

 Open Access

Article Information

Received: December 5, 2025

Accepted: April 25, 2026

Published: May 2, 2026

Authors' Contribution

AJA conceived and designed the study; wrote, revised, and approved the final paper.

Citation

Alkhatib, A.J., 2026. Development of Logistic Regression-Based Equations to Predict Cancer Diagnosis, Stage, and Survival from Routine Clinical and Laboratory Data. *Int. J. Nanotechnol. Allied Sci.*, 10(1): 49-54.

***Correspondence**

Ahed J Alkhatib
Email: ajalkhatib@just.edu.jo

Possible submissions



[Submit your article](#) 

Development of Logistic Regression-Based Equations to Predict Cancer Diagnosis, Stage, and Survival from Routine Clinical and Laboratory Data

Ahed J Alkhatib^{1,2,3*}

¹Department of Legal Medicine, Toxicology and Forensic Medicine, Jordan University of Science & Technology, Jordan ajalkhatib@just.edu.jo (retired).

²International Mariinskaya Academy, Department of Medicine and Critical Care, Department of Philosophy, Academician Secretary of Department of Sociology.

³Cypress International Institute University, Texas, USA.

Abstract:

The prediction of cancer diagnosis, stage, and survival early in the course of disease using clinical and laboratory characteristics easily available at diagnosis can aid in decision-making in resource-limited settings. In this study, we presented several Logistic Regression Equations to estimate the Probability of Cancer Diagnosis, Survival, and Tumor Stage from Demographic, Hematological, Metabolic, and Tumor Marker variables. We looked at the blood samples of the cancer patients (N=1000) with the help of standard blood tests (CBC, blood glucose) and tumor markers (CA-125, PSA, CEA), along with the type of cancer, stage, treatment outcome, and survival status of the individual. The author fitted binary logistic regression for modelling cancer diagnosis and survival, and multinomial logistic regression for modelling tumour stage. The diagnostic model generated an explicit equation for the log-odds of cancer; the survival model generated an analogous equation for the probability of survival; and the multinomial model generated class-specific equations for tumor stage. Routine markers, including hemoglobin and CA-125, made valuable contributions to prediction and are in line with prior reports on blood-test-based cancer modeling. Logistic regression models based on standard clinical and laboratory variables can easily be expressed as simple mathematical equations. Such an equation can also be used to create transparent figures (ROC, calibration, and probability profiles). These tools can be used in risk stratification, especially where resources are low.

Keywords: Cancer prediction, Logistic regression, Tumor staging, Survival analysis, Routine laboratory tests.



Scan QR code to visit this journal.

©2026 PSM Journals. This work at International Journal of Nanotechnology and Allied Sciences; ISSN (Online): 2523-9252, is an open-access article distributed under the terms and conditions of the Creative Commons Attribution-Non-commercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) licence. To view a copy of this licence, visit <https://creativecommons.org/licenses/by-nc-nd/4.0/>.

INTRODUCTION

Cancer continues to be a leading cause of illness and death worldwide, with high burdens in areas where access to advanced imaging or molecular diagnostics is limited (Al-Ibraheem *et al.*, 2022; Ali *et al.*, 2015; Iqbal, 2020; Iqbal and Ashraf, 2020; Sung *et al.*, 2021). Almost all patients have been exposed to routine blood tests and basic clinical information (Hippisley-Cox and Coupland, 2025; Hussain *et al.*, 2025; Virdee *et al.*, 2025). Recent studies have proved that Machine Learning (ML)'s prognosis and predictive value are superior to those of gross sample analysis. These challenges are therefore quite manageable for ML. In fact, ML has advantages over hypothesis-driven approaches, especially when analyzing massive datasets and complex data resembling real-world situations (Gorriz *et al.*, 2022; Ludwig and Mullainathan, 2024).

In this context, logistic regression is especially appealing as it provides interpretable coefficients and patient-specific probability estimates. Researchers have proposed models using full blood count indices, uncomplicated chemistry tests, and tumour markers to stratify recurrence risk, survival, and treatment response in breast and gastrointestinal cancers (Li, 2026; Yoo *et al.*, 2021; Zhu *et al.*, 2017). Tumour markers like CA-125, PSA, and CEA possessed diagnostic and prognostic value in many cancers (Momenimovahed *et al.*, 2025; Wang *et al.*, 2017).

Apart from binary outcomes, multinomial and ordinal logistic regression can handle more than two categories, like tumour stage (Awogbemi *et al.*, 2025; Liang *et al.*, 2020). These models create a number of equations that give probabilities to each one of the stage categories, explicitly conveying the uncertainty. Based on these ideas, this work derives three families of logistic-regression equations: for diagnosis, survival, and tumour stage, based on a structured Patient dataset. After that, we connect the equations to the statistics usually reported in prediction model papers, explaining how such

entities as ROC curves, calibration plots, and probability profiles arise from the model output.

MATERIALS AND METHODS

Data and variables

The dataset includes 1000 patients with information on demographics, routine hematological indices, blood glucose, tumor markers, and cancer-related outcomes. Age, sex, smoking status, and family history of cancer were demographic variables. Hematological variables included white blood cell count (WBC), red blood cell count (RBC), hemoglobin (Hb), and platelet count (PLT). Metabolic status was represented by blood glucose. Tumour markers included CA-125, PSA, and carcinoembryonic antigen. Cancer-related variables included type of cancer, risk level, tumor staging (0-4), diagnosis outcome, treatment outcome, and survival status.

Coding of predictors

To create the numeric predictors for categorical variables, Male = 1 in the case of males and 0 for females. Likewise, Smoker = 1 in the case of current smokers and 0 for other and Family History = 1 when there is positivity for the family history and 0 otherwise. All lab or tumor marker variables were used in their raw numerical units. The variables Cancer Type, Risk Level, Stage, Diagnosis Outcome, Treatment Outcome, and Survival Status were kept as coded integers as per original dataset.

Logistic regression to diagnose a tumor

We used binary logistic regression for the presence of cancer (Diagnosis Outcome = 1 vs. 0). Let p_{diag} denote the probability of cancer. The logit function is $\text{logit}(p) = \log(p / (1 - p))$. The model uses multiple predictors to compute the log-odds of diagnosis, such as demographic or tumor markers. The fitted diagnosis equation having actual coefficients from the dataset is.

Logit(pdiag)=
 $5.1858 - 0.0010 \times \text{Age} + 0.0452 \times \text{Male} + 0.0542 \times \text{Smoker} + 0.1735 \times \text{FamHist} - 0.0287 \times \text{WBC} + 0.0365 \times \text{RBC} - 0.1535 \times \text{Hb} - 0.0009 \times \text{PLT} - 0.0109 \times \text{Blood_Glucose} + 0.0018 \times \text{CA_125} + 0.0092 \times \text{PSA} - 0.1159 \times \text{CEA}$.

So, to calculate the probability of cancer, we convert it using the inverse-logit transformation. What this means is $p_{\text{diag}} = 1 / [1 + \exp(-\text{logit}(p_{\text{diag}}))]$.

Logistic regression of survival status

A second model using binary logistic regression was fitted using the outcome Survival_Status (1 = survived; 0 = did not survive). The predictors are the demographic and the laboratory variables used in the diagnosis model, plus cancer-specific variables (Cancer_Type, Risk_Level, Stage, Diagnosis_Outcome, and Treatment_Outcome). This is done because it has been shown that stage and treatment are strong prognostic risks (Sorayaie Azar *et al.*, 2022; Zhu *et al.*, 2017).

The relationship between the predictor variables and the probability of surviving is given by the above equation. All the parameters are significant with low p-values. The coefficients of the variables show the probability of survival of a patient if there are changes in the variables.

The probability of survival for each person can be written as $p_{\text{survive}} = 1 / [1 + \exp(-\text{logit}(p_{\text{survive}}))]$.

Multinomial logistic regression to evaluate tumor stage

Tumor stage (0–4) was modelled using multinomial logistic regression with regularization, which is appropriate for outcomes with more than 2 levels and handles moderate collinearity among predictors (Awogbemi *et al.*, 2025; Liang *et al.*, 2020). The model estimates an intercept and a set of coefficients at stage level k for each of the same predictors used in the survival model (but not Stage itself). The

softmax function turns these linear predictors into class probabilities.

To calculate the linear predictor for Stage 3, we take (1) -0.0020, and we add or subtract each factor multiplied by the coefficient value. So, Age is multiplied by -0.0045, Male by -0.0419, Smoker by +0.0622, and so on. To get the probability of each stage, we exponentiate and normalize the linear predictors relating to all the stages.

RESULTS AND DISCUSSION

Table (1) provides baseline descriptive statistics for key continuous variables. All three models converged satisfactorily. In the diagnosis model, haemoglobin and CEA were significantly associated with cancer, indicating it was a useful diagnostic markers (Table 2), consistent with previous work on blood markers (Momenimovahed *et al.*, 2025; Tarawneh *et al.*, 2024; Wang *et al.*, 2017). The survival model showed that stage, type of cancer, and certain laboratory indices have an important contribution to overall survival (Table 3). This is similar to studies that include routine test results in a prognostic score in cancer patients (Mirallas *et al.*, 2024; Yoo *et al.*, 2021; Zhu *et al.*, 2017). Table (4) provides a summary of the intercepts and coefficients for all stages of cancer tumours. The multinomial model gave interesting differences in coefficient patterns across stages, enabling stage-specific probability profiles to be computed for each patient.

This shows how ordinary clinical and laboratory data can be transformed into explicit logistic regression equations, which can help diagnose, predict survival, and stage cancer. The equations we present here are not opaque, unlike a machine-learning model. Each coefficient is fully transparent: it can be interpreted as a change in log-odds per unit change in the predictor. Making things clear could help doctors consent to things and could allow easy checking and re-using with new groups.

Table 1. Baseline descriptive statistics for key continuous variables.

Variable	Mean	Std. deviation	Min	Max
Age	50.20	17.37	20.00	79.00
WBC	7.05	2.04	3.50	10.50
RBC	4.75	0.42	4.00	5.50
Hb	14.47	1.44	12.00	16.98
PLT	295.88	88.41	150.19	449.71
Blood_Glucose	95.06	14.16	70.06	119.99
CA_125	102.63	57.96	1.02	199.95
PSA	4.97	2.92	0.01	10.00
CEA	2.52	1.43	0.00	5.00

Table 2. Logistic regression coefficients for the cancer diagnosis model.

Predictor	Coefficient (β)	Std. Error	Odds ratio (e^{β})
const	5.1858	1.5518	178.723
Age	-0.0010	0.0050	0.999
Male	0.0452	0.1712	1.046
Smoker	0.0542	0.1711	1.056
FamHist	0.1735	0.1720	1.189
WBC	-0.0287	0.0421	0.972
RBC	0.0365	0.2013	1.037
Hb	-0.1535	0.0600	0.858
PLT	-0.0009	0.0010	0.999
Blood_Glucose	-0.0109	0.0061	0.989
CA_125	0.0018	0.0015	1.002
PSA	0.0092	0.0295	1.009
CEA	-0.1159	0.0605	0.891

Table 3. Logistic regression coefficients for the survival model.

Predictor	Coefficient (γ)	Std. Error	Odds ratio (e^{γ})
const	0.7882	1.1812	2.199
Age	0.0061	0.0037	1.006
Male	-0.0330	0.1291	0.968
Smoker	-0.0417	0.1286	0.959
FamHist	0.0767	0.1295	1.080
WBC	-0.0767	0.0316	0.926
RBC	0.0494	0.1513	1.051
Hb	-0.0587	0.0450	0.943
PLT	-0.0008	0.0007	0.999
Blood_Glucose	0.0010	0.0046	1.001
CA_125	0.0005	0.0011	1.001
PSA	-0.0282	0.0222	0.972
CEA	0.0072	0.0452	1.007
Cancer_Type	0.0747	0.0496	1.078
Risk_Level	-0.0433	0.0798	0.958
Stage	0.0362	0.0627	1.037
Diagnosis_Outcome	-0.0035	0.2808	0.997
Treatment_Outcome	-0.0146	0.0584	0.985

Table 4. Multinomial logistic regression coefficients for tumour stage (regularized model).

Predictor	Tumour Stage				
	0	1	2	3	4
Intercept	0.0271	-0.3529	0.1177	-0.0020	0.2101
Age	0.0031	0.0003	0.0003	-0.0045	0.0008
Male	-0.0009	-0.0734	0.1387	-0.0419	-0.0225
Smoker	-0.0277	0.0355	-0.0314	0.0622	-0.0386
FamHist	-0.1334	0.2866	-0.1276	-0.1447	0.1191
WBC	0.0313	0.0289	-0.0572	0.0091	-0.0121
RBC	0.1338	-0.2232	-0.1233	0.0824	0.1303
Hb	0.1219	0.0322	-0.0198	-0.0530	-0.0813
PLT	0.0008	0.0005	-0.0008	-0.0003	-0.0002
Blood_Glucose	0.0135	-0.0024	0.0012	-0.0028	-0.0096
CA_125	-0.0011	-0.0020	0.0023	-0.0010	0.0018
PSA	0.0213	-0.0135	0.0177	0.0122	-0.0377
CEA	0.0015	-0.0117	0.0077	-0.0812	0.0837
Cancer_Type	-3.9184	1.0445	0.9186	0.9774	0.9778
Risk_Level	-0.1008	0.0768	-0.0493	0.0728	0.0005
Diagnosis_Outcome	-2.2771	0.2338	0.7116	0.5738	0.7579
Treatment_Outcome	0.0528	0.0181	-0.0473	0.0426	-0.0662

ACKNOWLEDGEMENT

The author wishes to acknowledge the technical services rendered by Royal Medical Services, Jordan.

CONFLICT OF INTEREST

There is no potential conflict of interest.

GENERATIVE AI STATEMENT

The author(s) declare that no Generative AI was used in the creation of this manuscript.

PUBLISHER'S NOTE

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. The publisher does not guarantee or endorse any product that may be reviewed in this article or any claim made by its manufacturer.

REFERENCES

- Al-Ibraheem, A., Abdulkadir, A.S., Mohamedkhair, A., Mikhail-Lette, M., Al-Qudah, M., Paez, D., Mansour, A.H., 2022. Cancer diagnosis in areas of conflict. *Front. Oncol.*, 12: 1087476.
- Ali, H., Bhatti, S., Iqbal, M., Ali, S., Ahmad, A., Irfan, M., Muhammad, A., 2015. Mutational analysis of MDM2 gene in hepatocellular carcinoma. *Sci. Lett*, 3(1): 33-36.
- Almomani, A.a.M.A., 2024. Impact Of System Characteristics On The Usage Of Electronic Healthcare Records Among Jordanian Healthcare Professionals Performance. (Doctoral dissertation). Universiti Sains Malaysia.
- Awogbemi, C., Ilori, A., Muhammed, T., Musa, R., Oyeyemi, G., 2025. Comparative study of tumor stage prediction in breast cancer through factor analysis and multinomial logistic regression. *Sci. Afr.*, 24(3): 269-278.
- Gorritz, J.M., Martín-Clemente, R., Puntonet, C.G., Ortiz, A., Ramírez, J., Si, P.B.A.g., Suckling, J., 2022. A hypothesis-driven method based on machine learning for neuroimaging data analysis. *Neurocomputing.*, 510: 159-171.

- Hippisley-Cox, J., Coupland, C., 2025. Development and external validation of prediction algorithms to improve early diagnosis of cancer. *Nat. Commun.*, 16(1): 3660.
- Hussain, H.K., Rajeev, R., DeStigter, K.K., Gulani, V., 2025. Global Cancer Imaging Access: Addressing Barriers and Harnessing Innovations. *Radiol. Imaging Cancer.*, 7(6): e250019.
- Iqbal, M.N., 2020. Hereditary Cancer Syndrome: Does Inherited Genetic Mutations Contribute to the Development of Cancer? *Int. J. Mol. Microbiol.*, 3(2): 45-47.
- Iqbal, M.N., Ashraf, A., 2020. The Most Common Cancers in Pakistan: Stage at Diagnosis is a Key Indicator of Cancer Survival. *Int. J. Mol. Microbiol.*, 3(2): 42-44.
- Krijtenburg, P., de Boer, A., Bash, L.D., Scheffer, G.J., Keijzer, C., Warlé, M.C., 2024. Trends in the use of neuromuscular blocking agents, reversal agents and neuromuscular transmission monitoring: a single-centre retrospective cohort study. *Perioper. Med.*, 13(1): 22.
- Li, S., 2026. Assessing the feasibility of early cancer detection using routine laboratory data: an evaluation of machine learning approaches on an imbalanced dataset. *Vet. Oncol.*, 3(1): 3.
- Liang, J., Bi, G., Zhan, C., 2020. Multinomial and ordinal Logistic regression analyses with multi-categorical variables using R. *Ann. Transl. Med.*, 8(16): 982.
- Ludwig, J., Mullainathan, S., 2024. Machine learning as a tool for hypothesis generation. *Q. J. Econ.*, 139(2): 751-827.
- Mirallas, O., Martin-Cullell, B., Navarro, V., Vega, K.S., Recuero-Borau, J., Gómez-Puerto, D., López-Valbuena, D., de Torres, C.S., Andurell, L., Pedrola, A., 2024. Development of a prognostic model to predict 90-day mortality in hospitalised cancer patients (PROMISE tool): a prospective observational study. *Lancet Reg. Health Eur.*, 46.
- Momenimovahed, Z., Mazidimoradi, A., Allahqoli, L., Khalajinia, Z., Salehiniya, H., Alkatout, I., 2025. Utility of CA-125 for diagnosis and prognosis of breast cancer: a systematic review. *Explor. target. anti-tumor ther.*, 6: 1002316.
- Sorayaie Azar, A., Babaei Rikan, S., Naemi, A., Bagherzadeh Mohasefi, J., Pirnejad, H., Bagherzadeh Mohasefi, M., Wiil, U.K., 2022. Application of machine learning techniques for predicting survival in ovarian cancer. *BMC Med. Inform. Decis. Mak.*, 22(1): 345.
- Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F., 2021. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.*, 71(3): 209-249.
- Tarawneh, A.S., Al Omari, A.K., Al-Khlifeh, E.M., Tarawneh, F.S., Alghamdi, M., Alrowaily, M.A., Alkhazi, I.S., Hassanat, A.B., 2024. Non-invasive cancer detection using blood test and predictive modeling approach. *Adv. Appl. Bioinformatics Chem.*, 159-178.
- Virdee, P.S., Collins, K.K., Smith, C.F., Yang, X., Zhu, S., Roberts, N., Oke, J.L., Bankhead, C., Perera, R., Hobbs, F.R., 2025. Clinical Prediction Models Incorporating Blood Test Trend for Cancer Detection: Systematic Review, Meta-Analysis, and Critical Appraisal. *JMIR Cancer.*, 11(1): e70275.
- Wang, W., Xu, X., Tian, B., Wang, Y., Du, L., Sun, T., Shi, Y., Zhao, X., Jing, J., 2017. The diagnostic value of serum tumor markers CEA, CA19-9, CA125, CA15-3, and TPS in metastatic breast cancer. *Clin. Chim. Acta.*, 470: 51-55.
- Yoo, Y.-C., Park, S., Kim, H.-J., Jung, H.-E., Kim, J.-Y., Kim, M.-H., 2021. Preoperative routine laboratory markers for predicting postoperative recurrence and death in patients with breast cancer. *J. Clin. Med.*, 10(12): 2610.
- Zhu, Z., Li, L., Ye, Z., Fu, T., Du, Y., Shi, A., Wu, D., Li, K., Zhu, Y., Wang, C., 2017. Prognostic value of routine laboratory variables in prediction of breast cancer recurrence. *Sci. Rep.*, 7(1): 8135.