbergh ve ark. (2022) bu zararı lojistik regresyon için gösterse de (6), dengeleme stratejilerinin parametrik olmayan bir model ailesi (GB) üzerindeki etkisinin, model×strateji matrisi, istatistiksel anlamlılık testi ve karar eğrisi analiziyle birlikte sistematik biçimde karşılaştırıldığı çalışmalar — özellikle Türkçe metodoloji yazınında — sınırlıdır. Bu çalışmanın özgün katkısı tam da bu boşluğu doldurmasıdır: (i) GB'nin eklenmesiyle model ailesinin genişletilmesi, (ii) karar eğrisi (net fayda) analiziyle klinik maliyetin niceliklenmesi, (iii) ayrım farklarının istatistiksel anlamlılık ve replikasyon-temelli belirsizlikle sınanması ve (iv) dengeleme sonrası yeniden kalibrasyonun düzeltici etkisinin gösterilmesi.

Bu çalışmanın amacı, kontrollü bir simülasyon ortamında, ESK-benzeri dengesiz bir sonuç için regülarize LR ile GB modellerini dört dengeleme stratejisi (dengeleme yok, SMOTE, sınıf ağırlıklandırma, azaltarak örnekleme) altında ayrım, kalibrasyon ve klinik net fayda açısından eş zamanlı olarak karşılaştırmaktır. İki temel hipotez sınanmıştır: (H1) Dengeleme, ayrım ölçütlerini anlamlı biçimde iyileştirmez; (H2) Dengeleme, ayrım pahasına olmasa bile kalibrasyonu ve net faydayı bozar ve bu bozulma iki model ailesinde farklı şiddette ortaya çıkar.

2. Gereç ve Yöntem

2.1. Çalışma tasarımı ve veri

Çalışma, kontrollü koşullarda yöntem karşılaştırması yapmaya olanak veren bir Monte Carlo simülasyon çalışması olarak tasarlanmıştır. Gerçek hasta verisi kullanılmamış; tüm veriler, bilinen üretim mekanizmasına sahip sentetik ESK-benzeri veri kümeleri olarak üretilmiştir. Bu tasarım, gerçek veride gözlemlenemeyen "doğru" risk yapısının bilinmesini ve dengeleme stratejilerinin etkisinin yanlılık katmadan izole edilmesini sağlar. Ana

senaryoda birbirinden bağımsız nsim=200 veri kümesi üretilmiş; her replikasyonda tüm model–strateji kombinasyonları için tam analiz boru hattı yeniden yürütülmüştür. Bu yaklaşım, tek bir örneğe özgü şansa bağlı değişkenliği ortadan kaldırır ve her performans ölçütü için replikasyonlar arası ortalama ile Monte Carlo standart hatasının (MCSE) raporlanmasına olanak verir. Tüm analizler R 4.6.1 ortamında yürütülmüş (7); yeniden üretilebilirlik için tohum (set.seed) her replikasyonda deterministik biçimde sabitlenmiştir.

2.2. Sentetik veri üretim mekanizması

Akut böbrek hasarı benzeri ikili bir sonucu temsil eden, n=4000 gözlemli ve 15 öngörücülü bir veri kümesi üretildi. Öngörücüler; demografik (yaş), vital (sistolik kan basıncı, kalp hızı, solunum hızı, ateş), laboratuvar (baseline kreatinin, üre azotu, sodyum, potasyum, lökosit, laktat, hemoglobin) ve komorbidite (Charlson indeksi, diyabet, sepsis göstergesi) değişkenlerini içerecek biçimde, gerçekçi dağılımlar ve değişkenler arası korelasyon (ör. üre azotunun kreatinine bağımlılığı) ile oluşturuldu. Sonuç olasılığı; doğrusal terimlerin yanı sıra **doğrusal olmayan** (solunum hızının eşik üstü etkisi) ve **etkileşim** (kreatinin × laktat) terimleri içeren bir lojistik doğrusal öngörücüden üretildi. Kesişim terimi, olay oranını yaklaşık %12'ye sabitleyecek biçimde köklendirme (uniroot) ile ayarlandı. Bu yapı, hem parametrik (LR) hem parametrik olmayan (GB) modellere adil bir zemin sağlar: doğrusal olmayan yapı GB lehine, doğru belirtilmiş olasılık yapısı LR lehine işler.

2.3. Dengeleme stratejileri

Dört strateji karşılaştırıldı: (1) **Dengeleme yok** (temel); (2) **SMOTE** — azınlık sınıfının k-en yakın komşu enterpolasyonuyla tam dengeye (1:1) çoğaltılması; (3) **Sınıf ağırlıklandırma** — azınlık sınıfına çoğunluk/azınlık oranında ağırlık verilmesi (LR'de gözlem ağırlığı, GB'de scale_pos_weight); (4) **Azaltarak örnekleme** — çoğunluk sınıfının azınlık düzeyine indirgenmesi. Veri sızıntısını önlemek için her strateji yalnızca dış çapraz doğrulama eğitim katmanlarına uygulanmış; test katmanları ham (dengelenmemiş) ve dokunulmamış bırakılmıştır.

SMOTE, recipes/themis uygulamasıyla standart biçimde çalıştırıldı; sentetik örnekler tüm öngörücüler için sürekli özellik uzayında k-en yakın komşu enterpolasyonuyla üretildi. Kategorik ve ikili öngörücüler (diyabet, sepsis göstergesi) ile sayım tipi Charlson indeksi de bu sürekli enterpolasyona tabi tutulmuş olup, karma-tipli değişkenlere özgü SMOTE-NC gibi bir varyant kullanılmamıştır. Bu ayrıntı, aşağıda (Sınırlılıklar) tartışılan genellenebilirlik kapsamı açısından önemlidir.

2.4. Modelleme yöntemleri

Regülarize LR: Elastik ağ cezası (α=0,5) ile lojistik regresyon; ceza parametresi λ, eğitim katmanı içi 5-katmanlı çapraz doğrulamayla (sapma ölçütü) seçildi (glmnet) [8]. LR, öngörücülerin yalnızca ana etkileriyle (etkileşim veya doğrusal-olmayan terim elle eklenmeden) belirtildi; böylece veri üretimindeki doğrusal olmayan eşik etkisi ve kreatinin×laktat etkileşimi model tarafından bilinmeyen yapı olarak kaldı ve esnek GB için potansiyel avantaj korundu. **Gradient boosting:** XGBoost (9); ağaç derinliği {2, 4} üzerinde eğitim katmanı içi 3-katmanlı çapraz doğrulama ile (AUPRC ölçütü, en çok 400 tur, 25 turluk erken durdurma) ayarlandı; öğrenme oranı 0,1, satır alt örneklemesi 0,8 ve sütun alt örneklemesi 0,8 olarak sabitlendi. GB'de dengesizlik düzeltmesi yalnızca scale_pos_weight üzerinden yapıldı (sınıf ağırlıklandırma stratejisinde sınıf oranına ayarlandı); gözlem ağırlıkları LR'ye özgü tutularak çift düzeltme önlendi.

2.5. Doğrulama şeması

Model başarımı 5-katmanlı iç içe (nested) çapraz doğrulama ile değerlendirildi. Dış döngü başarımı tarafsız biçimde tahmin ederken, iç döngü her dış eğitim katmanında hiperparametre seçimini gerçekleştirdi. İki model ailesi için aynı katman bölünmeleri (tabakalı) kullanılarak adil karşılaştırma sağlandı. Çapraz doğrulama katmanları her replikasyonda yeniden ve tabakalı olarak çizilmiştir; dolayısıyla katman-seçimi (bölünme) değişkenliği, replikasyonlar arası dağılıma ve MCSE'ye doğrudan yansımaktadır (sabit-katman tasarımının aksine bu değişkenlik dışlanmamıştır). Dengeleme stratejileri veri sızıntısını önlemek amacıyla yalnızca eğitim katmanlarına uygulanmış; test katmanları ham (dengelenmemiş) bırakılmıştır. Her gözlem için test katmanında üretilen olasılıklar birleştirilerek katman-dışı (out-of-fold) olasılık vektörü oluşturuldu ve tüm metrikler bu vektör üzerinden hesaplandı.

2.6. Değerlendirme metrikleri

Üç eksende değerlendirme yapıldı. **Ayrım:** AUROC ve dengesiz veride birincil ölçüt olan AUPRC (eğri altı kesinlik-duyarlılık alanı) (10). AUPRC, kesinlik-duyarlılık eğrisi altındaki alanın dikdörtgen (adım) kuralıyla — enterpolasyonsuz, sağ Riemann toplamıyla — hesaplanmasıyla elde edildi (Davis–Goadrich enterpolasyonu kullanılmadı); "beceri-yok" tabanı olay prevalansına (≈0,12) eşittir ve

değerler buna göre yorumlanmalıdır. **Kalibrasyon:** kalibrasyon kesişimi (calibration-in-the-large, ideal=0) ve eğimi (ideal=1) ile Brier skoru. Kalibrasyon kesişimi, tahmin edilen olasılıkların logit dönüşümü ofset alınıp yalnızca kesişim terimini kestiren lojistik modelle [y ~ 1 + offset(logit(p̂))]; kalibrasyon eğimi ise logit(p̂)'nin tek öngörücü olduğu lojistik modelin [y ~ logit(p̂)] eğim katsayısıyla hesaplandı. Kalibrasyon ayrıca **biçim (shape) düzeyinde**, esnek (loess) kalibrasyon eğrisinden türetilen Entegre Kalibrasyon İndeksi (ICI), E50 ve E90 ile de değerlendirildi (Austin & Steyerberg, 2019) (11); bu ölçütler, tahmin edilen olasılık ile esnek eğriden okunan gözlenen olasılık arasındaki mutlak farkın sırasıyla ortalaması, medyanı ve 90. yüzdeliğidir.

Eşik-bağımlı ölçütler: Matthews korelasyon katsayısı (MCC), F1, dengeli doğruluk, duyarlılık ve özgüllük; bu ölçütler, %50 eşiğinin nadir olaylı modellerde anlamsızlaştığı dikkate alınarak olay prevalansına eşit eşikte (0,12) hesaplandı.

Klinik fayda: karar eğrisi analizi (net fayda) (12), 0–0,40 eşik aralığında hesaplanıp klinik-akla-yatkın aralığa (≈0,05–0,20) odaklanılarak sunuldu. Öngörücü başına olay sayısı (EPV) yaklaşık 32'dir (≈480 olay / 15 öngörücü) ve model geliştirme için önerilen alt sınırların oldukça üzerindedir (13).

2.7. İstatistiksel analiz

Her metrik için birincil belirsizlik ölçütü replikasyonlar arası dağılımdan türetildi: ortalama, %2,5–%97,5 replikasyon yüzdelikleriyle %95 aralık ve Monte Carlo standart hatası (MCSE = replikasyonlar arası SD / √nsim). Ayrım farkları için birincil kanıt, her strateji içinde LR ile GB arasındaki AUROC farkının replikasyon dağılımıdır (ortalama, %95 aralık, MCSE). DeLong testi (14) yalnızca tamamlayıcı olarak — replikasyonlar arası ortalama Z ve medyan p ile — raporlandı; dört stratejili karşılaştırmalarda çoklu test için Bonferroni (α=0,0125) da bildirildi. Tek-veri belirsizliği tamamlayıcı olarak 1000 tekrarlı bootstrap ile de incelendi; ancak birleştirilmiş katman-dışı tahminlerde bootstrap'ın katman yapısındaki bağımlılığı yok saydığı ve belirsizliği bir miktar düşük tahmin edebileceği not edilmelidir; bu nedenle bootstrap birincil ölçüt olarak kullanılmamıştır.

Yeniden kalibrasyon Her model×strateji için, katman-dışı tahminler üzerinde tek adımlı kesişim (in-the-large) yeniden kalibrasyonu [y ~ 1 + offset(logit(p̂))] uygulandı; düzeltilmiş olasılıklarla kalibrasyon kesişimi, Brier, ICI ve net fayda yeniden hesaplandı. Katman-dışı tahminler model eğitiminden dışlandığından bu düzeltme model eğitimine göre sızıntısızdır; yeniden kalibrasyon kestirimi aynı katman-dışı tahminler üzerinde yapıldığından tek-parametreli bir küresel illüstrasyon olarak nitelendirilmelidir. Analizlerde glmnet, xgboost, pROC, recipes/themis ve ggplot2 paketleri kullanıldı.

3. Bulgular

3.1. Örneklem

Üretilen 200 veri kümesinde ortalama olay oranı %12,0 olarak gerçekleşti (hedef %12) ve amaçlanan dengesizlik düzeyini yansıttı. Her replikasyonda tüm modeller ve stratejiler aynı veri ve aynı çapraz doğrulama bölünmeleri üzerinde değerlendirildi.

3.2. Ayrım performansı

En yüksek ayrımı dengelemesiz regülarize LR sağladı: AUROC 0,706 (0,680–0,732) ve AUPRC 0,273 (0,236–0,312). AUPRC değerleri, prevalansa eşit "beceri-yok" tabanı (≈0,12) dikkate alınarak yorumlanmalıdır. Dengeleme stratejileri LR'nin ayrımını iyileştirmedi (AUROC tüm stratejilerde 0,700–0,706; AUPRC 0,265–0,273); LR'de dengeleme yok ile dengeleme stratejileri arasındaki AUROC farkları anlamlı değildi. GB için dengeleme ayrımı ya değiştirmedi ya da kötüleştirdi: dengelemesiz GB AUROC 0,681 (0,647–0,709) iken SMOTE altında 0,632 (0,598–0,666)'ye, AUPRC ise 0,242'den 0,197'ye geriledi. Bu sonuçlar H1 hipotezini (dengeleme ayrımı iyileştirmez) desteklemektedir (Tablo 1). Ayrım sonuçlarının yorumlanmasında, sonucun lojistik-doğrusal bir öngörücüden üretildiği ve dolayısıyla doğru-belirtilmiş LR'nin yapısal bir avantaj taşıdığı göz önünde bulundurulmalıdır; "dengelemesiz LR en yüksek ayrımı verdi" bulgusu bu tasarım tercihiyle ilişkilidir (bkz. Sınırlılıklar).

3.3. Kalibrasyon performansı

Kalibrasyon, stratejiler arasında en çarpıcı farkın ortaya çıktığı eksendir. Dengelemesiz modeller iyi kalibre idi: LR için kalibrasyon kesişimi −0,000 (−0,004–0,004), eğim 1,024 (0,977–1,078) ve ICI 0,007 (0,003–0,012); GB için kesişim 0,008, eğim 1,032 ve ICI 0,008. Buna karşın dengeleme stratejileri kalibrasyonu ciddi biçimde bozdu: LR'de kalibrasyon kesişimi yaklaşık −2,0'ye düştü (LR-SMOTE −2,011; LR-ağırlık −1,969; LR-azaltma −1,985), bu da modellerin riski sistematik olarak

aşırı tahmin ettiğini gösterir. Biçim düzeyinde de bozulma belirgindi: dengeleme stratejilerinde ICI 0,320–0,334, E90 ise 0,44–0,47 düzeyine yükseldi (dengelemesiz ICI ≈ 0,007'ye karşı). Brier skoru dengelemesiz 0,098–0,100'den dengelemeli stratejilerde 0,210–0,228'e yükseldi. GB-SMOTE kombinasyonu özel bir durum oluşturdu: kalibrasyon kesişimi görece ılımlı (−0,122) kalırken kalibrasyon eğimi 0,317'ye (0,242–0,402) düştü ve ayrım da geriledi (AUROC 0,632); yani GB-SMOTE'ta bozulma yalnız olasılık düzeyinde değil, sıralamada da ortaya çıkmıştır (Şekil 1, Şekil 2). Bu bulgular H2 hipotezini doğrulamaktadır.

3.4. Eşik-bağımlı ölçütler

Olay prevalansı eşiğinde (0,12) değerlendirildiğinde, dengelemesiz LR en yüksek F1 (0,310) ve dengeli doğruluk (0,649) değerlerini verdi. Dengeleme stratejileri, olasılıkları yukarı kaydırdığı için bu eşikte sınıflandırmayı iyileştirmedi, hatta F1 ve dengeli doğruluğu düşürdü (ör. LR-SMOTE F1 0,218, dengeli doğruluk 0,510). Dengelemeli GB kombinasyonlarında olasılıklar eşiğin bu denli üzerine kaydı ki tüm gözlemler pozitif sınıfa atandı; dengeli doğruluk 0,500'e indi ve MCC, karışıklık matrisinin bir sütunu boşaldığından (payda çarpanlarından biri sıfırlandığından) tanımsız hâle geldi. Bu sonuç, eşik-bağımlı ölçütlerin seçilen işletim noktasına güçlü biçimde bağımlı olduğunu ve dengelemenin "görünür" yararının büyük ölçüde eşik seçimi yapaylığından kaynaklanabileceğini göstermektedir.

3.5. Klinik net fayda (karar eğrisi analizi)

Karar eğrisi analizinde dengelemesiz LR tüm klinik-akla-yatkın eşik aralığında (≈0,05–0,20) en yüksek net faydayı sağladı (ör. 0,10 eşiğinde net fayda 0,0445; 0,20 eşiğinde 0,0160); bu aralıkta hiçbir dengeleme stratejisi LR'yi geçemedi. Dengelemesiz GB ikinci sırada, benzer ancak biraz düşük fayda gösterdi (0,10 eşiğinde 0,0409; 0,20 eşiğinde 0,0118). SMOTE uygulanmış GB en düşük net faydayı verdi ve yaklaşık 0,20 eşiğinden sonra "kimseyi tedavi etme" doğrusunun altına inerek negatife döndü (0,20 eşiğinde −0,0003); yani kalibre edilmediğinde bu eşiklerde klinik olarak zararlı hâle geldi (Şekil 4). Dengelemeli LR ise olasılık kaymasının doğrudan sonucu olarak daha erken, yaklaşık 0,12 eşiğinden itibaren "herkesi/kimseyi tedavi etme" doğrularının altına indi. Net fayda sıralaması, ayrım sıralamasıyla uyumlu fakat kalibrasyon bozulmasının klinik maliyetini açıkça ortaya koyacak biçimde belirgindir.

3.6. Dengeleme sonrası yeniden kalibrasyonun düzeltici etkisi

Gözlenen zararın ne ölçüde geri-döndürülebilir olduğunu değerlendirmek için, katman-dışı tahminlere tek adımlı kesişim yeniden kalibrasyonu uygulandı (Tablo 3; Şekil 5). Prevalans kaynaklı kesişim kayması tam olarak geri döndü: LR-SMOTE'ta kalibrasyon kesişimi −2,011'den ≈0,000'e, Brier skoru 0,215'ten 0,099'a ve ICI 0,320'den 0,012'ye geriledi; aynı düzelme LR-ağırlık, LR-azaltma ve GB-ağırlık/azaltma stratejilerinde de gözlendi (Brier ≈0,21–0,23'ten ≈0,10'a). Net fayda açısından da LR-SMOTE, yeniden kalibrasyonla dengelemesiz LR düzeyine geri döndü (0,20 eşiğinde −0,0788'den +0,0154'e). Buna karşın GB-SMOTE yalnızca kısmen düzeldi (Brier 0,114→0,112; ICI 0,070→0,066): bu kombinasyonun asıl hasarı kesişimde değil, kalibrasyon eğiminde ve sıralamada (eğim 0,317; AUROC 0,632) olduğundan, tek adımlı kesişim yeniden kalibrasyonu bu bozulmayı gideremez. Bu ikili bulgu, klinik reçeteyi keskinleştirir: prevalans kaynaklı zarar yeniden kalibrasyonla giderilebilir; ancak SMOTE'un GB'de yol açtığı sıralama bozulması yeniden kalibrasyonla da düzeltilemez.

3.7. Ayrım farkının replikasyon dağılımı (LR vs GB)

Her strateji içinde regülarize LR, GB'ye kıyasla daha yüksek AUROC verdi. Birincil kanıt olarak AUROC farkının (LR−GB) replikasyonlar arası dağılımı sunulur (Tablo 2): dengeleme yok için fark 0,025 (%95 aralık 0,009–0,046; MCSE 0,001), SMOTE için 0,072 (0,048–0,094; MCSE 0,001), sınıf ağırlıklandırma için 0,030 (0,014–0,048) ve azaltarak örnekleme için 0,033 (0,014–0,053). LR'nin üstünlüğü en belirgin biçimde SMOTE altında ortaya çıktı. Tamamlayıcı DeLong testi bu yönü desteklemiştir (replikasyonlar arası ortalama Z sırasıyla 3,70; 6,16; 4,05; 4,21; medyan p tüm stratejilerde <0,001); farkın α=0,05 eşiğinde anlamlı bulunduğu replikasyon oranı dengeleme yok için 187/200, SMOTE için 200/200, ağırlıklandırma için 195/200 ve azaltma için 194/200 olmuştur.

3.8. Duyarlılık analizi

Bulguların sağlamlığı, üç farklı olay prevalansı (%5, %12, %20) ve üç örneklem büyüklüğü (n=2000, 4000, 8000) altında, her hücrede 60 replikasyonla sınandı (Tablo 4, Şekil 6). "Dengeleme paradoksu" tüm senaryolarda tutarlı biçimde gözlendi: dengelemesiz modeller her koşulda iyi kalibre kaldı (kalibrasyon kesişimi ≈ 0; MCSE ≤ 0,002), buna karşın tüm dengeleme stratejileri kalibrasyon kesişimini belirgin biçimde negatife (aşırı tahmin

yönüne) kaydırdı. Önemli ve sezgisel bir örüntü olarak, dengesizlik arttıkça (prevalans düştükçe) kalibrasyon bozulması şiddetlendi: dengeleme stratejilerinde kalibrasyon kesişimi %5 prevalansta yaklaşık −2,9'a, %12'de −2,0'ye ve %20'de −1,4'e ulaştı. Bu değerler, tam dengelemenin (1:1) beklenen teorik kaymasıyla — logit(prevalans), yani gerçekleşen prevalanslar için sırasıyla ≈ −2,94; −2,00 ve −1,39 — MCSE bantları içinde yakın uyum göstermektedir (kesişim MCSE'leri ≤ 0,013; Şekil 6'da hata çubukları ±1,96·MCSE ve "×" ile gösterilen teorik değerler). Ayrım açısından dengeleme hiçbir senaryoda iyileşme sağlamadı; örneklem büyüklüğü arttıkça tüm modellerin ayrımı beklenildiği gibi hafifçe yükseldi (ör. LR dengeleme yok, %12: n=2000'de 0,699, n=8000'de 0,710; AUROC MCSE ≤ 0,003) ancak stratejiler arası sıralama değişmedi. GB-SMOTE'un kalibrasyon kesişimi (ör. %12'de −0,128) diğer dengeleme stratejilerine göre ılımlı kalsa da AUROC'u en düşük değere geriledi; bu, GB-SMOTE'ta bozulmanın kesişimden çok sıralamada ortaya çıktığı yönündeki ana bulguyla tutarlıdır. Bu sonuçlar, ana bulguların tek bir prevalans veya örneklem büyüklüğüne özgü olmadığını göstermektedir.

Tablo 1. Katman-dışı (out-of-fold) performans metrikleri; 200 replikasyonlu Monte Carlo tasarımından replikasyon-temelli ortalama (%2,5–%97,5 replikasyon aralığı). MCSE değerleri metinde ve ayrı özet dosyasında verilmiştir. ICI/E90: esnek (loess) kalibrasyon eğrisinden. Dengelemeli GB stratejilerinde tüm gözlemler pozitife atandığından MCC tanımsızdır.

Model	Strateji	AUROC	AUPRC	Brier	Kalib. eğim	Kalib. kesişim	ICI	E90	MCC	F1	Dengeli doğr.
LR	none	0,706 (0,680 – 0,732)	0,273 (0,236 – 0,312)	0,098 (0,091 – 0,104)	1,024 (0,977–1,078)	−0,000 (−0,004–0,004)	0,007 (0,003 – 0,012)	0,014 (0,006 – 0,027)	0,200	0,310	0,649
LR	smote	0,704 (0,676 – 0,730)	0,271 (0,234 – 0,308)	0,215 (0,206 – 0,222)	0,854 (0,799–0,909)	−2,011 (−2,107–−1,922)	0,320 (0,304 – 0,333)	0,468 (0,449 – 0,487)	0,042	0,218	0,510
LR	weight	0,705 (0,679 – 0,732)	0,273 (0,235 – 0,311)	0,214 (0,206 – 0,222)	1,039 (0,964–1,109)	−1,969 (−2,060–−1,887)	0,328 (0,313 – 0,341)	0,437 (0,420 – 0,456)	0,023	0,216	0,503
LR	down	0,700 (0,669 – 0,729)	0,265 (0,227 – 0,302)	0,218 (0,209 – 0,226)	1,032 (0,920–1,152)	−1,985 (−2,076–−1,893)	0,334 (0,318 – 0,347)	0,440 (0,416 – 0,463)	0,021	0,216	0,502
XGB	none	0,681 (0,647 – 0,709)	0,242 (0,205 – 0,282)	0,100 (0,093 – 0,107)	1,032 (0,873–1,231)	0,008 (−0,006–0,023)	0,008 (0,003 – 0,017)	0,016 (0,006 – 0,031)	0,178	0,295	0,631

Model	Strateji	AUROC	AUPRC	Brier	Kalib. eğim	Kalib. kesişim	ICI	E90	MCC	F1	Dengeli doğr.
XGB	smote	0,632 (0,598–0,666)	0,197 (0,169–0,240)	0,114 (0,105–0,121)	0,317 (0,242–0,402)	−0,122 (−0,200–−0,051)	0,070 (0,058–0,079)	0,153 (0,118–0,184)	tanımsız	0,263	0,593
XGB	weight	0,675 (0,640–0,708)	0,236 (0,193–0,278)	0,210 (0,197–0,221)	1,095 (0,902–1,312)	−1,871 (−1,962–−1,789)	0,321 (0,296–0,340)	0,406 (0,384–0,428)	tanımsız	0,216	0,502
XGB	down	0,666 (0,630–0,700)	0,222 (0,186–0,260)	0,228 (0,219–0,236)	0,948 (0,693–1,240)	−1,996 (−2,094–−1,906)	0,345 (0,330–0,361)	0,446 (0,403–0,488)	tanımsız	0,215	0,502

Tablo 2. Ayrım farkının replikasyon dağılımı: her strateji içinde regülarize LR ile GB arasındaki AUROC farkı (ortalama, %95 replikasyon aralığı, MCSE) ve tamamlayıcı DeLong özeti (replikasyonlar arası ortalama Z, medyan p, α=0,05'te anlamlı replikasyon oranı).

Strateji	AUROC (LR)	AUROC (GB)	AUROC farkı (%95 aralık)	MCSE	Ort. Z	Medyan p	Anlamlı (α=0,05)
none	0,706	0,681	0,025 (0,009–0,046)	0,001	3,70	$2{,}5\times10^{-4}$	187/200
smote	0,704	0,632	0,072 (0,048–0,094)	0,001	6,16	$7{,}0\times10^{-10}$	200/200
weight	0,705	0,675	0,030 (0,014–0,048)	0,001	4,05	$4{,}6\times10^{-5}$	195/200
down	0,700	0,666	0,033 (0,014–0,053)	0,001	4,21	$2{,}0\times10^{-5}$	194/200

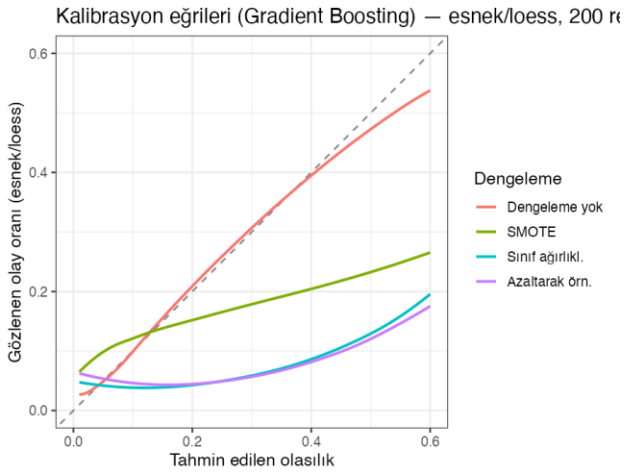
Tablo 3. (Yeniden kalibrasyon). Tek adımlı kesişim (in-the-large) yeniden kalibrasyonunun etkisi: ham ve yeniden kalibre kalibrasyon kesişimi, Brier ve ICI (200 replikasyon ortalaması). Prevalans kaynaklı zarar tam geri döner; GB-SMOTE'un sıralama hasarı kalır.

Model	Strateji	Kesişim (ham→recal)	Brier (ham→recal)	ICI (ham→recal)
LR	none	−0,000 → 0,000	0,098 → 0,098	0,007 → 0,007
LR	smote	−2,011 → 0,000	0,215 → 0,099	0,320 → 0,012
LR	weight	−1,969 → −0,000	0,214 → 0,098	0,328 → 0,007
LR	down	−1,985 → −0,000	0,218 → 0,099	0,334 → 0,008
XGB	none	0,008 → −0,000	0,100 → 0,100	0,008 → 0,008
XGB	smote	−0,122 → −0,000	0,114 → 0,112	0,070 → 0,066

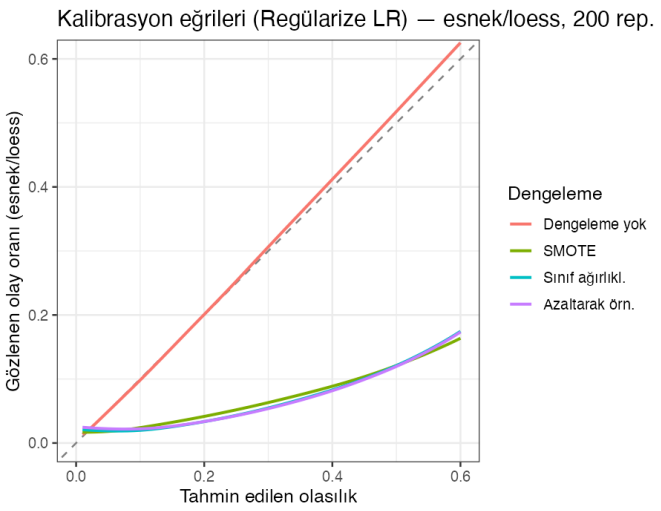
Model	Strateji	Kesişim (ham→recal)	Brier (ham→recal)	ICI (ham→recal)
XGB	weight	−1,871 → 0,000	0,210 → 0,101	0,321 → 0,010
XGB	down	−1,996 → −0,000	0,228 → 0,102	0,345 → 0,010

Tablo 4. (Duyarlılık). Farklı prevalans ve örneklem büyüklüklerinde (hücre başına 60 replikasyon) kalibrasyon kesişimi / AUROC; her hücre "kesişim / AUROC" biçiminde. Dengelemesiz (none) modeller tüm koşullarda iyi kalibre; dengeleme stratejilerinde kesişim prevalans düştükçe daha çok negatife kayar. Kesişim MCSE'leri ≤ 0,013, AUROC MCSE'leri ≤ 0,004 (bantlar Şekil 6'da). Teorik kesişim kayması = logit(gerçek prevalans): %5 için ≈ −2,94; %12 için ≈ −2,00; %20 için ≈ −1,39.

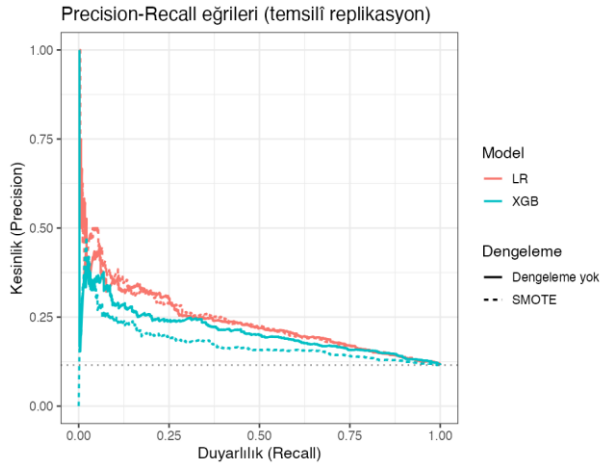
Senaryo	**Model**	**none**	**smote**	**weight**	**down**
n=4000, prev≈%5	LR	0,000 / 0,705	−2,991 / 0,701	−2,892 / 0,703	−2,933 / 0,691
n=4000, prev≈%5	GB	0,019 / 0,666	0,015 / 0,622	−2,718 / 0,655	−2,987 / 0,644
n=4000, prev≈%12	LR	0,000 / 0,709	−2,028 / 0,707	−1,983 / 0,708	−2,001 / 0,702
n=4000, prev≈%12	GB	0,008 / 0,683	−0,128 / 0,635	−1,880 / 0,680	−2,007 / 0,671
n=4000, prev≈%20	LR	0,000 / 0,705	−1,396 / 0,704	−1,373 / 0,705	−1,379 / 0,702
n=4000, prev≈%20	GB	0,004 / 0,687	−0,215 / 0,652	−1,308 / 0,685	−1,382 / 0,677
n=2000, prev≈%12	LR	0,000 / 0,699	−2,028 / 0,696	−1,972 / 0,698	−2,003 / 0,688
n=2000, prev≈%12	GB	0,021 / 0,665	0,025 / 0,626	−1,812 / 0,655	−2,039 / 0,646
n=8000, prev≈%12	LR	0,000 / 0,710	−2,013 / 0,709	−1,980 / 0,710	−1,989 / 0,708
n=8000, prev≈%12	GB	0,003 / 0,695	−0,262 / 0,645	−1,904 / 0,693	−1,989 / 0,687



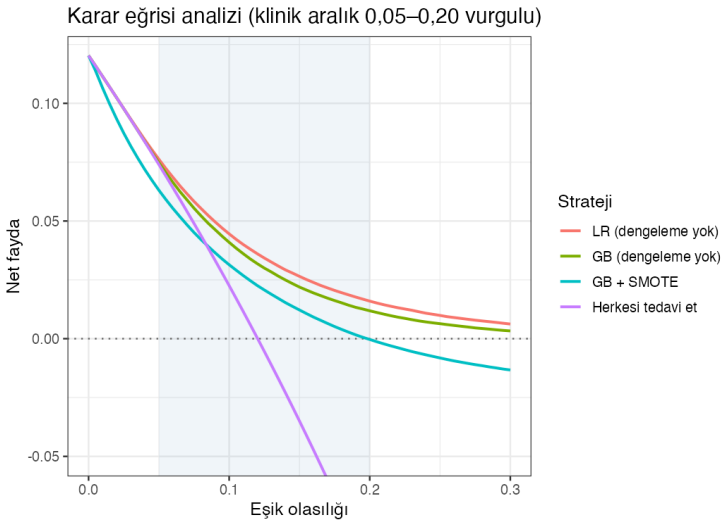
Şekil 1. *Gradient boosting için dört dengeleme stratejisi altında esnek (loess) kalibrasyon eğrileri (200 replikasyon ortalaması). Dengelemesiz model köşegene yakınken, dengeleme stratejileri köşegenin belirgin biçimde altında kalır (aşırı tahmin).*



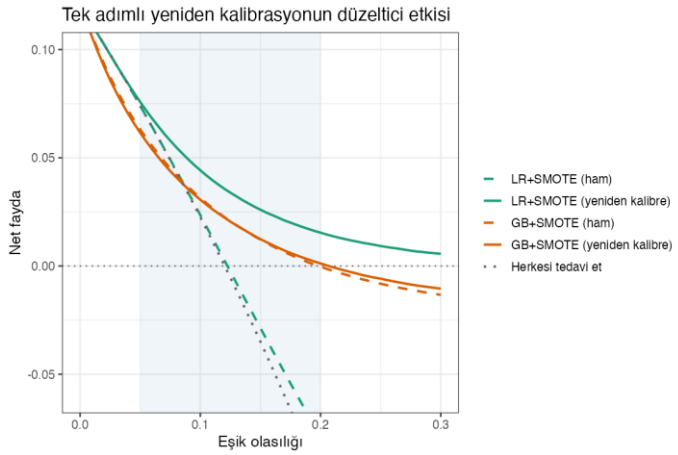
Şekil 2. *Regülarize LR için esnek (loess) kalibrasyon eğrileri (200 replikasyon ortalaması).*



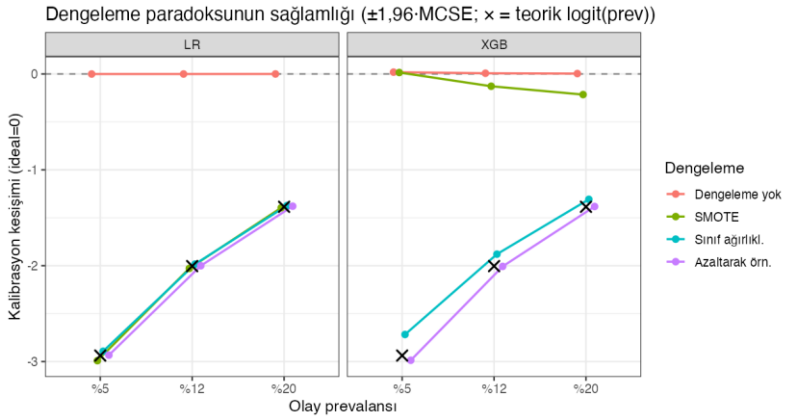
Şekil 3. *Dengelemesiz ve SMOTE koşullarında iki model için precision-recall eğrileri (temsilî replikasyon).*



Şekil 4. *Karar eğrisi analizi (net fayda), klinik-akla-yatkın aralık (0,05–0,20) vurgulu. Dengelemesiz LR tüm eşiklerde en yüksek net faydayı verir; SMOTE uygulanmış GB yaklaşık 0,20 eşiğinden sonra "kimseyi tedavi etme" doğrusunun altına inerek kalibre edilmediğinde klinik olarak zararlı hâle gelir.*



Şekil 5. *Tek adımlı kesişim yeniden kalibrasyonunun net faydaya etkisi. LR+SMOTE'ta net fayda dengelemesiz düzeye tam olarak geri döner; GB+SMOTE'ta ise sıralama hasarı nedeniyle yalnızca kısmen düzelir.*



Şekil 6. *Duyarlılık analizi: kalibrasyon kesişiminin prevalansa göre değişimi (model başına panel; hata çubukları ±1,96·MCSE). "×" işaretleri, tam dengelemenin beklenen teorik kaymasını [logit(prevalans)] gösterir ve gözlenen değerlerle yakın uyum içindedir.*

4. Tartışma

Bu çalışmada, kontrollü bir ESK-benzeri simülasyon ortamında, dengesiz sınıf problemine yönelik dört dengeleme stratejisinin regülarize LR ve GB üzerindeki etkisi ayrım, kalibrasyon ve klinik net fayda eksenlerinde karşılaştırılmıştır. Üç temel bulgu öne çıkmaktadır: (i) dengeleme ayrım gücünü iyileştirmemiş, GB'de belirgin biçimde kötüleştirmiştir; (ii) tüm dengeleme stratejileri kalibrasyonu ciddi biçimde bozarak modelleri sistematik aşırı tahmine yöneltmiştir; (iii) dengeleme klinik net faydayı azaltmış, en uç durumda (GB-SMOTE) modeli kalibre edilmediğinde klinik olarak zararlı hâle getirmiştir. Bu bulgular birlikte "dengeleme paradoksu"nu somutlaştırmaktadır: dengeleme, dengesiz veride sezgisel olarak yararlı görünse de, karar verme amaçlı olasılık tahmininde fayda yerine zarar üretebilir.

Bulguların altında yatan mekanizma açıktır ve niceliksel olarak öngörülebilir. Yeniden örnekleme ve sınıf ağırlıklandırma, eğitim verisindeki olay oranını yapay olarak yükseltir (tam dengelemede 1:1'e); sonuç olarak model, gerçek prevalansa göre çok daha yüksek olasılıklar üretmeyi öğrenir. Bu kaymanın beklenen büyüklüğü kapalı biçimde yazılabilir: eğitim olay oranı $\pi_{eğitim}$'e çekildiğinde kalibrasyon kesişimindeki beklenen değişim yaklaşık $\text{logit}(\pi_{gerçek}) - \text{logit}(\pi_{eğitim})$ kadardır; tam dengeleme ($\pi_{eğitim}=0{,}5$) için bu $\text{logit}(\pi_{gerçek})$'e indirgenir. Ayrım ölçütleri (AUROC, AUPRC) olasılıkların yalnızca sıralamasına duyarlı olduğundan bu kaymadan büyük ölçüde etkilenmez; ancak kalibrasyon ve net fayda olasılıkların mutlak düzeyine bağlı olduğundan ciddi biçimde bozulur. Nitekim tek

adımlı kesişim yeniden kalibrasyonunun bu zararı büyük ölçüde geri döndürmesi (Bölüm 3.6), mekanizmanın prevalans kaymasına dayandığını doğrudan doğrular. GB-SMOTE örneğinde ise kalibrasyon kaymasının yanı sıra ayrımın da gerilemesi ve yeniden kalibrasyonla düzelmemesi, sentetik azınlık örneklerinin GB'nin karar sınırlarını aşırı uyuma sürükleyerek olasılık sıralamasını da bozduğunu düşündürmektedir.

Bu sonuçlar, ayrım ve kalibrasyonun ayrı ve tamamlayıcı boyutlar olduğunu vurgulayan klinik tahmin modeli yazınıyla (ör. TRIPOD ve kalibrasyon hiyerarşisi üzerine çalışmalar) (2,15,16) ve sınıf dengesizliği düzeltmesinin risk tahmin modellerinde kalibrasyonu bozduğunu bildiren güncel bulgularla (6) uyumludur. Çalışmanın özgün katkısı, bu olguyu tek bir çerçevede dengeleme x model matrisi, istatistiksel anlamlılık, karar eğrisi analizi ve yeniden kalibrasyon kolu ile birlikte ve Türkçe biyoistatistik yazınında somut biçimde göstermesidir.

Pratik açıdan üç çıkarım önerilebilir. Birincisi, dengesiz klinik veride dengeleme bir "varsayılan" işlem olarak uygulanmamalı; gerekliliği, hedefin sıralama mı (ör. tarama önceliklendirmesi) yoksa kalibre olasılık mı (ör. risk iletişimi, eşik-temelli karar) olduğuna göre değerlendirilmelidir. İkincisi, dengeleme yine de kullanılacaksa, sonrasında bir yeniden kalibrasyon adımı (ör. Platt ölçekleme, isotonik regresyon veya kesişim yeniden tahmini) zorunlu görülmelidir; ancak bu çalışmanın gösterdiği gibi, yeniden kalibrasyon yalnızca olasılık düzeyi (kesişim) bozulmasını onarır, SMOTE'un GB'de yol açtığı sıralama hasarını gidermez. Üçüncüsü, regülarize LR, dengeleme olmadan dahi hem iyi kalibrasyon hem yüksek net fayda sağladığından, ESK temelli dengesiz problemlerde güçlü ve savunulabilir bir başlangıç (referans) modeli olarak konumlandırılmalıdır.

Eşik-bağımlı ölçütlere ilişkin bulgu da metodolojik olarak öğreticidir: aynı modelin F1 ve dengeli doğruluk değerleri, eşik %50'den olay prevalansına (%12) kaydırıldığında dengeleme lehine bulgudan aleyhine bulguya dönmektedir. Bu durum, eşik-bağımlı ölçütlerin tek başına model karşılaştırmasında kırılgan olduğunu ve eşik-bağımsız (ayrım, kalibrasyon) ile karar-teorik (net fayda) ölçütlerle birlikte raporlanması gerektiğini göstermektedir. Raporlamada, klinik tahmin modelleri için TRIPOD kılavuzunun (ve makine öğrenmesi bileşeni olan çalışmalarda TRIPOD+AI uzantısının) (15,16) izlenmesi ve ilgili kontrol listesinin ek olarak sağlanması önerilir.

Kapsam. Buradaki modeller konuşlandırılabilir tahmin araçları değil, bir mekanizmayı (dengeleme paradoksunu) kontrollü koşullarda gösteren metodolojik illüstrasyonlardır; bildirilen mutlak performans değerleri bu kapsamda yorumlanmalı, gerçek veride dış doğrulama ile desteklenmelidir.

Sınırlılıklar

Birincisi, çalışma sentetik veriye dayanır; üretim mekanizması gerçekçi seçilmiş olsa da gerçek ESK verisinin tüm karmaşıklığını (ölçüm hatası, yapısal eksiklik, zamansal kayma) yansıtmaz. İkincisi, sonucun lojistik-doğrusal yapıdan üretilmesi, doğru-belirtilmiş LR'ye yapısal bir avantaj sağlar; bu nedenle LR'nin ayrımdaki üstünlüğü, gerçek (yanlış-belirtilmiş) veride daha küçük olabilir — bulgular mutlak üstünlükten çok göreli örüntü (dengelemenin kalibrasyonu bozması) açısından yorumlanmalıdır. Üçüncüsü, gözlenen SMOTE zararı, sürekli özellik uzayındaki k-en yakın komşu enterpolasyonuna özgüdür; karma-tipli (kategorik/ikili) değişkenler içeren gerçek EHR verisine ve SMOTE-NC gibi varyantlara doğrudan genellenemez. Dördüncüsü, GB için hiperparametre uzayı hesaplama maliyeti gözetilerek sınırlı tutulmuştur (derinlik {2,4}); daha geniş bir arama GB'nin mutlak başarımını bir miktar değiştirebilir, ancak kalibrasyon kaymasının yönü prevalans mekanizması gereği korunması beklenir. Beşincisi, yeniden kalibrasyon kolu, katman-dışı tahminler üzerinde tek adımlı kesişim düzeltmesiyle küresel bir illüstrasyon olarak sunulmuştur; tam sızıntısız bir post-hoc kalibrasyon kolunun (eğitim katmanında kestirilip test katmanına uygulanan Platt/isotonik) niceliklenmesi ileri çalışmalara bırakılmıştır. Altıncısı, belirsizlik başlıca Monte Carlo replikasyonlarıyla nicelenmiş; tamamlayıcı bootstrap, birleştirilmiş katman-dışı tahminlerdeki katman bağımlılığını yok saydığından tek-veri belirsizliğini bir miktar düşük tahmin edebilir.

Gelecek çalışmalar

Bulguların gerçek ESK kohortlarında (ör. MIMIC-IV) dış ve zamansal doğrulaması; sızıntısız post-hoc yeniden kalibrasyon yöntemlerinin karar eğrisi üzerindeki düzeltici etkisinin tam olarak niceliklenmesi; aşırı dengesizlik (<%2) gibi uç koşulların ve SMOTE-NC benzeri karma-tipli varyantların incelenmesi; ve derin öğrenme modellerinin aynı çerçevede karşılaştırılması öncelikli araştırma yönleridir.

5. Sonuç

ESK temelli dengesiz sınıf probleminde, yaygın dengeleme stratejileri ayrım gücünü iyileştirmemiş; buna karşılık kalibrasyonu ve klinik net faydayı

belirgin biçimde bozmuştur. Karar verme amaçlı klinik tahmin modellerinde model başarımının yalnızca AUROC ile değerlendirilmesi yanıltıcıdır; kalibrasyon ve karar eğrisi analizi vazgeçilmez tamamlayıcılardır. Dengeleme kullanılacaksa mutlaka bir yeniden kalibrasyon adımı izlemelidir. Regülarize lojistik regresyon, dengeleme uygulanmadan, hem iyi kalibrasyon hem yüksek klinik fayda sağlayarak güçlü bir varsayılan model olmuştur.

Veri ve Kod Erişilebilirliği

Çalışmanın tüm verileri, bilinen bir üretim mekanizmasına sahip sentetik (simüle edilmiş) verilerdir; gerçek hasta verisi içermez. Simülasyon ve analiz boru hattının tamamı (veri üretimi, dengeleme, iç içe çapraz doğrulama, metrik/ICI/net fayda hesapları, yeniden kalibrasyon ve duyarlılık analizi), sabit tohumlu R betikleri olarak hazırlanmıştır ve tam olarak yeniden üretilebilir. Betikler, sessionInfo çıktısı ve üretilen tablo/şekiller bir genel depoda (GitHub/OSF) paylaşılacak; kabul aşamasında sürüm bir DOI (Zenodo) ile arşivlenerek kalıcı erişim sağlanacaktır.

KAYNAKLAR

1. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21(1):128-138.
2. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17(1):230.
3. He H, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng. 2009;21(9):1263-1284.
4. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. J Artif Intell Res. 2002;16:321-357.
5. Wynants L, Van Calster B, Collins GS, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ. 2020;369:m1328.
6. van den Goorbergh R, van Smeden M, Timmerman D, Van Calster B. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. J Am Med Inform Assoc. 2022;29(9):1525-1534.
7. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria; 2025.
8. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33(1):1-22.
9. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016:785-794.
10. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015;10(3):e0118432.
11. Austin PC, Steyerberg EW. The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. Stat Med. 2019;38(21):4051-4065.
12. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26(6):565-574.
13. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. BMJ. 2020;368:m441.
14. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837-845.
15. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med. 2015;162(1):55-63.
16. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378.