Evaluation of total HCC lifespan, including both clinically evident and pre-clinical development, using combined Network Phenotyping Strategy and Fisher information analysis.

Petr Pančoška[a,c], Lubomír Skála[b], Jaroslav Nešetřil[c], Brian I. Carr[d]

[a] Department of Medicine and Center for Craniofacial and Dental Genetics, University of Pittsburgh, Pittsburgh, PA, USA
[b] Department of Chemical Physics and Optics, Faculty of Mathematics and Physics, Charles University Prague, Czech Republic
[c] Computer Science Institute (IUUK) of Charles University Prague, Czech Republic
[d] IRCCS de Bellis National Medical Center, Castellana Grotte, Italy

Grant support: ERZ-CZ LL1201 (CORES) to P.P. and J.N. and NIH CA 82723 (BIC)

ABSTRACT: ***We previously showed that for prognostication of hepatocellular cancer (HCC) outcomes, disease parameters need to be considered within a total personal clinical context. This requires preserving the coherence of data values, observed simultaneously for each patient during baseline diagnostic evaluation. Application of Network Phenotyping Strategy provided quantitative descriptors of these patient coherences. Combination of these descriptors with Fisher information about a patient's tumor mass and the histogram of the tumor masses in the whole cohort, permitted estimation of the time that passed from disease onset until clinical diagnosis ($t_{baseline}$). We found faster growth of smaller tumors having total masses <70 (80% of cohort) which involved ~3 times more interacting cellular processes than were observed for slower growing larger tumors (20% of cohort) with total masses >70. Combining the clinical survival and $t_{baseline}$ normalized all HCC patients to a common 1045 days of mean total disease duration ($t_{baseline}$ plus post diagnosis survival). We also found a simple relationship between the baseline clinical status, $t_{baseline}$ and survival. Every difference between a patient baseline clinical profiles and special coherent clinical status (HL_1) reduced the above common overall survival by 65 days. In summary, we showed that HCC patients with any given tumor can best have their tumor biology understood, when account is taken of the total clinical and liver context, and with knowing the point in its total history when an HCC diagnosis is made. This ability to compute the $t_{baseline}$ from standard clinical data brings us closer to calculating survival from diagnosis of individual HCC patients.***

INTRODUCTION

Two general processes are thought to contribute to hepatocellular cancer (HCC) prognosis. They are 1: liver damage, monitored by indices such as plasma levels of bilirubin and transaminases such as serum glutamic oxaloacetic acid (AST), and 2: tumor aggressiveness, monitored by indices such as tumor size, tumor number, presence of portal vein thrombosis (PVT) and blood alpha-fetoprotein (AFP) levels (1,2). These two general processes may also affect one another. Non disease-specific factors such as gender and age can also influence HCC outcomes, suggesting that any individual disease parameter needs to be considered within a total personal clinical context. This context might even provide personalization for the prognostic meaning of these factors for each patient, given the individual patient pattern of the measured parameters. Thus, a given level of bilirubin or tumor diameter might have a different significance in different total clinical personal contexts.

We developed a new approach to clinical data processing (3,4), a Network Phenotyping Strategy (NPS), which allowed the conversion of the above qualitative statements about the importance of the complete clinical context for determination of personal prognostic significance for levels of the

parameters, into a quantitative prognostic scheme. Improvements in prognostication likely require finding new information in the standard clinical data. We previously demonstrated that this new information resides in the coherence of data values, observed simultaneously for one particular person during the standard clinical screening at the baseline, provides that additional clinical characteristic, and provides additional patient characterization into HCC subtypes (3). This has been possible in the current work because we have combined into one clinical profile of every patient, not only the observed individual levels of the liver function parameters and the tumor indices, but also combined these personal values into a pattern, storing the total clinical context for each patient by explicitly capturing all the relationships between all of the parameters together. Without any reference to the actual tumor sizes in this NPS processing, HCC patients were classified by their baseline data coherence patterns into 2 subgroups, which were found to have significantly different S ('smaller tumor') and L ('larger tumor') phenotypes (5). The on-line tool allowing to enter the clinical data and obtain the HCC S/L subtype classification is available at http://www.entromics.com/pnm/.

This insight provides the necessary complementary clinical information to the conventional HCC subtyping, because "closeness" or "difference" is not evaluated by the identity or difference of the respective parameter levels, but by the identity or difference between the relationships between the pairs of parameter levels for concrete patient and respective landmark patterns. For L-phenotype patients, hepatic inflammation and tumor factors contributed collectively to more aggressive L tumors, with parenchymal destruction and shorter survival. This was manifested in the simultaneous observation of the following parameter values, and their relationship dominated the five L-characteristic patterns: presence of PVT in the context of simultaneously high levels of tumor growth indicators (AFP and/or platelet levels higher than the above thresholds), together with alcohol related liver damage. NPS thus integrated the liver, tumor and basic demographic factors and processed the information how all these data were simultaneously present for every patient into characterization of the new HCC subtypes, with indication for important differences in the underlying tumor biologies (micro- and macro environment related). We found that patients in the L phenotype group had 1.5 x larger mean tumor masses relative to S, $p=6\times10^{-16}$. In addition, S-phenotype patients had statistically highly significantly 1.7 × longer mean survival, $p<10^{-15}$, compared to L-phenotype.

This indicated that the NPS diagnostic scheme provided detailed, survival-related "matching" of the patient's personal clinical profiles that might be relevant for the personal survival prognosis. By relating the NPS characteristics of our patients to their actual survival, we obtained statistically significant prognostic models. However, the difference ranges between personal survival and the one that was predicted for individual patients from parameters reflecting these typical trends spanned very large ranges (close to thousand days) to provide translationally relevant predictions. The large range of individual survivals of patients with comparable (or even identical) clinical contexts and diagnostic relationship patterns indicates that the baseline clinical pattern context is not the exclusive determinant of the survival.

In the current paper, we show that a combination of NPS characterization of the patient clinical profiles and coherences of the parameter observations in them with processing of information from tumor imaging, results in significant improvement in prognostication. One of the important reasons for problematic OVS prediction is the unknown $t_{baseline}$, the time that passed from the disease onset until the clinical baseline (diagnosis).

The tumor size (or "tumor mass", a descriptor computed as the product of the number of tumor nodules and the tumor size) is known to be associated with survival (6-9) in a qualitative sense (larger tumor means shorter survival). Unfortunately, again perhaps mostly due to uncertainty in $t_{baseline}$, no direct correlation between the tumor masses and OVS was observed that was significant enough to provide a

useful relationship model between the individual observed tumor mass and actual survival. Thus, the large tumor mass as poor prognosis predictor is again

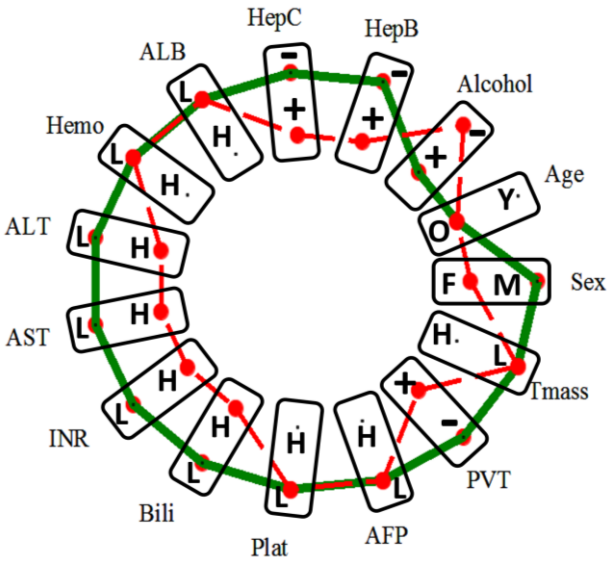

Fig. 1. Graphical representation of two clinical patterns as 15-partite graph. Symbols L and H indicate parameter values lower or higher than tertile-based thresholds (see text), + and – the presence and absence of observed property, O is older and Y younger than 55 years, M is male, F female gender. Green pattern represents the landmark pattern HL_1 of parameter value, used providing the best characterization of the OVS. To demonstrate the quantification of similarity between two relationship patterns, we added profile, for another patient, shown by red line (see text for direct clinical characteristics). This patients differs from HL_1 in 13 out of total 15 relationships between parameter values, and therefore $\delta(P_i,HL_1)=13$

just the "collective group property", which is useful "on average", but is difficult to personalize into a prediction with acceptable error range for an individual patient. We therefore used Fisher information formalism (8) to modify the tumor growth law, permitting estimation of personal $t_{baseline}$ from the patient's tumor mass and the histogram of the tumor masses in the whole cohort.

METHODS

1. Clinical patterns for survival prediction.

To characterize baseline clinical data relationship patterns for our 641 HCC patient cohort (see ref 3 for cohort description), we used the same clinical parameters as previously (3,4). To increase the detail with which the coherences in clinical parameter values are considered in subdividing patients into clinically matching subgroups, we processed every variable separately and added the tumor mass size explicitly to the pattern (unlike previously). The baseline data were therefore transformed by the NPS algorithm into 15-partite graphs, where each partition corresponded to one clinical parameter. The ranges or types of parameter values were dichotomized, using the same tertile-based thresholds for continuous variable levels as in our previous work (5). All patient baseline data were transformed into easy to interpret graphs of the personal relationship profiles. One of these graphs is shown by solid line (green) line in **Fig. 1**, representing male, older than 55 years, with self-reported alcoholism, without HepB and HepC antigens, albumin<4 g/dL, hemoglobin <15 g/dL, ALT <80 IU/L, AST <100 IU/L, INR <13, bilirubin <1.5 mg/dL, platelets $<200x10^3$/dL, AFP <29,000 ng/dL, no PVT and tumor mass <25.

These personal profiles were unified into complete study pattern (see Appendix, **Fig. A1**), so the information about the total frequencies of complete pairwise relationships between all variable levels, observed simultaneously for individual patients, can be processed by the NPS algorithm. We recovered seven reference landmark patterns (see Appendix) into which the study pattern is decomposed. These patterns were then used as the definition of clinical statuses, relative to which actual patient's relationship profiles were matched. Because every individual patient status is compared to the same landmark pattern, this permitted a simple summary characterization of the clinical profile subtypes by grouping together all patients with the same number of differences between their personal and landmark clinical profiles. An example of this quantitative

characterization of differences between two patterns is in **Fig. 1**, where the green (solid) profile is patient characterized above. The profile of another patient, shown in red (dotted), differs from the previous profile in 13 out of total 15 relationships used in the NPS, so the "clinical pattern distance" $\delta(P_7,HL_1)$ for

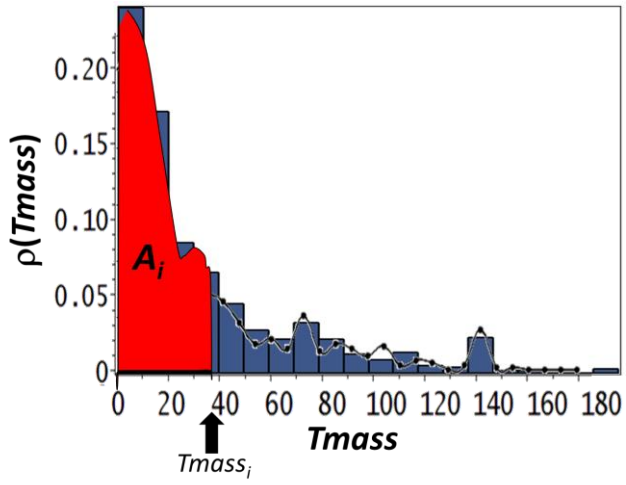

Fig. 2. Explanation of extracting the information about personal component $\boldsymbol{A_i}$ (red area) of $\boldsymbol{t_{baseline}}$ for patient with $\boldsymbol{Tmass_i}$=38 from the histogram of tumor masses in this study (blue bars), normalized to represent the $\rho(Tmass)$. Natural cubic spline interpolation and integration was used to obtain the numerical values of $\boldsymbol{A_i}$.

these two patients is 13. After computing the distances $\delta(P_i,HL_j)$ of all patient profiles (i=1,…641) from all seven landmarks (j=1,…,7), we tested the relationship of these new descriptors of the individual patient clinical profiles for the relationship to overall survivals. We used nonlinear (cubic polynomial) model function and SigmaPlot 11 software for this purpose. We then systematically monitored the statistical significance and goodness of fit (using residual standard deviation between predicted and actual survival) of these predictive models, which used distances $\delta(P_i,HL_j)$ from respective landmarks (both individually and in multivariable variants). We found that distances from HL_1 provide the optimal approximation.

2. Fisher information based estimation of time to baseline ($t_{baseline}$)

To improve the prognosis prediction from distances between each individual patient personal coherence clinical profiles from the HL_1 landmark profile, we applied Fisher information based processing of the tumor mass histogram. This allowed us to derive the analytical formula for the tumor growth law (10) and to use this law to estimate the time from the disease onset, $\boldsymbol{t_{baseline}}$, for each individual patient from his/her observed tumor mass (***Tmass***). Mathematical details are in the Appendix. There are several important aspects to this novel approach. Firstly, the clinical information that allows an estimation of the $\boldsymbol{t_{baseline}}$ is not just the tumor mass of individual patients, but instead includes the integrated probabilities, describing how probable it is to observe the individual patient's known tumor mass in the whole cohort. Practically, the personal patient contributions to the estimation of $t_{baseline}$ are these probabilities, computed as the personal partial areas ($\boldsymbol{A_i}$) of the tumor mass histogram for this study, integrated from zero (no tumor) to the actual tumor mass of every individual patient (see **Fig. 2**).

Secondly, the personal patient contributions were processed in a common (to all patients) disease context, characterizing HCC. These common HCC-specific parameters, needed for $\boldsymbol{t_{baseline}}$ estimate, were obtained from the tumor growth formula, derived by Fisher information processing (see Appendix). It is a power law,

$$\rho(\mathrm{Tmass}) = \tilde{E}^2 . T_{mass}^{2\gamma}. \qquad \text{Equation (1)}$$

Here $\rho(Tmass)$ are frequency values, obtained from normalized histogram of all 641 tumor masses in the study, $\boldsymbol{E}$ and $\boldsymbol{\gamma}$ are constants to be obtained by fitting this formula to the actual histogram of tumor masses.

Thirdly, the processing of the personal integrated area in the observed tumor mass histogram through the differential equation, based upon Fisher information resulted in the formula (1), relating the personal values, obtained by the above described

computational processing of the tumor mass

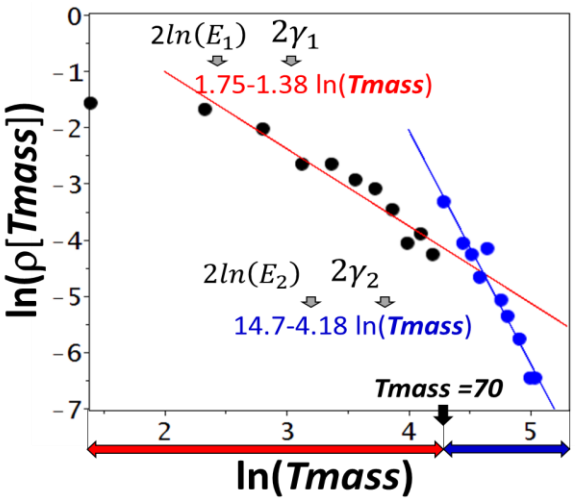

Fig. 3. Explanation of extracting the information about HCC-specific parameters of $t_{baseline}$ from tumor growth law (1), which indicated that log-log transformed ***Tmass*** histogram should be represented by linear functions. The two lines for tumor masses below (red) and above (blue) 70 were least-square fitted, resulting in the numerical values of E_j and γ_j as is shown.

histogram to $t_{baseline}$:

$$t_{baseline} = T_{max} \cdot e^{\left(\frac{ln\left[\frac{2\gamma \cdot A_i + A_i + E^2}{E^2}\right]}{(2\gamma+1)} \right)} \quad \text{Equation (2)}$$

In this equation, input parameters A_i, γ and E are described above and derivation of the optimal value T_{max} = 1200 days is described in the Appendix.

RESULTS:

Evidence for a power law for tumor growth.

We supposed that among many possible factors, a main contributor to the broad range of OVS (overall survival) is uncertainty in our knowledge of time for tumor development until diagnosis. From available parameters in the standard baseline clinical examination, tumor mass (***Tmass***) is one of the most important that can carry the time to baseline information. In our prognostic approach, we therefore decided to use the ***Tmass***, obtained from the baseline CT scan as the information from which we would estimate the time to the onset of tumor presentation.

As we did not have specific detailed data about our patients that would allow us to use current tumor growth models (11-13) (this would be the case for most standard clinical practice situations), we took advantage of Fisher information processing, which is a valid approach to obtain the relationship between tumor mass data from clinical radiology and time $t_{baseline}$ till appearance of the tumor mass (clinical presentation).

This new approach, which was shown to derive other natural laws from its fundamental principles, is applicable to the problem of tumor growth (10). We modified the results from ref 10 as shown in the Appendix. From a clinical point of view, in this approach we did not look for a "microscopic law" of tumor growth, but instead looked for the answer to the following question. Given the known distribution of probabilities of finding a tumor mass in the patient cohort, what is the law of tumor growth in time, that exactly reproduces that actual tumor mass distribution? The first result of the mathematical solution of this problem, which Fisher information-based formalism enabled us to find, is a power law, (see eq. 1 in Methods). This result indicates that there is a linear relationship between the log of the probability of observing a tumor mass and the log of the actual tumor mass.

Fig 3 provides experimental validation for the existence of this law, by examining actual clinical data from our 641 patients. It shows that when the tumor mass histogram (see **Fig. 2**) was converted into a log-log scale, then the transformed data obey the predicted law (eq. 1) of tumor growth. Additionally, it is clear that the two intersecting lines, needed to obtain the proper statistically best fit of the relationship, are likely to reflect two tumor-growth processes, one for ***Tmass*** ≤70, and the other for ***Tmass*** >70.

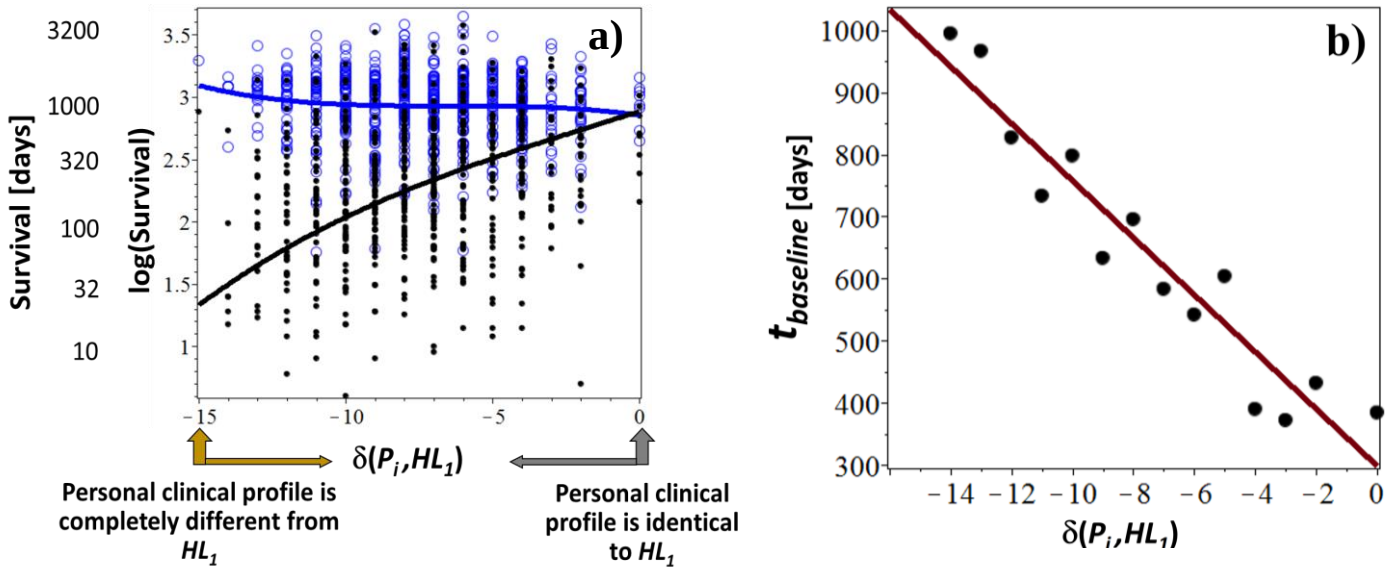

Fig. 4. a) Relationships between $\delta(P_i,HL_1)$, ***OVS*** (black points) and OVS_C (blue circles). The lines are least-square fits of the relationships by cubic model. 15 groups of patients (G_K, K=1,...,15) with matching differences $\delta(P_{iK},HL_1)$of their personal clinical relationship profiles form the vertical groups of points.
b) Linear relationship between the clinical profile differences and mean values $\langle t_{baseline} \rangle$ computed as average of individual values of $t_{baseline}$ for each patient group G_K.

Examination of the slopes of these two fitted lines shows that the larger tumors grew more slowly than the smaller tumors. This is consistent with the known sigmoid shape of growth curves of other biological populations (14,15). The Fisher method provides additional insight into the functional meaning of the parameters found in the **Fig. 3**. To show this, we used the definition of the slope γ in the power law as $\gamma_i = \left(\frac{\kappa_i}{1+\kappa_i}\right)$ (see detailed derivation in the Appendix), where in our case i=1 (for ***Tmass*** ≤ 70) and i=2 (for ***Tmass*** > 70) and $\boldsymbol{\kappa_i}$ are the numbers of growth related processes, contributing to the tumor masses observed. By using the estimated values of γ_1 and γ_2 from **Fig. 3**, we found that $\frac{\kappa_1}{\kappa_2} = 2.8$. Our Fisher entropy-based analysis thus showed that growth of these smaller tumors, which constituted about 80% of all tumors found in this cohort, involved on average ~3 times more interacting cellular processes than were observable for multiple, very large tumors with total masses above 70, observed for the remaining 20% of screened patients. This can be also interpreted as showing that smaller tumors are more sensitive to the overall clinical context of the patient (micro- and macro environmental factors) (16-18).

Estimation of $t_{baseline}$ from individual tumor masses

The above results justified our computing a corrected survival (**OVS**C) for each patient, using

$OVS_C = OVS + t_{baseline}$. We used eq. (2) with parameters obtained from **Fig. 3** together with the individual tumor masses to compute $t_{baseline}$ for each patient. These results were combined with the characterization of the patient clinical status, which we described quantitatively by the differences $\delta(P_i,HL_1)$ from the landmark pattern HL_1 (see **Fig. 1** and related text for clinical characterization of HL_1).

We were then able to derive the final optimal relationship between $\delta(P_i,HL_1)$ and OVS_C (corrected overall survival**)**. **Fig. 4a** shows a comparison between these relationships before (solid black points) and after correction for $t_{baseline}$ (blue circles).

We observed that OVS_C has a flat dependence on the $\delta(P_i,HL_1)$, with a common mean of ~1045 days. By contrast, we found (see **Fig. 4b**) a linear dependence between $\delta(P_i,HL_1)$ and mean $t_{baseline}$. The parameters of this simple linear relationship can be combined into the following practical observation:

A. The 300 days is the shortest mean $t_{baseline}$ typical for patients with the HL_1 baseline profile.

B. Every difference of patient's baseline clinical profile from HL_1 adds 65 days to the mean $t_{baseline}$.

These results together give a simple rule for ***OVS*** estimation (prognostication): the predicted ***OVS*** is $1045 - (65 \times |\delta(P_i,HL_1)|)$. Thus, if a patient has the HL_1 profile at baseline, then ***OVS*** is $1045 - (65 \times 0) = 1045$ days. As another extreme example, the largest $t_{baseline}$ is found for patients with profiles that have nothing in common with HL_1, so their $|\delta(P_i,HL_1)|=15$ and the corresponding $t_{baseline} = 15 \times 65 = 975$ days. Predicted survival for these "non-HL_1" patients is the shortest ***OVS*** = 1045 - 975, just 70 days and less.

Validation of clinical relevance of $t_{baseline}$ in the context of hepatitis status.

We examined differences in tumor masses for combinations of patient subgroups with different hepatitis status (Hep B, Hep C, Hep B+Hep C, neither) and we undertook this separately for the S and L phenotypes because of their quite different biologies and outcomes (4,19). We found that no differences in tumor masses were statistically significant for Hep B versus Hep C HCC patients (see examples in **Fig. 5a,b**). By contrast, examination of $t_{baseline}$ for the same hepatitis-type defined subgroups (**Fig. 5c,d**) resulted in significant differences, which were more pronounced in S than in L phenotype patients.

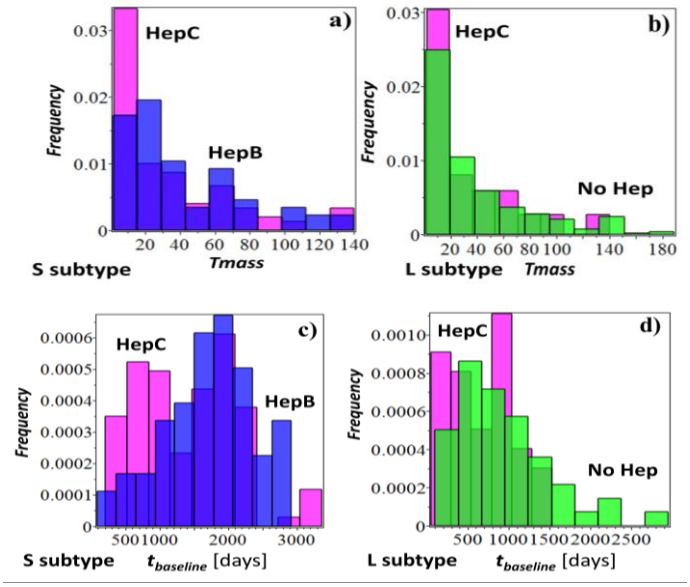

Fig. 5. Comparisons of distributions of tumor masses (**a** and **b**) which show no statistically significant differences and $t_{baseline}$ values (**c** and d) for S (left panels) and L (right panels) HCC phenotypes.

DISCUSSION

We previously developed an NPS-based classification system for HCC patients, using only blood based common hematologic and biochemical parameters, that resulted in the identification of 2 HCC phenotypes, labeled S and L, that differed significantly in their median tumor masses and survival. We attempted to relate descriptors of personal clinical profiles available by NPS to survival, but we found that for each group of patients with matching NPS clinical profiles, the survival range was very wide. In order to improve the survival prediction for an individual patient, in the current study, we have added tumor mass from baseline radiological measurements, to our prediction algorithm. We did this, because we suspected that aspects of tumor mass reflected previously unaccounted for characteristics of tumor biology that strongly impacted survival calculations. The most important of these characteristics that is the focus of this study, is the unknown period of tumor evolution till clinical diagnosis, that we have called $t_{baseline}$. We found that the estimate of $t_{baseline}$ requires two

components, one being disease-specific (HCC characteristics) and the other being individual patient characteristic. The HCC specific factors are common for all patients, which exhibited interesting heterogeneity, defining two categories of HCC tumors, those $\leq$ and > ***Tmass*** of 70.

The advantage of this Fisher information based approach, which makes it more powerful than conventional statistical analysis, is that in addition to tools that describe the clinical data, we obtained mechanistic insights into the disease biology and processes.

We attempted to find independent validation of $t_{baseline}$ relevance by relating these new data obtained for our patients against known facts about the processes and factors which were available in the data. Thus, for example we examined distributions of $t_{baseline}$ within subgroups of HCC patients with different hepatitis backgrounds. We found that while the tumor mass distributions were not significantly different between the hepatitis B and C and hepatitis negative subgroups, there were significant differences in time to baseline ($t_{baseline}$) when we analyzed these differences separately for S and L phenotypes. For S phenotype, we found that the significant differences were observed for hepatitis B only versus hepatitis C only patient subgroups, in which, for patients with hepatitis B, $t_{baseline}$ corresponded to longer tumor growth and thus to larger tumors compared to hepatitis C based HCC, in accordance to independent observations noted elsewhere (20-22).

Our Fisher information-based analysis also showed that growth of the smaller tumors with tumor masses < 70, which constituted about 80% of all tumors found in the total cohort, involved on average ~3 times more interacting cellular processes than were observable for multiple, very large tumors with total masses above 70, observed for the remaining 20% of screened patients. This information was obtained from the ratio of the power law constants γ_1 and γ_2, governing the two tumor growth processes, discovered by our analysis of the observed ***Tmass*** distribution. This also indicated that earlier stages of the tumor growth are more sensitive to the overall clinical context of the patient, which can be interpreted in terms of micro- and macro-environmental factors (16-18). This increased growth rate of small compared to large tumors, is similar to the growth rates in other biological populations, such as cells in culture of plants in a defined area. In those instances, explanations have included increased competition in dense populations for nutrients, oxygen or light, as well as for the phenomenon called contact inhibition amongst normal, but not cancer cells. Tumor growth is generally considered to be a reflection of the balance between growth and death processes. Slowing tumor growth can be attributed to changes in cell cycle time, nutrient availability and reduced growth fraction, amongst other factors (23). Our model describes these processes quantitatively by showing that the balance between growth and death in tumor cells is proportional to the number of interacting tumorigenic processes and is inversely proportional to time. This quantitative consideration of the totality of these multiple complex processes thus permits a more realistic estimate of $t_{baseline}$ and explains why the tumor mass alone cannot predict time till baseline diagnosis.

Our results are thus consistent with two hypotheses, explaining the existence of 2 tumor patient populations, identified by the different tumor growth rates (**Fig. 3**). A simpler hypothesis is that all small tumors reflect at least two HCC populations, viz those small tumors that will be always small tumors and another population that are the precursors to large HCCs. We postulate that the small precursors of small tumors are subject to two types of growth influence. They are endogenous factors (growth factors and oncogenes), as well as micro-environmental factors. By contrast, we hypothesize that large tumors are likely to be mainly driven by endogenous factors, such as growth factor gene products (24). A more complex hypothesis for the 2 types of small tumors, supposes that the precursors for the larger tumors may have 2 phases of growth: an initial phase that is mainly microenvironment driven, and a second, endogenous phase, which predominates when the tumors have reached a certain mass.

The personal component of the $t_{baseline}$ is understandable as the total probability that a patient with a given HCC tumor mass will be diagnosed radiologically at baseline evaluation. Equation 2 showed how to combine these personal factors with the HCC-specific tumorigenic parameters, derived from the complete baseline data and encompassing in their values above processes, into $t_{baseline}$ prediction. With the knowledge of $t_{baseline}$, we were then able to adjust the survival estimate derived from the tumors masses plus the personal differences of the patient personal profiles from calculated HL_1 profile, to give a corrected estimate of survival time from clinical diagnosis (baseline). Some examples of this corrected calculation are shown in the Results.

Our new analysis revealed that on average a patient with HCC has a disease length of 1045 days. This has 2 components, a $t_{baseline}$ during which the tumor develops before diagnosis and the clinically evident survival from clinical diagnosis. For the clinically evident part of the survival, our analysis provided a quantitative relationship between the patient clinical profile descriptors at diagnosis (baseline) and OVS. Thus, a tumor of any given mass that is seen and measured radiologically, can only have its biology and natural history understood, both in terms of the total clinical and liver context, but also with knowledge of the point it is at in its natural history. This is the real value of our ability to compute the time in the evolution of any patient's tumor till its baseline clinical evaluation ($t_{baseline}$) from standard clinical screening data.
The additional insights provided in the current work bring us closer to more realistically calculating survival from diagnosis of real HCC patients in the total clinical setting.

APPENDIX

1. Clinical profile patterns for survival prediction.

Fig. A1 shows the union of coherence relationship profiles, constructed from the screening baseline data for all 641 patients in the study. Resulting study graph is shown in (**Fig. A1**), with the edges, representing the simultaneous co-occurrence of pair of variable values, connected by the line in the graph. Edges are weighted by the total co-occurrence frequencies (these frequencies are visualized by the proportional thicknesses of the lines). This graph was then decomposed by the NPS algorithm into seven subgraphs (**Fig. A1**), which we call heterogeneity landmarks, HL_1, HL_2, …, HL_7.

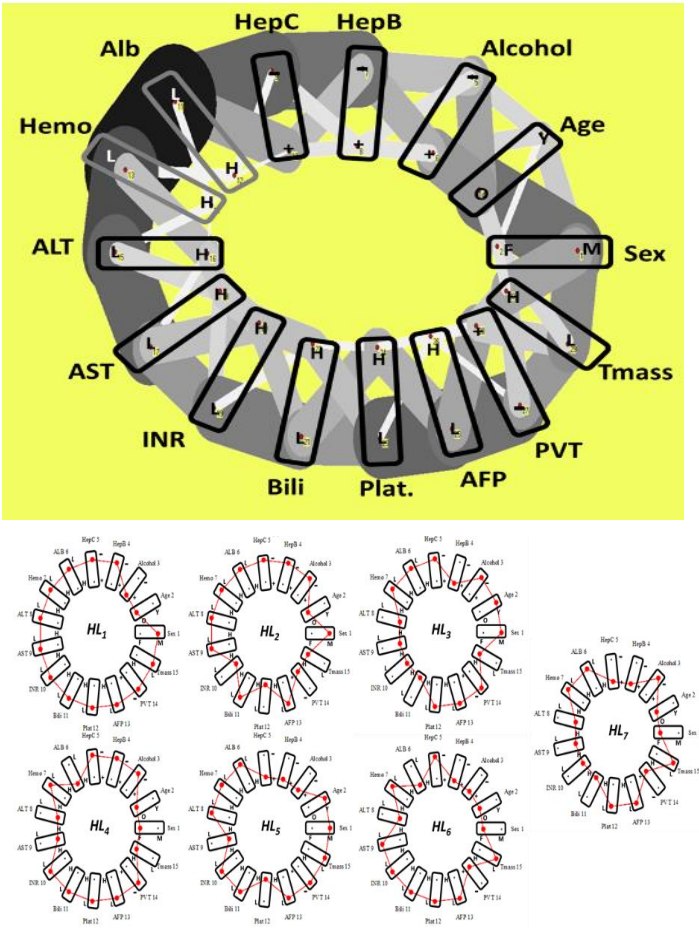

Fig. A1. Left panel: 15-partite study graph (symbols as in **Fig. 1**) with edge width proportional to the co-occurrence frequencies of the relationships between the connected parameter levels for all 641 patients of this study. **Right panel:** Decomposition of the study graph into landmark subgraphs HL_1-HL_7. The condition for this decomposition is that in HL_j, all co-occurrence frequencies are identical, which results in the independence of all binary relationships in the clinical profile patterns, captured by the respective landmarks.

2. Fisher information and tumor growth

What is Fisher information? It is an information descriptor, allowing derivation of law(s), governing the phenomenon (such as hepatocellular cancer and consequent liver tumor growth process) generating the data (such as tumor mass, a number and size of tumor nodules, observed in liver by CT scan). Advantage of this approach for clinical applications is that the only requirement for its use is that we assume, that clinical data are carrying the information we want to estimate in the experimentally observable parameters (time to disease onset and the prognosis of patient's survival from the diagnosis examination, given his/her baseline clinical status and tumor mass). While more widely known information descriptor, Shannon entropy, characterizes how well that (clinical) information is transmitted into the data (in the presence of noise), Fisher information $\boldsymbol{J_F}$ characterizes how well we can estimate from the (clinical) data the function-related parameters, responsible for the disease-related information content.

Another important feature of this information theory tool was identified by showing that Fisher information $\boldsymbol{J_F}$, describing the observability of a (disease related) parameter ϑ is in the special way equivalent to Kullback-Leibler entropy KL (see ref. 10):

$$J_F \sim -\frac{2}{\Delta\vartheta^2} KL\big(\rho(\vartheta_n), \rho(\vartheta_n + \Delta\vartheta)\big)$$

This relationship best explains the "local" character of the Fisher information (entropy): In contrast to global descriptors as Shannon entropy, which integrate the information over complete distribution of observed data (signals), the Fisher information (applied for better clarity to our tumor growth analysis) quantifies how the information about the patient's tumor, characterized by the time-dependent law changes, when the tumor grows by a small increment from a specific tumor mass (ϑ_n) to a new size $(\vartheta_n + \Delta\vartheta)$. This has very important consequence for the ability to derive physically, biologically and clinically relevant laws from the Fisher information. As the tumor mass increase $\Delta\vartheta$ can be selected to be very small (in the limit allowing integration actually infinitesimally small), we can use simple relationships between the functional and clinical parameters, entering into the derivation of the laws, and allowing in the final result estimating the "hidden" information, such as time to disease onset, $\boldsymbol{t_{baseline}}$, from the tumor mass. While in the global picture, such relationships can be very complicated, non-linear etc., in the Fisher information processing we deal only with small change $\Delta\vartheta$ of the processed data variability. In this way, it is fully justified to use simple relationships between the observed and "hidden" parameters (generally valid recipe is using just the first terms of the Taylor expansion of this complex relationship, resulting in the proportionality etc.). In this way, we mathematically correctly decompose that complicated relationship into piece-wise linearized series of relationships for consecutive steps of the tumor growth and then use calculus to generalize that discrete representation into the final law.

Another advantage of processing the clinical data via the Fisher information is the consequence of the fact that (by direct enumeration), the $\boldsymbol{J_F}$ of the normal (Gaussian) distribution is equal to the constant, $1/\sigma$, which is the variance (width) of this special distribution. This shows, that important features of the data, which Fisher information studies and quantifies, are the underlying biological and clinically relevant processes that are responsible for non-random features and biases in the distributions of the collected clinical data. This indicates, why standard statistical techniques, with parameters derived from the (axiomatic) assumption, that data are independently and identically normally randomly distributed cannot capture the functional information in the data, while Fisher information is actually actively "pursuing" and fully exploiting the part of the clinical data, that are non-random, because of underlying functional processes, internal to the patient's biosystem (cell system). We will show below, that this separation of the internal (disease-related) and external (tumor micro- and macro-environment in the liver, CT-scan experimentation, etc.) is needed for finding the fundamental a priori optimization principle, allowing to derive the result in a closed form of personal

formulae, converting the patient's tumor mass into the estimate of the time to disease onset (and also many other derived laws in all fields of science – see ref 10).

For application to our concrete case of tumor growth, we follow the derivation in ref. 10, which we then extend to estimation of time to disease onset estimation from the patient's tumor mass. What follows is the outline of the derivation, which summarizes just steps and resulting formulae, relevant for clinical interpretation (detailed step by step derivation can be found at

http://www.entromics.com/content/time-disease-onset-fisher-information :

Define the probability density $\rho(Tmass, t)$ of finding HCC tumor of tumor mass = ***Tmass*** at some time *t*, measured from the disease onset. The $\rho(Tmass, t)$ can be found experimentally from the properly normalized histogram of baseline tumor masses in a study/screening (see **Fig. 2**).

For the purposes of the (formal) mathematical derivations, define the following (formal) quantities:

$$\psi^2 = \rho(Tmass, t)$$

$$\frac{d\psi}{dt} = \dot{\psi} \quad \Rightarrow \dot{\psi}^2 = \left(\frac{\partial\psi}{\partial t}\right)^2$$

Define external observed Fisher information *J*:

$$J = 4\int j(\rho(Tmass), t)dt = 4\int \left(\frac{\partial\psi}{\partial t}\right)^2 dt = 4\int \dot{\psi}^2 dt = j$$

Define internal Fisher information *I*:

$$I = 4\int i(\rho(Tmass), t)dt = i$$

Formulate the problem of balancing the external and internal Fisher information to achieve optimal estimation of parameter $\boldsymbol{t_{baseline}}$ from the observed values of ***Tmass***:

$$J - \kappa I = 4\int \dot{\psi}^2 dt - 4\kappa \int i(\rho(Tmass), t)dt = j - \kappa i = 0$$

with $\kappa \neq 0$.

The result

$$j - \kappa i = 0 \qquad \text{(A1)}$$

is quantitative formulation of two basic conditions that start the analysis:

- First, $\kappa \neq 0$ reflects the fact that the internal observed Fisher information i (tumor mass) is a result of multiple interacting (cellular) processes and what we see is their overall effect.
- Second, $j = \kappa i$ reflects the requirement, that in the extraction of the desired parameter (time to disease onset) from the $\rho(Tmass)$-dependency on ***Tmass*** we want to lose minimal information about the internal biology of the tumor, encoded in κi.

These two conditions have to be complemented by the recipe, showing how to optimize the integrals, which define the functionals j and κi, so the two above conditions are actually met simultaneously. This recipe is derived in the functional optimization theory and is called Euler-Lagrange equation. Applied this general equation to j and i results in

$$\frac{d}{dt}\left(\frac{\partial(j-i)}{\partial\dot{\psi}}\right) = \frac{\partial(j-i)}{\partial\psi} \qquad \text{(A2)}$$

Solving (A1) and (A2) simultaneously results in the following differential equation for i

$$\sqrt{\kappa}\frac{\partial i}{\partial t} + \sqrt{i}\frac{\partial i}{\partial\psi}(1+\kappa) = 0 \qquad \text{(A3)}$$

Solution of this equation can be found in terms of the amplitudes of probability density for tumor masses:

$$\psi(t) = E.t^{\left(\frac{\kappa}{1+\kappa}\right)} = E.t^{\gamma} \qquad \text{(A4)}$$

Here E and γ are constants to be determined from study data. E is the HCC-specific "amplitude", summarizing all non-tumor mass tumorigenesis processes and $\gamma = \left(\frac{\kappa}{1+\kappa}\right)$ is the rate constant, describing the probability gradient rate of finding certain tumor mass in the patient's cohort at baseline (when CT scan was taken). Note (for interpretation purposes) that γ is solely determined by κ, the parameter, which is proportional to the number of all

INTERACTING intercellular biological processes, influencing the tumor growth. Now because of local dependence of Fisher information on the variables, it is possible to use Taylor expansion of the (in general complicated) relationship between time to disease onset, $\boldsymbol{t_{baseline}}$, and ***Tmass*** and use its first term. This means that for small changes in tumor masses along the HCC progression, we have ***Tmass*** proportional to disease duration. By integration, we convert this locally linearized proportionality into the (nonlinear) global relationship. Thus, eq. (A4) can be re-written as

$$\rho(Tmass) = \psi^2(Tmass) = \tilde{E}^2.Tmass^{2\gamma} \quad \text{(A5)}$$

Still, both sides of (4) are (continuous and general) functions of time. To personalize this general result, we therefore need to express explicitly the change of $\psi(t)$ and $E.t^{\gamma}$ with time and integrate the resulting formulae up to the time, when the patient's tumor mass was observed in clinic. To facilitate direct comparison with tumor mass histogram, we use $\psi^2 = \rho(Tmass, t)$:

$$\int_{t=0}^{Tmass=Tmass_{baseline}} \rho(Tmass) = E^2 \int_{t=0}^{t=t_{baseline}} t^{2\gamma} dt \quad \text{(A6)}$$

Carrying the integration, we obtain the following functions of the (personal) upper integration limit $t_{baseline}$:

$$\int_{t=0}^{Tmass=Tmass_{baseline}} \rho(Tmass) = A_i \quad \text{(A7)}$$

$$E^2 \int_{t=0}^{t=t_{baseline}} t^{2\gamma} dt = \frac{E^2\left(t_{onset}^{(2\gamma+1)}-1\right)}{(2\gamma+1)} = A_i \quad \text{(A8)}$$

The personal parameter A_i is obtained by integrating the (normalized) histogram of the study tumor masses, up to the value found for a patient. We can then solve the last equation for $t_{baseline}$, obtaining the final formula:

$$t_{baseline} = T_{max}\,.\,e^{\left(\frac{ln\left[\frac{2\gamma.A_i+A_i+E^2}{E^2}\right]}{(2\gamma+1)}\right)} \quad \text{(A9)}$$

To improve the prognosis prediction from distances between patient's personal coherence clinical profiles from the (selected) landmark profiles, we applied Fisher information based processing of the tumor mass histogram, allowing us to derive the analytical formula for the tumor growth law (ref. 10) and to use this law to estimate the time from the disease onset, $\boldsymbol{t_{baseline}}$, for every individual patient from his/her observed tumor mass (***Tmass***). The first step in applying these theoretical results to actual data, was to verify that the actual ***Tmass*** histogram is compatible with the "power law" $\rho(Tmass) = \tilde{E}^2.T_{mass}^{2\gamma}$. By taking the logarithm of this result, we have

$$ln\big(\rho(Tmass)\big) = ln\big(\tilde{E}^2\big) + 2\gamma.ln(Tmass) \quad \text{(A10)}$$

indicating that if the (normalized) tumor mass histogram is presented in the log-log format, the logarithm of $\rho(Tmass)$ should be linear function of logarithm of ***Tmass***, with $ln\big(\tilde{E}^2\big)$ being equal to intercept and 2γ to the slope of that relationship.

The second step is the actual derivation of the time $\boldsymbol{t_{baseline}}$, which is needed to observe the tumor mass for any given patient. In eq. (A9), the values of γ and E^2 are obtained from the least squares fits of the law (A4) to the histogram of the observed tumor masses and the derivation of A_i which is the partial integral of the histogram of observed tumor masses is explained in the main text (see **Fig. 2**).

Fig. A2 shows how the eq. (A9) is actually used to compute the $\boldsymbol{t_{baseline}}$. In **Fig. 2** we demonstrated how is the value of $\boldsymbol{A_i}$ for one patient obtained as the partial area of the histogram of observed tumor masses, computed from zero to the actual patient's tumor mass at baseline. **Fig. A2** shows the plot of $t_{baseline}$ (in relative units) as the function of $\boldsymbol{A_i}$, computed from the continuous tumor masses within the actual observed range, where we explicitly consider the presence of two tumor growth rates for tumors with masses below and above 70 .

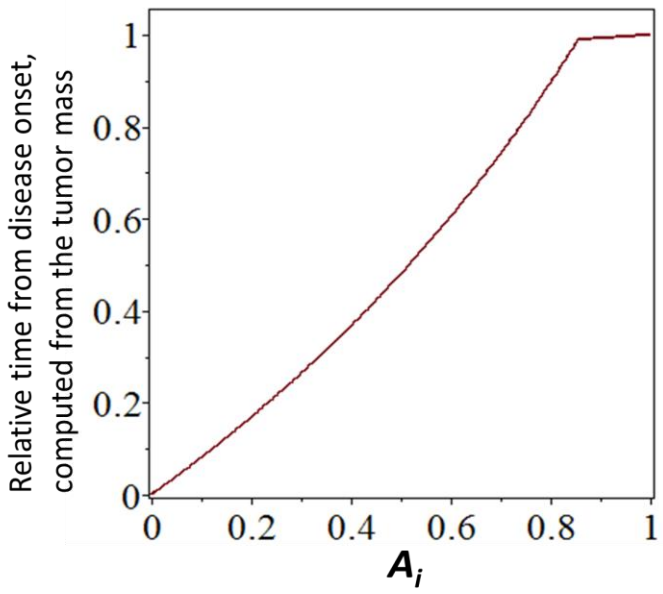

Fig. A2 Plot of the dependence of $t_{baseline}$ on A_i, as derived in eq. (A9) with T_{max}=1, which explicitly considered the presence of two growth rates for smaller and large tumors.

As the eq. (A9) determines the $t_{baseline}$ only in the

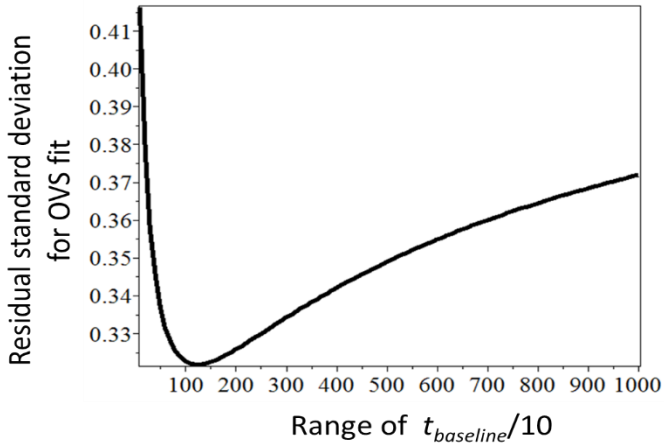

Fig. A3: Plot of the residual standard deviations of OVS_C – $\delta(P_i,HL_1)$ fits for systematically varied values of T_{max}. The optimal value is in the minimum of this curve, at T_{max}= 1200.

relative units, the last step of the converting of the Tmass data into $t_{baseline}$ is finding the common range, T_{max}, which will convert the relative time units in eq. (A9) to actual days. This range is a constant, with a value, that will reproduce best the survival prognosis computed from the coherence descriptors $\delta(P_i,HL_1)$ of our patients at the baseline. To find the value of this constant matching that criterion, we systematically varied its value from 0 to 10 000 days with 10 day increment. For each of these values of T_{max}, and for every patient, we corrected the patient's survival by the value $T_{baseline}=T_{max}.\ t_{baseline}$: $OVS_C=OVS-T_{baseline}$ and fitted the resulting set of 641 OVS_C values by $\log(OVS_C) = b_3 \times \delta(P_i,HL_1)^3 + b_2 \times \delta(P_i,HL_1)^2 + b_1 \times \delta(P_i,HL_1) + q$. **Fig. A3** shows the residual standard deviations of the fits OVS_C – $\delta(P_i,HL_1)$ for all considered values of T_{max}. The optimal prediction of the survival was found in the minimum of this curve, for T_{max} = 1200 days.

REFERENCES

1. Carr BI, Pancoska P, Branch RA. Tumor and liver determinants of prognosis in unresectable hepatocellular carcinoma: a large case cohort study. Hepatol Int. 2009;4: 396-405

2. Zhang JF, Shu ZJ, Xie CY, Li Q, Jin XH, Gu W, Jiang FJ, Ling CQ. Prognosis of unresectable hepatocellular carcinoma: comparison of seven staging systems (TNM, Okuda, BCLC, CLIP, CUPI, JIS, CIS) in a Chinese cohort. PLoS One. 2014; 9(3):e88182. doi: 10.1371

3. Pancoska P, Carr BI, Branch RA. Network-based analysis of survival for unresectable hepatocellular carcinoma. Semin Oncol. 2010; 37:170-81

4. Pancoska P, Lu SN, Carr BI. Phenotypic Categorization and Profiles of Small and Large Hepatocellular Carcinomas. J Gastrointest Dig Syst. 2013; Suppl 12. pii: 001.

5. Carr BI, Guerra V, Pancoska P. Thrombocytopenia in relation to tumor size in patients with hepatocellular carcinoma. Oncology. 2012; 83:339-45

6. Njei B, Rotman Y, Ditah I, Lim JK. Emerging trends in hepatocellular carcinoma incidence and mortality. Hepatology. 2014 Aug 20. doi: 10.1002

7. Yip VS, Gomez D, Tan CY, Staettner S, Terlizzo M, Fenwick S, Malik HZ, Ghaneh P, Poston G.
Tumour size and differentiation predict survival after liver resection for hepatocellular carcinoma arising from non-cirrhotic and non-fibrotic liver: a case-controlled study. Int J Surg. 2013; 11:1078-82

8. Marelli L, Grasso A, Pleguezuelo M, Martines H,

Stigliano R, Dhillon AP, Patch D, Davidson BR, Sharma D, Rolles K, Burroughs AK. Tumour size and differentiation in predicting recurrence of hepatocellular carcinoma after liver transplantation: external validation of a new prognostic score.
Ann Surg Oncol. 2008; 15:3503-11
9. Liu C, Duan LG, Lu WS, Yan LN, Xiao GQ, Jiang L, Yang J, Yang JY. Prognosis Evaluation in Patients with Hepatocellular Carcinoma after Hepatectomy: Comparison of BCLC, TNM and Hangzhou Criteria Staging Systems. PLoS One. 2014 Aug 18; 9(8):e103228
10. Frieden BR Science from Fisher Information. A Unification, Cambridge University Press, Cambridge, New York, 2nd edition, 2004, chapters 1, 14, 15. ISBN 0 521 81079 5
11. Ribba B, Holford NH, Magni P, Trocóniz I, Gueorguieva I, Girard P, Sarr C, Elishmereni M, Kloft C, Friberg LE. A review of mixed-effects models of tumor growth and effects of anticancer drug treatment used in population analysis. CPT Pharmacometrics Syst Pharmacol. 2014 May 7;3:e113. doi: 10.1038/psp.2014.12.
12. Wang Z, Butner JD, Kerketta R, Cristini V, Deisboeck TS. Simulating cancer growth with multiscale agent-based modeling. Semin Cancer Biol. 2014 May 2. pii: S1044-579X(14)00049-2
13. Masoudi-Nejad A, Bidkhori G, Hosseini Ashtiani S, Najafi A, Bozorgmehr JH, Wang E. Cancer systems biology and modeling: Microscopic scale and multiscale approaches. Semin Cancer Biol.: 2014 Mar 18. pii: S1044-579X(14) 00039 X.
14. Gyllenberg M, Webb GF. Quiescence as an explanation of Gompertzian tumor growth.
Growth Dev Aging. 1989 Spring-Summer; 53(1-2):25-33.
15. Crispen PL, Viterbo R, Boorjian SA, Greenberg RE, Chen DY, Uzzo RG. Natural history, growth kinetics, and outcomes of untreated clinically localized renal tumors under active surveillance.
Cancer. 2009 ; 115:2844-52
16. Hernandez-Gea V, Toffanin S, Friedman SL, Llovet JM. Role of the microenvironment in the pathogenesis and treatment of hepatocellular carcinoma. Gastroenterology. 2013; 144:512-27
17. Carr BI, Guerra V. HCC and its microenvironment. Hepatogastroenterology. 2013; 60:1433-7
18. Tu T, Budzinska MA, Maczurek AE, Cheng R, Di Bartolomeo A, Warner FJ, McCaughan GW, McLennan SV, Shackel NA. Novel aspects of the liver microenvironment in hepatocellular carcinoma pathogenesis and development. Int J Mol Sci. 2014; 15:9422-58
19. Carr BI, Pancoska P, Giannini EG, Farinati F, Ciccarese F, Ludovico Rapaccini G, Di Marco M, Benvegnù L, Zoli M, Borzio F, Caturelli E, Chiaramonte M, Trevisani F; Italian Liver Cancer Group. Identification of two clinical hepatocellular carcinoma patient phenotypes from results of standard screening parameters. Semin Oncol. 2014; 41:406-14
20. Barazani Y1, Hiatt JR, Tong MJ, Busuttil RW. Chronic viral hepatitis and hepatocellular carcinoma. World J Surg. 2007; 31:1243-8.
21. . Dohmen K1, Shigematsu H, Irie K, Ishibashi H. Comparison of the clinical characteristics among hepatocellular carcinoma of hepatitis B, hepatitis C and non-B non-C patients. Hepatogastroenterology. 2003; 50:2022-7
22. Hiotis SP, Rahbari NN, Villanueva GA, Klegar E, Luan W, Wang Q, Yee HT. Hepatitis B vs. hepatitis C infection on viral hepatitis-associated hepatocellular carcinoma. BMC Gastroenterol. 2012; 12:64.
23. Bassukas ID, Maurer-Schultze B. Mechanism of growth retardation of the adenocarcinoma EO 771. Radiat Environ Biophys. 1987; 26:125-41
24. Chan VT, McGee JO. Cellular oncogenes in neoplasia. J Clin Pathol. 1987; 40:1055-63. Review