Phenotypic categorization and profiles of Small and Large hepatocellular carcinomas

Petr Pancoska PhD[1], Brian I. Carr MD, FRCP, PhD[2] and Shenh-Nan Lu MD,[3]

[1]Center for Craniofacial and Dental Genetics, University of Pittsburgh, Pittsburgh, PA, USA,

[2]Department of Liver Tumor Biology IRCCS de Bellis, National Institute for Digestive Diseases, Castellana Grotte , BA, Italy;

[3]Division of Gastroenterology, Department of Internal Medicine, Chang Gung Memorial Hospital, Kaohsiung Medical Center, Chang Gung University, Kaohsiung, Taiwan;

Running title: Large and Small HCC subsets

Key Words: HCC, tumor mass, portal vein thrombosis, AFP

Author contributions: Carr: idea, writing; Lu, clinical data collection Pancoska, statistical analysis

Abbreviations: HCC, hepatocellular carcinoma; AFP, alpha-fetoprotein; PVT, portal vein thrombosis, Hgb, hemoglobin, plts, platelets, WBC, white blood count; INR, international normalized ratio; CAT, computerized axial tomogram phy; HBV, hepatitis B virus; HCV, hepatitis C virus; L, large; S, small.

Disclosures: none; Grant support: none

Correspondence: Brian I. Carr MD, FRCP, PhD

IRCCS 'S. de Bellis', via Turi 27, 70013 Castellana Grotte (BA), Italy

Tel. 39 080 4994603; Fax. 39 080 4994313

E-mail: brianicarr@hotmail.com

Abstract

We used a database of 4139 Taiwanese HCC patients to take a new approach (Network Phenotyping Strategy) to HCC sub-set identification. Individual parameters for liver function, complete blood count, portal vein thrombosis, AFP levels and clinical demographics of age, gender and hepatitis or alcohol consumption, were considered within the whole context of complete relationships, being networked with to all other parameter levels in the entire cohort. We identified 4 multi-parameter patterns for one tumor phenotype of patients and a separate 5 multi-parameter patterns to characterize another tumor phenotype of patterns. The 2 sub-groups were quite different in their clinical profiles. The means of the tumor mass distributions in these phenotype sub-groups were significantly different, one associated wither larger (L) and the other with smaller (S) tumor masses. These significant differences were seen systematically throughout the tumor mass distributions. Essential and common clinical components of L-phenotype patterns included simultaneously high levels of AFP, low platelet levels plus presence of portal vein thrombosis. S included higher levels of liver inflammatory parameters. The 2 different parameter patterns of L and S-sub-groups suggest different mechanisms; L, possibly involving tumor-driven processes and S associated more with liver inflammatory processes.

Introduction

The prognosis and choice of treatments in patients who have hepatocellular carcinoma (HCC) has long been recognized to depend both on tumor factors as well as liver factors and was the basis for the first published classification scheme of Okuda (1). This is because HCCs usually arise in a liver that has been chronically diseased (hepatitis or cirrhosis from hepatitis or other causes, or both) (2-6). Many more complex classification and prognostication schemes have since been published, all of which take these 2 broad categories of factors into account, and patients can die either from their tumor growth or from their liver failure. However, there are additional layers of complexity that need to be taken into consideration. Thus, quite large HCCs can arise in surprisingly normal (non-cirrhotic liver). Furthermore, many small HCCs do not seem to grow further. Thus, some small HCCs stay small and others are precursors of larger HCCs. Since a patient can present at any random part of their HCC disease growth process, it is usually difficult to know at what point in their disease process they have been diagnosed. Given the suspicion that the diagnosis of HCC carries within it several or multiple sub-sets of disease, we recently used a tercile approach, to identify HCC sub-sets at the extreme wings of an HCC patient cohort that had been ordered according to tumor size and then trichotomized into tumor size terciles (7,8). We found that on the extreme terciles, there was a relationship between plasma platelet numbers and HCC size. This likely reflected that small HCCs arising in cirrhotic liver for which thrombocytopenia is a surrogate (9) with portal hypertension and a larger tumor size tercile without thrombocytopenia. However, it still left the central part of the tumor/disease continuum uncharacterized and unordered into sub-sets. Furthermore, we also showed a relationship between blood alpha-fetoprotein (AFP) levels, a marker of HCC growth, and blood total bilirubin levels, in a large part of the cohort (10). This led support, as has evidence of others (11-13), that HCC may not only arise and grow in a cirrhotic milieu, but may even depend on signals from that micro-environment for its growth and other biology. Given this unsatisfactory clinical HCC heterogeneity, it seems that ‘one size fits all’ doesn’t work for individual prognostic factors, probably because of the absence of significant sub-subset patient separation. In addition, some parameters such as AFP can be elevated in either small or large HCCS.

We reasoned that attempts to extract new information cannot rely only on standard clinical data, but rather upon processing relationships between given data. In this report, we have taken a different approach to identify phenotypically different HCC patients groups. We first transformed the raw clinical screening data into a new form, considering in full the individual parameters within the whole context of complete relationships to all other parameter levels. After this transformation, individual parameters were not treated as single entries into the analysis, but were each considered as a parameter within the whole clinical context (liver function tests, presence of cirrhosis or hepatitis, inflammation and different manifestations of tumor growth-size, number of tumor nodules, presence of PVT), with considerations of age and gender.

Methods

Patient clinical data. Clinical practice data, recorded within Taiwanese HCC screening program, was prospectively collected on newly-diagnosed HCC patients and entered into a database that was used for routine patient follow-up. Data included: Baseline CAT-scan characteristics of maximum tumor diameter and number, presence or absence of PVT; Demographics (gender, age, alcohol history, presence of hepatitis B or C); Complete blood counts (hemoglobin, platelets, INR); blood AFP and routine blood liver function tests, (total bilirubin, AST and ALT, albumin) –see Table I. The retrospective analysis was done under a university IRB-approved analysis of de-identified HCC patients.

Patient profiles. We developed a Network Phenotyping strategy (NPS), a graph-theory based approach (10), allowing personalized processing of complex phenotypes, with explicit consideration of functional parameter correlations and interdependencies. NPS was applied here to integrate the data of all 4139 HCC patients.

Table I. Demographic and clinical characterization of patients:

		Number	Percent in study		Number	Percent in study
Gender	female	1033	24.9%	male	3106	75.1%
Age	Younger<55year	1432	34.6%	Older>55year	2707	65.4%
Alcohol	[-]	2950	71.2%	[+]	1189	28.7%
HBV	[-]	2010	48.6%	[+]	2129	51.4%
HCV	[-]	2520	60.9%	[+]	1619	39.1%
PVT	[-]	3187	77%	[+]	952	23%
AFP	low<200	2671	64.5%	high>200	1468	35.5%
Bilirubin	low<1.2	2618	63.2%	high>1.2	1521	36.8%
ALT	low<40	1693	40.9%	high>40	2476	59.1%
AST/ALT	low <1.0	1344	32.5%	high>1.0	2795	67.5%
albumin	low<3.0	951	23.0%	high>3.0	3188	77.0%
hemoglobin	low<13.0	2236	54.0%	high>13.0	1903	46.0%
platelets	low<= 125	1603	38.7%	high> 125	2536	61.3%
INR	low<1.0	1355	32.7%	high>1.0	2784	67.3%

Number of tumors	Patients/percent of cohort	Size range [cm]	Mean [cm]	Median [cm]	std. deviation [cm]
1	2147 (51.9%)	1-27.7	4.9	3.2	4.1
2	575 (13.9%)	1-22.2	4.4	3.3	3.3
3	178 (4.3%)	1-15	3.9	3.0	2.5
>3	1239 (30%)	0.9 – 26.0	7.9	8.5	4.3

There were no missing data in this data set. Individual patient profiles were created, in which each of 15 parameters was assessed in the context of all the other parameters for that same patient and processed by NPS approach. The technical details of NPS are presented in the Appendix. Here we summarize the concrete steps and their results:

Step 1. To reduce the complexity of the relationships that needs to be considered in the analysis, we considered correlations between blood liver function and hematological parameters. Out of 8 liver function parameters, we found 4 unique pairs that showed the most correlated and significant trends in their values. Some of these 4 were also strongly correlated in our previous work (10). While the selection of the 4 parameter pairs with the strongest correlations amongst all >20,000 possible was done using just a maximal cut mathematical algorithm (14), these 4 unique pairs were inter-related through established underlying functional processes: total blood bilirubin/prothrombin

time (a measure of liver function), SGOT/SGPT (a measure of liver inflammation) and AFP and blood platelet counts (reflections of tumor growth) (7).

Step 2. We continued by transforming the original patient data into a form of "levels". This step unified the demographic (categorical) parameters with liver function (real value) parameters needed for consideration of their inter-relationships within directly clinically interpretable framework. Considering the established practice in HCC diagnostics (15,16), we determined 'high' and 'low' levels of each individual parameter

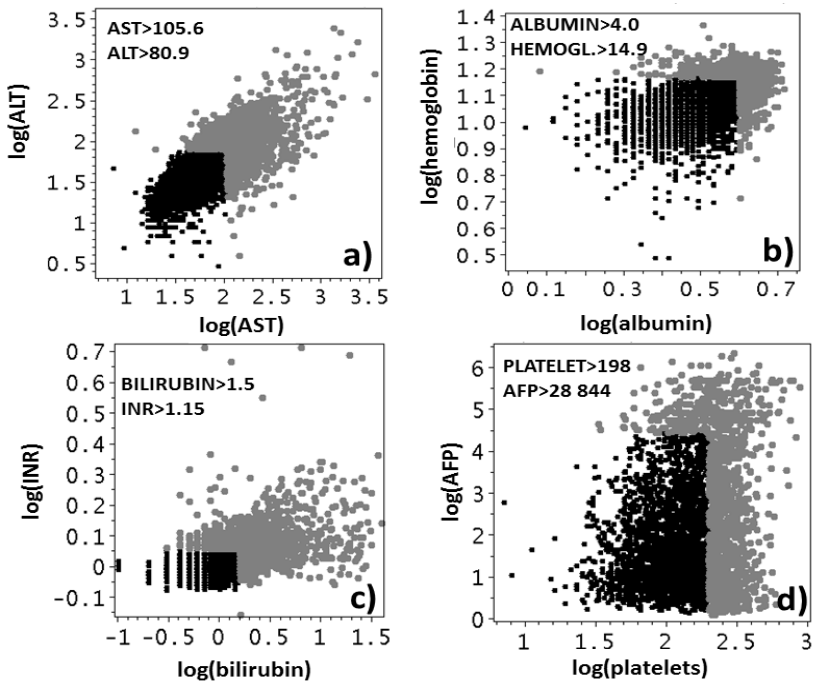

Figure 1. Set of four parameter pairs that have the highly correlated trends. One point represents a patient, in gray we show the upper tertile of patients identified with "high" levels of both parameters. In black are 2/3 of patients with "low" parameter levels. The boundary between the high and low levels is defined by the two threshold values indicated in every picture.

using a tercile-based dichotomization. For gender, reported alcoholism, evidence for hepatitis B and/or C and presence or absence of PVT the dichotomization was natural. For the other parameters, we tested several alternatives (50%:50%, quartiles) but found that tercile dichotomization with 2/3 of patients with the lowest parameter levels designated as “Low” phenotype and 1/3 of patients with the highest parameter levels designated as “High” phenotype was optimal for further processing. For age the “old” tercile was separated from the lower 2 “young” terciles by 55 years (17) . For the four significantly correlated parameter pairs, we used the two-thresholds that separate High from Low phenotypes, as shown in Fig. 1. This resulted in clinically familiar value cutoffs, such as bilirubin of 1.5 mg/dl, AST 200 IU/l and ALT 105 IU/l.

Step 3. Using actual data for each patient, an individual clinical profile was created by connecting all the actual parameter high, low, + and - levels (Fig. 2) into a representation of their complete networked relationships. In Fig.2 example, profiled

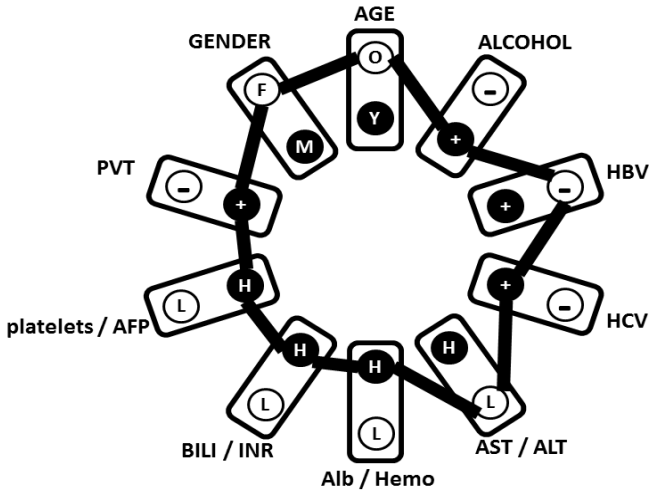

Figure 2. Example of 10-partite individual clinical profile of a patient. F=female, M=male, O=age>55, Y age<55 years, +/- presence/absence of indicated parameter. H/L – correlated parameter pair levels, shown in **Fig. 1**. The patient’s clinical profile is recovered from this scheme by following the black line. BILI = bilirubin, Alb=albumin, Hemo=hemoglobin

patient is an older female, reporting alcoholism, diagnosed with HCV but not HBV, with AST<105 and ALT<80 IU/l, albumin > 4.0g/l, hemoglobin >15, bilirubin >1.5mg/dl, INR >1.2, platelets >200 x 10^{-9}/l, AFP>29,000 ng/ml and presence of PVT. All these 4139 individual profiles were unified into a single schema (Appendix), that carries new information about co-occurrence frequencies of all parameter levels.

Step 4. We found a simpler structure in the networked HCC clinical data for this cohort. The schema was completely decomposed into only 19 reference profiles C_1-C_{19} (Appendix). These reference clinical profiles had to have identical co-occurrence frequencies between all the parameter levels. This ensured the independency of the results on the parameter ordering in the clinical profile: re-arranging the sections in Figure 1 will generate identical data for subsequent steps. C_1-C_{19} collect the information about the most frequent relationship co-occurrences of various parameter levels. C_1-C_{19} thus serve as idealized clinical statuses.

Step 5: The 4139 individual profiles were then compared in turn to each of the 19 reference profiles and the total number (0-10) of mismatches in the relationships they describe were recorded as differences d_1-d_{19} between the profiles.

Step 6. We next used logistic multiple regression (18) with variable selection algorithm (SigmaPlot11), using patient's 19 differences d_1-d_{19} as independent variables, to predict whether an individual had a tumor mass (product of maximum tumor diameter and number of tumor nodules) smaller than 5.5 (1826 individuals, 44%) of larger (2313 subjects, 56%).

Results

Only the differences between patient actual clinical profiles and 9 reference clinical profiles out of 19 contributed significantly to the tumor mass classification. Of these, small differences (<6) from the 5 reference clinical profiles (C_1,C_3,C_6,C_8 and C_{16}) resulted in high odds for a S-phenotype tumor mass and small differences (<6) from the 4 reference clinical profiles (C_5,C_9,C_{12} and C_{18}) resulted in the high odds for L-phenotype tumor mass. This logistic regression model correctly predicted 70% of the tumor mass categories in a 10-fold cross-validation (ROC area 0.78). We used the logistic regression equation to identify 2034 patients as S-sub-group and 2105 patients as L-sub-group. The distributions of tumor mass in these 2 sub-groups had their means, (13.0 for L and 4.4 for S) significantly statistically different, $p = 10^{-240}$, t-test. The significant differences were also seen systematically throughout the L and S tumor mass distributions. The Kaplan-Meier formalism (Fig. 3) have shown with strong statistical significance that patients in the L-sub-group had a 2-4 fold greater odds of having a larger tumor. Equivalently, once the patient has been categorized in S or L-

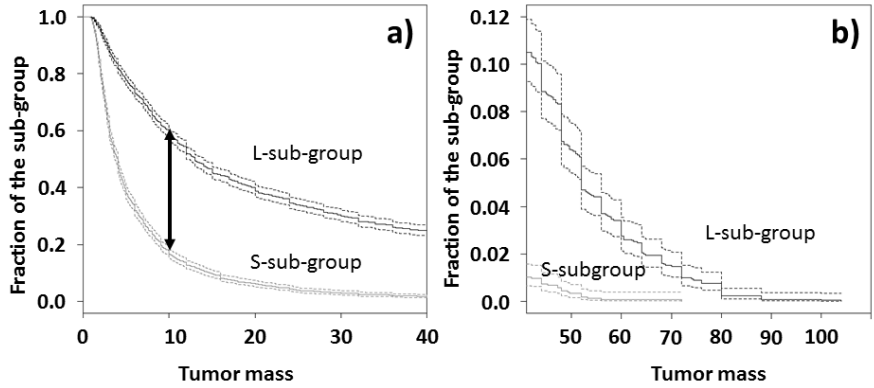

Figure 3. Modified Kaplan-Meier characterization of odds for a given the tumor mass in S- and L-subgroups. The dotted lines are 95% confidence intervals for respective odds curves. Double arrow in **a)** is 3-fold odds difference. Note different scales for the largest tumor masses in **b)**.

sub-group, then the odds of finding a given tumor mass were ~3 fold higher in the L-sub-group, compared with S-phenotype patients. This indicated that our findings are independent of specific choice of tumor mass threshold in optimization of the logistic regression classification model. The main result thus far was that liver function tests and patient demographic descriptors identified S and L-phenotypic groups with strongly statistically significant separation of their tumor masse distributions.

Logistic regression identified L-phenotype-associated reference profiles C_5, C_9, C_{12} and C_{18}, having in common high platelets/high AFP levels, accompanied by the presence of PVT and self-reported chronic alcohol consumption. The S-phenotype had 2 associated sub-groups of reference clinical profiles: C_1, C_3 and C_6 and C_8 and C_{16}. The former had in common: low platelets/low AFP and absence of PVT. The latter had in common low AST/low ALT, high albumin/high hemoglobin, low bilirubin/low INR and high platelets/high AFP.

Table 1. Summary of S- and L-associated networks of reference parameter levels

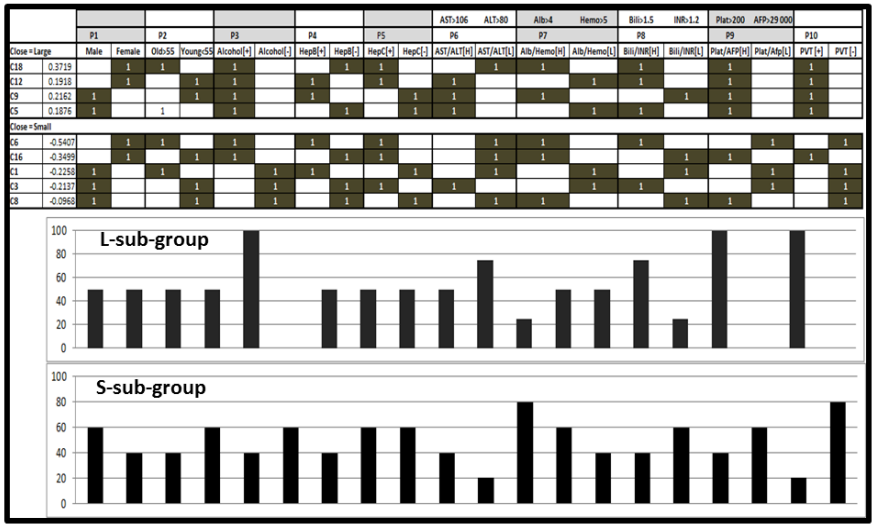

Table legend: Top two panels: columns P1-P10 correspond to 10 parts of the clinical profile shown in Fig.2. The two sub-columns indicate one of the two levels for the respective

parameters. The actual levels for given reference clinical profile are shown by "**1**" in black field. Left columns: Reference profile ID and coefficient in the classification logistic regression model. Top: reference clinical profiles associated to L-subgroup, bottom: reference clinical profiles associated to S-subgroup. Bottom two panels: percentage of commonality of all reference clinical profile levels in respective sections P1-P10.

L-phenotype associated reference profiles (C_5,C_9) were male-related and C_{12},C_{18} were female-related. C_9 described younger and C_5 described older (>55 years) patients. For the female-associated profiles, C_{12} described younger and C_{18} older patients. S-phenotype associated reference profiles C_6 (young) and C_{16} (older) were for female patients and C_1 (older), C_3 and C_8 (older) for male patients.

Trends between individual parameters and tumor mass in the S/L- sub-groups. The networked characteristic profiles for the L-sub-group are more homogeneous than those in the S-sub-group. We examined whether there were significant differences in typical parameter values for the same tumor mass that might be found in each of the S/L-sub-groups.

We used a moving average filtering (Fig.3) where any tumor mass is characterized by the average of the clinical parameter values of 61 patients with the closest tumor masses (9). We examined these trends in AFP (reflective of tumor growth) and platelet values and found increasing AFP and platelet counts with increasing tumor mass in S- and L-sub-group with different rates and magnitudes (Fig. 3a). L-sub-group displayed a pattern of AFP/platelet level oscillations that were not observed in the S- sub-group. Importantly, these L-phenotype unique oscillations were characteristic for the same tumor masses in both AFP and platelet trends (Fig. 3a)**.**

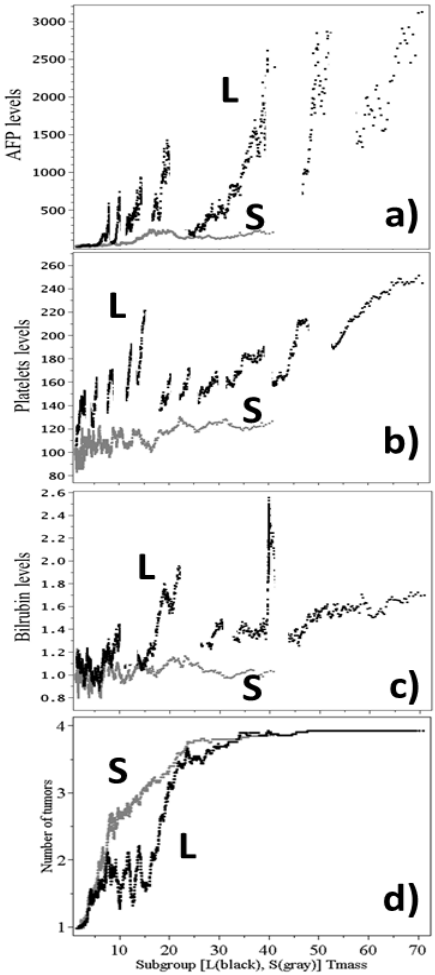

Figure 4. Typical parameter levels as function of tumor mass in S- (gray) and L-subgroups (black). Typical values are results of moving average processing (see Methods) **a)** AFP, **b)** platelets, **c)** bilirubin, **d)** number of tumors.

The analysis of typical tumor-mass-related bilirubin level changes also showed differences in the 2 sub-groups (Fig. 3b). In the S-sub-group there was a shallow bilirubin increase as the tumor mass increased. In the L-sub-group, oscillations were found below tumor mass 20, which were not seen in the S-sub-group. The oscillations in bilirubin levels in the L-phenotype cohort occurred at the same tumor masses as those in AFP and platelet values in the L-sub-group. Additionally, there was a steady increase in bilirubin levels for increasing tumor mass beyond 20.

Examination of AST/ALT trends showed that they were steady in the S-sub-group and at higher levels than in the L-sub-group in the smallest tumors of equivalent mass <30 (Fig.3c). In the L-sub-group, there was a steady increase in AST/ALT levels as tumor mass increased above 30.

The mechanisms underlying the oscillations did not seem to have an obvious explanation from clinical practice. However, possible clues came from analysis of the number of tumor nodules typical for the given tumor mass (Fig 3d). We processed the data for tumor numbers in the same way as for other parameters and we found that (Fig. 3e) oscillations in tumor numbers corresponded to spikes in the parameter trends, especially seen for tumor mass <30.

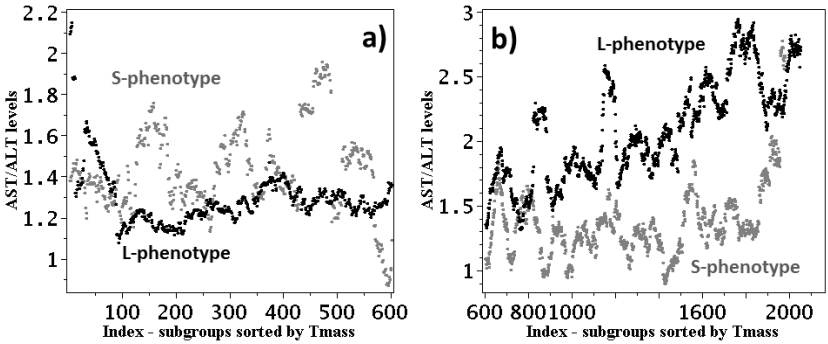

Figure 5. Typical levels of AST/ALT ratio for S- and L-subgroups, rank-ordered by tumor mass. **a)** first 600 patients with tumor masses 1-2.5 in S-subgroup and 1-7 in L-subgroup. **b)** remaining patients with larger tumor masses. Gray – S-subgroup, black – L-subgroup

Discussion

A database was constructed from a large number of newly diagnosed HCC patients, was used to characterize patient profiles, from which developed an algorithm that classified the patients into S- and L-subgroups. This characterization with significant difference in tumor masses was not obtained using raw data. Thus, the added information about trends and inter-dependencies of parameter values in a total parameter context was crucial for successful classification. The liver function and blood parameters were treated in 4 pairs, with unique significant correlations having a direct relationship to liver properties. These pairs were intuitively inter-related through established underlying functional processes: total blood bilirubin/prothrombin time (a measure of liver function), SGOT/SGPT (a measure of liver inflammation) and AFP

/blood platelet counts (an estimate of tumor growth) (7). We therefore processed 10 components of the patient profiles (Methods). Decomposition of unified schema of 4139 profiles into 19 different reference profiles, from which only 9 contributed significantly to the assessment of tumor mass outcome revealed simpler structure of HCC clinical information.

Because the 9 reference clinical profiles were idealizations of the S and L-clinical phenotypes, an individual patient characterization involved a quantitative description of how close the actual pattern of relationships between parameter values were for an individual patient from all significant reference clinical profile patterns. In this approach, the single value of a parameter cannot change the classification. It was the majority of the parameter relationships matching the S or L-associated patterns that determined the classification. The patients with either of these 2 clinical profile patterns had distributions of their tumor masses with significantly different means. Despite these differences in means, each phenotype had a wide range of tumor masses. Nevertheless, we have shown (Fig.3) that there were always 2-4 times higher odds for larger tumors in L than in S phenotype group. The differences in the tumor mass trends in the two sub-groups were significantly separated, showing the efficiency of our approach.

In the 2 phenotype groups, there was much greater homogeneity in the characteristic parameter patterns in L than in S. The rate of change for typical parameter values per unit change of tumor mass was always significantly higher for L-phenotype patients compared to S-phenotype patients, excepting the AST/ALT ratio, which was higher in S for tumor masses below 10 than for the same size tumors in L. One possible interpretation of these observations is that in S-phenotype patients, small tumors are associated with processes producing higher levels of the inflammatory markers, AST/ALT. We hypothesize that this might reflect the inter-connectedness of hepatic inflammation with tumor growth in the small tumors in this phenotype group. In L-subgroup, the simplest explanation of parameter levels oscillations might be consideration of the number of tumor nodules that composed the tumor mass. A relationship between platelet numbers and tumor size was recently reported (7). Low platelets were interpreted to be a consequence of the portal hypertension that is

secondary to liver fibrosis. We found that most small HCCs in 2 large western cohorts occurred in the presence of thrombocytopenia, whereas the largest tumors occurred in patients with significantly higher, but normal platelet values.

By contrast, in the L-phenotype, the AST/ALT only really increased as the tumor masses became quite large. This may reflect the parenchymal liver damage that occurs when a large tumor develops and replaces underlying liver. In addition, in the L-phenotype, but not in S, several additional liver parameters showed oscillations in their typical values, as the tumor mass increased. We found a relationship between these oscillations and the numbers of tumors (Fig 3). Given the lesser association of changes in inflammatory markers in the L-phenotype, we consider that other factors, likely tumor-related, may be more important on the growth of these tumors. Such factors likely include genetic drivers of HCC cell growth. In the L-phenotype, the observed higher levels of various parameter contributions from multiple nodules to the total parameter levels could be additive.

Acknowledgement: PP was supported in part by ERC-CZ LL1201 program CORES.

References

1. Okuda, K., Nakashima, T., Kojiro, M., Kondo, Y. and Wada, K. (1989) Hepatocellular carcinoma without cirrhosis in Japanese patients. *Gastroenterology*, **97**, 140-146.
2. Venook, A.P., Papandreou, C., Furuse, J. and de Guevara, L.L. (2010) The incidence and epidemiology of hepatocellular carcinoma: a global and regional perspective. *The oncologist*, **15 Suppl 4**, 5-13.
3. Kumada, T., Toyoda, H., Kiriyama, S., Sone, Y., Tanikawa, M., Hisanaga, Y., Kanamori, A., Atsumi, H., Takagi, M., Arakawa, T. *et al.* (2010) Incidence of hepatocellular carcinoma in patients with chronic hepatitis B virus infection who have normal alanine aminotransferase values. *Journal of medical virology*, **82**, 539-545.
4. Trevisani, F., D'Intino, P.E., Caraceni, P., Pizzo, M., Stefanini, G.F., Mazziotti, A., Grazi, G.L., Gozzetti, G., Gasbarrini, G. and Bernardi, M. (1995) Etiologic factors and clinical presentation of hepatocellular carcinoma. Differences between cirrhotic and noncirrhotic Italian patients. *Cancer*, **75**, 2220-2232.
5. Rosa, J.C., Chaves, P., de Almeida, J.M. and Soares, J. (1995) [Hepatocellular carcinoma. Rare forms of presentation]. *Acta medica portuguesa*, **8**, 243-245.
6. Lok, A.S., Seeff, L.B., Morgan, T.R., di Bisceglie, A.M., Sterling, R.K., Curto, T.M., Everson, G.T., Lindsay, K.L., Lee, W.M., Bonkovsky, H.L. *et al.* (2009) Incidence of hepatocellular carcinoma and associated risk factors in hepatitis C-related advanced liver disease. *Gastroenterology*, **136**, 138-148.

7. Carr, B.I., Guerra, V. and Pancoska, P. (2012) Thrombocytopenia in relation to tumor size in patients with hepatocellular carcinoma. *Oncology*, **83**, 339-345.
8. Carr, B.I., Guerra, V., De Giorgio, M., Fagiuoli, S. and Pancoska, P. (2012) Small hepatocellular carcinomas and thrombocytopenia. *Oncology*, **83**, 331-338.
9. Lu, S.N., Wang, J.H., Liu, S.L., Hung, C.H., Chen, C.H., Tung, H.D., Chen, T.M., Huang, W.S., Lee, C.M., Chen, C.C. *et al.* (2006) Thrombocytopenia as a surrogate for cirrhosis and a marker for the identification of patients at high-risk for hepatocellular carcinoma. *Cancer*, **107**, 2212-2222.
10. Pancoska, P., Carr, B.I. and Branch, R.A. (2010) Network-based analysis of survival for unresectable hepatocellular carcinoma. *Seminars in oncology*, **37**, 170-181.
11. Hoshida, Y., Villanueva, A., Kobayashi, M., Peix, J., Chiang, D.Y., Camargo, A., Gupta, S., Moore, J., Wrobel, M.J., Lerner, J. *et al.* (2008) Gene expression in fixed tissues and outcome in hepatocellular carcinoma. *The New England journal of medicine*, **359**, 1995-2004.
12. Leonardi, G.C., Candido, S., Cervello, M., Nicolosi, D., Raiti, F., Travali, S., Spandidos, D.A. and Libra, M. (2012) The tumor microenvironment in hepatocellular carcinoma (review). *International journal of oncology*, **40**, 1733-1747.
13. Utsunomiya, T., Shimada, M., Imura, S., Morine, Y., Ikemoto, T. and Mori, M. (2010) Molecular signatures of noncancerous liver tissue can predict the risk for late recurrence of hepatocellular carcinoma. *Journal of gastroenterology*, **45**, 146-152.
14. Diestel, R. (2010) *Graph theory*. 4th ed. Springer, Heidelberg ; New York.
15. Yang, J.D., Sun, Z., Hu, C., Lai, J., Dove, R., Nakamura, I., Lee, J.S., Thorgeirsson, S.S., Kang, K.J., Chu, I.S. *et al.* (2011) Sulfatase 1 and sulfatase 2 in hepatocellular carcinoma: associated signaling pathways, tumor phenotypes, and survival. *Genes, chromosomes & cancer*, **50**, 122-135.
16. Tanaka, K., Sakai, H., Hashizume, M. and Hirohata, T. (2000) Serum testosterone:estradiol ratio and the development of hepatocellular carcinoma among male cirrhotic patients. *Cancer research*, **60**, 5106-5110.
17. Carr, B.I., Pancoska, P. and Branch, R.A. (2010) HCC in older patients. *Digestive diseases and sciences*, **55**, 3584-3590.
18. Hall, M., Eibe, F., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I.H. (2009) The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, **11**.

Supplementary Material

Phenotypic categorization and profiles of Small and Large hepatocellular carcinomas

Petr Pancoska[1] and Brian Carr[2] and Shenh-Nan Lu[3]

As there were no missing data in the original dataset, all statistical computations were done considering the full set of 4139 data points for all subjects.

Step 1: In the first step, we considered correlations between clinical (blood liver function test and hematological) parameters. This has two important ramifications. First, in a formal, statistical sense any significant inter-correlation between individual parameter values represents a simplification of the analysis complexity (dimensionality reduction, we take into account that any two significantly correlated parameter pairs carry similar information). Second, any such significant correlation might indicate a functional relationship of the underlying processes. Therefore, building the NPS transformation of a study data, which explicitly considers such relationships, allows a simplified and more clinically intuitive processing of the extensive patient information such as in this study.

For logarithmically transformed values of eight plasma hematological and liver test parameters (AFP, total bilirubin, ALT, AST, INR, albumin, platelets, hemoglobin) we computed all $(8^2-8)/2 = 28$ pairwise correlations and characterized the extent of (linear) proportionality between all possible parameter pairs by 28 linear regression coefficients. These 28 correlation coefficients were arranged into an 8x8 symmetrical correlation matrix. This matrix represents at the same time an adjacency matrix, defining a complete weighted graph (clique). In this clique, 8 liver test parameters are completely connected by strengths of their 28 co-linearities, quantified by the pairwise correlation coefficients.

We then used maximal cut graph theory theorem (1,2), defining a non-parametric algorithm, finding in this complete weighted correlation graph the set of 8/2 = four liver test parameter pairings, that are unique: all correlations in these pairs are individually significant (using the statistical significance test for the correlation coefficient on 0.99

significance level.) More importantly, these four selected pairs also had the absolutely largest sum of their respective 4 pairwise correlation coefficients out of all possible 20, 475 selections of such 4 pairs. Because no other information but the complete correlation matrix between the full sets of actual parameter values was used in this analysis step, the resulting unique pairing represented very important information encoded in the actual data set. It was therefore interesting to formulate the possible functional mechanisms and factors underlying these four highly significantly co-linear parameter pairings.

Step 2. The next step in the construction of the NPS, was transformation of the original patient clinical data and their conversion into a common form of "levels". This step was necessary to be able to unify the demographic (categorical) parameters with liver function (real value) parameters and allow for consideration of their inter-relationships within a common, explicit, quantitative but still directly clinically interpretable framework. Considering the established practice in HCC diagnostic evaluation, we found that a common approach to determine levels of each individual parameter can be tercile-based dichotomization. For gender, reported alcoholism, evidence for hepatitis B and C and presence or absence of portal vein thrombosis (PVT) the dichotomization was natural. For the rest of parameters, we tested several alternative dichotomizations (50%:50%, quartiles) but found that tercile dichotomization with 2/3 of patients with the lowest parameter levels designated as "Low" phenotype and 1/3 of patients with the highest parameter levels designated as "High" phenotype was optimal for further processing (see below). Also, the resulting thresholds, splitting the 4139 patients into 2759 subjects in "low" sub-group and 1380 patients into a "high" sub-group were comparable with values used in established HCC classification schemes. For age the "old" tercile was separated from the lower 2 "young" terciles by 55 years. For the four parameter pairs that were found as significantly correlated in step 1, we used two-threshold method, as shown in Fig. 1

For these pairs, the upper tercile is defined by the 2 individual thresholds of the individual parameters that separate High from Low values. This resulted in clinically familiar value cutoffs, such as bilirubin of 1.5 mg/dl, AST 200 IU/l and ALT 105 IU/l.

Another advantage of this approach, seen in text Fig. 1, is that the outliers from these trends, with the very high values, do not impact significantly the values of the tercile-based thresholds. We have also shown that inclusion or exclusion of these outliers in all trends, both individually and simultaneously from analysis did not change the final NPS representation of patient's data and, more importantly, did not affect the essential steps in further processing. Thus, our NPS approach is very robust and optimally sensitive to coherent as well as insensitive to stochastic components of raw HCC data.

Step 3. Using actual data for each patient, an individual clinical profile was created by connecting the actual parameter high, low, + and - levels for that individual (text Fig. 2). We then created an individual patient profile for each of the 4139 patients. All these individual profiles were then unified into a single schema. In this unified cohort of profiles, each line in each patient's profile, representing the relationship between the actual parameter levels in individual profiles is counted as a contribution of that specific individual parameter relationship to collective NPS schema, In this way, we incorporated into the study schema the new information about frequencies of co-occurrences of all respective parameter levels in a simple intuitive way. The co-occurrence frequencies

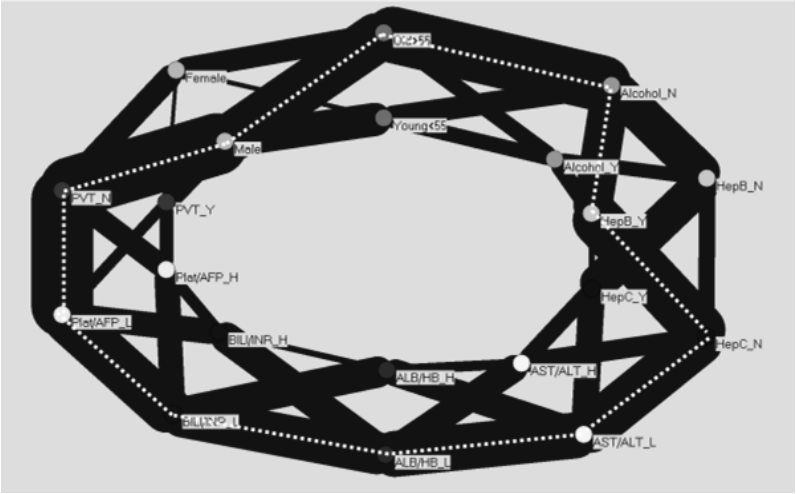

Figure S1. Single schema, generated by unifying all 4139 individual patient clinical profiles from the complete database. Line thicknesses are proportional to co-occurrence frequencies of various parameter levels. This graph is the simplest possible representation of full context network of all parameter level relationships in the study. White dotted line follows the most frequent parameter level combination.

are represented graphically by the thickness of the line connections between levels of each parameter (Fig. S1). Thus, as an illustrative example, we can directly infer from Fig. S1 (following the dotted edges) that the majority of the cohort has low platelets, low AFP levels and no portal vein thrombus and also have more males than females. The full information of text Table 1 is represented by this schema. The next step that we took ensured that the data transformation was not dependent on the order in which the parameters inter-connected in this unified profile.

Step 4. Using a greedy algorithm (extracting sequentially the co-occurrence relationship profiles that has the highest frequency in the dataset and repeating this process until there was no relationship left), the unified profile was decomposed into 19 reference

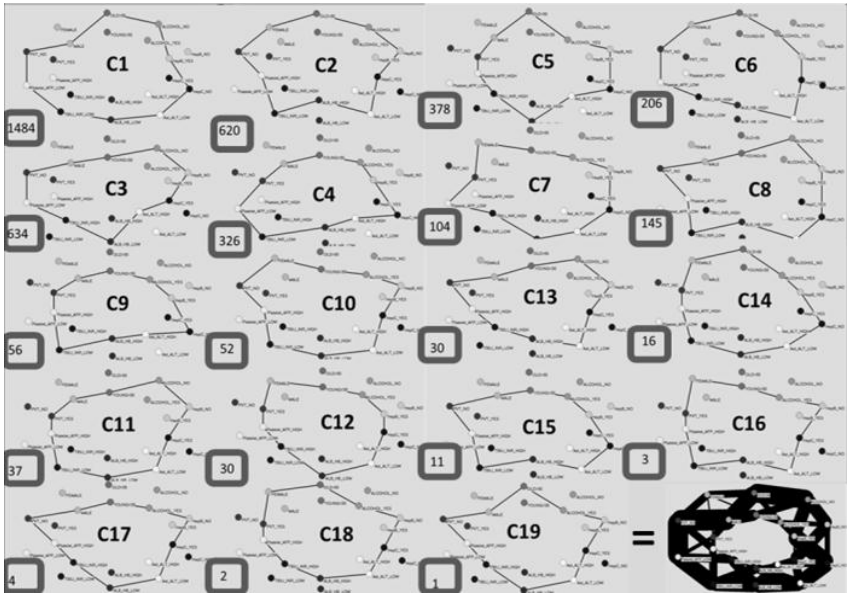

Figure S2. Full decomposition of the single schema of all networked context relationships in the study data into reference clinical profiles C1-C1. Numbers in rectangles are multiplicities of every reference clinical profile in the decomposition. By dividing these multiplicities by 4139, the total number of patients in the study, we obtain for each reference profile the relative frequencies of its occurrence in the HCC study.

profiles, in such a way that all co-occurrence frequencies between all the parameter levels in a reference profile were identical. This generated unique, data-determined reference clinical profiles with equal frequencies of co-occurrences of all parameter pair levels – see Fig. S2.

The identity of co-occurrence frequencies also ensured that the resulting reference profiles and the data transformation, that uses them as "triangulation points" in the space of all clinical data relationships, are independent of parameter ordering in the scheme. From the statistical point of view, it can be shown that the pair-wise relationships that constitute the reference clinical profiles are also unique by being independent of each other.

The 4139 individual profiles were then compared in turn to each of the 19 idealized reference profiles and the differences were recorded between the relationships in an individual actual patient profile to those in each of the 19 reference profiles. We had 10 parameters or their pairwise trend constructs in the processed data and also on each reference profile (4 liver function blood pairs and 6 other parameters). Therefore, each individual profile could differ in 0, 2, 3, 4, 5, 6, 7, 8, 9 or 10 relationships from those, recorded in each reference clinical profile. The number of these mismatches between an individual patient clinical profile and an idealized reference profile defined the "distance" of the actual individual patient's clinical state from the idealized reference clinical state. These 19 distances localized the precise position of each patient in the clinical landscape of HCC. This transformation of raw data into 19 distances contained information on trends and co-occurrences in terms of their clustering by closeness to the reference clinical profiles, serving as the "triangulation points" in the clinical data relationship landscape. A formal advantage for further statistical processing was also observed: namely, that the distributions of 19 distance vector components for all patients were Gaussian, while raw data often exhibited multimodal, nonsymmetrical histograms/distributions.

Step 5. We next used logistic multiple regression, using 19 distances of the individual patient's clinical profiles from each of the 19 reference clinical profiles as independent variables, to predict whether an individual had a tumor mass larger or smaller than 5.5.

Using the variable selection algorithm for logistic multiple regression (a combination of forward and backward variable selection methods (SigmaPlot 11)) revealed that not all 19 distances were relevant for identifying tumor mass and only 9 of them contributed significantly. Of these, small distances from 5 reference clinical profiles (C_1,C_3,C_6,C_8 and C_{16}) resulted in high odds for a 'S' phenotype tumor mass, while small distances from 4 reference clinical profiles (C_5,C_9,C_{12} and C_{18}) resulted in high odds for 'L' phenotype tumor mass. This optimal logistic regression model correctly predicted 70% of the tumor mass categories in a 10-fold cross-validation (ROC area was 0.78).

With this ability to recognize two significantly different tumor phenotypes, we used the logistic regression equation as a tool to identify these two clinical phenotypes, S and L. There were 2034 patients identified by that optimized predictive regression model with S-phenotype clinical profiles and 2105 patients as having L-phenotype clinical profiles. When the actual distributions of tumor mass in these 2 groups were compared, their means (which were 13.0 for L and 4.4 for S) were significantly statistically different using the standard t-test on two samples (unequal variances), $p = 10^{-240}$. As the classification of all patients into these two clinical phenotype sub-groups is based on closeness to the respective reference clinical profiles in the S and L category, the closer an individual patient profile was to all the 5 S-phenotype associated or all the 4 L-phenotype associated idealized reference clinical profiles, the larger the odds that patient's tumor mass would smaller or larger.

For more detailed analysis of the tumor mass differences in the two different phenotype groups, we used the R-implementation of Kaplan-Meier analysis in the "survival" package. The motivation of this approach was that the large number of patients in our data set (and in the two S- and L-phenotype sub-categories) allowed us to consider their tumor masses as snapshots of the growth of an idealized tumor in different clinical contexts. Thus, the respective tumor masses in the two clinical phenotype sub-categories were considered as independent variables in KM processing. In this model, no censoring is required. Results indicated that the odds to find the same tumor masses in the S- and L-phenotype groups are significantly different with separation close to 10x of the 95% confidence intervals.

For moving average processing of the dependencies of the clinical parameters on the tumor masses, we used a dedicated program (Maple 12). Patients were rank-ordered separately in the S- and L-phenotype subgroups. The first tumor mass that is characterized in the presented plots was the tumor mass in the 31st position in these rank-orderings. The program retrieved and averaged the clinical parameter values and the tumor masses of patients with 30 closest smaller tumor masses and 30 larger tumor masses to the mass 31 in the rank ordering. These two typical values were plotted as the first point of the processed relationship. The window was then moved to the 32 tumor mass in the rank ordering and the process continued throughout the respective tumor mass intervals in the S- and L-phenotype sub-categories.

REFERENCES

1. Diestel, R. (2010) *Graph theory*. 4th ed. Springer, Heidelberg ; New York.
2. Matoušek, J.i. and Nešetřil, J. (2009) *Invitation to discrete mathematics*. 2nd ed. Oxford University Press, Oxford ; New York.