Synthetic RNA-seq Cohorts: A Feasibility Study for Privacy-Aware Collaboration on Sensitive Genomic Data

Aditya Nanda and Ashish Patel

Executive Summary

Why this matters: Sharing patient-level transcriptomic data across teams and partners is mired in compliance and friction because RNA-seq counts (i.e., gene expression data) can pose substantial re-identification risks, especially when paired with clinical metadata.

What we tested: We ran a structured feasibility evaluation of dbTwin's proprietary, high-fidelity synthetic bulk RNA-seq cohort generator on a real bulk RNA-seq cohort (sepsis vs. control), assessing fidelity across differential expression (DE), machine learning (ML), and pathway analysis, as well as record-level similarity to assess re-identification risk.

Outcome: dbTwin generated synthetic cohorts that faithfully reproduced key analytical signals while maintaining record-level diversity from real samples, confirming that high-utility, privacy-safe data sharing is achievable without re-engineering existing omics workflows.

Background & Problem

Clinico-genomic datasets are essential to clinical and translational research, yet are among the most tightly governed assets and least accessible in the modern drug development lifecycle. Institutional Review Boards (IRBs), Data Use Agreements (DUAs), and regulatory guidance restrict redistribution of omics records because genetic signals remain identifying even after standard de-identification. Leading firms, including Novartis and Roche, treat genomic data as a category that often cannot be shared at all.

The practical cost is real: programs scientifically ready to collaborate are delayed by months of contracting and legal review. Synthetic data shifts the question from *"Can we redact enough?"* to *"Can we share a biology-aware synthetic twin?"*

Data & Generation Method

Input: A publicly available clinical RNA-seq cohort with 56 sepsis and 80 non-sepsis control samples, spanning ~27k gene features, provided by Decode Health.

Generation: dbTwin used its proprietary non-deep-learning generative approach for high-dimensional clinico-omics data that preserves cohort statistics, ML performance, and pathway-level expression patterns.

Output: A synthetic cohort with the same schema, dimensions, and cohort distribution, ready to drop into DE, ML, and pathway pipelines.

Validation: Comparative DE and ML analyses between real and synthetic cohorts, along with downstream biological interpretation, were led by Decode Health.

Validation Results

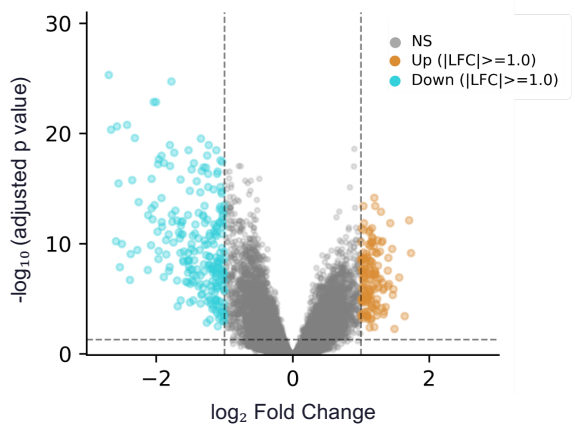

Figure 1. *Differential expression on real cohort (sepsis vs. non-sepsis). 368 genes meet moderate stringency thresholds ($padj < 0.05$, $|\log_2 FC| \geq 1$, $baseMean \geq 10$). See Fig. 3 for gene counts.*

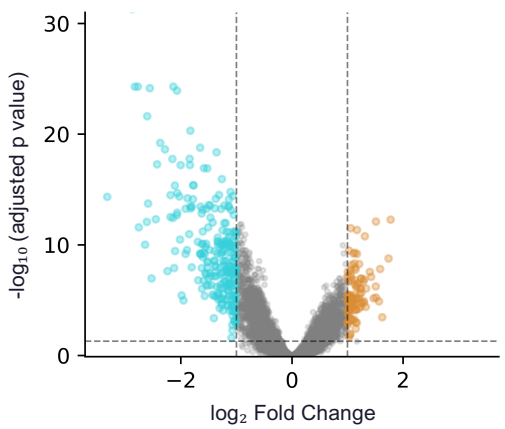

Figure 2. *Differential expression on synthetic cohort. 352 genes meet moderate stringency thresholds, with **300 (81.5%) overlapping** the real DE set. log_2FC and $-log_{10}(padj)$ correlate with real at **Pearson's r = 0.95 and 0.92**, respectively.*

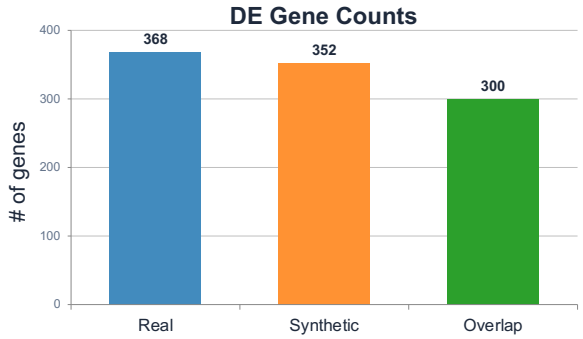

Figure 3. *DE gene counts (moderate stringency): 368* ***real****, 352* ***synthetic****, and 300* ***overlapping*** *(81.5%) genes meeting* $padj < 0.05,\ |\log_2 FC| \geq 1,\ baseMean \geq 10.$

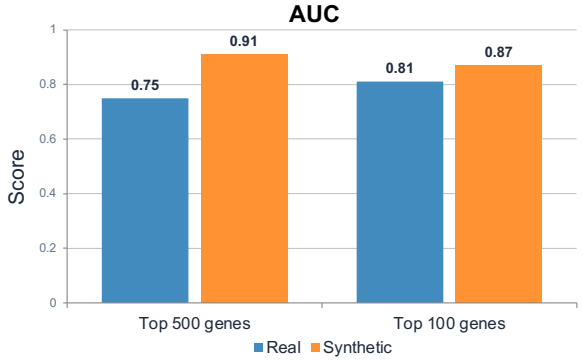

Figure 4. ***Synthetic****-trained ML models classify sepsis vs. control samples as accurately as* ***real****-trained models, confirming that predictive performance is preserved.*

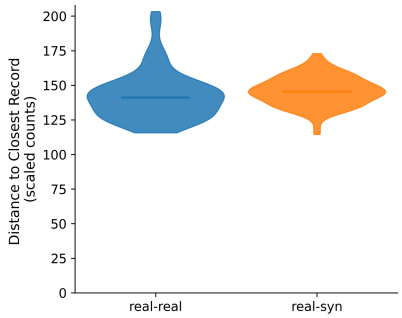

Figure 5. *Record-level distance to closest record on scaled counts. The* ***real-synthetic*** *distribution (median = 145.0) tracks* ***real-real*** *(median = 140.7), confirming that synthetic cohorts are unique and not memorized versions of real transcriptomic signatures.*

Key Takeaways

- **No workflow re-engineering required.** Synthetic cohorts share schema, dimensionality, and label structure with the source data, dropping directly into existing DE, ML, and pathway pipelines without modification.
- **dbTwin cohort preserves key DE and ML performance metrics.** At moderate stringency, **368 real** and **352 synthetic** DE genes overlap by **300 genes (81.5%)** with 100% directional concordance, and $\log_2$FC effect sizes are tightly concordant on overlapping genes (Spearman's ρ = 0.99). Across all genes, $\log_2$FC and adjusted p-value also show strong agreement (r = 0.92 and 0.95). ML models trained on synthetic data (top 100 and 500 gene features) achieved higher AUC than real datasets.
- **Key disease pathways recapitulated in dbTwin cohort.** Among the **300 overlapping DE genes**, enrichment converged on biologically relevant immune pathways — most prominently **Immune System** ($p = 2.81 \times 10^{-5}$, $FDR = 0.013$) and **Neutrophil degranulation** ($p = 7.36 \times 10^{-5}$, $FDR = 0.017$), with additional enrichment for STAT3 signaling and TP53-regulated programs — preserving sepsis biology alongside statistical fidelity.
- **Distance-to-record tests show low re-identification risk.** Distance-to-closest-record analysis confirms that the generator does not memorize individual samples. Synthetic records are no closer to their nearest real neighbor than real samples are to each other (median distance: 145.0 real-to-synthetic vs. 140.7 real-to-real), supporting compliant cross-partner data sharing.

Next Steps: This study demonstrates feasibility on a single public sepsis bulk RNA-seq cohort. The immediate next step is applying the same pipeline to proprietary clinico-transcriptomic datasets, where governance and compliance constraints make synthetic data generation most valuable. We are running scoped pilots to answer a concrete question: can a partner share a synthetic version of their bulk RNA-seq cohort that preserves ML, DE, and pathway statistics while meeting their data governance requirements? A typical pilot takes **less than 7 business days** and delivers a full utility and privacy-readiness report. If you have a bulk RNA-seq dataset you'd like to explore, reach out to the dbTwin team at www.dbtwin.com/team.

Acknowledgements

The authors thank the team at Decode Health for providing the source RNA-seq cohort and the analytical contributions to the work reported here.