**Supplementary Information**

**I. ‘FAIRness’ of implementation**

Table S1: Assessment of implementation according to FAIR Guiding Principles.

|  |  |  |
| --- | --- | --- |
| **The FAIR Guiding Principles** | | **20K Challenge manuscript implementation** |
| To be Findable | |  |
| F1. | (meta)data are assigned a globally unique and persistent identifier | (Meta)data are assigned a Uniform Resource Identifier (URI) (i.e. stable URLs). |
| F2. | data are described with rich metadata (defined by R1 below) | All data are described with rich metadata according to public ontologies (e.g., NCI thesaurus (NCIt), Radiation Oncology Ontology (ROO)) which is machine-readable (RDF). |
| F3. | metadata clearly and explicitly include the identifier of the data it describes | All metadata includes the identifier of the data it describes via *subject-predicate-object* triples (RDF). |
| F4. | (meta)data are registered or indexed in a searchable resource | (Meta)data are registered in local namespaces that are searchable via published SPARQL queries [1]. |
| To be Accessible | |  |
| A1. | (meta)data are retrievable by their identifier using a standardized communications protocol | (Meta)data are stored according to the RDF protocol and retrievable by their identifier using the standardized SPARQL protocol and the queries provided. |
| A1.1 | the protocol is open, free, and universally implementable | The RDF and SPARQL protocols are open and free. |
| A1.2 | the protocol allows for an authentication and authorization procedure, where necessary | The data is stored locally and only available for locally executed algorithms due to the sensitivity of the data. The algorithms are signed by the researchers and verified based on their signature prior to their deployment at the local data sites. |
| A2. | metadata are accessible, even when the data are no longer available | Metadata is stored similarly as the data (RDF) and available when data is no longer available. |
| To be Interoperable | |  |
| I1. | (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation | (Meta)data is mapped to RDF using public ontologies and shares the same representation. |
| I2. | (meta)data use vocabularies that follow FAIR principles | The open ontologies (NCIt, ROO) adhere to FAIR. |
| I3. | (meta)data include qualified references to other (meta)data | The ontologies used (NCIt, ROO) have many crosslinks to other ontologies. |
| To be Reusable | |  |
| R1. | meta(data) are richly described with a plurality of accurate and relevant attributes | (Meta)data are described according to the public ontologies. |
| R1.1. | (meta)data are released with a clear and accessible data usage license | Due to sensitivity of the data only the metadata is publicly available. The data are not public and only accessible by the respective data owners. Algorithms can be deployed after approval by the respective data owners. |
| R1.2. | (meta)data are associated with detailed provenance | (Meta)data are assigned a Uniform Resource Identifier (URI). |
| R1.3. | (meta)data meet domain-relevant community standards | The ROO is the most comprehensive ontology in radiation oncology and the NCIt is the most comprehensive in the cancer domain. |

**II. Description of ADMM**

ADMM decomposes the optimization problem underlying logistic regression (finding regression coefficients that maximize the log-likelihood of all training data) into an iterative optimization: each site computes regression coefficients that optimize a trade-off between maximizing the log-likelihood for the site’s local data and a degree of agreement with the network consensus (a combination of the regression coefficients determined at all sites). This trade-off includes a penalty for disagreeing with this consensus. At the master, the sets of site-specific regression coefficients are combined to a new consensus and a new disagreement penalty value is determined. This consensus and the new penalty are then returned to each site to again optimize site-specific coefficients (the trade-off between maximizing log-likelihood and agreement with consensus changes because of the new consensus and disagreement penalty). This iterative procedure is repeated until the discrepancy between the sites’ local coefficients and the consensus, as well as the change in the consensus solutions over iterations is sufficiently small.

**III. Missing data imputation**

First, the missing values are logically induced from the permitted combinations of T, N, M, and overall stages. For example, a patient diagnosed in 2011 with N0M0 and overall stage IIA but missing T can only have T2b according to TNM edition 7.

If the logical imputation is ambiguous because multiple imputation results are possible, the missing values are imputed probabilistically based on a subset. This subset contains patients that are:

- treated at the same site,  
- from the training cohort,  
- within the time interval corresponding to the selected AJCC TNM edition (see below),  
- matching the available variables of the patient.

Note that this subset also contains patients for which missing values have been logically imputed so that probabilistic imputation is also feasible for sites E and H which miss some variable for all patients. The empirical probability of each T, N, M, and overall stage combination observed in this patient subset is computed and one of these combinations is randomly sampled according to the computed empirical probabilities. For example, a patient diagnosed in 2013 with T1aN0 and overall stage IV but missing M can be imputed with M1a or M1b according to TNM edition 7. If there are 30 patients with T1aN0M1a & overall stage IV and 70 patients with T1aN0M1b & overall stage IV diagnosed starting 2010 and before 2018, the missing M value is imputed by 1a with probability 0.3 and by 1b with probability 0.7. This probabilistic imputation procedure assumes variables to be missing at random which is a simplifying assumption in routine clinical care data.

To decide on an AJCC TNM edition for a given patient in probabilistic imputation, the most recent TNM edition meeting two criteria is selected:

* it was effective before or in the patient’s year of diagnosis,
* it yields a complete imputation.

The modeling choice to also use preceding editions takes into account the possibility that the treating physician has not yet adopted the newest AJCC TNM cancer staging edition.

The following official effective dates for AJCC TNM cancer staging editions are used [2]:

* Edition 1: 1978 – 1983
* Edition 2: 1984 – 1988
* Edition 3: 1989 – 1992
* Edition 4: 1993 – 1997
* Edition 5: 1998 – 2002
* Edition 6: 2003 – 2009
* Edition 7: 2010 – 2017
* Edition 8: 2018 – present

**IV. Discussion of published survival prediction models**

Published NSCLC two-year survival prediction models report AUCs of, for example, 0.68-0.77 [3,4]. Comparing the presented model’s performance with published models is difficult for multiple reasons:

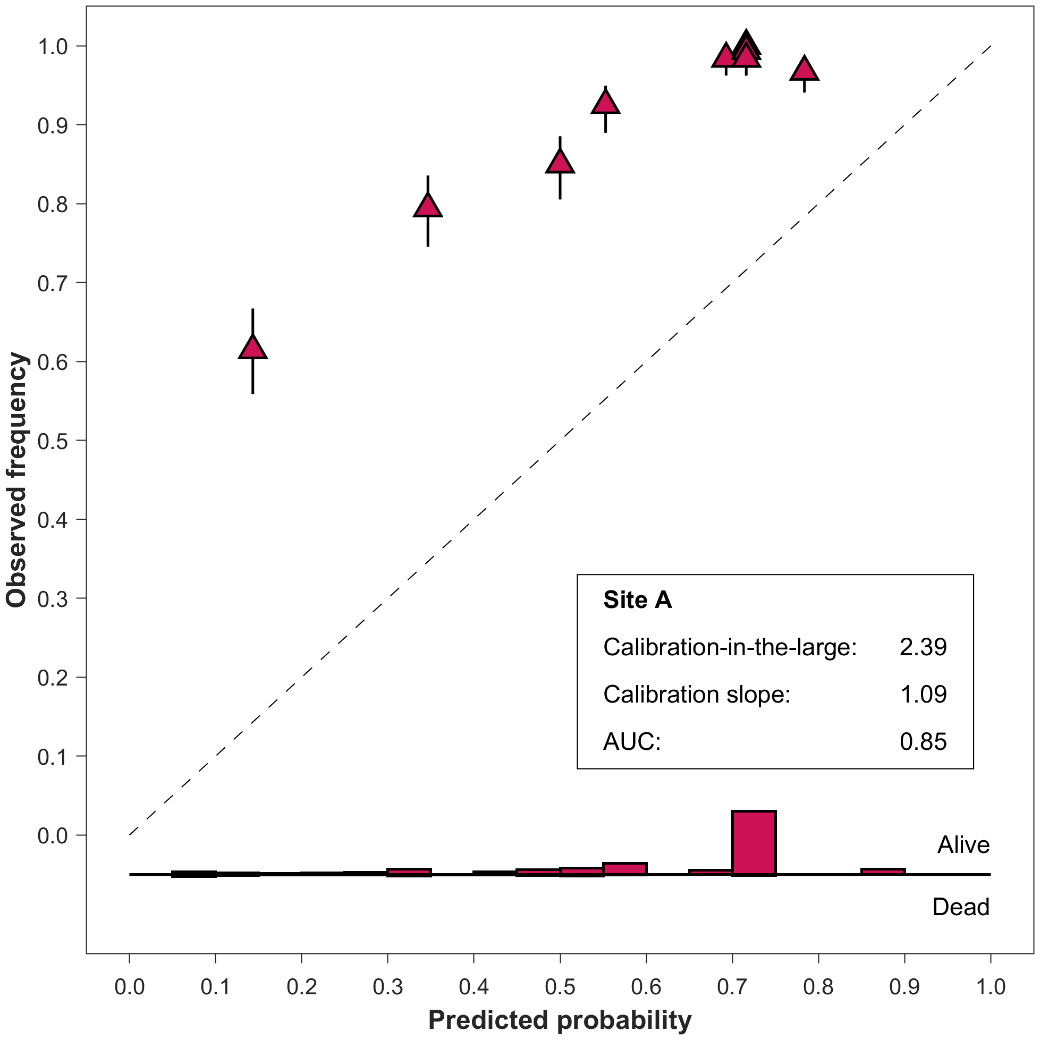
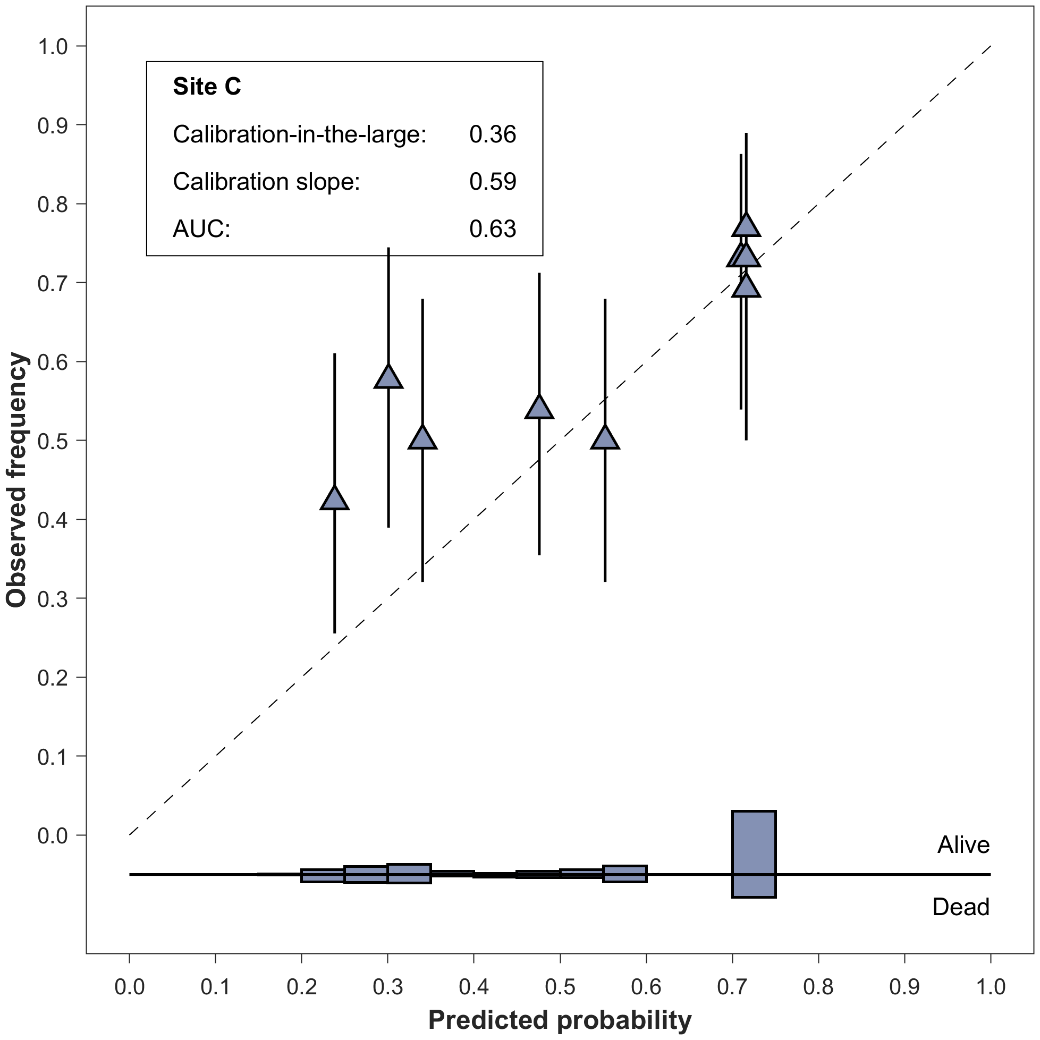
* inclusion criteria: patient inclusion is restricted to treatment with curative intent [4] or different treatment techniques [5],
* methodology: Cox regression models [6-8] or early mortality [9] predictions are not directly comparable to two-year survival predictions,
* performance estimates: sizes of validation cohorts vary across studies, causing different degrees of variability in the performance estimates, therefore rendering comparison unreliable.

The presented model is trained and validated on patients exhibiting all NSCLC stages, including stage IV patients who are generally not treated with curative intent, have the worst prognosis, and are least likely to survive two years after diagnosis (the two-year survival probability is approximately 10% according to the seventh edition AJCC TNM cancer staging manual [10]). Their bad prognosis is easily predicted but published studies mostly do not include stage IV NSCLC patients. Therefore, the presented model’s estimated two-year survival prediction performance is expected to be higher than for published models.

**V. Figures and tables**

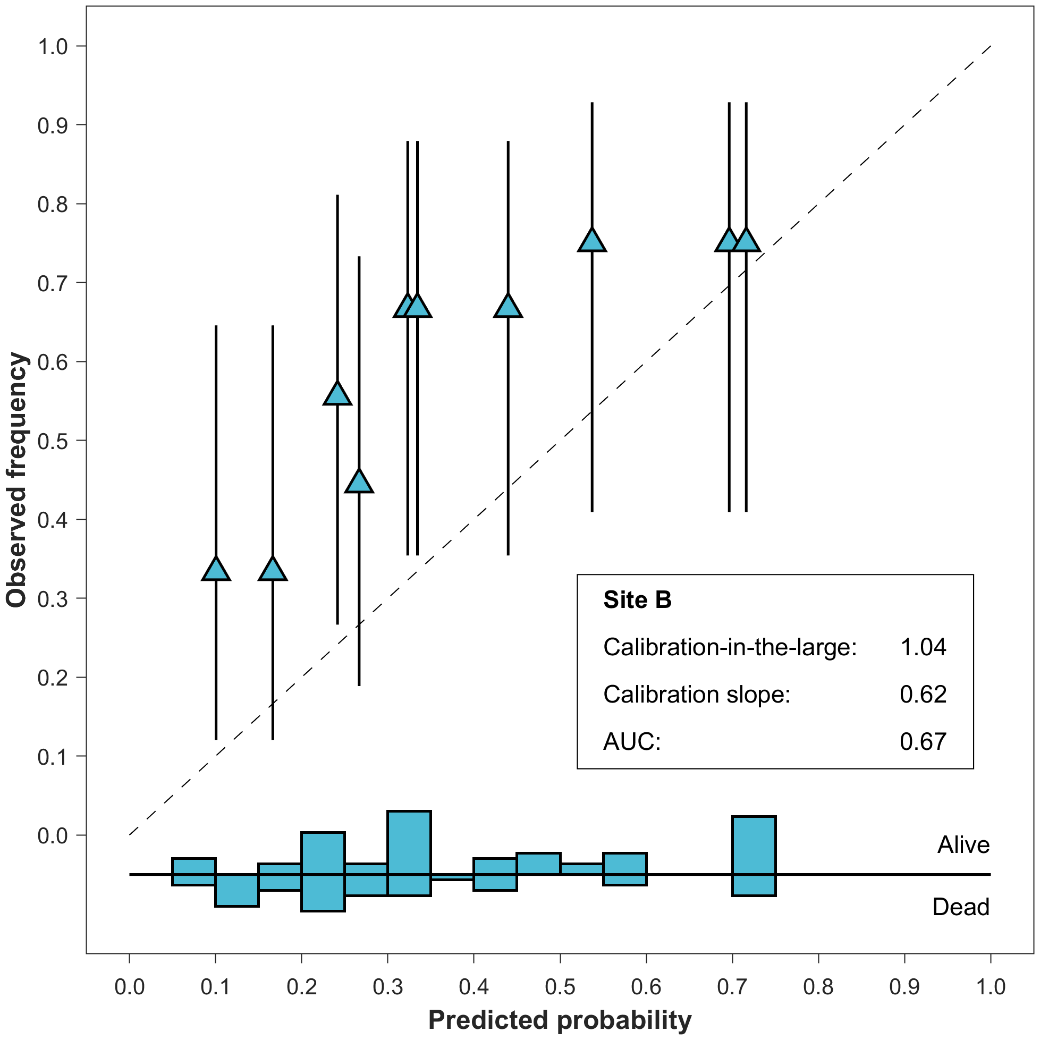
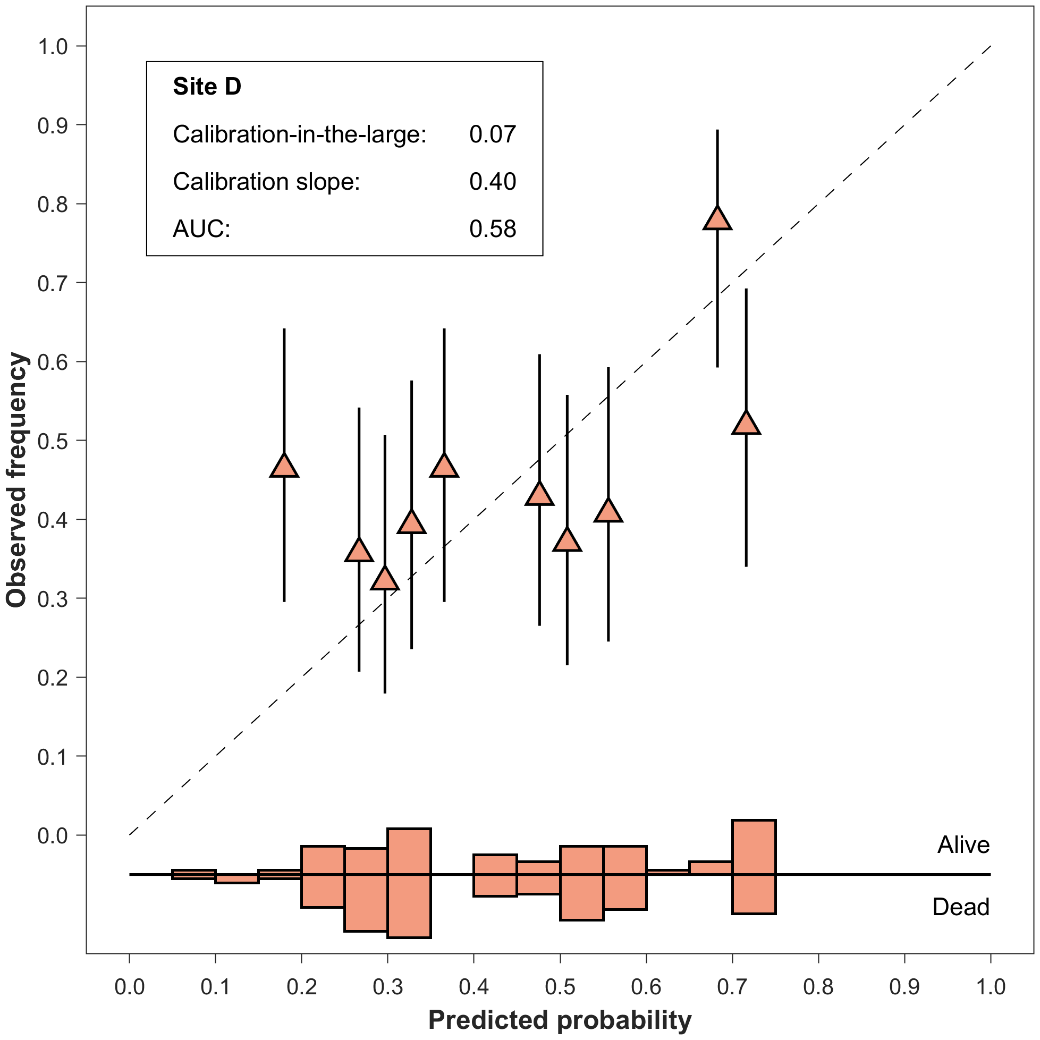
(a)

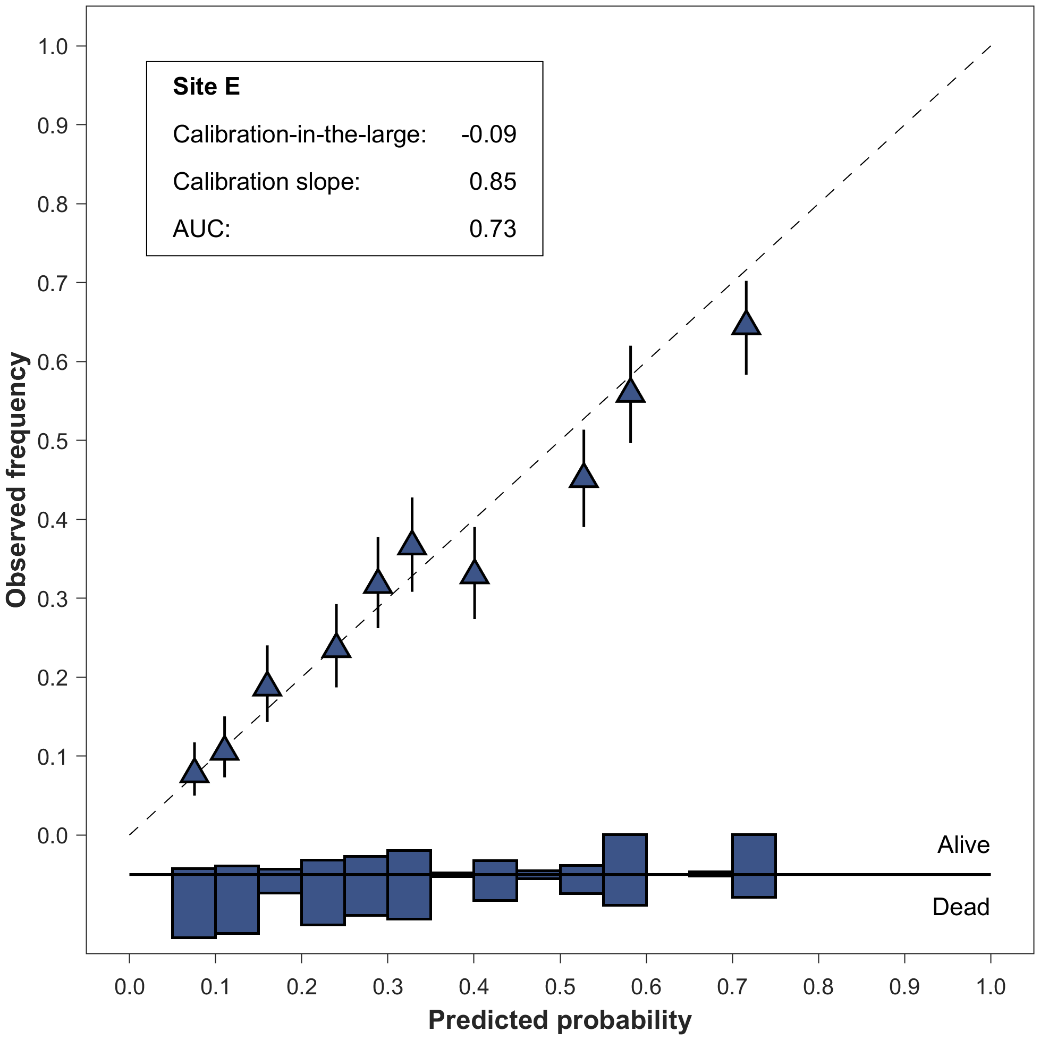
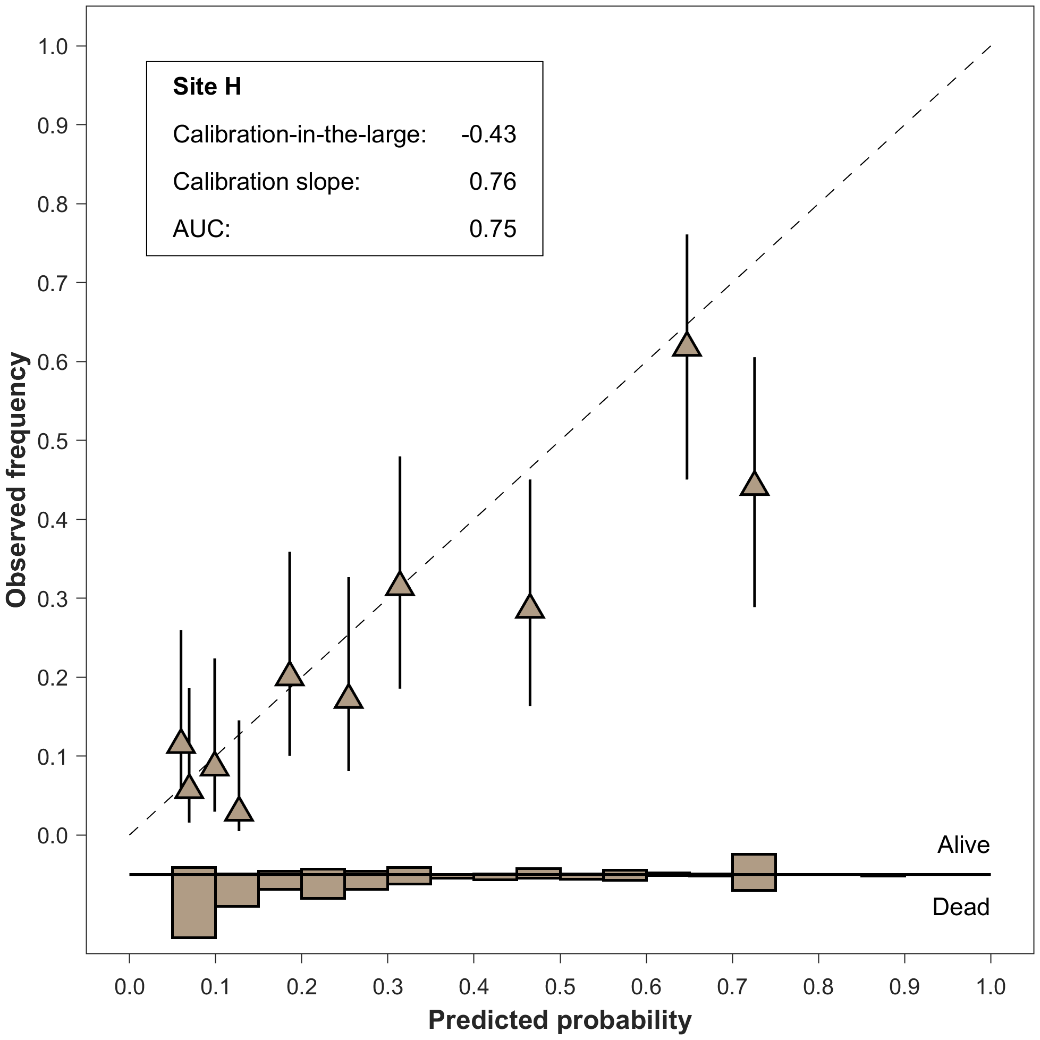
(c)

(d)

(b)

(f)

(g)

(e)

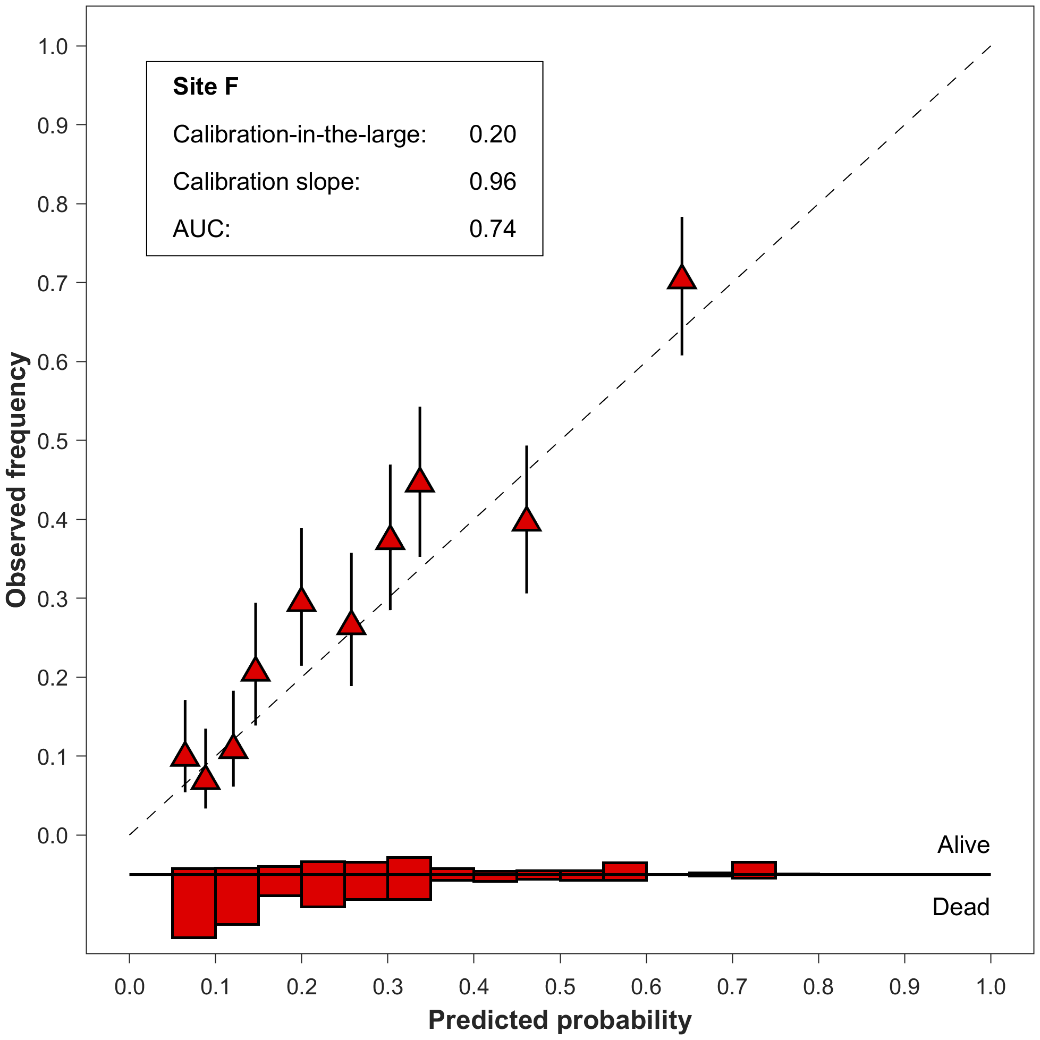


Figure S1: Calibration plots of the validation data for all sites excluding site G (Figure 5d). AUC: area under the receiver operating characteristic curve.

Table S2: Patient counts with stages IA, IB, IIA, IIB, IIIA, IIIB, IV in the validation cohort and corresponding model performance per site for the presented model and the AJCC TNM cancer staging edition 7 survival probabilities [10]. Survival probabilities for stage 0 and Occult are not available in the reference. The corresponding patients were thus excluded. Patients not staged according to edition 7 in the validation cohort were also excluded. AUC: area under the receiver operating characteristic curve. CI: confidence interval.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Site** | **Validation cohort patient counts (stage IA, IB, IIA, IIB, IIIA, IIIB, IV)** | **Model performance** | | | | |
| **Logistic regression** | | **AJCC edition 7** | |  |
| **AUC** | **95%-CI** | **AUC** | **95%-CI** | **Δ AUC** |
| Site A | 2803 | 0.87 | [0.84, 0.89] | 0.86 | [0.84, 0.89] | 0.00 |
| Site B | 87 | 0.67 | [0.54, 0.78] | 0.69 | [0.56, 0.80] | -0.02 |
| Site C | 131 | 0.54 | [0.43, 0.64] | 0.52 | [0.42, 0.61] | 0.02 |
| Site D | 273 | 0.59 | [0.52, 0.66] | 0.59 | [0.53, 0.65] | 0.00 |
| Site E | 2455 | 0.73 | [0.70, 0.74] | 0.71 | [0.69, 0.73] | 0.01 |
| Site F | 939 | 0.73 | [0.69, 0.76] | 0.72 | [0.68, 0.75] | 0.01 |
| Site G | 878 | 0.71 | [0.67, 0.75] | 0.71 | [0.66, 0.74] | 0.01 |
| Site H | 341 | 0.76 | [0.69, 0.82] | 0.77 | [0.69, 0.81] | -0.01 |
| **Total** | 7907 |  |  |  |  |  |

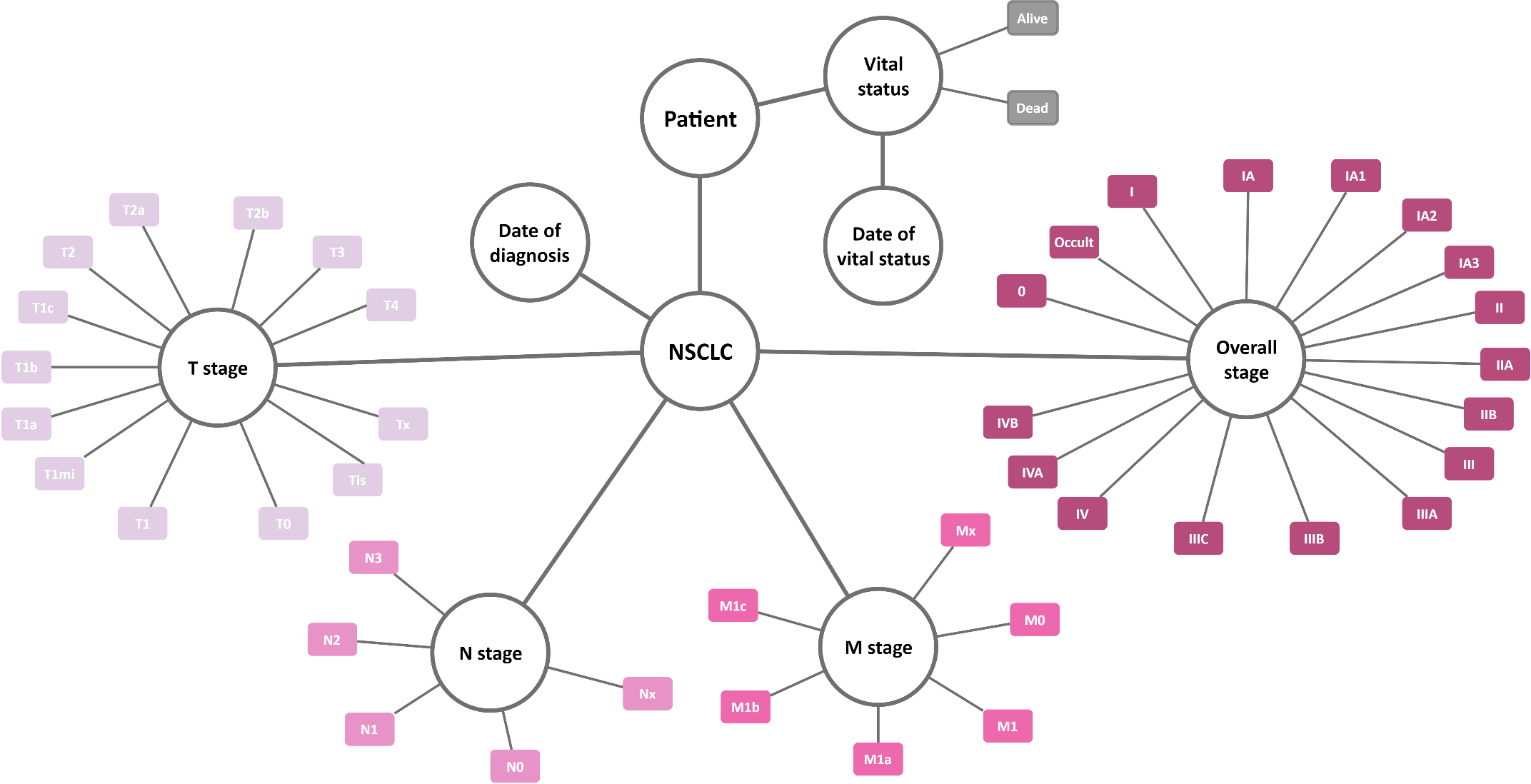
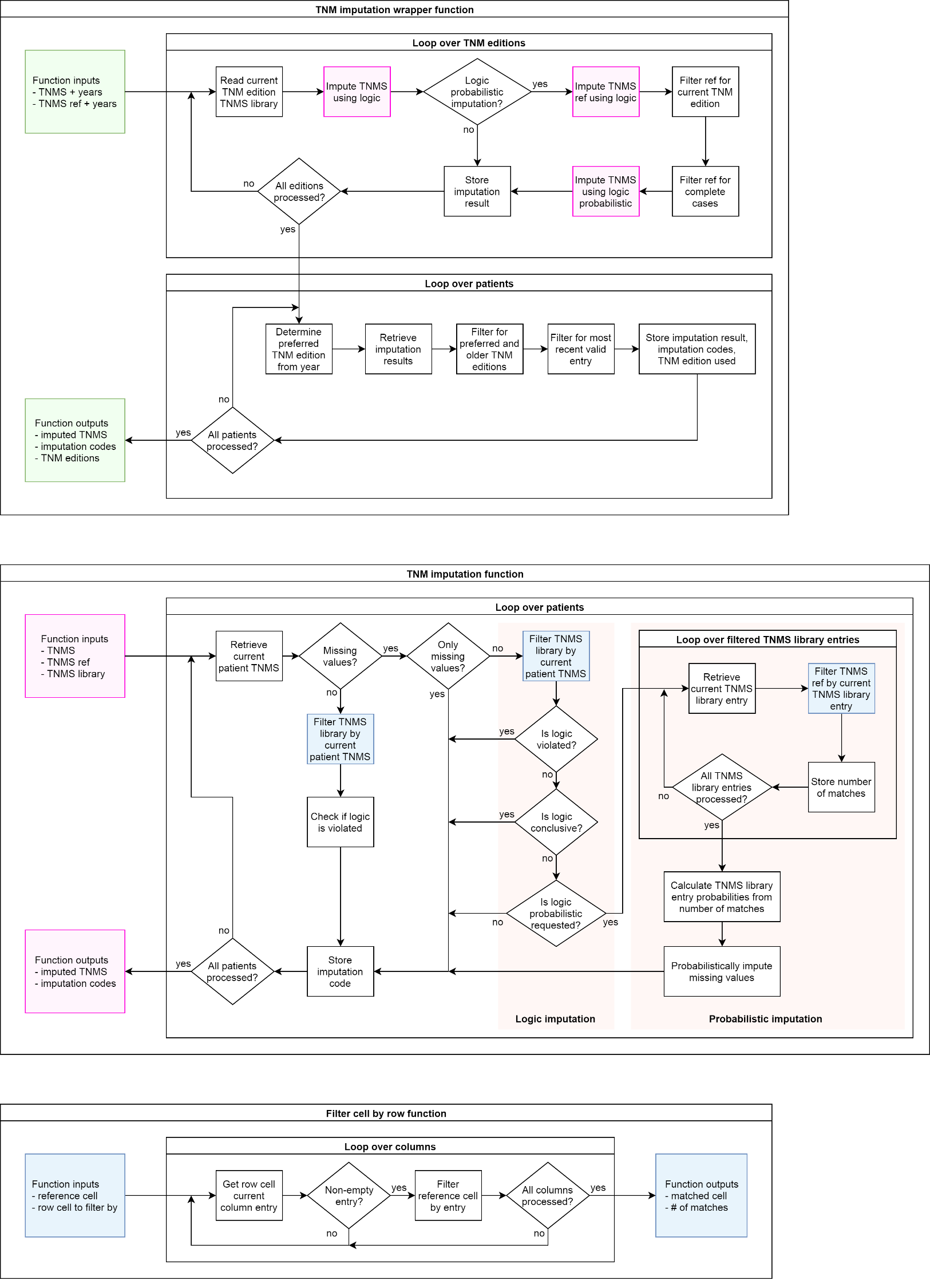


Figure S2: A graphical representation of the data model employed in the distributed learning network.



(a)

(c)

(b)

Figure S3: Imputation process. The TNM imputation wrapper function (a) is the outermost function which uses the TNM imputation function (b) and the Filter cell by row function (c) as subfunctions. The wrapper function has two input groups: data for the patients that are to be imputed and data for patients that act as the reference for probabilistic imputation. For both input groups, T, N, M, overall stage, and diagnosis year per patient are needed. TNM: cancer staging system based on tumor size (T), lymph node involvement (N) and metastasis (M). TNM edition: one of eight released TNM cancer staging system editions effective since 1978 and in non-overlapping time periods. TNMS: combination of TNM and cancer stage (S) for a patient (can contain missing values) or in the TNMS library (complete cases). Years: year of diagnosis corresponding with time of TNM staging, and used to determine the currently effective TNM edition. TNMS ref: reference patient TNMS combinations to be used for logic probabilistic imputation. TNMS library: library of valid combinations of TNM and cancer stages according to a specific TNM edition. Logic imputation: imputation of a missing TNMS value according to a single conclusive combination in the TNMS library. Logic probabilistic imputation: imputation of a missing TNMS value according to multiple inconclusive combinations in the TNMS library and their respective probabilities of occurrence in the TNMS reference cell. Imputation code: patient specific codes to indicate if the TNMS entries follow the TNM edition logic and the type of imputation performed (if any).

**References**

[1] Deist TM, Dankers FJWM, Ojha P, Scott Marshall M, Janssen T, Faivre-Finn C, et al. Code for: Distributed learning on 20 000+ lung cancer patients – The Personal Health Train. 2018. https://github.com/RadiationOncologyOntology/20kChallenge (accessed September 12, 2018).

[2] Editions of the AJCC Cancer Staging Manual n.d. http://cancerstaging.org/references-tools/deskreferences/Pages/default.aspx (accessed September 12, 2018).

[3] Jayasurya K, Fung G, Yu S, Dehing-Oberije C, De Ruysscher D, Hope A, et al. Comparison of Bayesian network and support vector machine models for two-year survival prediction in lung cancer patients treated with radiotherapy. Med Phys 2010;37:1401–7. doi:10.1118/1.3352709.

[4] Dehing-Oberije C, Yu S, De Ruysscher D, Meersschout S, Van Beek K, Lievens Y, et al. Development and external validation of prognostic model for 2-year survival of non-small-cell lung cancer patients treated with chemoradiotherapy. Int J Radiat Oncol Biol Phys 2009;74:355–62. doi:10.1016/j.ijrobp.2008.08.052.

[5] Mao Q, Xia W, Dong G, Chen S, Wang A, Jin G, et al. A nomogram to predict the survival of stage IIIA-N2 non-small cell lung cancer after surgery. J Thorac Cardiovasc Surg 2018;155:1784-1792.e3. doi:10.1016/j.jtcvs.2017.11.098.

[6] Carvalho S, Troost EGC, Bons J, Menheere P, Lambin P, Oberije C. Prognostic value of blood-biomarkers related to hypoxia, inflammation, immune response and tumour load in non-small cell lung cancer – A survival model with external validation. Radiotherapy and Oncology 2016;119:487–94. doi:10.1016/j.radonc.2016.04.024.

[7] Oberije C, De Ruysscher D, Houben R, van de Heuvel M, Uyterlinde W, Deasy JO, et al. A Validated Prediction Model for Overall Survival From Stage III Non-Small Cell Lung Cancer: Toward Survival Prediction for Individual Patients. Int J Radiat Oncol Biol Phys 2015;92:935–44. doi:10.1016/j.ijrobp.2015.02.048.

[8] Chaddad A, Desrosiers C, Toews M, Abdulkarim B. Predicting survival time of lung cancer patients using radiomic analysis. Oncotarget 2017;8:104393–407. doi:10.18632/oncotarget.22251.

[9] Jochems A, El-Naqa I, Kessler M, Mayo CS, Jolly S, Matuszak M, et al. A prediction model for early death in non-small cell lung cancer patients following curative-intent chemoradiotherapy. Acta Oncol 2018;57:226–30. doi:10.1080/0284186X.2017.1385842.

[10] Edge SB, American Joint Committee on Cancer, editors. AJCC cancer staging manual. 7th ed. New York: Springer; 2010.