Risk stratification systems for surgically treated localized primary Gastrointestinal Stromal Tumors (GIST). Review of literature and comparison of the three prognostic criteria: MSKCC Nomogramm, NIH-Fletcher and AFIP-Miettinen

Giulio Belfiori*, Massimo Sartelli**, Luca Cardinali*, Cristian Tranà**, Raffaella Bracci***, Rosaria Gesuita****, Cristina Marmorale*

*Division of General Surgery, Ospedali Riuniti, Polytechnic University of Marche, Ancona, Italy
**Department of Surgery, Macerata Hospital, Macerata, Italy
***Clinica di Oncologia Medica, Ospedali Riuniti, Polytechnic University of Marche, Ancona
****Center of Epidemiology, Biostatistics and Medical Information Technology, Department of Biomedical Sciences and Public Health, School of Medicine, Polytechnic University of Marche, Ancona, Italy

Pervenuto in Redazione Maggio 2014. Accettato per la pubblicazione Settembre 2014
Correspondence to: Giulio Belfiori MD, P.zza G. Galilei 9, 60100 Collemarino, Ancona, Italy (e-mail: giulio_belfiori@live.it)

Risk stratification system for surgically treated localized primary Gastrointestinal Stromal Tumors (GIST). Review of literature and comparison of the three prognostic criteria: MSKCC Nomogram, NIH-Fletcher and AFIP-Miettinen

PURPOSE: The discovery of Imatinib mesylate (Gleevec®) has revolutionized the treatment of GIST, increasing disease-free survival (DFS) after complete surgical resection of a primary localized GIST and extending overall survival in metastatic disease. The definition of an accurate prognostic system is critical for the therapeutic decision making process. In literature, there are three main prognostic criteria F/NIH consensus, AFIP standards and modified NIH standards. In recent years were added various risk identification methods applying mathematical calculation model, including MSKCC risk nomogram, Rossi nomogram and Joensuu high Hotline Dengjun. Despite all these attempts, it seems that the recurrence risk probability still cannot be predicted accurately. The aim of our study was to assess and compare the real ability of these prognostic instruments in our single-centre clinical experience, and to define if the use of the MSKCC nomogram can bring benefits in the therapeutic decision.

METHODS: All data regarding 37 GISTs who underwent surgical resection from 1996 to 2011 in our institution were retrospectively reviewed. We selected only primary GIST without metastatic disease who underwent a radical resection (R0) but no other therapy. The literature data concerning GISTs prognostication criteria were reviewed. All patients were classified according to the three prognostic criteria (NIH, AFIP and Nomogram MSKCC) and the three instruments were compared with the Kaplan-Meier method. Then we compared the three criteria for their c-index value and we assessed the performance of the nomogram with the calibration test.

RESULTS: We observed 9 recurrences (24%) with an average time to relapse of 43 months; the median follow-up was 65 months. In the study selected sample occurred 5 relapses. The probability of relapsing after radical surgery was 7.9% (95% CI 0 - 17.3) at 2 years and 13.3% at 5 years (95% CI 0 - 26.4). The C-Index of the three risk assessment tools was 0.93 (95% CI 0.83-1) for the Nomogram at 5 years, 0.86 (95% CI 0.76-0.95) for the NIH risk criteria and 0.88 (95% CI 0.74-1) for the AFIP risk criteria. The calibration analysis of the nomogram showed an overestimating trend both at 2 and 5 years.

CONCLUSION: MSKCC nomogram seems to perform better than NIH, NIH modified and AFIP in our sample and can be used in clinical practice to predict the risk of recurrence, being especially helpful for the therapeutic decision making since it is at the same time simple to use and accurate. As showed from calibration, MSKCC doesn’t seem to neglect relapses, even though it is not impeccable in predicting the DFS. Among the 2 older criteria AFIP was more precise than NIH, but considering size in not linear way represented a limit in comparison with the MSKCC Nomogram. All the three risk assessment tools considered are capable to predict recurrence in high-risk GISTs while they performed worse in those with lower risk. MSKCC nomogram main limit remains the not linear consideration of mitotic count.

KEY WORDS: GIST, Localized, Prognostic criteria, Recurrence, Surgery
Background

GIST (Gastrointestinal Stromal Tumours) is an heterogeneous group of neoplasms of the gastrointestinal tract that originates from the precursors of the interstitial cells of Cajal 1,3,4. Even though Gastrointestinal stromal tumors (GIST) are the most common mesenchymal neoplasm of the intestinal tract (80% of all mesenchymal tumours), they represent only 1% of all gastrointestinal tumours 2. GISTs occur more frequently in the stomach (65%), in the small bowel (25%), in the colon-rectum (10-15%) and in the esophagus (less than 1%) 3. We can also find in only 1-2% of cases 4 a type of primitive GIST not associated to gastrointestinal tract, called E-GIST (extra-gastrointestinal GIST). The most important characteristic of GIST is the expression of the protein c-KIT , that can be shown, by immunohistochemical assay, using the antigen CD117 5,34. More than 80% of GIST are KIT positive at immunohistochemistry. GIST driving mechanisms of growth is due to a mutation of KIT gene that codifies a tyrosine kinase membrane receptor that favors tumor growth 30,34. More than 80% of GISTs present a mutation of c-KIT, 10% a mutation of PDGFRA 6 (a receptor very similar to KIT), and the other 5-10% doesn't present any mutation and is called for this reason “wild-type” 7. GISTs have an incidence of 10-15 cases/million of persons per year and a prevalence of 129 cases per million of persons. This proportion of prevalence to incidence is justified from the long clinical course of this disease , that is approximately of 10-15 years. Despite the existence of GIST has been hypothesized before, they were described for the first time from Mazur and Clark in 1983 1, and they were widely recognized from the scientific community only after the discovery of the c-KIT mutation from Hirota in 1998 8. Only 4 years later (February 2002) in USA Imatinib use was approved for treatment of metastatic GISTs. In following years the use of Imatinib was extended in adjuvant and neoadjuvant setting30,32. So the point is that GISTs have the worst. The AFIP system is separated in several classes, any class has its probability of recurrence according to the observation made in the AFIP's population. The third most used risk assessment tool is the MSKCC Nomogram, published in 2009 17, it assigns for any of the factors mentioned a score , and the sum of the three scores corresponds to a prediction of 2-year and 5-year recurrence free survival (RFS) 17. According to Gold et. al, the nomogram provided a better prediction of the likelihood of recurrence for individual patients as validated in three databases MSKCC (n=127), GEIS (n=212), Mayo clinic (n=148). This difference wasn’t statistical significant. According to these Authors the difference with the commonly used staging criteria (AFIP and NIH) is that they stratify patients into a few broad groups, instead of considering variables in a linear way, as the Nomogram does. The limit of this Nomogram is that it doesn’t considers mitotic count in a linear way. An attempt to overcome this limit was made by an Italian group in 2011 18. They tried to develop a new Nomogram that considers the parameters in linear mode. This nomogram was assessed in a sample of 526 patients, from 25 Italian institutes, and provides for each patient the overall survival(OS) at 10 years. The authors 18 reported that they calculated the OS because they didn't have complete and accurate information about the recurrences. So in comparison with the MSKCC this new Nomogram was designed on a larger sample , but it calculates the OS , that is less important than RFS in the therapeutic decision-making process of GIST because they have a long clinical course. In 2012 Joensuu et al. 19 suggested a new non linear risk assessment system based on prognostic maps and compared it with the previous systems. This scheme was based on tumor site, rupture and on size and mitotic count in a continuous non linear way. These authors 19 demonstrated that these maps are appropriate for the estimation of individualized outcomes but they suggested that the modified NIH classification is the best cri-
Methods

All patient surgically treated in our Department (Clinica Chirurgica, Umberto I Hospital, University of Ancona, Italy) from 1996 until 2011 with an histological diagnosis of GIST. All reports were viewed by expert pathologists in the field of GISTs (I. B). The histological diagnosis was made using immunohistochemical staining for CD117 and/or DOG-1 and, in case of doubt it was made a molecular biology analysis for kit or PDGFRα mutation. The mitotic index was determined by counting the number of mitotic figures per 50 high power fields (HPFs). Size measurements were performed by the pathologist, either before or after formalin fixation. All data were collected by the patients clinical notes and by contacting them by phone or examining them during the follow up. During follow-up, we analyzed the incidence of disease recurrence. RFS was defined as the time from patient surgery to the development of tumor recurrence. A database was created in order to analyze all the data. The database included pathology information (such as site, size, mitotic count, histological type, histological grading, margins of resection, percentage of necrosis, intra-operative rupture, immunohistochemistry, molecular analysis) and clinical characteristics (symptoms, date of diagnosis, type of metastasis, date of recurrence, other diseases, kind of therapy). Also in this database, patients were classified according to their tumors risk of recurrence calculated with the NIH and AFIP criteria. We also classified the patient according to the modified NIH criteria, but in our sample coincided with the standard NIH classification. So the study data referring for NIH are overlapping with the NIH modified one. Any class of AFIP and NIH corresponds to a probability of recurrence reported in literature and validated in population studies. At the same time, we calculated the MSKCC nomogram scores for each patient, every score corresponding to a probability of RFS.

The follow up of patients was performed with chest and abdominal computer tomography (CT) every 6 months for patients with intermediate/high risk and every year for patients with very low/low risk. However, CT scans were repeated earlier whenever clinically indicated, on the discretion of oncologists. We had a population of 37 GIST, from this population of patients we composed a champ including only the patients with a primary localized, not metastatic GIST at diagnosis who underwent a radical surgery (R0). No patient had adjuvant or neoadjuvant therapy with imatinib (still not in indication), some patients had imatinib treatment if relapsing with metastatic disease.

Statistical Tools

The probability of recurrence was estimated using the survival analysis, with the Kaplan-Meier method. The probability was assessed by stratifying the observations for some factors of interest: tumors dimension, mitotic count, tumors site according to NIH and AFIP criteria. The comparison between the curves was performed using the Log Rank test. Multivariable analysis wasn’t possible because of the small sample size. The ability to predict the recurrence of the Nomogram (at 5 years), of NIH (long time) and AFIP (long time) was evaluated using the estimation of the C-Index (concordance index) with a confidence interval of 95%. We assumed that the time of 5 years can be considered as a long time observation of GISTs recurrence, since previous studies showed that from 60% to 100% of GISTs relapse within the first 2 years. RFS (Recurrence Free Survival), performed by the Nomogram is defined as the complement to one of the relapse probability at 2 and 5 years from the surgery. The C-index is the area under the ROC curve (Receiving Operating Characteristic Curve). Furthermore, an analysis of the calibration of the Nomogram was conducted by comparing the probability not developing recurrence obtained by stratifying subjects according to the probability predicted by the nomogram of Recurrence free survival (RFS) at 2 and 5 years after surgery. For all analyses the significance level was set at 5%.
Results

ANALYSIS OF OUR DATABASE

From 1996 until 2011, thirty-seven GIST patients underwent a surgical operation in our clinic. The mean follow-up of these patients was 65 months. 51% of these patients were male and 49% were female. The mean age at diagnosis was 64.25 years. Only the 19% was metastatic at diagnosis while the remaining 81% had localized disease. Forty percent of the diagnoses were incidental while the remaining 60% came to the attention of physicians because of symptoms. GISTs site was in 51% of the cases stomach, in 32% small bowel, in 10% duodenum and in 7% rectum. Microscopically, 62% showed a spindle morphology, 14% of cases were epithelioid, and 24% of cases were mixed. The 43% of GIST were high grade and 16% presented intrallesional necrosis. Almost all patients underwent a radical surgery; only 4 patients had a positive margins (R1) and in 3 cases an intraoperative rupture occurred. Regarding the risk of recurrence, according to the NIH-FLETCHER criteria, we can divide them into high-risk 45% (17), intermediate risk 16% (6), and very low risk 23% (8). Until now 8 patients (20%) died but only 4 died of the disease. The percentage of patients with recurrence after surgery was 24% (9) (2 of them underwent an R2 margin resection since they had peritoneal metastases at diagnosis). Overall, 11 patients (30%) have metastasized: 5 patients (45%) at liver, 2 (18%) to the peritoneum and 4 (37%) to both (liver and peritoneum). Five patients (45%) underwent surgery for metastases.

ANALYSIS OF STUDY SAMPLE

The selected sample consists of 27 patients with primary localized GIST at diagnosis, who didn't underwent adjuvant or neoadjuvant therapy with imatinib until they relapsed. Fourteen were males (51%) and 13 females (49%). Their mean follow-up was 68 months. The average age of patients at diagnosis was 65.5 years. Regarding their localization: 16 (60%) were gastric GIST, 9 ileal/jejunal (33%) and 2 duodenal (7%). Overall 6 patients died and, 3 of them died of disease. There were five recurrences (18%), including 4 with an average time...
of less than 5 years (46.4 months). We proceeded classifying these patients according to the three prognostic tool: NIH, NIH modified, AFIP and MSKCC Nomogram. The NIH standard and NIH modified distinguish 4 risk classes, that are overlapping: high risk n= 9 (33%), intermediate risk n= 3 (11%), low risk n= 7 (26%) and very low risk n= 8 (30%). Instead according to the AFIP criteria we identified 5 classes: high risk n= 6 (22%), moderate risk n= 5 (19%), low risk n= 4 (15%), very low risk n= 3 (11%), no risk n= 8 (30%) and a finally a class where the risk is unknown n= 1 (3%). All relapses occurred in high risk group, only 1 patient with a jejunal GIST with 5cm diameter and mitotic count equal to 3, classified as NIH low risk (2,4% risk of recurrence) and AFIP risk category 2 (low risk - 4,3% risk of recurrence) with a Nomogram score equal to 70 (so 75% of RFS at 5 years) experienced a recurrence 108 months after radical surgery. The results of the Nomogram are not continuous data and are divided into classes that correspond to a percentage of RFS which was calculated for each patient individually. The following graphic (Fig. 1) shows the results of the survival analysis with the Kaplan-Meier method. The probability of relapsing after radical surgery was 7.9% (95% CI 0 - 17.3) at 2 years and 13.3% at 5 years (95% CI 0 - 26.4).

**STATISTICAL ANALYSIS**

The survival analysis showed a statistically significant difference in the probability of developing the relapse according to the size of the tumor. The probability of relapse was significantly higher for subjects with tumor...
size larger than 5 cm (p = 0.017). The probability of relapse was significantly more likely for subjects with a number of mitoses > 5 (p = 0.001). No statistically significant difference was found when the probability of developing a recurrence was evaluated as a function of the site of the tumor. The probability of relapse was significantly greater for subjects with a high level risk according to the AFIP criteria.

The predictive ability of the MSKCC nomogram, measured by the C-Index and evaluated in all subjects was equal to 0.9 (95% CI 0.74-1) and 0.93 (95% CI 0.83-1) respectively for the score of the nomogram to 2 and 5 years. The C-Index for the NIH risk criteria was 0.86 (95% CI 0.76-0.95) and for the AFIP risk criteria was 0.88 (95% CI 0.74-1). Figure 7 shows the values of C-Index and the confidence intervals to 95% respectively of the whole sample instead of the NIH that were 33% and 73% respectively (Fig. 1). We observed in our sample a high accuracy of both criteria, NIH and AFIP in predicting the recurrence in high-risk classes. In particular, we compared the class of higher risk of the 2 systems with the remaining classes of each system (see Fig 5 and 6 for NIH and AFIP). In both graphs, has been shown a statistically significant difference (p < 0.001) in the occurrence of relapse in favor of the high risk classes of both systems. But the percentage of patients classified at high risk for the classification AFIP were only the 22% (6) of the whole sample instead of the NIH that were 33% of total (9). From these data we can assume that the class of high-risk for the AFIP criteria is probably is more selective. On the other hand, we must observe that both these criteria have in common some limits which concerns small lesions, less than three centimeter, that are frequent, but not without risks (they can give metastases and be fatal).9-11 Another example of limitation of these prediction tools is the cut-off of 5 mitoses, which poses a clinical problem, related to the significance of a single mitosis (from 5 to 6), which can radically alter the risk of recurrence calculation and consequently the indication for adjuvant treatment. This can't be ignored also considering the greater importance of the mitotic count on prognosis as evidenced by a 2006 study of Bearzi et al.27 of 158 cases and Miettinen et al.28 in 2004.

The MSKCC-nomogram tries to overcome these limits and to substitute these rigid schemes with non-dogmatic parameters and has a simple clinical use. We tried to verify the accuracy of the nomogram in our sample and compare it with that of the other two (NIH and AFIP). In order to compare the accuracy in predicting the recurrence of the three prognostic systems we calculated the c-index for each one. For the Nomogram the c-index was calculated only for the percentage of RFS at 5 years, this time it was considered sufficiently extended for the terms of RFS and survival but also adverse effects. So the Hot topic is to separate the subject of high risk patient who are likely to benefit from adjuvant therapy from those who will do just as well without it. Actually patients are selected according to the risk of recurrence of their disease. As previously said this risk depends on three main factors that are well defined.36 These factors are also confirmed in our sample in which there is a statistically significant difference in the occurrence of relapse in GIST larger than 5 cm (p = 0.001) and in GIST that have more than 5 mitosis (p = 0.017) per 50 HPF. Analyzing the correlation between site and relapse, we found a controversial result as we revealed in our sample a greater number of relapses in gastric GISTs (Fig. 4). This observation, isn't statistically significant (p = 0.27), and is probably due to the limited size of our sample and to the presence in it of a certain number of large, histologically epitheloid and with high mitotic count gastric GISTs. In our sample, the probability of recurrence at 2 and 5 years was 7.9% and 13.3% respectively (Fig. 1). We observed in our sample a high accuracy of both criteria, NIH and AFIP in predicting the recurrence in high-risk classes. In particular, we compared the class of higher risk of the 2 systems with the remaining classes of each system (see Fig 5 and 6 for NIH and AFIP). In both graphs, has been shown a statistically significant difference (p < 0.001) in the occurrence of relapse in favor of the high risk classes of both systems. But the percentage of patients classified at high risk for the classification AFIP were only the 22% (6) of the whole sample instead of the NIH that were 33% of total (9). From these data we can assume that the class of high-risk for the AFIP criteria is probably is more selective. On the other hand, we must observe that both these criteria have in common some limits which concerns small lesions, less than three centimeter, that are frequent, but not without risks (they can give metastases and be fatal).9-11 Another example of limitation of these prediction tools is the cut-off of 5 mitoses, which poses a clinical problem, related to the significance of a single mitosis (from 5 to 6), which can radically alter the risk of recurrence calculation and consequently the indication for adjuvant treatment. This can't be ignored also considering the greater importance of the mitotic count on prognosis as evidenced by a 2006 study of Bearzi et al.27 of 158 cases and Miettinen et al.28 in 2004.

The MSKCC-nomogram tries to overcome these limits and to substitute these rigid schemes with non-dogmatic parameters and has a simple clinical use. We tried to verify the accuracy of the nomogram in our sample and compare it with that of the other two (NIH and AFIP). In order to compare the accuracy in predicting the recurrence of the three prognostic systems we calculated the c-index for each one. For the Nomogram the c-index was calculated only for the percentage of RFS at 5 years, this time it was considered sufficiently extended for the
comparison with the other two systems that are based on long-term follow-up data studies (see methods). As can be seen in Fig. 7, even though the confidence intervals are almost overlapping (especially for the AFIP and the nomogram), in our sample we observed a greater ability of the nomogram to predict diseases relapse. This observation reinforces the Authors’ one, that reported that the nomogram has a not statistically significant difference with AFIP but resulted to have a higher accuracy in comparison with AFIP too. The same observation is also made by Naoki Tianimine too, in 2012 in a single centre study of 60 patients with 10% of recurrences. To further investigate the ability of the nomogram we also carried out the analysis of the calibration at 2 at 5 years, as can be seen in the Figure 8 and 9 respectively. This analysis shows an overestimation of recurrence of events by the nomogram both at 2 and 5 years. It also identifies an event which falls precisely above the bisector of the axes, to signify that the Nomogram is ideal. This figure is, however, most likely due to the effect of case, since the small sample size. The overestimation of the calibration data is not certain, but even though the nomogram probably isn’t optimal in calculating the RFS it rarely neglects the prediction of a relapse. There is a case in our experience that constitutes an example of the advantage of nomogram towards NIH and AFIP; is a Jejunal GIST patient with a 4.9 cm tumor with a mitotic count equal to 3, that relapsed after 108 months. According to the three systems, his relapse risk were 2.4% for NIH, 4.3% for AFIP and 25% within 5 years (RFS=75) for the nomogram. So we can deduce that the MSKCC nomogram use in clinical practice is safe and precise as well as convenient and easy. This Nomogram main limitation is that although it uses a linear classification for size, it uses the same dichotomic classification, of AFIP and NIH , for the mitotic count. It has been proven that when we give the right importance to the mitotic count, the size of the lesion (especially small intestine versus stomach) loses its statistical significance. In order to overcome this problem there are already ongoing studies for nomograms that consider the mitotic count as a linear parameter, as it is in real life. An example is the Rossi nomogram that considers the mitotic count in a continuous way. We didn’t validate this system in our sample because it calculates the overall survival that is less important than the RFS in a pathology with such a long clinical course as GIST. We didn’t consider the Joensuu high Hotline Dengjun while it doesn’t provide advantages in accuracy despite is more complicated to use than the others systems. To sum up we can state that the MSKCC Nomogram is a safe and efficacious tool for the stratification of GISTs risk of recurrence in our ordinary clinical practice. For the future we expect new prognostic schemes that use the mitotic count in a linear way and assigns to each prognostic factor the adequate importance, especially for the mitotic count.

Riassunto

L’avvento dell’Imatinib mesilato (glivec) ha rivoluzionato la terapia dei GIST, apportando un aumento della sopravvivenza libera da malattia dopo resezione chirurgica completa di un GIST a localizzazione primitiva (RFS: Recurrence Free Survival). La definizione di un sistema prognostico accurato è fondamentale per decidere quali pazienti sottoporre a tale trattamento. In letteratura, esistono attualmente vari sistemi prognostici di riferimento in grado di predire la probabilità di recidiva, tra cui: NIH-FLETCHER, AFIP-MIETTINEN standard e modificato. A questi che sono i più diffusamente utilizzati, di recente si sono aggiunti altri metodi che utilizzano modelli matematici o no, come il Nomogramma del MSKCC, Nomogramma di Rossi ed il Joensuu high hotline Degjin. Nonostante tutti questi tentativi la storia naturale dei GIST rimane ancora non completamente nota e controversa e non è ancora possibile predire le recidive con una accuratezza assoluta. Lo scopo del nostrostudio è stato quello di trovare quale sistema è più accurato e pratico per essere utilizzato nella nostra pratica clinica. Particolare attenzione è stata posta al Nomogramma del MSKCC, che è stato per primo confrontato con i NIH-Fletcher ed AFIP-Miettinen. Sono stati analizzati retrospektivamente i dati riguardanti 37 GIST operati presso il nostro istituto dal 2002 al 2012 e da questi sono stati selezionati 27 GIST a localizzazione primitiva, completamente resecati e non trattati con l’imatinib ne prima ne dopo l’intervento, sui quali è stato eseguito il confronto. Le conclusioni sono state che il nomogramma MSKCC è un metodo prognostico pratico, sicuro e valido, probabilmente più del NIH e AFIP e può essere utilizzato nella pratica clinica per predire il rischio di recidiva, specialmente nella pianificazione della strategia terapeutica, anche se non è un metodo ottimale per calcolare il tempo di sopravvivenza libera da recidiva. Il limite del Nomogramma del MSKCC sta nel valutare il fattore mitosi in maniera non lineare. Comunque tutti i criteri prognostici considerati (NIH, AFIP, Nomogramma MSKCC ) hanno dimostrato una grande capacità nel predire le recidive nelle classi ad alto rischio mentre presentano dei limiti per quelle a basso rischio.

References


