Supporting Information

Extending Deep Cox Survival Models with Case-Cohort Risk Set Sampling

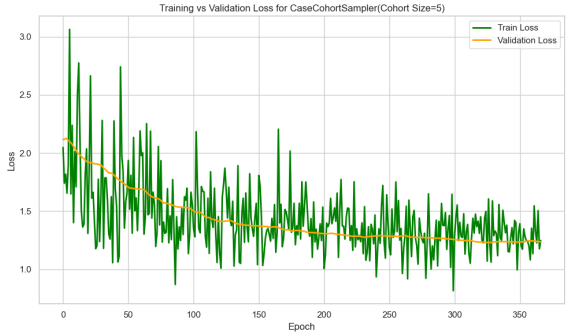
Michael P. Falk Justine B. Nasejje Marvin N. Wright Wende C. Safari
Pierre-Philippe Dechant Raeesa M. Docrat Rukia Nuermaimaiti

Mini-batch training on real-world data (METABRIC)

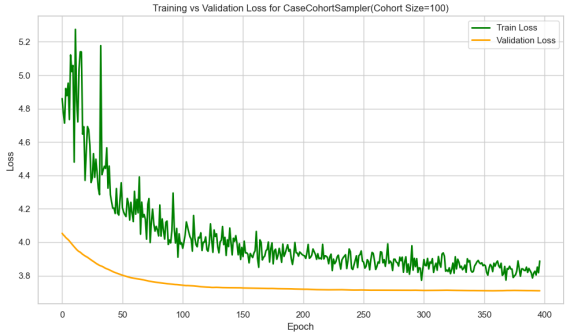
The convergence behavior and structural trade-offs observed on the simulated dataset largely carry over to the real-world METABRIC dataset, providing additional empirical support for the effectiveness of mini-batch risk-set sampling in this setting. As shown in Figure 1, the overall optimization dynamics mirror the synthetic experiments, though real-world feature heterogeneity introduces distinct nuances in training stability. Specifically, real-world data exhibits noticeably higher high-frequency variance (jitter) in the *training loss* curves across mini-batches due to empirical feature dimensionality and variance. Despite this training volatility, the *validation loss* remains remarkably smooth and well-behaved across all sampling strategies, steadily decreasing toward a stable loss floor without evidence of severe overfitting.

The relative performance across sampling schemes remains consistent with theoretical expectations. The *Full-Risk* and *Nested Case-Control* samplers both demonstrate exceptional robustness when applied to METABRIC data. In particular, the Nested Case-Control sampler ($m = 5$) maintains a clean, highly efficient convergence trajectory, confirming that fixed-control sampling provides sufficiently informative risk-set approximations even in high-dimensional empirical data.

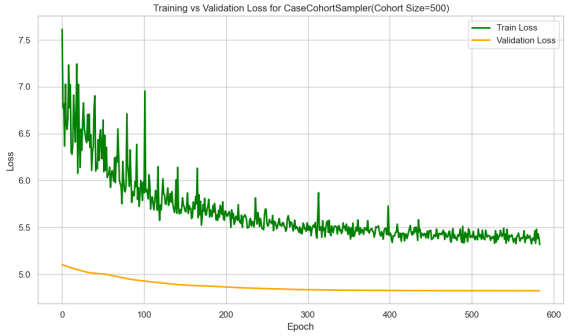
In contrast, the trade-off regarding subcohort size in the *Case-Cohort Sampler* is strikingly evident in the real-world setting. At a very small subcohort size of 5 (Figure 1a), batch-restricted risk sets suffer from severely impoverished sample diversity, inducing extreme gradient noise that causes the training loss to oscillate wildly between 0.8 and 3.1 (With a negative linear trend). However, progressively scaling the subcohort size to 100 (Figure 1b) and 500 (Figure 1c) enriches the risk-set representation per mini-batch. This significantly dampens training loss volatility, stabilizes parameter updates, and produces a far smoother optimization path.



(a) Real-world case-cohort loss curve (subcohort size = 5, batch size = 200).



(b) Real-world case-cohort loss curve (subcohort size = 100, batch size = 200).



(c) Real-world case-cohort loss curve (subcohort size = 500, batch size = 1000).

Figure 1: Mini-batch training dynamics on the real-world METABRIC dataset across varying case-cohort subcohort and batch size configurations.

Memory resource usage

We observed a substantial reduction in memory consumption when using sampling techniques compared to the full-risk-set approach during full-batch training. As shown in Figure 2 (panels 2a and 2b), the full-risk-set model consumes up to approximately 13 GB of RAM during training (reaching 13,000 MB by Fold 5). By contrast, sampling strategies with small sample sizes ($k = 5$) maintain significantly lower memory footprints: the nested case-control sampler (panels 2c and 2d) peaks at roughly 7.5 GB, while the case-cohort sampler (panels 2e and 2f) stays below 3.6 GB.

Impact of Subcohort and Control Size

Increasing the subcohort or control size from 5 to 100 demonstrates the scalability limits of these sampling techniques. For the nested case-control sampler with 100 controls, memory usage increases substantially, reaching a peak of approximately 12.1 GB in Fold 5—nearly approaching the consumption of the full-risk model. Similarly, expanding the case-cohort subcohort size to 100 increases its peak memory usage from 2.28 GB to roughly 7.65 GB. This indicates that while sampling reduces memory overhead, large control allocations diminish the memory benefits relative to full-batch full-risk training.

Notebook Memory Accumulation

We also observed that executing experiments sequentially within an interactive notebook environment (IPython/Jupyter) results in persistent memory allocations across k-folds. Later folds consistently begin from a higher baseline memory footprint, even after clearing explicit execution caches. To account for this baseline drift, per-fold consumption is evaluated by comparing the delta between the start and end memory levels of each fold, confirming that smaller sample configurations consistently require less incremental memory per fold.

Mini-Batch Memory and Execution Dynamics

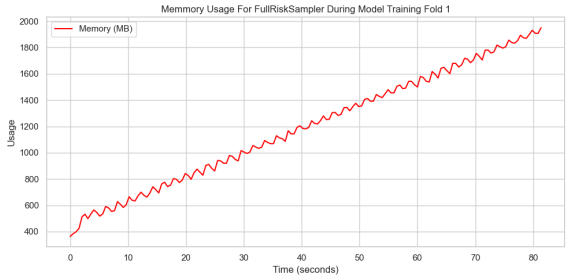
In contrast to full-batch optimization, mini-batch training results in a substantially lower and more stable memory footprint across the sampling methods and control sizes evaluated. As shown in Figure 3, memory consumption remains confined to a narrow range between 730 MB and 854 MB across all k-folds, regardless of whether a sampling method is applied.

Invariance to Subcohort and Control Size

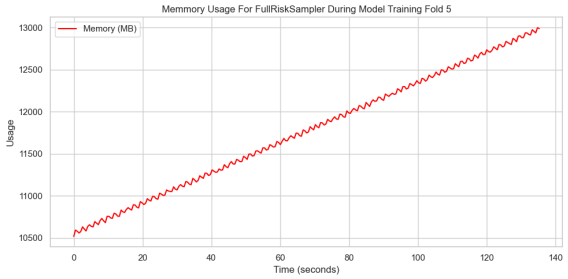
Under mini-batch execution, expanding the subcohort or control size (k) from 5 to 100 yields virtually no increase in peak memory consumption. For example, during Fold 5 training of the nested case-control sampler, peak memory usage remains virtually identical at approximately 838 MB for $k = 5$ and 836 MB for $k = 100$. A similar pattern holds for the case-cohort sampler, where peak memory stabilizes at approximately 845 MB for $k = 5$ and 843 MB for $k = 100$.

Trade-off Between Memory and Computational Runtime

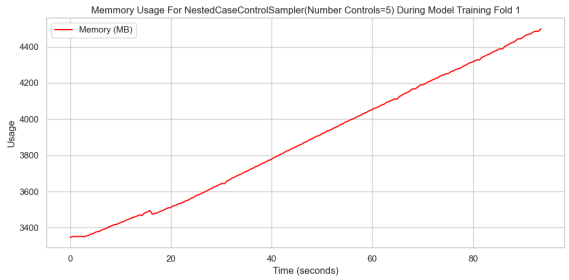
While mini-batching reduces differences in RAM requirements across sampling strategies, the computational cost of increasing the control size (k) manifests in wall-clock execution time rather than memory allocation. For instance, increasing k from 5 to 100 in the nested case-control sampler extends Fold 1 training time from 21 seconds to 48 seconds. Notably, the mini-batch full-risk model exhibits the lowest absolute RAM utilization ($\sim$577 MB in Fold 5), as mini-batching avoids the high-dimensional risk-set matrix calculations required during full-batch optimization.



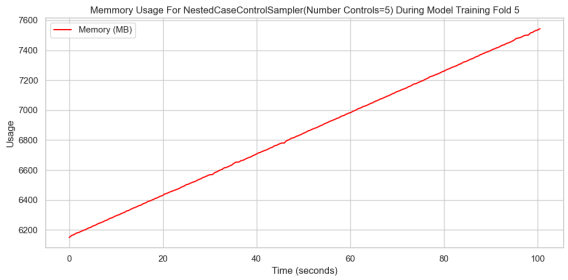
(a) Full-risk-set memory trace, fold 1.



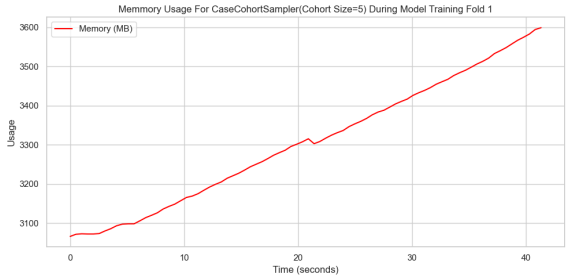
(b) Full-risk-set memory trace, fold 5.



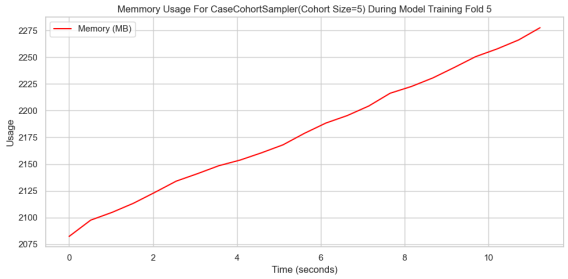
(c) Nested case-control memory trace (controls = 5), fold 1.



(d) Nested case-control memory trace (controls = 5), fold 5.

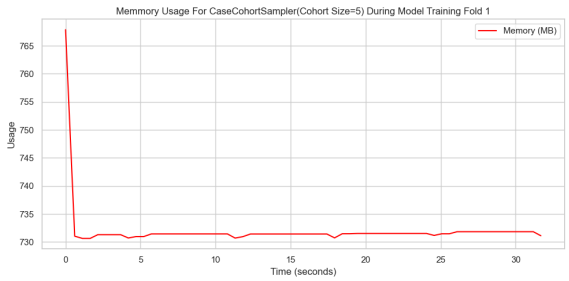


(e) Case-cohort memory trace (subcohort size = 5), fold 1.

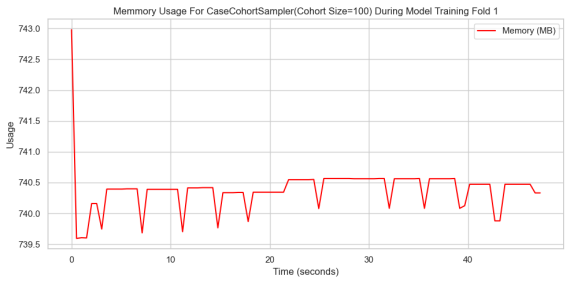


(f) Case-cohort memory trace (subcohort size = 5), fold 5.

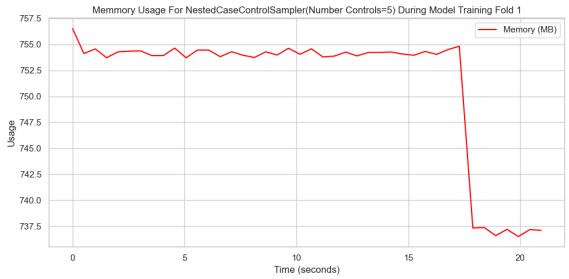
Figure 2: Memory usage traces across initial and final k-folds for full-risk and standard sampler configurations ($k = 5$).



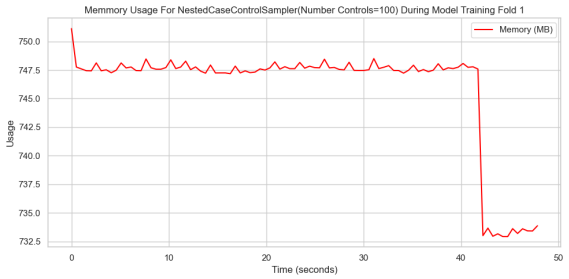
(a) Case-cohort ($k = 5$), fold 1.



(b) Case-cohort ($k = 100$), fold 1.



(c) Nested case-control ($k = 5$), fold 1.



(d) Nested case-control ($k = 100$), fold 1.

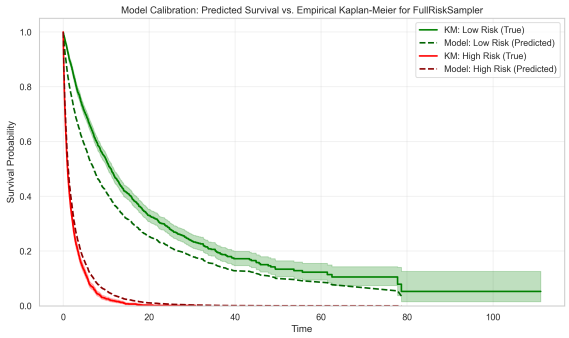
Figure 3: Representative mini-batch memory consumption traces illustrating bounded memory footprint across $k = 5$ and $k = 100$ control sizes.

Synthetic Survival Curves

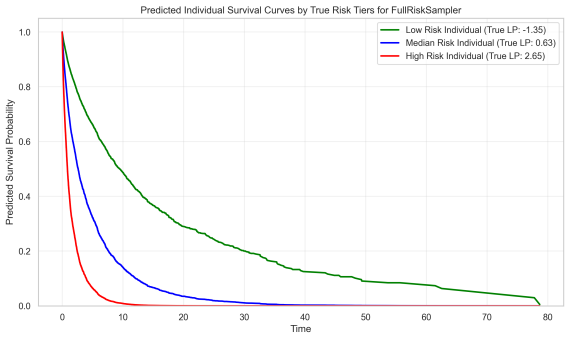
To understand the practical impact of down-sampling on clinical interpretability, we visualized the predicted survival curves generated by the models on the synthetic dataset. We examine therefore the full-batch execution regime, specifically comparing the full-risk baseline against the case-cohort sampler at extreme subcohort sizes ($k = 5$ and $k = 1000$).

Full-Batch Execution: Full-Risk Baseline

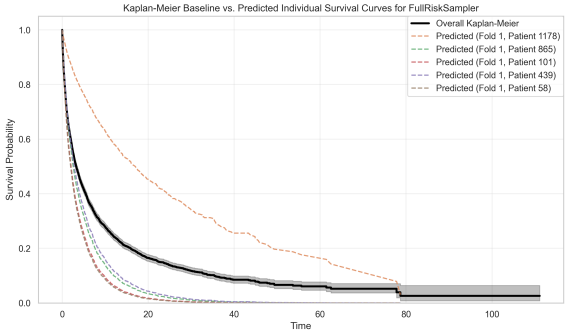
The full-risk sampler serves as the benchmark for survival prediction in this experiment. As shown in Figure 4a, the model shows close visual agreement; the predicted survival probabilities for both high- and low-risk cohorts tightly track the empirical Kaplan-Meier (KM) true risk curves. Furthermore, the model effectively stratifies individuals by risk tier (Figure 4b), generating proportional and visually distinct survival trajectories. When comparing individual predictions against the overall population baseline (Figure 4c), the predicted curves show a realistic, well-distributed spread across the entire timeline.



(a) Model calibration comparing empirical Kaplan-Meier vs. predicted survival (Baseline, Full-Batch).



(b) Predicted individual survival curves stratified by true risk tiers (Baseline, Full-Batch).

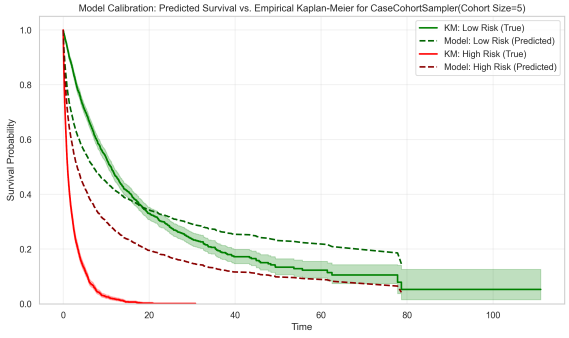


(c) Overall Kaplan-Meier baseline vs. predicted individual survival curves (Baseline, Full-Batch).

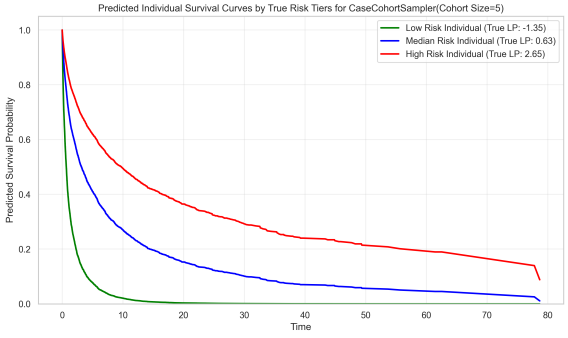
Figure 4: Full-risk baseline evaluations under the full-batch regime.

Full-Batch Execution: Case-Cohort ($k = 5$)

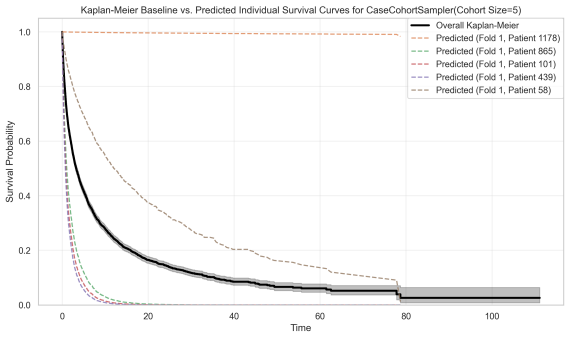
When the case-cohort subcohort size is heavily restricted to $k = 5$, the resulting survival predictions degrade significantly. Figure 5a reveals that the predicted low-risk survival curve deviates substantially from the empirical Kaplan–Meier curve. While the model still forces separation between risk tiers (Figure 5b), the curves are highly compressed toward the origin, with the low-risk prediction incorrectly plummeting to near-zero probability by $t = 20$. This distortion is echoed in the individual plots (Figure 5c), where the model shows substantially reduced heterogeneity in the predicted survival curves, artificially dragging all predictions downwards.



(a) Model calibration comparing empirical Kaplan-Meier vs. predicted survival (Case-Cohort Size=5, Full-Batch).



(b) Predicted individual survival curves stratified by true risk tiers (Case-Cohort Size=5, Full-Batch).

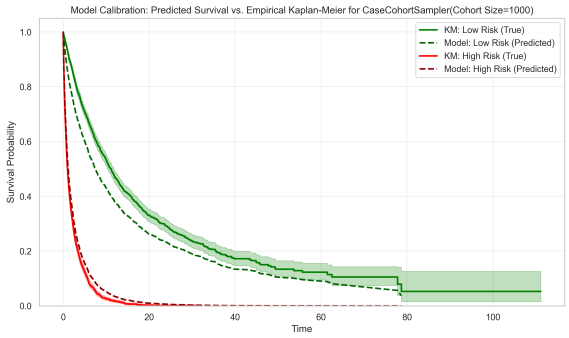


(c) Overall Kaplan-Meier baseline vs. predicted individual survival curves (Case-Cohort Size=5, Full-Batch).

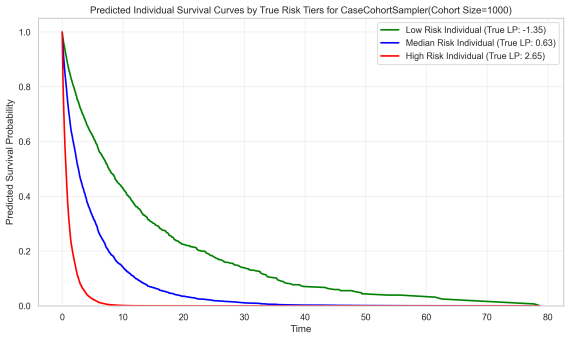
Figure 5: Case-cohort evaluations with subcohort size $k = 5$ under the full-batch regime.

Full-Batch Execution: Case-Cohort ($k = 1000$)

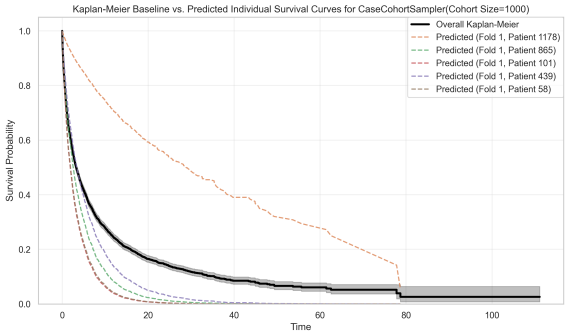
As the subcohort size approaches the scale of the full dataset, the predictive capacity of the case-cohort sampler recovers. At $k = 1000$, the model regains accurate calibration (Figure 6a), correctly matching the true underlying KM curves. The individual risk tiers are once again distinctly separated and realistically scaled (Figure 6b), and the individual survival trajectories properly encapsulate the variance of the true overall baseline (Figure 6c), effectively mirroring the performance of the full-risk sampler.



(a) Model calibration comparing empirical Kaplan-Meier vs. predicted survival (Case-Cohort Size=1000, Full-Batch).



(b) Predicted individual survival curves stratified by true risk tiers (Case-Cohort Size=1000, Full-Batch).



(c) Overall Kaplan-Meier baseline vs. predicted individual survival curves (Case-Cohort Size=1000, Full-Batch).

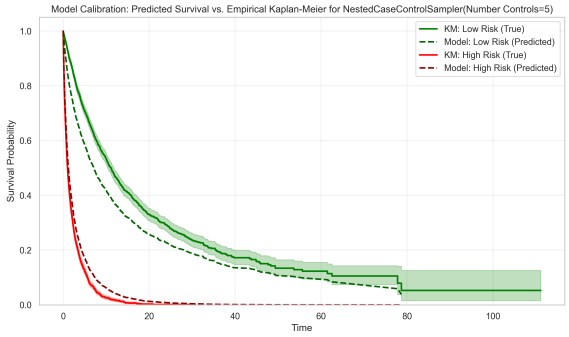
Figure 6: Survival analysis results for Case-Cohort Size=1000 (Full-Batch).

Full-Batch Execution: Nested Case-Control

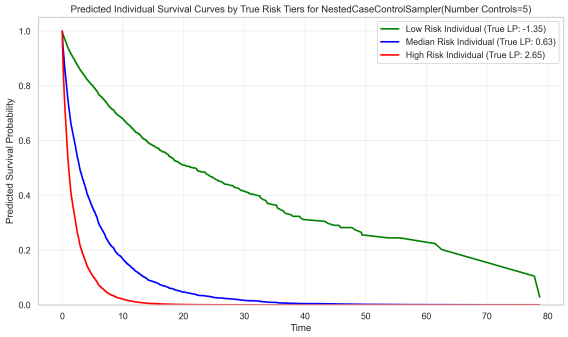
In stark contrast to the case-cohort method, the nested case-control sampler maintains remarkable structural stability and predictive accuracy even at the most extreme down-sampling levels. Figure 7a demonstrates that with a minimal configuration of just 5 controls per event, the model remains highly calibrated. The predicted high- and low-risk survival probabilities track the true empirical Kaplan-Meier curves closely, without the severe underestimation seen in the equivalent case-cohort model.

The model successfully preserves proportional and realistic separation between distinct risk tiers (Figure 7b), avoiding the downward compression observed previously. Furthermore, the individual survival predictions (Figure 7c) exhibit a healthy, realistic variance around the overall baseline that closely mirrors the full-risk model's behavior.

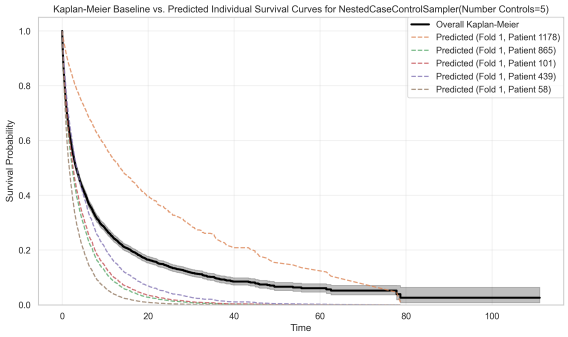
Crucially, this robust performance is not an anomaly of this specific size. Our evaluations show that across the entire range of tested control sizes—from 5 up to 100 samples—the predicted survival curves remain virtually identical. The nested case-control method effectively anchors to the full-risk baseline regardless of how tightly the control pool is restricted, making it an exceptionally stable choice for survival estimation under full-batch execution.



(a) Model calibration comparing empirical Kaplan-Meier vs. predicted survival (Nested Case-Control Controls=5, Full-Batch).



(b) Predicted individual survival curves stratified by true risk tiers (Nested Case-Control Controls=5, Full-Batch).

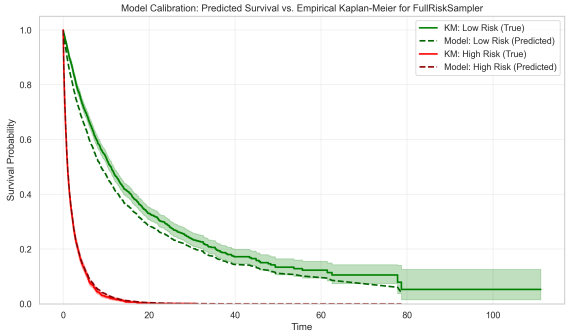


(c) Overall Kaplan-Meier baseline vs. predicted individual survival curves (Nested Case-Control Controls=5, Full-Batch).

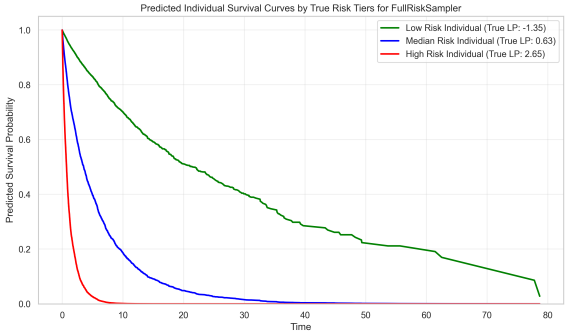
Figure 7: Survival analysis results for Nested Case-Control Controls=5 (Full-Batch).

Mini-Batch Execution: Full-Risk Baseline

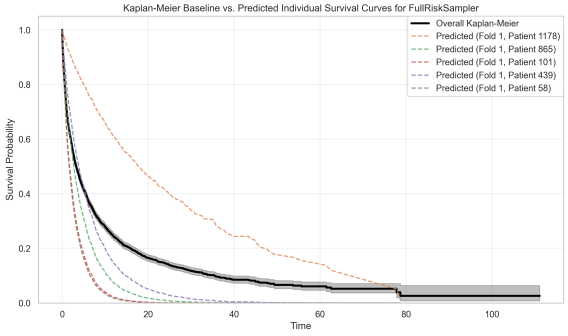
Transitioning to the mini-batch execution regime, we first establish the performance of the full-risk sampler. As expected, the baseline model demonstrates excellent predictive fidelity. Figure 8a shows tight calibration between the predicted risk curves and the empirical Kaplan-Meier baseline. The risk tiers remain distinct and highly proportional (Figure 8b), and individual survival trajectories exhibit a realistic distribution around the overall population curve (Figure 8c). This confirms that the baseline model's strong predictive capacity holds across different batching paradigms.



(a) Model calibration comparing empirical Kaplan-Meier vs. predicted survival (Baseline, Mini-Batch).



(b) Predicted individual survival curves stratified by true risk tiers (Baseline, Mini-Batch).

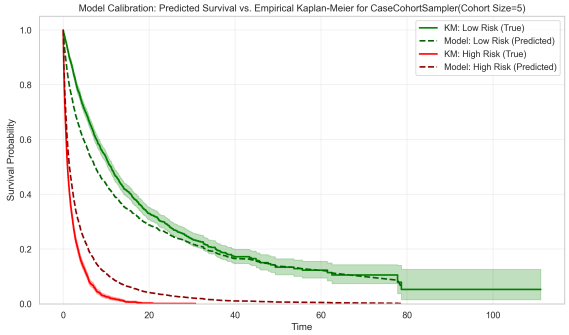


(c) Overall Kaplan-Meier baseline vs. predicted individual survival curves (Baseline, Mini-Batch).

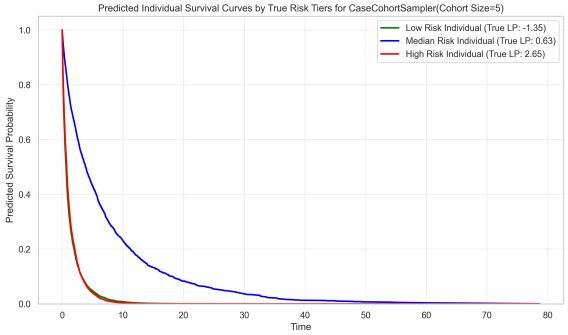
Figure 8: Survival analysis results for Baseline (Mini-Batch).

Mini-Batch Execution: Case-Cohort ($k = 5$)

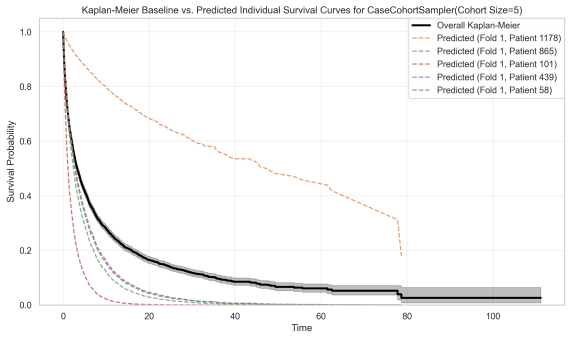
When applying extreme down-sampling using the case-cohort method ($k = 5$) under a mini-batch regime, we observe a degradation in performance, though interestingly, it is notably less catastrophic than its full-batch counterpart. While Figure 9a still reveals some underestimation in long-term survival probability, the curve does not entirely collapse to zero as it did in the full-batch setting. The separation of risk tiers (Figure 9b) and individual curve variance (Figure 9c) exhibit noticeable compression, indicating that a subcohort size of 5 remains too restrictive to capture the full population heterogeneity.



(a) Model calibration comparing empirical Kaplan-Meier vs. predicted survival (Case-Cohort Size=5, Mini-Batch).



(b) Predicted individual survival curves stratified by true risk tiers (Case-Cohort Size=5, Mini-Batch).

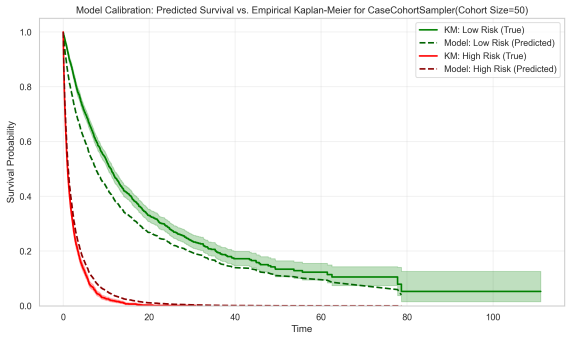


(c) Overall Kaplan-Meier baseline vs. predicted individual survival curves (Case-Cohort Size=5, Mini-Batch).

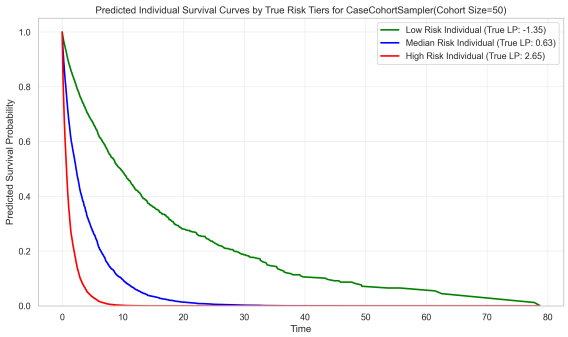
Figure 9: Survival analysis results for Case-Cohort Size=5 (Mini-Batch).

Mini-Batch Execution: Case-Cohort ($k = 50$)

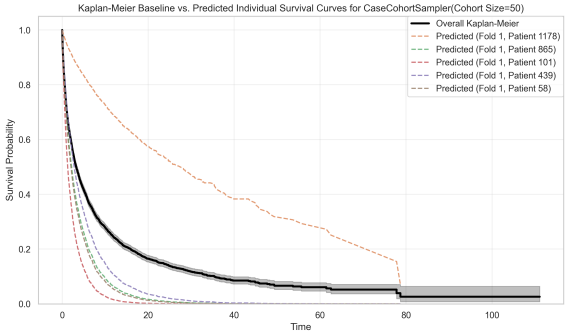
A striking dynamic emerges when marginally increasing the case-cohort size under the mini-batch regime. Unlike the full-batch scenario—which required scaling up to a massive $k = 1000$ to recover performance—the mini-batch approach allows the case-cohort sampler to effectively match the baseline with a remarkably modest size increase to just $k = 50$. As demonstrated in Figure 10a, the predicted survival curves at $k = 50$ show close visual agreement with those of the full-risk baseline. The compression observed at $k = 5$ is substantially reduced, with improved risk-tier stratification (Figure 10b) and recovery of the spread in individual predicted survival trajectories (Figure 10c). These results illustrate the improved performance of case-cohort sampling at larger subcohort sizes under mini-batching.



(a) Model calibration comparing empirical Kaplan-Meier vs. predicted survival (Case-Cohort Size=50, Mini-Batch).



(b) Predicted individual survival curves stratified by true risk tiers (Case-Cohort Size=50, Mini-Batch).



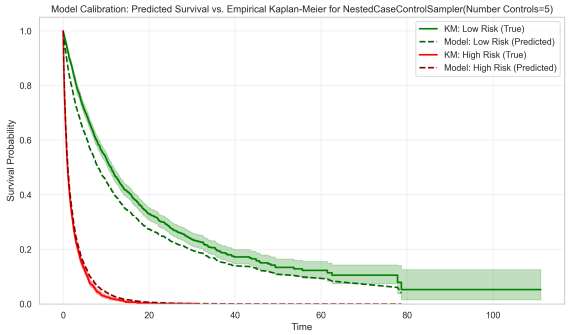
(c) Overall Kaplan-Meier baseline vs. predicted individual survival curves (Case-Cohort Size=50, Mini-Batch).

Figure 10: Survival analysis results for Case-Cohort Size=50 (Mini-Batch).

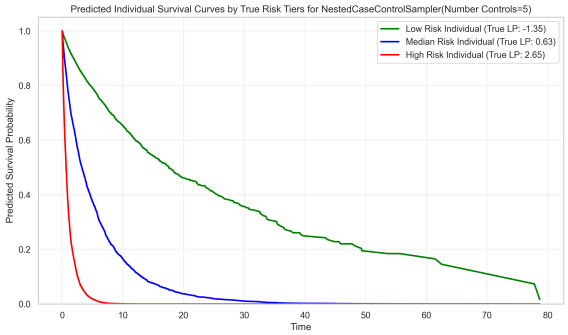
Mini-Batch Execution: Nested Case-Control

Finally, we examine the nested case-control sampler under the mini-batch regime with a heavily restricted control set ($k = 5$). Continuing the trend observed in the full-batch experiments, the nested case-control approach proves exceptionally resilient to extreme down-sampling.

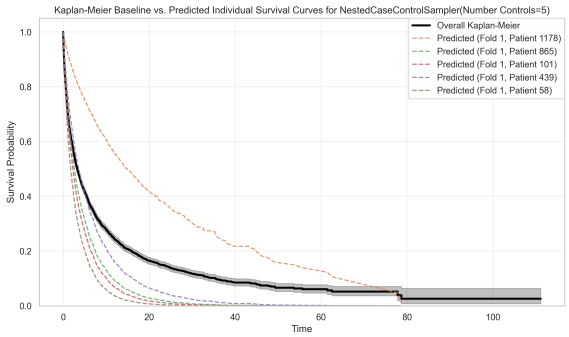
As shown in Figure 11a, the model calibration is virtually indistinguishable from the full-risk baseline, with the predicted survival probabilities tracking the empirical Kaplan-Meier curves tightly across both risk groups. The risk tiers exhibit sharp, well-defined proportional separation (Figure 11b), and the individual survival curves maintain a realistic, healthy spread around the overall population baseline (Figure 11c). This demonstrates that the nested case-control sampler successfully retains its structural integrity and baseline-equivalent performance regardless of whether it is trained in a full-batch or mini-batch environment.



(a) Model calibration comparing empirical Kaplan-Meier vs. predicted survival (Nested Case-Control Controls=5, Mini-Batch).



(b) Predicted individual survival curves stratified by true risk tiers (Nested Case-Control Controls=5, Mini-Batch).



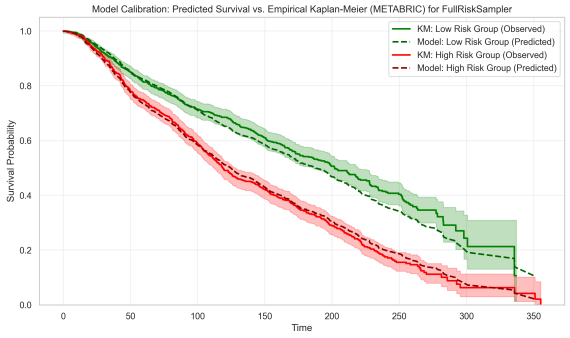
(c) Overall Kaplan-Meier baseline vs. predicted individual survival curves (Nested Case-Control Controls=5, Mini-Batch).

Figure 11: Survival analysis results for Nested Case-Control Controls=5 (Mini-Batch).

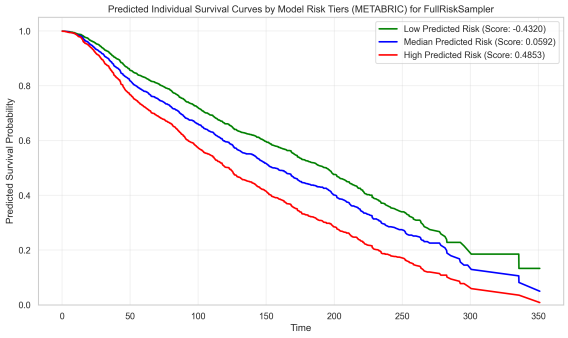
Real-World Survival Curves: Full-Batch Regime

Full-Batch Execution: Full-Risk Baseline

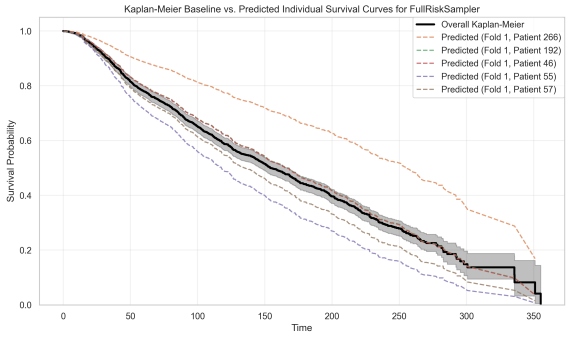
Evaluating the models on real-world clinical data (METABRIC) under full-batch training establishes the baseline predictive performance. As shown in Figure 12a, the full-risk sampler achieves tight calibration, closely tracking empirical Kaplan-Meier curves across both low- and high-risk strata throughout the entire follow-up period. Stratification by predicted risk tiers (Figure 12b) demonstrates clear, monotonic separation among low (score -0.4320), median (score 0.0592), and high (score 0.4853) risk groups. Individual predicted survival trajectories (Figure 12c) exhibit a natural, wide spread around the population-level Kaplan-Meier curve, capturing underlying patient heterogeneity accurately.



(a) Model calibration comparing empirical Kaplan-Meier vs. predicted survival (Baseline, Full-Batch).



(b) Predicted individual survival curves stratified by true risk tiers (Baseline, Full-Batch).

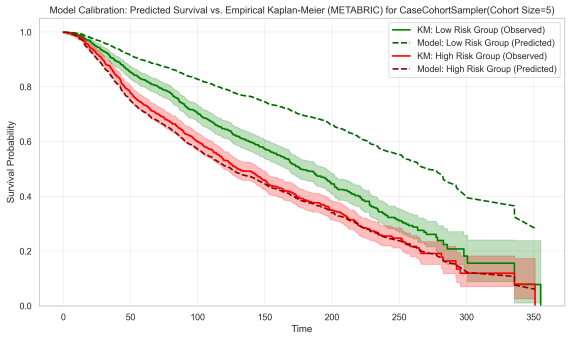


(c) Overall Kaplan-Meier baseline vs. predicted individual survival curves (Baseline, Full-Batch).

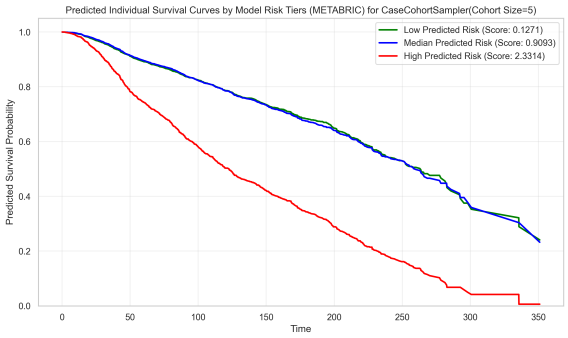
Figure 12: Full-risk baseline evaluations on METABRIC under the full-batch regime.

Full-Batch Execution: Case-Cohort ($k = 5$)

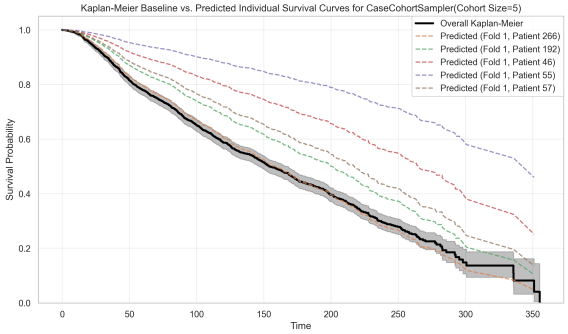
Applying severe subcohort down-sampling ($k = 5$) under full-batch execution leads to notable performance degradation on real-world data. Calibration analysis (Figure 13a) reveals significant overestimation of survival probabilities for the low-risk subgroup compared to empirical Kaplan-Meier observations. Furthermore, risk-tier stratification (Figure 13b) suffers from compression: predicted curves for low-risk (score 0.1271) and median-risk (score 0.9093) individuals overlap almost completely across the early timeline. Individual survival curves (Figure 13c) also display restricted trajectory variance, indicating that a subcohort size of $k = 5$ in full-batch training lacks sufficient risk set information to model complex real-world clinical dynamics.



(a) Model calibration comparing empirical Kaplan-Meier vs. predicted survival (Case-Cohort Size=5, Full-Batch).



(b) Predicted individual survival curves stratified by true risk tiers (Case-Cohort Size=5, Full-Batch).

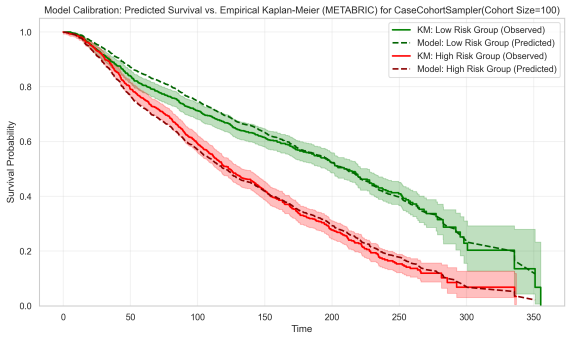


(c) Overall Kaplan-Meier baseline vs. predicted individual survival curves (Case-Cohort Size=5, Full-Batch).

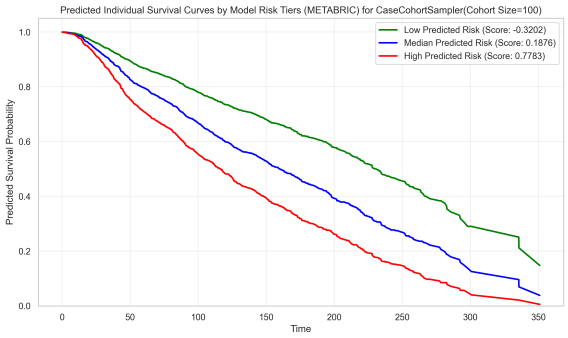
Figure 13: Case-cohort evaluations with subcohort size $k = 5$ on METABRIC under the full-batch regime.

Full-Batch Execution: Case-Cohort ($k = 100$)

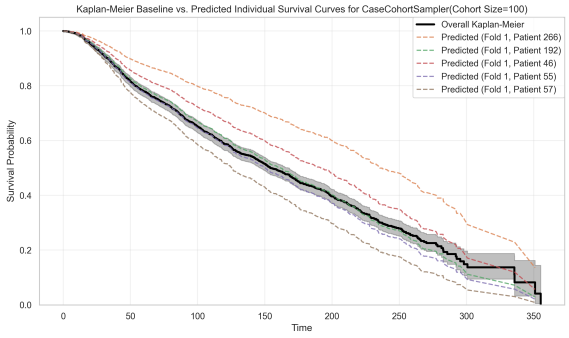
Increasing the subcohort size to $k = 100$ resolves the calibration drift and tier compression observed at $k = 5$. As illustrated in Figure 14a, calibration re-aligns tightly with the empirical Kaplan-Meier curves for both risk strata. Risk tier separation and individual curve trajectories (Figure 14c) fully recover the fidelity and diversity seen in the full-risk baseline, confirming that full-batch training requires large subcohort sizes to maintain accuracy on real-world datasets.



(a) Model calibration comparing empirical Kaplan-Meier vs. predicted survival (Case-Cohort Size=100, Full-Batch).



(b) Predicted individual survival curves stratified by true risk tiers (Case-Cohort Size=100, Full-Batch).

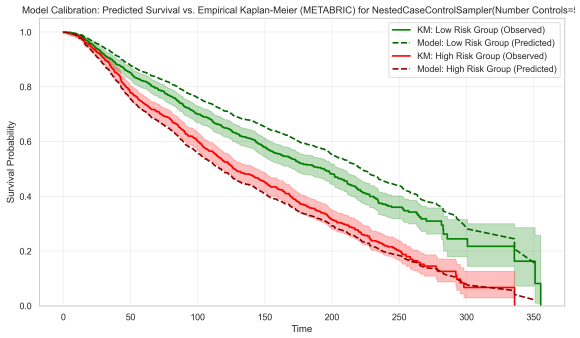


(c) Overall Kaplan-Meier baseline vs. predicted individual survival curves (Case-Cohort Size=100, Full-Batch).

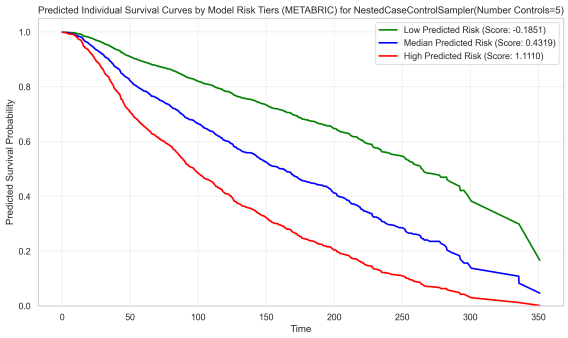
Figure 14: Case-cohort evaluations with subcohort size $k = 100$ on METABRIC under the full-batch regime.

Full-Batch Execution: Nested Case-Control ($m = 5$)

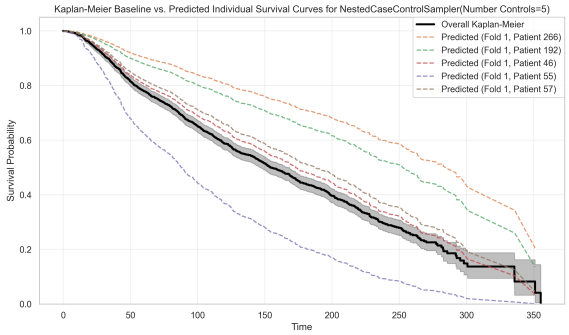
In contrast to the performance degradation observed with small subcohort sizes ($k = 5$) in case-cohort sampling, nested case-control sampling with $m = 5$ controls per risk set maintains strong predictive fidelity under full-batch execution. Calibration analysis (Figure 15a) demonstrates that predicted survival curves track empirical Kaplan-Meier observations closely for both risk strata, with only slight overestimation in the low-risk group across intermediate follow-up times ($t \in [50, 250]$). Risk-tier stratification (Figure 15b) yields clear, monotonic separation among low (score -0.1851), median (score 0.4319), and high (score 1.1110) risk groups without experiencing early curve compression. Furthermore, individual predicted survival trajectories (Figure 15c) exhibit broad, natural dispersion around the overall Kaplan-Meier baseline, confirming that $m = 5$ risk-set controls effectively capture patient-level heterogeneity under full-batch training.



(a) Model calibration comparing empirical Kaplan-Meier vs. predicted survival (Nested Case-Control Controls=5, Full-Batch).



(b) Predicted individual survival curves stratified by true risk tiers (Nested Case-Control Controls=5, Full-Batch).

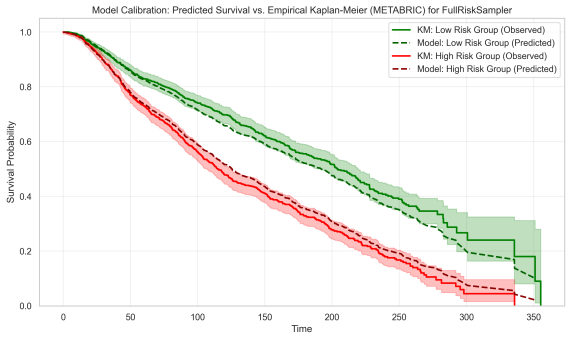


(c) Overall Kaplan-Meier baseline vs. predicted individual survival curves (Nested Case-Control Controls=5, Full-Batch).

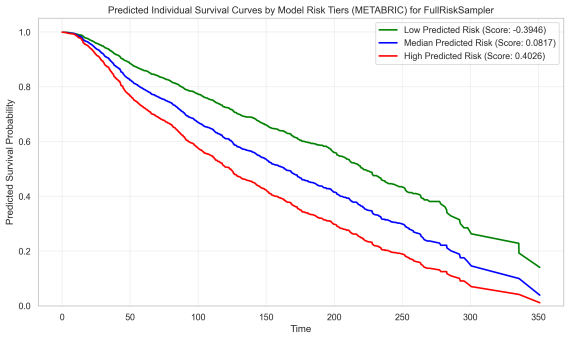
Figure 15: Nested case-control evaluations with $m = 5$ controls on METABRIC under the full-batch regime.

Mini-Batch Execution: Full-Risk Baseline

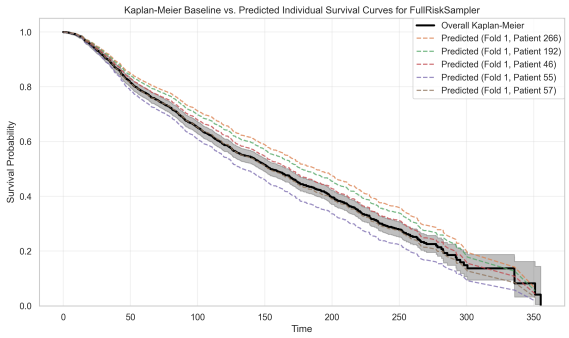
Evaluating the models on real-world clinical data (METABRIC) under mini-batch training shows that stochastic optimization maintains high predictive fidelity. As shown in Figure 16a, the full-risk sampler under mini-batch execution achieves strong calibration, closely tracking empirical Kaplan-Meier curves for both low- and high-risk strata across the follow-up timeline. Stratification by risk tiers (Figure 16b) yields clear, monotonic separation among low (score -0.3946), median (score 0.0817), and high (score 0.4026) risk groups. Individual predicted survival trajectories (Figure 16c) display a natural, wide dispersion around the overall Kaplan-Meier baseline, effectively capturing patient heterogeneity under mini-batch optimization.



(a) Model calibration comparing empirical Kaplan-Meier vs. predicted survival (Baseline, Mini-Batch).



(b) Predicted individual survival curves stratified by true risk tiers (Baseline, Mini-Batch).

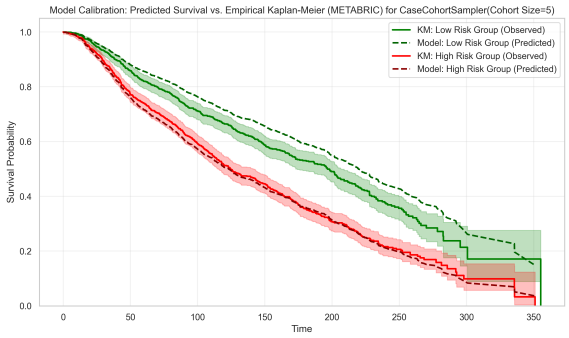


(c) Overall Kaplan-Meier baseline vs. predicted individual survival curves (Baseline, Mini-Batch).

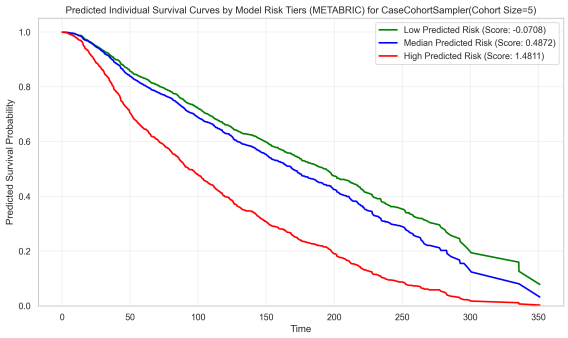
Figure 16: Full-risk baseline evaluations on METABRIC under the mini-batch regime.

Mini-Batch Execution: Case-Cohort ($k = 5$)

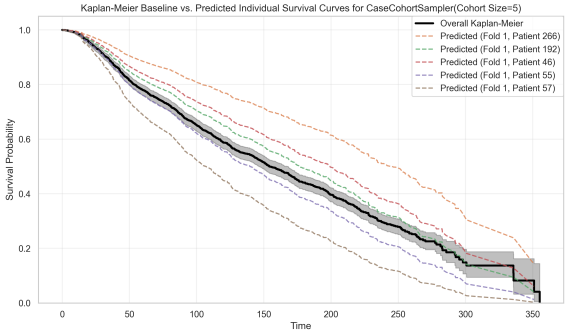
Under severe subcohort down-sampling ($k = 5$), mini-batch execution exhibits some calibration degradation due to limited risk-set representation within individual batches; however, it vastly outperforms its full-batch counterpart at the same subcohort size, reinforcing that case-cohort sampling operates far more effectively in a mini-batch setting. Calibration analysis (Figure 17a) reveals consistent overestimation of low-risk survival probabilities, with predicted curves lying above the empirical Kaplan-Meier confidence interval. Risk-tier stratification (Figure 17b) produces low (score -0.0708), median (score 0.4872), and high (score 1.4811) risk curves, where low- and median-risk trajectories exhibit noticeable proximity early on before diverging. Individual survival curves (Figure 17c) reflect restricted trajectory variation around the population baseline, illustrating the residual limitations of severe subcohort down-sampling despite mini-batch optimization providing a substantial improvement over full-batch training.



(a) Model calibration comparing empirical Kaplan-Meier vs. predicted survival (Case-Cohort Size=5, Mini-Batch).



(b) Predicted individual survival curves stratified by true risk tiers (Case-Cohort Size=5, Mini-Batch).

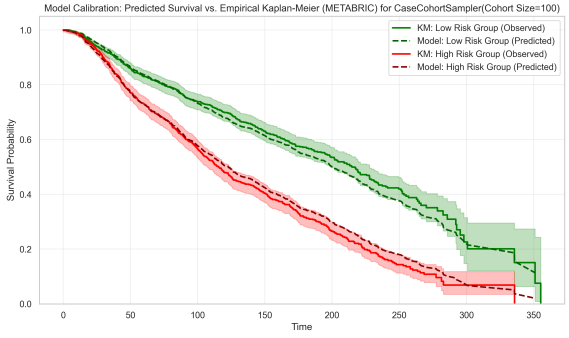


(c) Overall Kaplan-Meier baseline vs. predicted individual survival curves (Case-Cohort Size=5, Mini-Batch).

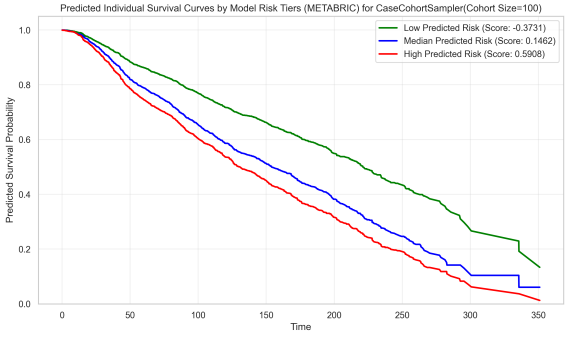
Figure 17: Case-cohort evaluations with subcohort size $k = 5$ on METABRIC under the mini-batch regime.

Mini-Batch Execution: Case-Cohort ($k = 100$)

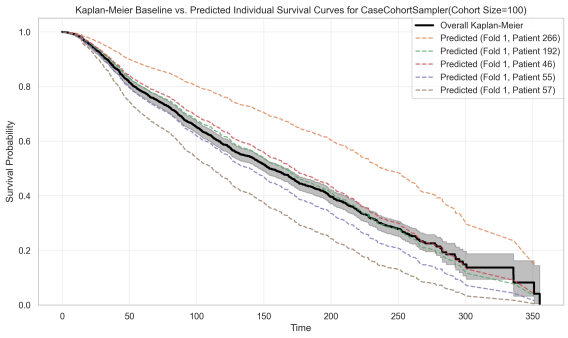
Increasing the subcohort size to $k = 100$ substantially improves model performance under mini-batch training. Mini-batch optimization combined with $k = 100$ provides sufficient stochastic variation to eliminate calibration bias (Figure 18a), bringing predicted survival curves tightly back within empirical Kaplan-Meier bounds. Risk-tier stratification (Figure 18b) recovers clean, monotonic separation among low (score -0.3731), median (score 0.1462), and high (score 0.5908) risk groups. Individual predicted survival trajectories (Figure 18c) also display full, natural dispersion across the population timeline.



(a) Model calibration comparing empirical Kaplan-Meier vs. predicted survival (Case-Cohort Size=100, Mini-Batch).



(b) Predicted individual survival curves stratified by true risk tiers (Case-Cohort Size=100, Mini-Batch).

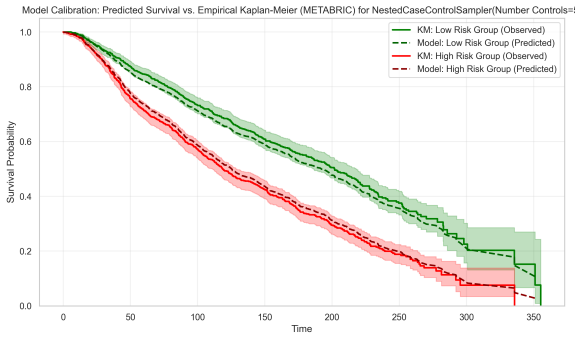


(c) Overall Kaplan-Meier baseline vs. predicted individual survival curves (Case-Cohort Size=100, Mini-Batch).

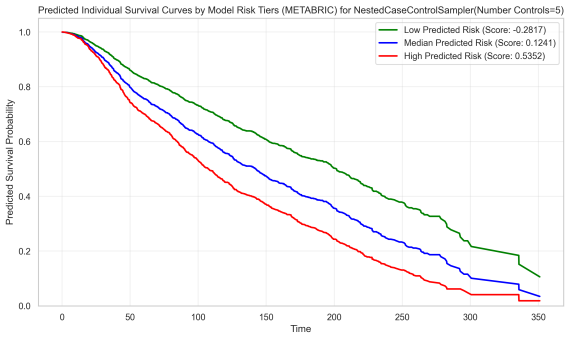
Figure 18: Case-cohort evaluations with subcohort size $k = 100$ on METABRIC under the mini-batch regime.

Mini-Batch Execution: Nested Case-Control ($m = 5$)

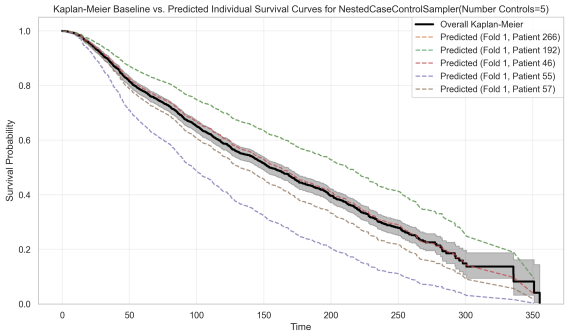
In contrast to case-cohort sampling at $k = 5$, nested case-control sampling with just $m = 5$ controls per event demonstrates remarkable robustness and accuracy in the mini-batch regime. Calibration analysis (Figure 19a) shows nearly ideal alignment between predicted survival curves and empirical Kaplan-Meier estimates, with both low- and high-risk strata remaining well within their respective confidence intervals across the full timeline. Risk-tier stratification (Figure 19b) yields clear, monotonic separation between low (score -0.2817), median (score 0.1241), and high (score 0.5352) risk cohorts. Individual predicted trajectories (Figure 19c) exhibit broad, realistic dispersion around the overall Kaplan-Meier baseline, confirming that matching controls directly to failure times effectively preserves local risk-set dynamics even in small stochastic mini-batches.



(a) Model calibration comparing empirical Kaplan-Meier vs. predicted survival (Nested Case-Control Controls=5, Mini-Batch).



(b) Predicted individual survival curves stratified by true risk tiers (Nested Case-Control Controls=5, Mini-Batch).



(c) Overall Kaplan-Meier baseline vs. predicted individual survival curves (Nested Case-Control Controls=5, Mini-Batch).

Figure 19: Nested case-control evaluations with $m = 5$ controls on METABRIC under the mini-batch regime.