The Dimensional Validity Bound: Structural Limits of Clinical AI Evaluation in Multimorbidity

Ian Todd

Sydney Medical School, University of Sydney, Sydney, Australia

itod2305@uni.sydney.edu.au

December 10, 2025

Abstract

Objective: Clinical decision support systems (CDSS) are typically evaluated using aggregate metrics like AUC, assuming the inference problem is well-posed across the patient population. We identify a structural failure mode where this assumption breaks: when patient effective dimensionality (D_{eff}) exceeds the model's representational capacity by a critical ratio.

Materials and Methods: We derive the "Dimensional Validity Bound" from information-theoretic principles (Fano's inequality) and validate it using simulation studies of coupled observer-system dynamics. We then apply this framework to 425,216 ICU admissions from the MIMIC-IV database, stratifying patients by multimorbidity burden to measure the relationship between physiological complexity and model performance.

Results: Simulations confirm a critical threshold at dimensional ratio $r \approx 0.3$, below which inference becomes unstable. In MIMIC-IV, we observe a paradoxical U-

shaped performance curve: models perform best on low-complexity (healthy) and high-complexity (stereotyped/collapsed) patients, but fail in the "moderate complexity" regime ($D_{\text{eff}} \approx 37.5$). This "Zone of Maximum Entropy" represents a systematic blind spot where aggregate AUC remains high (0.835) despite structural posterior instability.

Discussion: Current evaluation protocols that report aggregate metrics mask validity failures in moderate-complexity populations. We propose that D_{eff} be reported as standard metadata for clinical datasets.

Conclusion: Statistical learning in healthcare faces structural limits defined by patient complexity. Recognizing the dimensional validity bound is essential for safe CDSS deployment in multimorbid populations.

Keywords: clinical decision support; machine learning evaluation; high-dimensional inference; multimorbidity; model validity

1 Background and Significance

Despite the proliferation of high-performance algorithms in clinical medicine, real-world deployment has been plagued by reliability failures that aggregate validation metrics failed to predict. High-profile examples—the degradation of the Epic Sepsis Model in external validation, the limited generalizability of IBM Watson for Oncology, the systematic underperformance of COVID-19 prognostic models—suggest a gap between validation performance and deployment reality. Current explanations focus on dataset shift, label leakage, or implementation failures. We propose a more fundamental, structural cause: the **dimensional mismatch** between low-dimensional clinical models and high-dimensional patient physiology.

Clinical prediction is widely assumed to be a "capacity" problem—solvable with more data and deeper networks. However, information theory suggests that when the entropy of a system (the patient) exceeds the channel capacity of the observer (the model) by a critical margin, inference becomes not just difficult, but structurally ill-posed (Fano's inequality).

In this study, we formalize this limit as the **Dimensional Validity Bound**.

Objectives. We (1) formalize the dimensional validity bound for clinical prediction as a function of patient effective dimensionality and model capacity, (2) identify three mechanisms by which validity fails below this bound (projection error, noise amplification, hypothesis space explosion), and (3) empirically test these predictions in a large ICU cohort stratified by multimorbidity.

2 Introduction

The remarkable success of statistical learning in healthcare has created an implicit assumption: given sufficient data, appropriate algorithms, and careful validation, clinical prediction problems are solvable in principle. Performance may plateau, but the problem itself remains well-posed. ROC curves measure discrimination, calibration plots measure reliability, and the gap between current performance and perfect prediction represents reducible error.

We demonstrate that this assumption fails in a specific, identifiable regime. When the effective dimensionality of the system being predicted exceeds the representational capacity of the observer by a critical ratio, the inference problem becomes *structurally ill-posed*. This is not a limitation of current methods—it is an information-theoretic impossibility. No amount of additional data, algorithmic sophistication, or computational resources can rescue validity in this regime, because the problem itself has become undefined.

The failure mode is subtle. Standard metrics continue to produce numbers. Models can be trained, validated, and deployed. AUC values appear reasonable. But these metrics are measuring the wrong thing: they report how well the model separates outcomes *within its low-dimensional projection* of reality, not whether that projection preserves the structure needed for valid inference.

We formalize this as the **dimensional validity bound**: the critical ratio of observer to system dimensionality below which statistical learning becomes unreliable. The bound

emerges from applying Fano's inequality to the inference problem: when the entropy of the system state exceeds the observer's channel capacity, error probability has a non-vanishing lower bound regardless of the inference strategy employed.

Three distinct mechanisms drive failure below the validity bound:

1. Projection error. When $D_{\text{observer}} < D_{\text{system}}$, the model necessarily performs lossy compression. Distinctions that exist in the full state space collapse to indistinguishable representations. This is mathematically unavoidable—no algorithm can preserve information it lacks capacity to represent.

2. Noise amplification. Higher-dimensional system states amplify variance when projected onto lower-dimensional models. Small perturbations in unmeasured dimensions propagate as large, unpredictable fluctuations in predictions. The model experiences this as irreducible aleatoric uncertainty.

3. Hypothesis space explosion. As system dimensionality grows, the number of distinguishable states scales exponentially. Bayesian priors become unstable because no training set can adequately cover the hypothesis space. The posterior becomes dominated by prior assumptions rather than observed evidence.

We validate these predictions in clinical medicine—a domain where the dimensional validity bound has immediate practical consequences. Patients are high-dimensional dynamical systems; clinical decision support algorithms operate in low-dimensional feature spaces. The mismatch is structural, and the consequences are predictable.

3 Theory

3.1 Effective Dimensionality

We adopt the participation ratio as our measure of effective dimensionality [1]:

$$D_{\text{eff}} = \frac{\left(\sum_i \lambda_i\right)^2}{\sum_i \lambda_i^2} \tag{1}$$

where λ_i are eigenvalues of the covariance matrix. The participation ratio is widely used in neuroscience and physics as a measure of manifold dimensionality, where it has been shown to correlate with system degrees of freedom across domains from neural population activity to quantum systems [1]. We use it here because it is basis-invariant and captures both heavy-tailed and flat spectra with a single scalar. This metric measures how many dimensions contribute meaningfully to variance—a system with one dominant principal component has $D_{\text{eff}} \approx 1$; a system with variance spread uniformly across n dimensions has $D_{\text{eff}} \approx n$.

3.2 The Dimensional Validity Bound

Let an observer operate in D_{obs} effective dimensions while the system state occupies D_{sys} effective dimensions. Define the dimensional ratio:

$$r = \frac{D_{\text{obs}}}{D_{\text{sys}}} \tag{2}$$

Dimensional validity principle. *Under standard assumptions on observation noise and coupling dynamics, stable inference requires a dimensional ratio r above a critical value r^*. Below this threshold, the expected tracking error satisfies:*

$$\mathbb{E}[\text{error}] \geq f\left(\frac{D_{\text{sys}}}{D_{\text{obs}}}\right) \tag{3}$$

for some monotonically increasing function f, and the inference problem becomes ill-posed. We do not attempt to identify f analytically; here it serves to emphasize that the lower bound on error grows monotonically with dimensional mismatch. The critical value $r \approx 0.3$ emerges from the coupled oscillator simulations in Section 3.2 and should be interpreted as an order-of-magnitude characteristic of the phase transition, not a universal constant.

Information-theoretic motivation. The bound follows from Fano's inequality applied to the inference problem. Consider a system state X with entropy $H(X) \propto D_{\text{sys}}$, where D_{sys} is the effective dimensionality defined in Section 2.1—higher-dimensional states have expo-

nentially more distinguishable configurations. Let $|\mathcal{X}|$ denote the number of distinguishable system states in the relevant observation window. An observer with representational capacity $C \propto D_{\text{obs}}$ cannot reliably infer X when $C < H(X)$. Specifically, Fano's inequality guarantees:

$$P(\text{error}) \geq \frac{H(X) - C - 1}{\log |\mathcal{X}|} \tag{4}$$

Since $H(X)$ scales with D_{sys} (as measured by the participation ratio), increasing system dimensionality at fixed observer capacity necessarily increases the lower bound on error. The critical value $r^* \approx 0.3$ is estimated from dynamical systems simulations (Section 3) rather than analytically derived, and represents an order-of-magnitude threshold.

This is analogous to Shannon's channel capacity: it defines a regime where no inference strategy can succeed, regardless of algorithmic sophistication.

3.3 Why Standard Metrics Fail

ROC curves and AUC measure a model's ability to rank instances by predicted probability within its *projected* representation of the system. When projection error is severe, this ranking preserves little of the structure present in the full state space.

Consider: a model operating in 10 dimensions facing a 100-dimensional system ($r = 0.1$) will collapse many distinct system states to the same model representation. Spectrally, this occurs because the observer retains only the top D_{obs} principal directions; the tail of the eigenvalue spectrum acts as unmeasured latent variation, producing systematic conflation of distinct states. The model can achieve good separation *among states it can distinguish*, while systematically failing on states it conflates. Aggregate AUC averages over both, producing a number that conceals the structural failure.

The problem is worse in the intermediate complexity regime. Very low-complexity systems (low D_{sys}) are genuinely tractable. Very high-complexity systems often collapse into stereotyped attractors that paradoxically simplify prediction. The *moderate* complexity

regime—where D_{sys} is high but the system hasn't collapsed—is where the validity bound bites hardest.

We term this the **"zone of maximum entropy"**: the regime where posterior instability is maximal and standard metrics are most misleading.

4 Simulation Studies

We conducted three simulation studies corresponding to the three failure mechanisms.

4.1 Diagnostic Cascade: Hypothesis Space Explosion

When an observer performs multiple tests, each with false-positive rate ϵ, the probability of at least one spurious positive is:

$$P(\text{cascade}) = 1 - (1 - \epsilon)^m \tag{5}$$

We simulated 10,000 systems undergoing 1–50 independent tests with $\epsilon = 0.05$. Results confirm:

- 10 tests: 40% cascade probability
- 14 tests: 50% threshold
- 20 tests: 64% cascade probability
- 30 tests: 79% cascade probability

Theoretical predictions matched simulation precisely ($r^2 > 0.999$). This illustrates hypothesis space explosion: as test count grows, the number of distinguishable outcome patterns explodes exponentially.

4.2 Coupled Oscillator Dynamics: Projection Error

We modeled the observer-system relationship as coupled oscillator dynamics, where the system occupies a high-dimensional state space ($D_{\text{sys}} = 100$) and the observer operates in a lower-dimensional representation space ($D_{\text{obs}} \in \{5, 10, 20, 30, 50, 70, 100\}$).

This is not merely metaphor. The mathematics of coupled dynamical systems—synchronization thresholds, stability conditions, information transfer rates—apply to any system where one entity must track another through noisy observations.

Critical findings:

- $r < 0.1$: Near-complete coupling failure (stability < 0.3)
- $r \approx 0.3$: Critical threshold; steep stability increase
- $r > 0.5$: Approaching stable coupling
- $r = 1.0$: Full dimensional matching; stable synchronization

Models operating with $D \sim 10$ features facing systems with $D \sim 100$ effective dimensions are in the unstable regime.

4.3 Classifier Degradation: Noise Amplification

We generated synthetic data with 50 features, 10 informative, and varying complexity levels. Random forest classifiers were evaluated via 5-fold cross-validation.

Results confirmed a **U-shaped** relationship between classifier performance and system complexity:

Table 1: Classifier performance by system complexity (simulation)

Complexity	N	D_{eff}	AUC (mean ± SD)
Low	1000	41.0	0.91 ± 0.02
Moderate	1000	32.9	**0.61 ± 0.03**
High	1000	24.9	0.85 ± 0.04

The moderate complexity stratum shows worst performance—the zone of maximum entropy where posterior instability peaks.

5 Case Study: Clinical Prediction in High-Dimensional Patients

To test whether the dimensional validity bound manifests in real-world high-dimensional systems, we analyzed mortality prediction in the MIMIC-IV critical care database [2].

5.1 Why Clinical Medicine?

Clinical medicine provides an ideal test case for the dimensional validity bound:

- **Genuine high-dimensionality:** Patients are biological systems with genomic, proteomic, metabolic, and physiological state spaces far exceeding typical model dimensionality.
- **Systematic dimensional mismatch:** Clinical decision support algorithms operate in ∼10–50 dimensional feature spaces facing patients with $D_{\text{eff}} \sim 30$–100.
- **Known failure patterns:** High-profile clinical AI failures (Epic Sepsis Model, IBM Watson for Oncology) exhibit signatures consistent with dimensional validity violations.
- **Measurable complexity gradients:** Multimorbidity provides a natural axis of complexity variation within a single population.

5.2 Methods

We analyzed 425,216 adult hospitalizations from MIMIC-IV v3.1, stratifying by multimorbidity burden (ICD code count): Low (1–8), Moderate (9–15), High (16+). For each stratum, we

computed D_{eff} from clinical features and evaluated gradient boosting classifier performance for mortality prediction via 5-fold cross-validation.

Ethics approval. This study used publicly available, de-identified data from the MIMIC-IV database [2]. In accordance with institutional policy, analysis of these data is exempt from human subjects review and did not require informed consent.

5.3 Results

Table 2: Dimensional validity bound in clinical data (MIMIC-IV, N=425,216)

Multimorbidity	N	D_{eff}	Mortality %	AUC (95% CI)
Low	159,211	44.3	0.3	0.867 (0.858–0.876)
Moderate	128,979	37.5	1.1	**0.835 (0.832–0.839)**
High	137,026	29.5	5.8	0.859 (0.851–0.867)

Notably, the D_{eff} range observed in MIMIC (29.5–44.3) closely matches the simulation parameters (24.9–41.0 in Table 1), providing a consistency check between the synthetic and real-world analyses.

Two findings confirm theoretical predictions:

1. Effective dimensionality decreases with multimorbidity. Counter to naive intuition ("sicker patients are more complex"), D_{eff} drops from 44.3 to 29.5 as multimorbidity increases. This supports the complex systems view that illness represents loss of physiological complexity [3]—diseased systems collapse into stereotyped attractors as organ coupling increases (cardiorenal syndrome, hepatorenal syndrome).

2. Classifier performance shows the predicted U-shape. Worst AUC (0.835) occurs in the *moderate* multimorbidity stratum—the zone of maximum entropy. Low-multimorbidity patients have high D_{eff} but sparse outcomes (strong priors dominate). High-multimorbidity patients have low D_{eff} (collapsed dynamics are quasi-deterministic). The moderate regime has high D_{eff} *and* uncertain outcomes—maximum posterior instability.

This reverses standard assumptions about where models struggle. The dimensional valid-

ity bound bites hardest not in the sickest patients but in the *moderate* complexity transition zone that aggregate AUC systematically conceals.

6 Discussion

6.1 When Aggregate Metrics Conceal Structural Failure

The dimensional validity bound identifies a regime where standard evaluation metrics become unreliable—not because of insufficient data or poor algorithms, but because the inference problem itself is ill-posed.

Consider what aggregate AUC actually measures: the probability that a randomly chosen positive instance ranks higher than a randomly chosen negative instance *within the model's representation*. When projection error is severe, many distinct system states map to the same representation. The model may achieve excellent separation among distinguishable states while systematically failing on conflated states. Aggregate AUC reports the former; validity depends on the latter.

The zone of maximum entropy is particularly insidious. It is invisible to aggregate metrics, occurs in an intermediate regime that intuition misses, and represents the population where model errors may be most consequential.

A critical distinction must be drawn between the *ontological* loss of degrees of freedom and the *epistemic* loss of information due to projection. In high-dimensional probability theory, complex systems often appear simple because macroscopic observables concentrate around the mean—this is a measurement artifact (concentration of measure). In contrast, the attractor collapse observed in multimorbidity represents a *physical reduction in the system's accessible state space* [4]. The patient has not merely become hard to measure; they have structurally lost the capacity for physiological variation. Both phenomena produce apparent simplicity from high complexity, but the mechanisms differ: concentration of measure places the simplification in the observation, while attractor collapse places it in the system itself.

Aggregate metrics cannot distinguish between these cases.

6.2 Implications for Model Deployment

Three practical recommendations follow:

1. Compute D_{eff} at runtime. Before generating predictions, estimate system complexity. When $r < 0.3$, flag that the validity bound may be exceeded.

2. Stratify validation by complexity. Report performance separately for low, moderate, and high complexity strata. Aggregate AUC should not be the sole basis for deployment decisions.

3. Design appropriate abstention. In regimes where the validity bound is violated, models should abstain rather than predict with false confidence.

6.3 Implications for Evaluation Protocols

Current evaluation practice assumes that a model validated on a representative sample will perform comparably on deployment data. The dimensional validity bound shows this assumption fails when complexity distributions differ—even if marginal distributions match.

We propose that evaluation protocols include:

- Mandatory complexity stratification in validation reports
- Explicit statement of the complexity range where validity is demonstrated
- Prospective monitoring for complexity drift in deployed systems

6.4 Limitations

Our simulations make simplifying assumptions. The critical threshold $r \approx 0.3$ is empirically estimated and should be interpreted as order-of-magnitude guidance—what we expect to be robust is the *existence* of such a threshold and the qualitative shape of the stability curve,

not the specific numerical value. Clinical validation uses a single database (MIMIC-IV) with limited feature sets. The zone of maximum entropy location may vary by domain.

6.5 Relation to Prior Work

The dimensional validity bound extends several established research threads. The "curse of dimensionality" literature [5] identifies challenges in high-dimensional statistical estimation, but typically frames these as practical limitations requiring more data or better algorithms. We show that beyond a critical ratio, the problem becomes structurally ill-posed—additional data cannot rescue validity. This structural limit connects to our prior work on measurement thresholds and the sub-Landauer domain in biological systems [6]. Recent work on double descent and effective dimension in overparameterized models [7] reveals non-monotonic generalization behavior; the zone of maximum entropy represents an analogous non-monotonicity in inference validity itself, occurring at the observer-system boundary rather than in model capacity. Our critical threshold is also conceptually related to inference phase transitions identified in statistical physics [8], where sharp boundaries separate regimes of tractable and intractable inference—though our bound arises at the observer-system interface rather than from computational complexity.

The complex systems view that illness represents loss of physiological complexity [4, 9] provides the biological substrate for our findings: diseased systems collapse into lower-dimensional attractors, paradoxically simplifying prediction. This connects the dimensional validity bound to the broader literature on physiological dynamics, where reduced heart rate variability, EEG entropy loss, and similar signatures mark pathological transitions. Our contribution is to formalize when this dimensional collapse makes standard evaluation metrics unreliable.

7 Conclusion

Statistical learning has limits that are not merely practical but structural. When the effective dimensionality of the system being predicted exceeds the observer's representational capacity by more than approximately threefold, inference becomes ill-posed and standard evaluation metrics become misleading.

The dimensional validity bound is not a call for pessimism but for appropriate humility. In domains where models operate well within the bound, current practice is sound. In domains where the bound may be violated—biological systems, social networks, complex physical systems—we need different evaluation protocols, explicit validity checks, and appropriate abstention mechanisms. Because many real-world systems exhibit high intrinsic dimensionality—ecological food webs, financial markets, climate dynamics, immune networks—domains previously assumed to be "data-hungry but tractable" may, in fact, lie beyond the regime where statistical learning is well-posed.

The zone of maximum entropy represents a systematic blind spot in current evaluation practice. It is invisible to aggregate metrics, occurs in an intermediate complexity regime, and may contain the instances where model errors are most consequential. Recognizing this regime is the first step toward addressing it.

Acknowledgments

None.

Funding

This research did not receive specific funding.

Author Contributions

I.T. conceived the study, developed the theoretical framework, performed simulations and data analysis, and wrote the manuscript.

Declaration of Interests

The author declares no competing interests.

Data and Code Availability

MIMIC-IV is available via PhysioNet (`https://physionet.org/content/mimiciv/`). Simulation and analysis code is available at `https://github.com/todd866/clinical-validity-bounds`.

References

[1] John P. Cunningham and Byron M. Yu. Dimensionality reduction for large-scale neural recordings. *Nature Neuroscience*, 17(11):1500–1509, 2014. doi: 10.1038/nn.3776.

[2] Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10:1, 2023. doi: 10.1038/s41597-022-01899-x.

[3] Alan A. Cohen, Luigi Ferrucci, Tamàs Fülöp, Dominique Gravel, Nan Hao, Andres Kriete, Morgan E. Levine, Lewis A. Lipsitz, Marcel G. M. Olde Rikkert, Andrew Rutenberg, Nicholas Stroustrup, and Ravi Varadhan. A complex systems approach to aging biology. *Nature Aging*, 2:580–591, 2022. doi: 10.1038/s43587-022-00252-6.

[4] Lewis A. Lipsitz and Ary L. Goldberger. Loss of 'complexity' and aging: Potential

applications of fractals and chaos theory to senescence. *JAMA*, 267(13):1806–1809, 1992. doi: 10.1001/jama.1992.03480130122036.

[5] Visar Berisha, Chelsea Krantsevich, P. Richard Hahn, et al. Digital medicine and the curse of dimensionality. *npj Digital Medicine*, 4:153, 2021. doi: 10.1038/s41746-021-00521-5.

[6] Ian Todd. The limits of falsifiability: Dimensionality, measurement thresholds, and the sub-Landauer domain in biological systems. *BioSystems*, 258:105608, 2025. doi: 10.1016/j.biosystems.2025.105608.

[7] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019. doi: 10.1073/pnas.1903070116.

[8] Lenka Zdeborová and Florent Krzakala. Statistical physics of inference: thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016. doi: 10.1080/00018732.2016.1211393.

[9] Ary L. Goldberger, Luís A. N. Amaral, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, C.-K. Peng, and H. Eugene Stanley. Fractal dynamics in physiology: Alterations with disease and aging. *Proceedings of the National Academy of Sciences*, 99:2466–2472, 2002. doi: 10.1073/pnas.012579499.