

Evaluating Clinical Alignment of AI Explanations in Text-Based Immunotherapy Response Prediction for NSCLC

A. Ferrarin¹, T. Arangoa², B.V. Gonzalez³, L. Mazzeo^{1,4}, **V. Peppoloni**¹, G. Leone¹, C. Licciardello⁴, B. Guirges¹, L.M. Invernizzi¹, P. Ambrosini¹, R. Romanò^{1,4}, C. Silvestri¹, A. Venditelli¹, S. Natangelo¹, M. De Ponti¹, M. Meazza Prina¹, A. Spagnoletti¹, L. Provenzano^{1,4}, C. Cavalli¹, G. Montelatici¹, T.S. Cassano¹, A.D. Dumitrascu¹, F. Cusumano¹, M. Borraccino¹, R.M. di Mauro¹, E. Benassi¹, G. Vitello¹, M. Fink¹, F. Baggio¹, E. Zito¹, E. Sottotetti¹, P. Solli¹, A.R. Filippi¹, V. Bartolomeo¹, G. Leuzzi¹, G. Corrao¹, T. Beninato¹, M. Ganzinelli¹, M. Occhipinti¹, M. Brambilla¹, C. Proto¹, F. Trovò⁴, M.C.Garassino⁵, F.M. de Braud¹, G. Lo Russo¹, A. Pedrocchi⁴, P. Baili¹, V. Miskovic^{1,4}, F. Corso¹, A. Prelaj¹

1. Fondazione IRCCS Istituto Nazionale dei Tumori di Milano, Milan/IT , 2. Hospital Universitario de La Princesa, Madrid/ES , 3. Hospital Universitario Virgen de la Victoria, Málaga/ES , 4. Politecnico di Milano, Milan/IT , 5. University of Chicago, Chicago/IL/USA
Correspondence: alberto.ferrarin@istitutotumori.mi.it

INTRODUCTION

Reliable predictors of immunotherapy response in **metastatic NSCLC** remain limited. AI models trained on **clinical text** may help identify predictive factors, but whether their explanations align with **clinical reasoning** is unclear.

We evaluated the **interpretability** of an explainable AI system trained to predict immunotherapy response from clinical reports.

METHODS

Design · Apollo 11 (NCT05550961) 32 pz · mNSCLC · IO alone, any line · first evaluation → last report pre-IO **20,773 sentences** (notes, lab, records)

Annotation - each sentence scored **-2 to +2**:

Score	Meaning	Evidence
-2 / +2	Strong progression / response	Literature-backed predictors
-1 / +1	Mild progression / response	Clinically meaningful, limited literature
0	Neutral	-

→ **Sign** = **direction** · **magnitude** = **evidence**

AI score · continuous per sentence (+ → response, - → progression)

Agreement · AI score direction vs mean oncologist annotation · inter-physician reliability: **Cohen's κ**

Lab value analysis: lab report sentences analysed **separately**, using **reference ranges**

Stats · below/in/above-range vs in-range → two-sided **Mann–Whitney U** (min n/group) · effect size: **Cohen's d**

RESULTS

κ = 0.96

Inter-physician agreement

Cohen's kappa on 20,773 independently double-annotated sentences - validates the **ground truth** used to evaluate the AI.

Laboratory report sentences were analyzed **separately**: their regular format allowed reliable sentence segmentation and comparison against **objective reference ranges** (below/in/above-range) - an **independent benchmark**, not dependent on physician annotation.

64.6% / 54.2%

AI-human concordance (response- / progression-oriented sentences)

% of physician-flagged sentences where AI score direction **matched**. **Higher for response-oriented text** than progression-oriented.

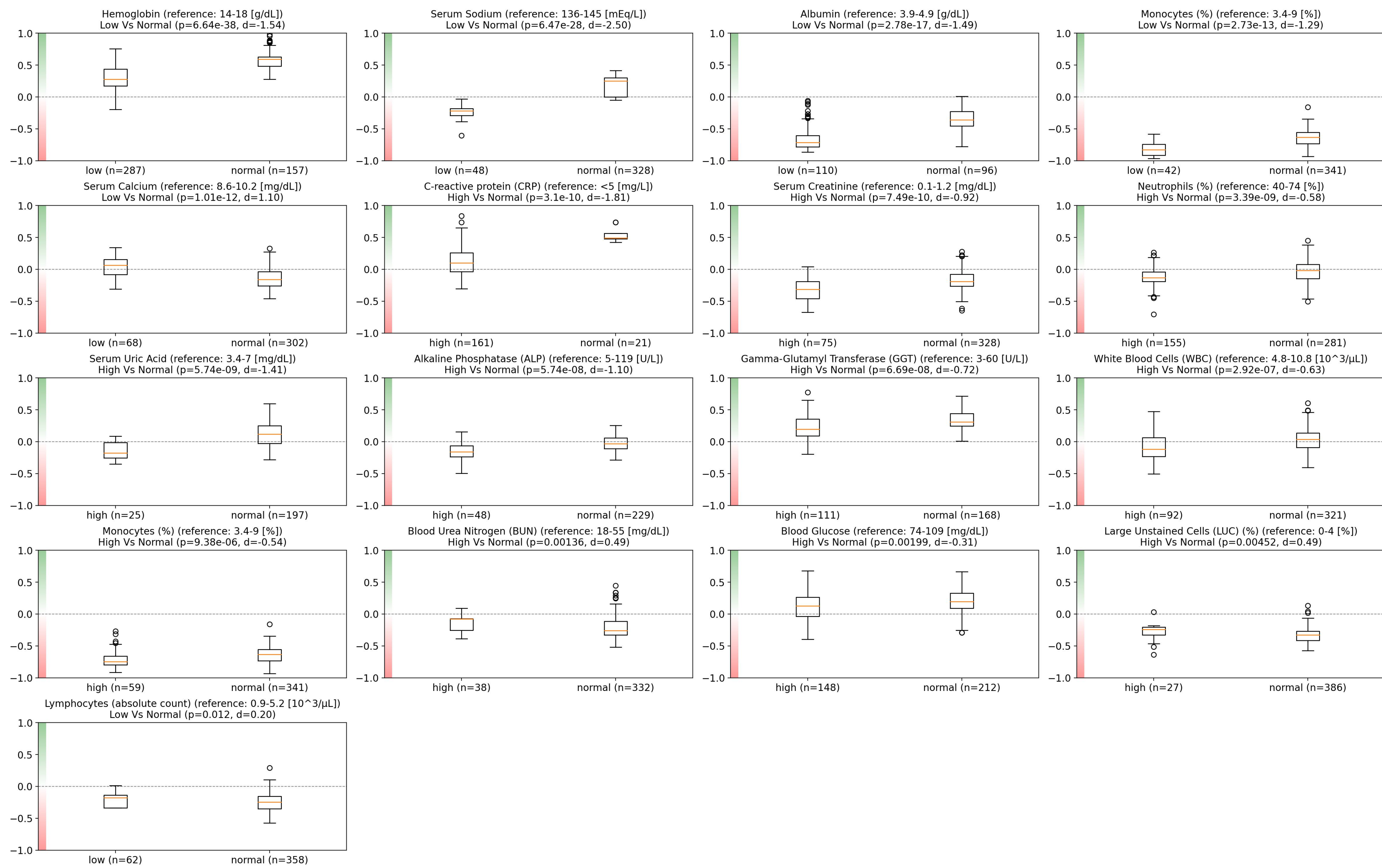


Figure 1. AI score by lab value range (below/above vs in-range). **Significant differences** in AI scores across lab categories (**p<0.05**), supporting **coherent behavior** on structured lab data.

CONCLUSION

- Limited free-text alignment, high coherence on structured labs** → Targeted improvement needed for narrative text.
- Digital Twin Platform**: Enables *in silico* trial simulation and virtual randomization (IO vs Chemo vs Chemo-IO) for personalized decision-making (*see Fig. 2*).

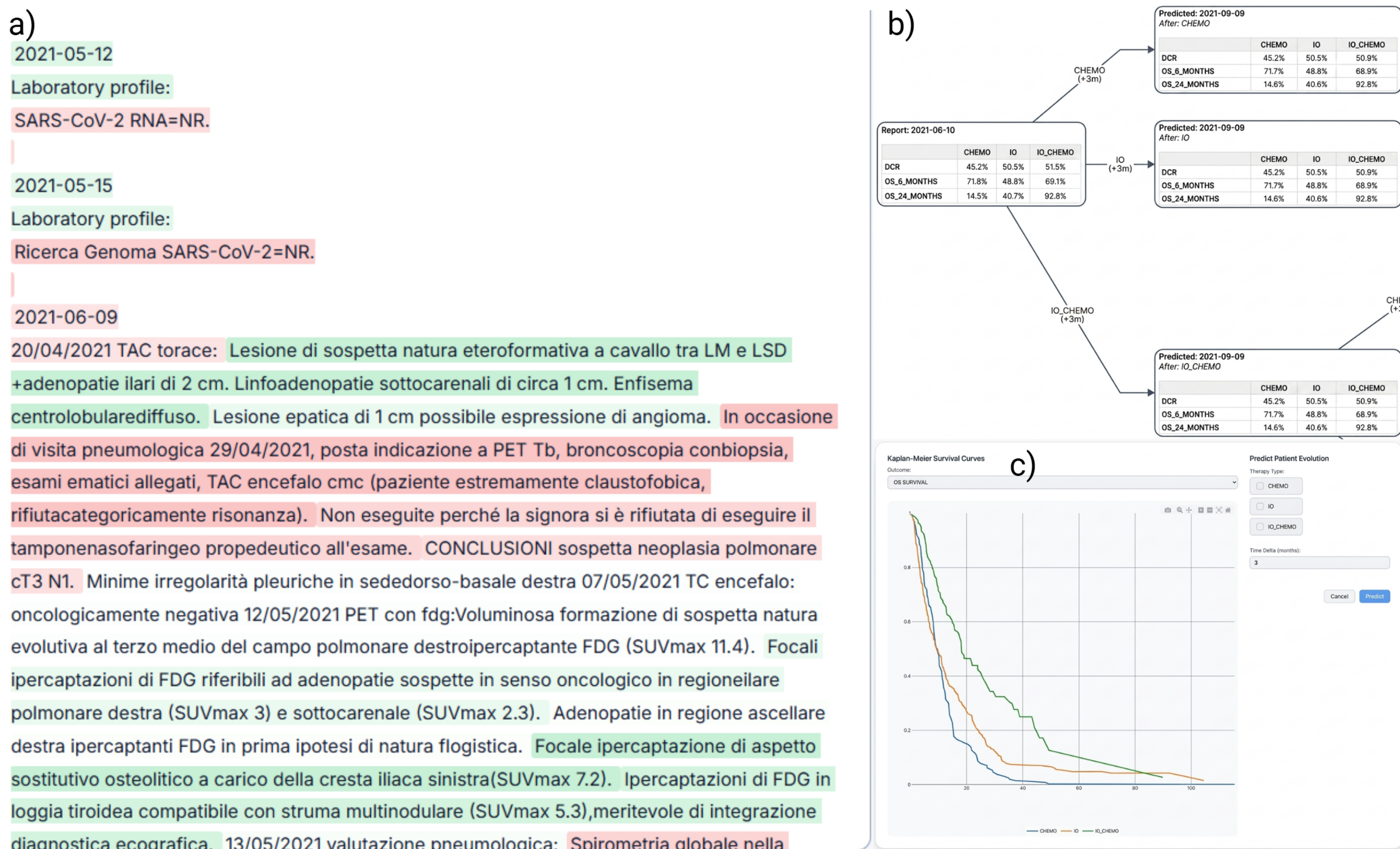


Figure 2. Digital Twin platform user interface. **a)** Sentence-level explainable AI attributions on the longitudinal patient timeline (green = response-oriented, red = progression-oriented). **b)** *In silico* trial simulation tree predicting treatment outcomes across therapy strategies. **c)** Kaplan-Meier overall survival curves for individual patient decision support

ACKNOWLEDGEMENTS



AI Use Disclosure

Claude (Anthropic) was used solely for English language polishing and text editing. The authors reviewed and edited all content and take full responsibility for the scientific accuracy of the presentation.



LET'S CONNECT!