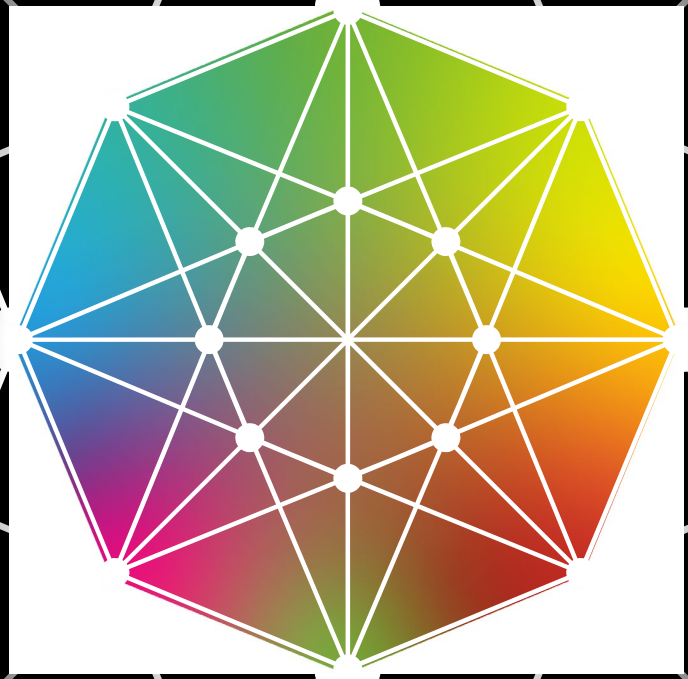
16TH ITU ACADEMIC CONFERENCE

ITU KALEIDOSCOPE
GENEVA2026

AI and frontier technologies for good

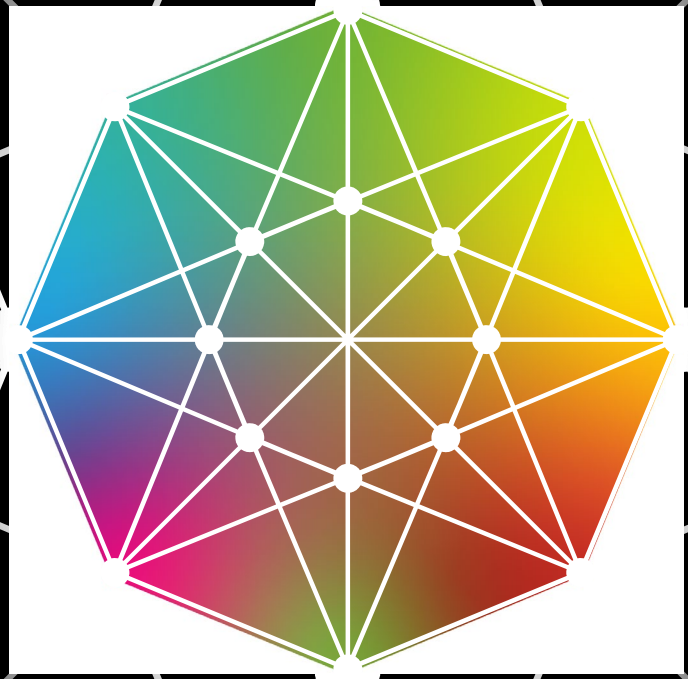
7-9 July 2026
Geneva, Switzerland

Hosted by AI for Good Global Summit



When 99.98% Is Too Good To Be True

July 2026



Patrick Sello

Department of Computer Science & Faculty of Community and Health Sciences
University of Western Cape
ISAT Research Group
South Africa

Session #S3.3

Authors:
Antoine Bagula | Landry Mbale | Olasupo Ajayi | Patrick Sello | Jennifer Chipps | Sean Broomhead | Petra Brysiewicz | Damian Clarke

Presentation Roadmap

01

The Problem: An Unrealistic Accuracy Paradox

02

Methodology: An Evolution of Labelling Functions

03

Experimental Results: Uncovering True Performance

04

Application: Telemedicine Integration & Workflow

05

Compliance: ITU Standards Alignment

06

Conclusions & Key Takeaways

Why This Research Matters

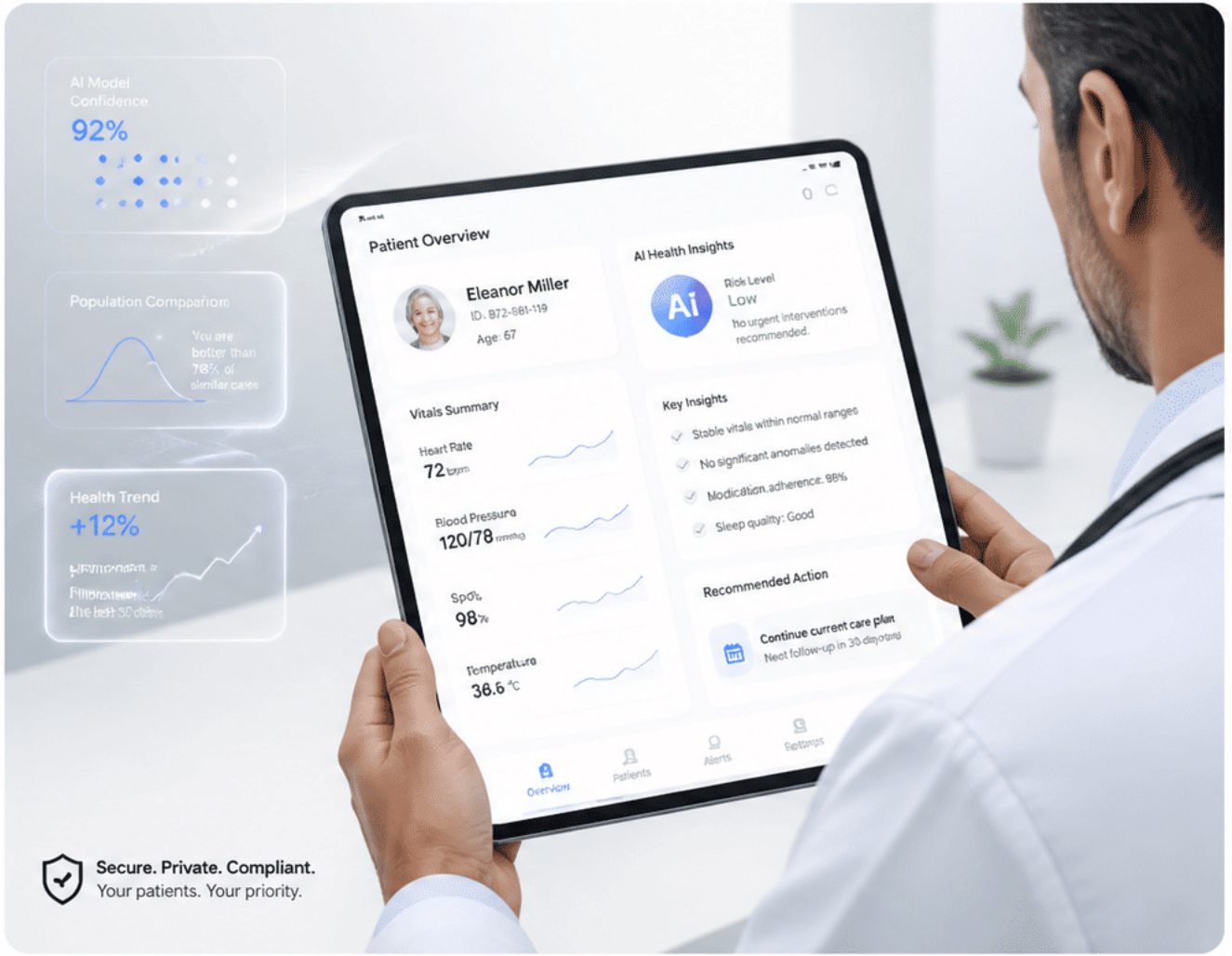
The Stakes in Clinical AI

As AI becomes more integrated into healthcare, its reliability is paramount. This is especially true in critical applications like telemedicine and clinical decision support.

Embodied AI
increasingly supports clinical decisions.

Telemedicine
requires trustworthy AI for remote patient care.

Incorrect predictions
can directly and negatively affect patient outcomes.



The 99.98% Accuracy Paradox

Our initial model (V1) achieved a near-perfect accuracy score. However, this performance was too good to be clinically plausible, raising immediate red flags.

01 Observed Accuracy

The model reported a 99.98% success rate in identifying postoperative complications.

02 The Problem

Such high accuracy is unheard of in real-world clinical settings, suggesting a fundamental flaw in the model's learning process.

03 Our Hypothesis

We suspected rule-induced target leakage, where the AI learns from artifacts in the data labeling process itself, not the underlying clinical reality.

Methodology: The Evolution of Labelling Functions

To solve the overfitting problem, we developed three successive versions of our labelling function, each adding a new layer of sophistication to move from simple keywords to true clinical context.

V1: Myopic Labelling

- Uses simple keyword ratios (complication vs. stability lexicons).
- Easy to interpret.
- Highly vulnerable to leakage and shortcut learning.

V2: Context-Aware

- Adds negation detection (e.g., ignores 'no fever').
- Validates temporal scope (e.g., distinguishes past from present).
- Introduces severity adjustment.

V3: Weighted Scoring

- Groups terms into five semantic categories.
- Applies weighted contributions based on clinical significance.
- Achieves improved calibration and discrimination.

How Simple Labelling Fails: Clinical Examples

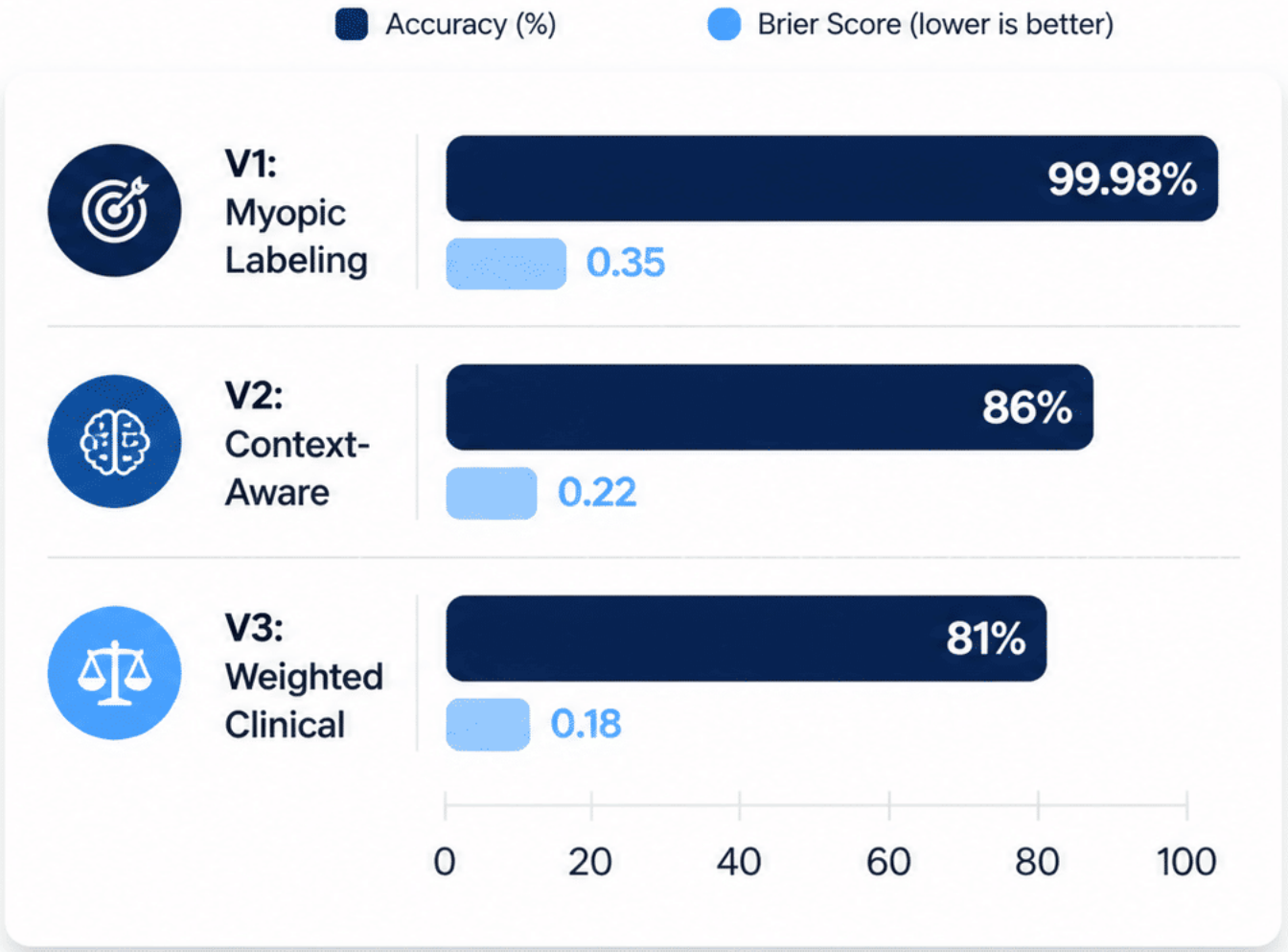
The flaws in the V1 model become obvious when examining how it interprets real clinical notes compared to the more advanced V3 model.

Clinical Note Text	V1 (Myopic) Interpretation	V3 (Contextual) Interpretation
'Patient reports no fever.'	**FALSE POSITIVE:** Sees 'fever' and flags a complication risk.	**CORRECT:** Detects 'no' and correctly identifies a stable sign.
'History of infection in 2021.'	**FALSE RISK:** Sees 'infection' and flags current risk.	**CORRECT:** Understands the historical context and does not escalate.
'Patient has a high-grade fever.'	**CORRECT (by chance):** Flags a risk.	**CORRECT (by design):** Flags a risk and applies the strongest escalation due to 'high-grade'.

Experimental Results: Accuracy vs. Reality

While V1's accuracy was highest, its Brier score (a measure of calibration) was terrible, indicating its confidence was misplaced.

V3 has lower accuracy but a much better Brier score, making it far more reliable for clinical use.



Key Takeaway

High accuracy alone can be misleading without proper calibration.

Clinical Insight

V3 offers the best balance of accuracy and calibration for reliable clinical use.

Key Finding: Calibration Matters More Than Accuracy

01

Accuracy is Misleading

Near-perfect accuracy was a sign of methodological failure (shortcut learning), not success.

02

Calibration Reveals Truth

Metrics like the Brier score, which measure how well-calibrated a model's confidence is, provided a much truer picture of performance.

03

V3 Offers the Best Balance

The V3 model, despite lower raw accuracy, provides the best balance of sensitivity and reliability for real-world clinical deployment.

Application: A Trustworthy Telemedicine Workflow

1. **Clinical Note Ingestion:**
   Raw text data is fed into the system.

2. **Context-Aware Risk Estimation (V3):**
   The model analyses notes for true complication risk.

3. **Adaptive Thresholds:**
   The system flags high-risk patients based on calibrated confidence scores.

4. **Human Clinical Oversight:**
   Clinicians are alerted to review high-risk cases, ensuring AI assists, not replaces, human expertise.

Alignment with ITU Standards

Our research and proposed framework align with key International Telecommunication Union (ITU) focus groups and standards for AI in health, ensuring our work contributes to a globally recognised foundation for trustworthy AI.

FG-AI4H
AI for Health

H.810/H.841
Frameworks for digital health services

FG-HS4AI
Health and safety testing of AI

FG-NET-2030
Future networks and trustworthiness

Conclusions & Final Takeaway

Summary of Findings

An anomalously high accuracy **(99.98%)** is a clear indicator of methodological failure, not success.

Trustworthy **weak supervision** is essential to prevent models from learning misleading shortcuts.

For clinical AI, **calibration** and **reliability** are far more important than raw accuracy scores.

The Core Message

"Trustworthy embedded AI depends not only on sophisticated models, but on trustworthy supervision mechanisms."

Thank you!

ITU KALEIDOSCOPE GENEVA 2026