








## Article

# Multi-Institutional Evaluation of Pathologists' Assessment Compared to Immunoscore

Joseph Willis <sup>1</sup>, Robert A. Anders <sup>2</sup>, Toshihiko Torigoe <sup>3</sup> , Yoshihiko Hirohashi <sup>3</sup> , Carlo Bifulco <sup>4</sup>, Inti Zlobec <sup>5</sup>, Bernhard Mlecnik <sup>6,7,8,9</sup> , Sandra Demaria <sup>10</sup> , Won-Tak Choi <sup>11</sup> , Pavel Dundr <sup>12</sup>, Fabiana Tatangelo <sup>13</sup> , Annabella Di Mauro <sup>13</sup> , Pamela Baldin <sup>14</sup>, Gabriela Bindea <sup>6,7,8</sup> , Florence Marliot <sup>6,7,8,15</sup>, Nacilla Haicheur <sup>6,7,8,15</sup>, Tessa Fredriksen <sup>6,7,8</sup>, Amos Kirilovsky <sup>6,7,8,15</sup>, Bénédicte Buttard <sup>6,7,8</sup>, Angela Vasaturo <sup>6,7,8</sup>, Lucie Lafontaine <sup>6,7,8</sup>, Pauline Maby <sup>6,7,8</sup>, Carine El Sissy <sup>6,7,8,15</sup>, Assia Hijazi <sup>6,7,8</sup>, Amine Majdi <sup>6,7,8</sup>, Christine Lagorce <sup>6,7,8,16</sup>, Anne Berger <sup>6,7,8,17</sup>, Marc Van den Eynde <sup>18</sup> , Franck Pagès <sup>6,7,8,15</sup>, Alessandro Lugli <sup>5</sup>  and Jérôme Galon <sup>6,7,8,\*</sup>

- <sup>1</sup> Department of Pathology, UH Cleveland Medical Center, Cleveland, OH 44106, USA; josephe.willis@uhhospitals.org
- <sup>2</sup> Pathology Department, John Hopkins, Baltimore, MD 21287, USA; rander54@jhmi.edu
- <sup>3</sup> Department of Pathology, Sapporo Medical University School of Medicine, Sapporo 060-8556, Japan; torigoe@sapmed.ac.jp (T.T.); hirohash@sapmed.ac.jp (Y.H.)
- <sup>4</sup> Department of Pathology and Molecular Genomics, Providence Portland Medical Center, Portland, OR 97213, USA; carlo.bifulco@providence.org
- <sup>5</sup> Institute of Pathology, University of Bern, 3008 Bern, Switzerland; inti.zlobec@pathology.unibe.ch (I.Z.); alessandro.lugli@pathology.unibe.ch (A.L.)
- <sup>6</sup> INSERM, Laboratory of Integrative Cancer Immunology, 75006 Paris, France; bernhard.mlecnik@crc.jussieu.fr (B.M.); gabriela.bindea@crc.jussieu.fr (G.B.); florence.marliot@aphp.fr (F.M.); nacilla.haicheur@aphp.fr (N.H.); tessa.fredriksen@crc.jussieu.fr (T.F.); amos.kirilovsky@gmail.com (A.K.); benedicte.buttard@crc.jussieu.fr (B.B.); angela.vasaturo@ultivue.com (A.V.); lucie.lafontaine@crc.jussieu.fr (L.L.); mabpau@gmail.com (P.M.); carineelsissy@hotmail.com (C.E.S.); assia.hijazi@sorbonne-universite.fr (A.H.); amine.majdi.pro@gmail.com (A.M.); christine.lagorce@aphp.fr (C.L.); aberger@ghsj.fr (A.B.); franck.pages@egp.aphp.fr (F.P.)
- <sup>7</sup> Centre de Recherche des Cordeliers, Sorbonne Université, Université Paris Cité, 75006 Paris, France
- <sup>8</sup> Equipe Labellisée Ligue Contre le Cancer, 75006 Paris, France
- <sup>9</sup> Inovation, 75005 Paris, France
- <sup>10</sup> Department of Pathology, Weill Cornell Medicine, New York, NY 10021, USA; szd3005@med.cornell.edu
- <sup>11</sup> Department of Pathology, University of California, San Francisco, CA 94143, USA; won-tak.choi@ucsf.edu
- <sup>12</sup> Institute of Pathology, First Faculty of Medicine, Charles University, General University Hospital in Prague, 12808 Prague, Czech Republic; pavel.dundr@vfn.cz
- <sup>13</sup> Department of Pathology, Istituto Nazionale Tumori IRCCS Fondazione G. Pascale, 80131 Napoli, Italy; f.tatangelo@istitutotumori.na.it (F.T.); annabella.dimauro@istitutotumori.na.it (A.D.M.)
- <sup>14</sup> Department of Pathology, Cliniques Universitaires St-Luc, Institut de Recherche Clinique et Experimentale (Pole GAEN), Université Catholique de Louvain, 1348 Brussels, Belgium; pamela.baldin@uclouvain.be
- <sup>15</sup> Immunomonitoring Platform, Laboratory of Immunology, AP-HP, Assistance Publique-Hopitaux de Paris, Georges Pompidou European Hospital, 75015 Paris, France
- <sup>16</sup> Department of Pathology, AP-HP, Assistance Publique-Hopitaux de Paris, Georges Pompidou European Hospital, 75015 Paris, France
- <sup>17</sup> Digestive Surgery Department, AP-HP, Assistance Publique-Hopitaux de Paris, Georges Pompidou European Hospital, 75015 Paris, France
- <sup>18</sup> Institut Roi Albert II, Department of Medical Oncology, Cliniques Universitaires St-Luc, Institut de Recherche Clinique et Experimentale (Pole MIRO), Université Catholique de Louvain, 1030 Brussels, Belgium; marc.vandeneynde@uclouvain.be
- \* Correspondence: jerome.galon@crc.jussieu.fr; Tel.: +33-1-44-27-90-85



**Citation:** Willis, J.; Anders, R.A.; Torigoe, T.; Hirohashi, Y.; Bifulco, C.; Zlobec, I.; Mlecnik, B.; Demaria, S.; Choi, W.-T.; Dundr, P.; et al. Multi-Institutional Evaluation of Pathologists' Assessment Compared to Immunoscore. *Cancers* **2023**, *15*, 4045. <https://doi.org/10.3390/cancers15164045>

Academic Editors: M. Walid Qoronfleh and Nader Al-Dewik

Received: 29 June 2023

Revised: 31 July 2023

Accepted: 8 August 2023

Published: 10 August 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

**Simple Summary:** This study aims to compare the performance of the standardized consensus Immunoscore (IS) digital pathology assay to an evaluation of the immune response via visual examination of hematoxylin–eosin (H&E) slides and CD3+/CD8+ stained slides, achieved by expert pathologists. Herein, we report the evaluation of 540 stained images by multi-institutional pathologists to determine the concordance between pathologist assessment before and after training. The results show that the IS assay outperformed expert pathologists' T-score evaluation in the clinical

setting. This reveals the potential of the IS as an immune pathology tool, critical for reproducible quantitative analysis of tumor-infiltrated immune cells. These findings can contribute to a better diagnosis, allowing one to stratify cancer patients into reliable prognostic groups, based on the immune parameters quantified by IS. This work will likely impact the management of colon cancer patients as it raises the importance of the implementation of digital pathology in cancer diagnosis to provide appropriate personalized therapeutic decisions.

**Abstract:** Background: The Immunoscore (IS) is a quantitative digital pathology assay that evaluates the immune response in cancer patients. This study reports on the reproducibility of pathologists' visual assessment of CD3+- and CD8+-stained colon tumors, compared to IS quantification. Methods: An international group of expert pathologists evaluated 540 images from 270 randomly selected colon cancer (CC) cases. Concordance between pathologists' T-score, corresponding hematoxylin–eosin (H&E) slides, and the digital IS was evaluated for two- and three-category IS. Results: Non-concordant T-scores were reported in more than 92% of cases. Disagreement between semi-quantitative visual assessment of T-score and the reference IS was observed in 91% and 96% of cases before and after training, respectively. Statistical analyses showed that the concordance index between pathologists and the digital IS was weak in two- and three-category IS, respectively. After training, 42% of cases had a change in T-score, but no improvement was observed with a Kappa of 0.465 and 0.374. For the 20% of patients around the cut points, no concordance was observed between pathologists and digital pathology analysis in both two- and three-category IS, before or after training (all Kappa < 0.12). Conclusions: The standardized IS assay outperformed expert pathologists' T-score evaluation in the clinical setting. This study demonstrates that digital pathology, in particular digital IS, represents a novel generation of immune pathology tools for reproducible and quantitative assessment of tumor-infiltrated immune cell subtypes.

**Keywords:** immunoscore; digital pathology; colon cancer; tumor microenvironment; prognostic markers; risk stratification; T cell; anatomopathology

## 1. Introduction

The AJCC/UICC-TNM classification system based on anatomic pathology evaluation of tumors provides useful yet incomplete prognostic information [1]. New ways to classify cancer focusing on tumor cells have only shown modest prediction accuracy and limited clinical usefulness [1,2]. However, an extensive literature review demonstrated a favorable prognostic impact of the pre-existing adaptive immune cells infiltrating tumors [1,3–12]. In colorectal cancer (CRC), we showed a correlation between the in situ densities of adaptive immune cells at the center of the tumor (CT) and the invasive margin (IM) with patients' survival [3,8,12–14]. A meta-analysis of the literature revealed the prognostic value of immune cells and that cytotoxic CD8+ T-cell enrichment was associated with a good prognosis in 97% of the studies [15]. We showed that cytotoxic and memory T cells were predictive of clinical outcome in early-stage CRC (I/II). We further showed that histopathologic-based prognostic factors of CRC are associated with the state of the local immune reaction [8]. The assessment of CD8+ cytotoxic T lymphocytes in combined tumor regions provided an indicator of tumor recurrence beyond that of the AJCC/UICC-TNM staging [16–18]. This immune response was defined by the "Immunoscore" (IS) [15,19–21].

An international IS consortium quantified the pre-existing immunity on stage I/II/III CC patients by using the first worldwide recognized and standardized consensus IS assay. The results established the consensus IS as a powerful and robust immune classifier to predict patient's prognosis [22]. A meta-analysis on more than 10,000 CC patients confirmed that the consensus IS provided a reliable estimate of the recurrence risk [23]. Its clinical utility was further reinforced by publications demonstrating the prognosis value of IS in four independent cohorts of stage III CC patients, including two randomized phase 3 clini-

cal trials [24,25], and its predictive value in response to chemotherapy [24,26]. The clinical utility of IS in Stage II CC patients was validated in multiple cohorts [14,22,27–32]. The immune response measured with the consensus IS was introduced as essential and desirable diagnostic criteria for CRC in the latest (5th) edition of the WHO Digestive System Tumors classification. Moreover, IS was introduced in the 2020 European and 2021 Pan-Asian adapted European Organization for Medical Oncology (ESMO) Clinical Practice Guidelines for gastrointestinal cancers to refine the prognosis and, thus, adjust the chemotherapy decision-making process [33,34]. Therefore, it is of the utmost importance to compare the performance of the standardized IS consensus performed with digital pathology to an evaluation of the immune response through visual examination of hematoxylin–eosin (H&E) slides or via a visual examination of CD3+ and CD8+-stained slides by expert pathologists.

## 2. Materials and Methods

### 2.1. Immunostaining Evaluation

An international group of 10 expert gastrointestinal (GI) pathologists, half from the USA and half from Europe and Japan, evaluated stained CD3+ and CD8+ slides ( $n = 540$ ) from 270 randomly selected full resections of CC cases (cohort demographic distribution and characteristics are presented in Supplementary Table S1). Pathologists performed a semi-quantitative visual assessment (T-score) of CD3+ and CD8+ and reported results for all cases blinded from IS results. Each pathologist evaluated the same 270 cases, before training (unsupervised evaluation) and after training (supervised evaluation). Pathologists' visual assessment and training were conducted according to previously described methods [35]. Before training, all pathologists reported CD3+ staining, CD8+ staining and the overall T-score of each patient into 3 categories (High, Intermediate or Low). All pathologists had the same reference slides ( $n = 12$ ), representing cases with known IS (High, Intermediate or Low). Images with CD3 and CD8 densities corresponding to High, Intermediate and Low cut points in CT and IM regions were provided. For the supervised evaluation, training of the pathologists was performed by providing 12 cases at the cutoff values for IS. Then, all pathologists reported their results accounting for several parameters into three categories (High, Intermediate or Low): CD3+ cell density in CT and IM of the tumor, overall CD3+ cell density, CD8+ cell density in CT and IM, overall CD8+ cell density and overall T-score for each patient. The concordance or discordance for the 10 independent T-score evaluations on the 270 cases and the concordance with the IS were evaluated for two (High, Low) and three categories (High, Intermediate, Low).

### 2.2. Immune Cell Infiltration Evaluation on H&E Slides

H&E slide evaluation for tumor-infiltrating lymphocytes (TILs) was performed by 11 independent evaluators on the same 270 representative CC cases. Each evaluation was performed on the same 270 cases, and each evaluator had the same reference slides (3 representative H&E slides for each IS category). The 11 independent evaluations of TIL were performed in CT and IM regions separately, and the TIL categorization was reported into two- and three-category IS for each case. The concordance or discordance for the 11 independent evaluations of TIL on 270 H&E slides and the concordance with the IS were evaluated for two- and three-category IS.

### 2.3. Immunohistochemistry

For each patient, a pathologist selected a tumor block containing CT and IM regions. Two consecutive tissue paraffin sections of 4  $\mu\text{m}$  were processed for single immunohistochemistry staining with CD3 and CD8 antibodies, followed by DAB substrate (3,3'-diaminobenzidine) in the presence of peroxidase (HRP) enzyme, according to a previously described protocol [22]. Digital slides were obtained with a 20 $\times$  magnification and a resolution of 0.45  $\mu\text{m}$ /pixel.

#### 2.4. Image Analysis

The stained CD3 and CD8 cell densities were determined in CT and IM regions using a specially developed IS<sup>®</sup> analyzer software (INSERM/Veracyte, Marseille, France). The mean and the distribution of the staining intensities were monitored, providing an internal quality control of each slide.

#### 2.5. IS Determination

For each case, CD3 and CD8 densities in CT and IM regions were converted into percentiles, as previously described [22]. The mean of the four percentiles obtained (two markers, two regions) was calculated and translated into the IS scoring system. IS categories were previously defined independently of clinical data [22]. These pre-defined categories were used herein: mean percentiles 0–25%, >25–70%, and >70–100% for IS Low, Intermediate and High, respectively. Additional analyses were performed with the pre-defined two-category IS: Low (0–25%) and Intermediate + High (25–100%). Repeatability Evaluation of IS method was performed according to previously described protocols [22,35].

#### 2.6. Statistics

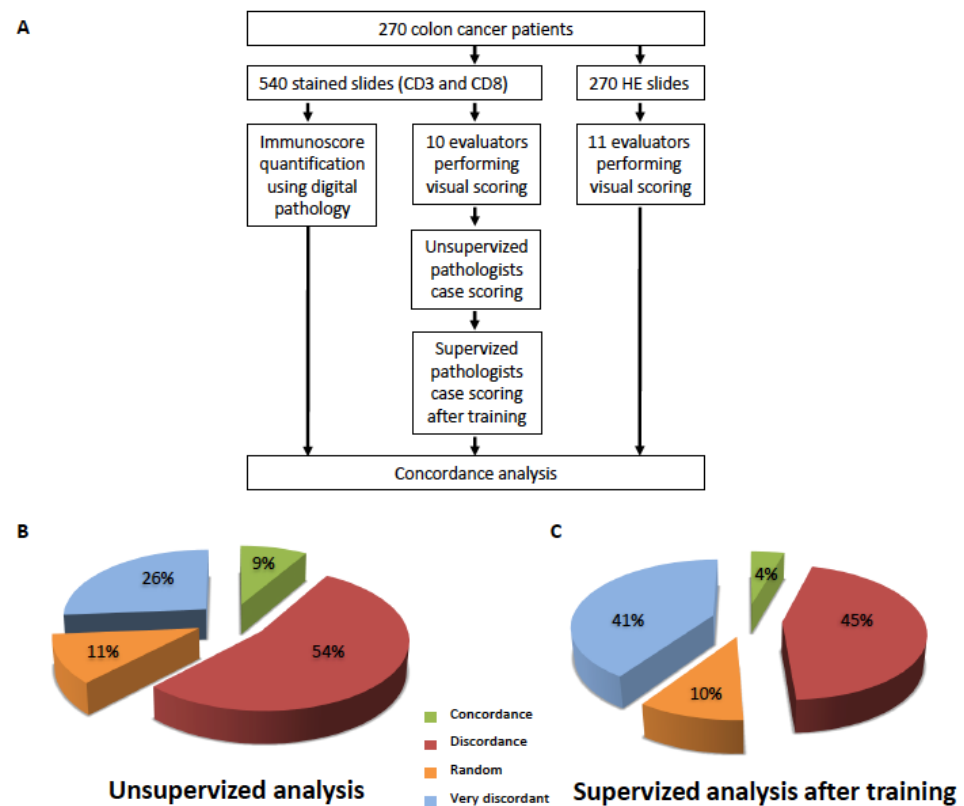
Statistical analysis was used to explore the following types of concordance: between individual pathologist's T-score assessment (from CD3 and CD8 staining) and IS for all cases ( $n = 270$ ), for the subset of cases around the clinical Low (25th percentile) IS cut point ( $n = 54$ ) and for the subset of cases around the High (70th percentile) IS cut point ( $n = 54$ ), before and after training, inter-pathologist agreement with visual assessment of T-score, among three repeated IS quantifications ( $n = 50$ ) and between 11 visual evaluations of TIL (from H&E slides) and IS for all cases ( $n = 270$ ). The Cohen's Kappa coefficient was used to evaluate agreement of IS results between the two rating methods, IS and pathologists' T-score, and between IS and TIL (H&E evaluation). The Fleiss's Kappa coefficient test, an extension of the Cohen's Kappa, was used to compute the agreement between multiple observers' assessment. In accordance with McHugh [36], the level of agreement was categorized according to the Kappa values as: none (0–0.20), minimal (0.21–0.39), weak (0.40–0.59), moderate (0.60–0.79), strong (0.80–0.90) and almost perfect (>0.90%). A negative Kappa indicated that there was less agreement than would be expected by chance given the marginal distributions of ratings. Observers Needed to Evaluate Subjective Tests (ONEST) analysis was used to visualize the change in overall percent agreement as a function of the number of observers, as previously described [37]. High discordance amongst observers is found when the plateau begins at a higher number of observers and occurs at a low overall percent agreement. Ethical, legal and social implications were approved by an ethical review board from île de France (#0912082).

### 3. Results

CC samples from 270 randomly selected representative cases were stained for CD3 and CD8, with the consensus IS being computed (Supplementary Figure S1). The consensus IS was established using the published pre-defined cut points [22] to convert CD3 and CD8 immune densities into percentiles and IS categories (Low, Intermediate, High). IS Low, Intermediate and High represented 33%, 49% and 18% of the cohort, respectively (Supplementary Figure S1).

CD3- and CD8-stained slides (540 images) were given to 10 pathologists blinded to the IS results (Supplementary Figure S2). Each pathologist evaluated the CD3+ and CD8+ cells on the whole slide (unsupervised analysis). The 10 independent evaluations of stained slides were performed in CT and IM regions separately. CD3 and CD8 stains were reported for each patient according to the pathologist's visual expertise into three T-score categories (Low, Intermediate or High). Pathologists were then trained with 12 reference images of known IS values at the IS cut points (see Methods for details). After training, the pathologists re-evaluated the 540 images of CD3 and CD8 stains and reported their semi-quantitative T-scores once again for each patient (Figure 1A, Supplementary Figure S1).

Concordance between pathologists and concordance between pathologists and the consensus IS obtained with digital pathology were then analyzed.



**Figure 1.** Schematic representation of the experimental design: 270 colon cancer patients from the international SITC cohort were selected for this study. For each patient, 2 consecutive whole slide samples were stained for CD3 and CD8 and 1 whole slide sample was stained using hematoxylin–eosin (H&E). CD3+ and CD8+ T cells of those stained slides were analyzed via either digital pathology (IS) or visual assessment by ten pathologists (T-score) before and after supervised training. In parallel, eleven pathologists were given H&E slides to visually assess the density of tumor-infiltrating immune cells in tumor tissue stained with H&E. Concordance between pathologists’ T-score and the digital IS was evaluated for two- and three-category IS. The concordance between the evaluation of the tumor-infiltrating immune cells on the corresponding H&E slides and the digital IS was also analyzed (A). Concordance analysis between individual pathologist’s T-score assessment. Pathologist’s T-score was evaluated based on two- (Low, High) and three-category IS (Low, Intermediate and High). Semi-quantitative evaluation of whole slide images for CD3+ and CD8+ cells was performed by pathologists blinded to IS results, before and after training. Pathologists’ disagreement was defined as the percentage of non-concordant cases for which at least one pathologist assessment was different from others, before (B) and after (C) supervised training. Results fall into four concordance levels: concordant (all pathologists agreeing on scoring), discordant (1 to 4 pathologists not agreeing with others), very discordant (one patient being scored High, Intermediate or Low) and random (5 pathologists with a T-score and 5 pathologists with another T-cell score).

### 3.1. Disagreement between Pathologists’ Visual Evaluation of CD3- and CD8-Stained Slides

Concordance between pathologists’ evaluation of CD3- and CD8-stained slides was analyzed (Figure 1B,C). Pathologists’ disagreement was defined as the percentage of non-concordant cases for which at least one pathologist’s assessment was different from others. Disagreement was observed in the vast majority of cases before training (94%) and after training (95%). Indeed, concordance between all pathologists was found for less than 9% and 4% of patients before (Figure 1B) or after training (Figure 1C), respectively. Discordance (one to four pathologists not agreeing with others) was observed in 54% and 45% of cases

before and after training, while a random T-score classification (five pathologists with one T-score and five pathologists with another T-score) was found in 11% and 10% of patient samples before and after training, respectively. Strikingly, 26% and 41% of patient samples before and after training, respectively, were very discordantly scored by pathologists, with the same case being classified as High, Intermediate and Low. Of note, a detailed analysis among IS Low, Intermediate, and High categories revealed that IS Intermediate was the least concordant (Supplementary Figure S3). Overall, this suggests that training had no effect on the pathologist's ability to properly classify cases, independently of IS categories (Figure 1B,C).

### 3.2. Disagreement between Pathologists' Visual Evaluation of CD3- and CD8-Stained Slides and IS Digital Pathology

We then aimed to compare concordance and discordance between pathologists and IS digital pathology analyses. For this matter, all cases were sorted from lowest IS (in blue) to highest IS (in red), before training (Figure 2A) and after training (Figure 3A). Heatmaps revealed a trend for a correlation between each pathologist's evaluation and IS quantification (from left to right). However, heatmaps also revealed major discrepancies between pathologists (from top to bottom). Indeed, each case was classified as concordant, discordant, random or very discordant. Without training, pathologists' disagreement was observed in the vast majority of the cases compared to digital pathology IS quantification (Figure 2B). After training, similar results were obtained with 95.5% of non-concordant cases (Figure 3B). In fact, concordance between all pathologists was only found in 8.6% and 4.5% of cases before and after training, respectively. Discordance was found in 54.1% and 44.8% of patients, while a random T-score classification was found in 11.6% and 10.0% of patients before and after training, respectively. Moreover, 26.5% and 40.7% of patients before and after training, respectively, were very discordantly scored by pathologists (Figures 2B and 3B).

**Table 1.** Cohen's Kappa statistical analysis highlighting agreements between pathologists' T-score and the reference IS for 270 colon cancer patients, before and after training. Each comparison was performed for the pathologist's classification (T-score) versus IS for the same sample and measured via Cohen's Kappa for all 270 patients and for patients around the 20% cut-points Low ( $n = 54$ ) and High ( $n = 54$ ) (cf. Figures 2 and 3). \*\* Kappa: worse than random (negative Kappa scores), none (0–0.2), weak (0.4–0.59), moderate (0.6–0.79), strong (0.8–0.9) and almost perfect (>0.9).

	Lo vs. Int vs. Hi */Classification (3 Groups) **				Lo vs. Int + Hi */Classification (2 Groups) **			
	Supervised		Unsupervised		Supervised		Unsupervised	
Pathologist	Kappa	Concordance	Kappa	Concordance	Kappa	Concordance	Kappa	Concordance
All patients (270 pts)								
1	0.350	minimal	0.378	minimal	0.486	weak	0.471	weak
2	0.413	weak	0.469	weak	0.550	weak	0.574	weak
3	0.361	minimal	0.394	minimal	0.391	minimal	0.496	weak
4	0.005	none	0.381	minimal	0.058	none	0.444	weak
5	0.448	weak	0.385	minimal	0.579	weak	0.508	weak
6	0.566	weak	0.461	weak	0.642	moderate	0.565	weak
7	0.258	minimal	0.396	minimal	0.282	minimal	0.486	weak
8	0.465	weak	0.421	weak	0.568	weak	0.517	weak
9	0.420	weak	0.526	weak	0.574	weak	0.593	weak
10	0.456	weak	0.270	minimal	0.520	weak	0.328	minimal
20% around 25% low (54 pts)								
Pathologist	Kappa	Concordance	Kappa	Concordance	Kappa	Concordance	Kappa	Concordance
1	0.001	none	0.054	none	0.051	none	0.038	none
2	0.002	none	0.189	none	0.024	none	0.189	none
3	0.092	none	0.146	none	0.092	none	0.146	none

Table 1. Cont.

	Lo vs. Int vs. Hi */Classification (3 Groups) **				Lo vs. Int + Hi */Classification (2 Groups) **			
	Supervised		Unsupervised		Supervised		Unsupervised	
4	−0.106	worse than random	−0.013	worse than random	0.000	none	−0.024	worse than random
5	0.238	minimal	0.152	none	0.262	minimal	0.152	none
6	0.329	minimal	0.071	none	0.329	minimal	0.071	none
7	0.020	none	−0.012	worse than random	−0.059	worse than random	0.008	none
8	0.089	none	0.316	minimal	0.112	none	0.329	minimal
9	0.040	none	0.156	none	0.091	none	0.203	none
10	0.167	none	−0.120	worse than random	0.151	none	−0.120	worse than random

20% around 70% high (54 pts)				
Pathologist	Kappa	Concordance	Kappa	Concordance
1	0.186	none	0.057	none
2	0.047	none	0.182	none
3	0.149	none	0.037	none
4	0.063	none	0.019	none
5	0.191	none	0.123	none
6	0.139	none	0.082	none
7	−0.104	worse than random	0.177	none
8	0.031	none	0.286	minimal
9	0.157	none	0.199	none
10	0.305	minimal	−0.026	worse than random

\* Each comparison is done for the pathologist’s classification vs. the Gold Standard Immunoscore for the same sample and measured by Cohen’s Kappa. \*\* Kappa: worse than random (negative Kappa), none (0–0.2), minimal (0.21–0.39), weak (0.4–0.59), moderate (0.6–0.79), strong (0.8–0.9), and almost perfect (>0.9).

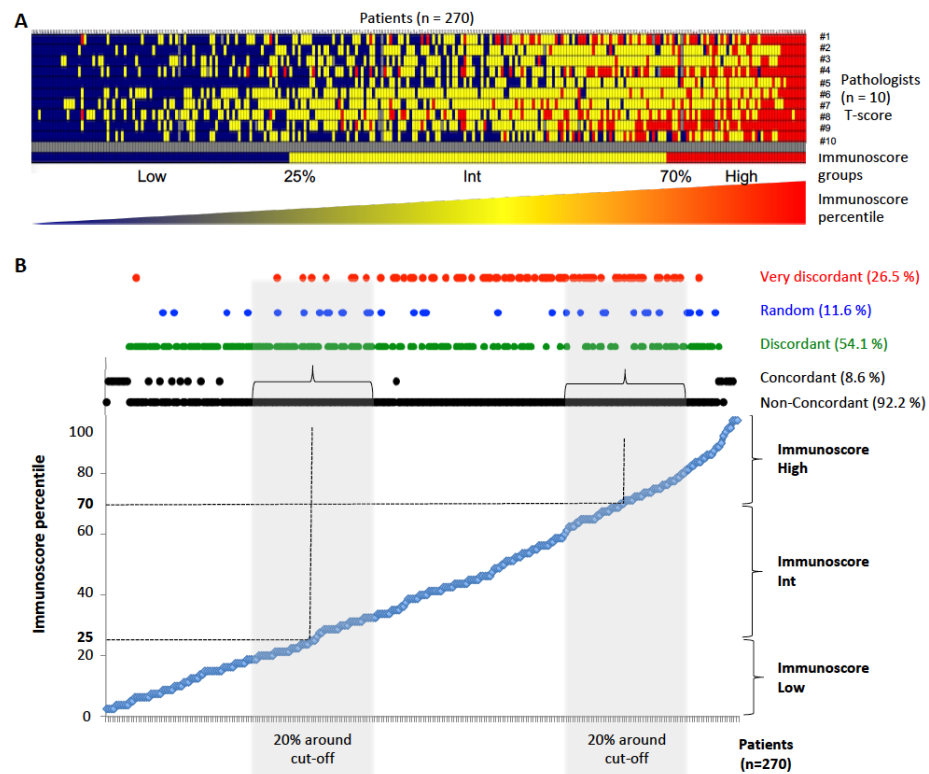
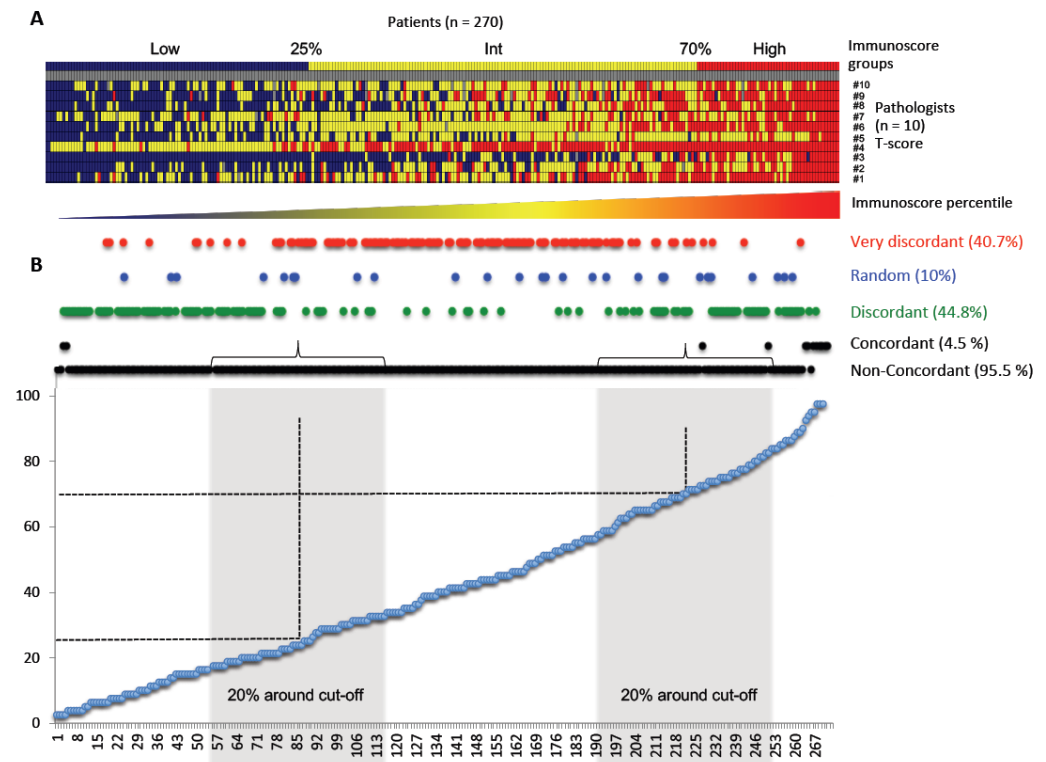


Figure 2. Concordance analysis between pathologists’ T-score and IS before training. (A) Heatmap

representing plotted data for each pathologist blinded to digital pathology IS results before training. Ten pathologists (#1 to #10) evaluated the 270 patients to attribute their T-scores. Patients were illustrated from lowest (blue) to Intermediate (yellow), to highest (red) IS. **(B)** Graph displaying concordance for each IS percentile (<25% = IS Low. >70% = IS High. 25% < IS Intermediate < 70%) and for each concordance level before training: concordant, discordant, random and very discordant. Non-concordant results group everything bare concordant results; 20% around cut points was also used for the statistical Cohen's Kappa results for concordance before training (cf. Table 1).



**Figure 3.** Concordance analysis between pathologists' T-score and IS after training. **(A)** Heatmap representing plotted data for each pathologist blinded to digital pathology IS results after training. Ten pathologists (#1 to #10) evaluated the 270 patients to attribute their T-scores. Patients were illustrated from lowest (blue) to Intermediate (yellow), to highest (red) IS. **(B)** Graph gathering concordance for each IS percentile (<25% = IS Low. >70% = IS High. 25% < IS Intermediate < 70%) and for each concordance level after training: concordant, discordant, random and very discordant. Non-concordant results group everything bare concordant results; 20% around cut points was also used for the statistical Cohen's Kappa results for concordance after training (cf. Table 1).

When evaluating concordance within each IS group (Low, Intermediate or High), concordant cases were mostly seen within the 5% lower and higher end of the Low and High IS categories, whereas non-concordance was observed across all categories. Indeed, discordant cases were spread across a large spectrum of IS (from 5% to 95% percentile) before training, whereas they were vastly associated with IS Low and High groups after training (Figures 2B and 3B). On the other hand, very discordant cases were found for a broad range of IS, both before (Min–Max 6–84%, median 55% percentile) and after training (Min–Max 6–88%, median 41% percentile) (Figures 2B and 3B).

Overall, these data suggest that pathologists are accurate in categorizing patients amongst the 5% with the lowest and highest T-cell infiltration (Low, High IS) but are not accurate enough for the rest, leaving behind the vast majority of patients (92.2% and 95.5% before and after training, respectively) (Figures 2B and 3B).



### 3.3. Concordance between Pathologists' T-Score Evaluation and Consensus IS Using Digital Pathology

The agreement between pathologists' classification and the two- or three-category IS was evaluated via Cohen's Kappa statistical analysis (Table 1). Without previous training, the agreements between pathologists' evaluation for CD3 and CD8 staining classification and the reference IS assessment of 270 CC cases were weak (Table 1). Indeed, the mean Cohen's Kappa was 0.498 (minimum and maximum agreements were (0.32, 0.59)) for the two-category IS (Low, High) and 0.408 (0.27, 0.52) for the three-category IS (Low, Intermediate or High). Similarly, after training, data showed a mean Kappa of 0.465 (0.282, 0.642) for the two-category IS and 0.374 (0.005, 0.566) for the three-category IS.

Furthermore, analysis of the 20% of CC cases around the IS clinical cut points (25%-Low, 70%-High) resulted in even lower concordance, with overall disagreement rates over 99%, both before and after training. This suggested that the pathologists' patient classification did not improve after training (Table 1; Figures 2B and 3B). Before training, Cohen's Kappa index for all pathologists versus IS for the 20% of cases around the IS cut points revealed no concordance with the mean Kappa in two categories of 0.10 (min/max  $-0.12/0.33$ ). Similar results, showing no concordance, were observed for both cut points (25%: Low, 70%: High), when grouping into two or three categories, before and after training (Table 1). We also analyzed the mean Cohen's Kappa index for all pathologists and the Cohen's Kappa index for each pathologist within subgroups of CC patients. The pathologist's T-score classification and its concordance with IS were evaluated for T1–T2, T3, T4 and T3–T4 subgroups, for patients with or without mucinous colloid type and for different grades of tumor differentiation. All mean Cohen's Kappa index values ranged between minimal and weak concordance ( $K = 0.21\text{--}0.59$ ) (Supplementary Table S2).

Interestingly, pathologists would change their classification in 41.8% of cases after training, with a mean percentage gain of cases correctly classified averaging  $-4.2\%$  (worse after training than before) (Figure 4A). This highlights the fact that pathologists often changed categories but were still inaccurate in the categorization of patients after training.



**Figure 4.** Mean changes of categories after training in pathologists' supervised visual evaluation compared to IS. (A) The left histogram represents the mean percentage of changes in categories (Low, Intermediate, High) after training in supervised visual evaluation. The right histogram represents the mean percentage gain (+) or loss (−) of correctly classified cases after training. (B) Average proportions of concordance between pathologists and IS before and after training.

### 3.4. Comparison of Individual Pathologist's Supervised Visual Assessment after Training to IS

The overall agreement rate was defined by the mean percentage of cases for which all pathologists' evaluations were in accordance with the reference IS for the 270 CC patients. Only 8.6% and 4.5% of cases were concordant with the IS before and after training, respectively. Representative images of CD3 staining evaluated by a pathologist compared to the IS are illustrated together with clinical information, (T, N, M, MSI, number of lymph nodes, recurrence and death), including the whole tumor at low magnification and six high-magnification fields (Supplementary Figure S5). Examples of patients with very discordant evaluation by pathologists (Supplementary Figure S5A–C), one extreme case of Low IS

(IS = 2.5%) with concordant evaluation (Supplementary Figure S5D), and one extreme case of High IS (IS = 97.5%) with concordant evaluation (Supplementary Figure S5E) are provided.

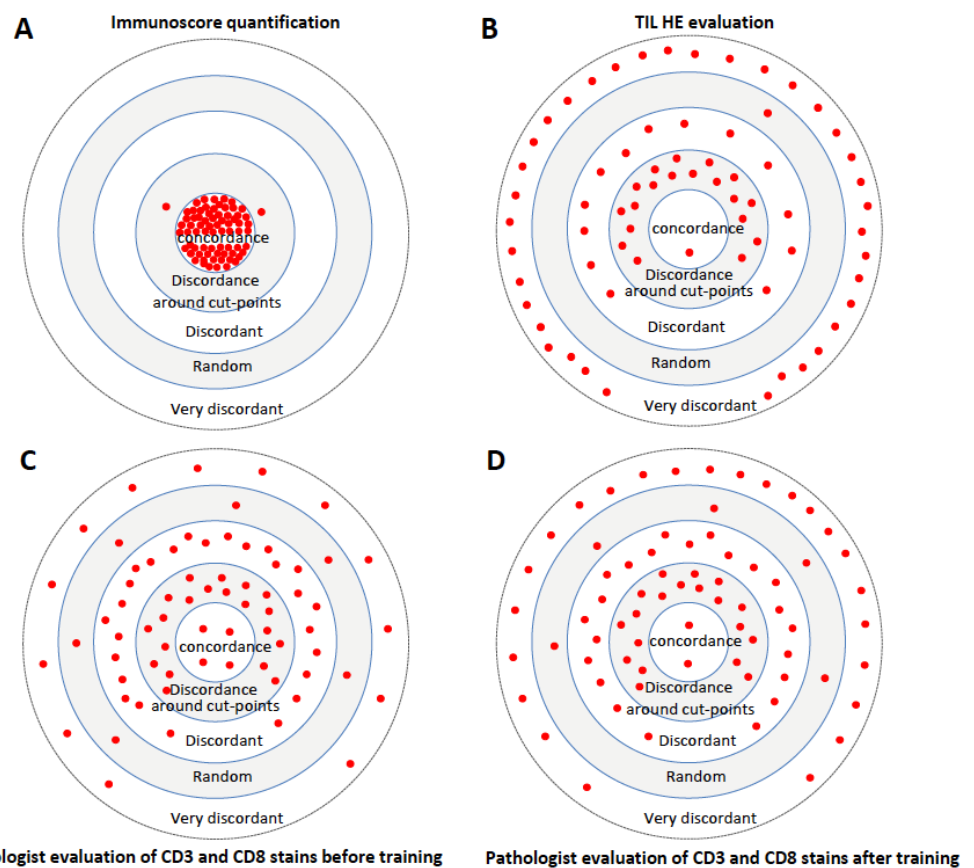
The mean percentage of cases concordant before and after training with IS for each pathologist was only 41.8% (Type 1) (Figure 4A). The average proportion of Type 2 disagreement (discordant classification before and after training) was 19, and a high disagreement rate between the pathologists' evaluation and the reference IS was observed before (37% of disagreement) and after (41% of disagreement) training (Figure 4B). After training, no gain in agreement was observed, but many cases (22%) correctly reported before training were reported incorrectly after training (Figure 4B). Indeed, 18% of cases were concordant after but not before training, but 22% of cases concordant before training were not concordant any longer after training.

### 3.5. Comparison of TIL Evaluation on H&E Slides to T-Score Evaluation and to Digital IS

Target plots illustrated the proportion of evaluation with concordance, discordance around cut points, discordance, random cases and very discordant cases for IS quantification (Figure 5A), TIL evaluation on H&E slides (Figure 5B), pathologists' evaluation of CD3 and CD8 stains before training (Figure 5C) and pathologists' evaluation of CD3 and CD8 stains after training (Figure 5D). TIL evaluation on H&E (Low, Intermediate or High) showed only 4% concordance between 11 evaluators, 51% of discordant cases and 45% of very discordant non-conclusive cases. IS quantification using digital immune pathology was more reproducible than visual evaluation of H&E slides or CD3+- and CD8+-stained slides.

ONEST analysis was used to determine the minimum number of evaluators needed to estimate concordance between several readers [37]. ONEST plots showed decreasing overall percent agreement as the number of observers increased, reaching a low plateau of 0.25 at ten observers for T-score in two categories and of <0.1 agreement for T-score in three categories (Supplementary Figure S4).

Finally, target plots illustrated almost perfect concordance (Cohen's Kappa  $K > 0.9$ ) in the reproducibility of IS, in two- or three-category IS, using digital pathology quantification (Figure 5). In contrast, no concordance (Cohen's Kappa  $K < 0.25$ ) was observed between TIL evaluated on H&E slides and IS. A weak or minimal concordance (Cohen's Kappa  $K < 0.5$ ) was observed between pathologists' visual evaluation of stained CD3 and CD8 slides, both before and after training and with known IS cases. No concordance (Cohen's Kappa  $K < 0.12$ ) was observed between pathologists' visual evaluation of stained CD3 and CD8 slides, both before and after training and with the 20% of cases around the IS cut-point categories (Supplementary Figure S6). The clinical utility of IS is illustrated with a treatment and surveillance TNM-IS decision tree. In stage II, IS High with low to no recurrence, clinicians could consider surgical resection only and low-intensity surveillance, in contrast to IS-Low patients. Overall, IS could impact treatment decision making between 23 and 48% and could impact surveillance decision making for 48% of the patients with stage II colon cancer (Supplementary Figure S7A). In stage III, IS could impact treatment decision making for 55% of the patients with stage III colon cancer. Visual evaluation of T-score by a pathologist would lead to 70% of cases being non-concordant, leading to inappropriate treatment and surveillance (Supplementary Figure S7B).



**Figure 5.** Target plot visualizations of concordance between pathologists' evaluation of T-score and IS, before and after training. Proportion of evaluation with concordant, discordant around cut points, discordant, very discordant and random cases for IS quantification (A), immune-infiltrating lymphocyte evaluation on hematoxylin–eosin (H&E) slides (B), pathologists' evaluation of CD3 and CD8 stains before training (C) and pathologists' evaluation of CD3 and CD8 stains after training (D). Each dot illustrates 4 patients.

## 4. Discussion

### 4.1. Reproducibility of IS (Specificity, Sensibility, Kappa and Concordance)

Multiple analyses and meta-analyses have highlighted the role of T lymphocytes and cytotoxic T cells having a major influence on patient survival [15,18,20,21,38–41]. The immune response, as measured by IS, was introduced for the first time in the latest 5th edition of the WHO Digestive System Tumors as “essential and desirable diagnostic criteria for colorectal cancer”. In addition, IS was introduced in the 2020 European and 2021 Pan-Asian adapted European Organization for Medical Oncology (ESMO) Clinical Practice Guidelines for gastrointestinal cancer to refine the prognosis and to adjust the chemotherapy decision-making process [33,34]. As previously documented, analytical validations of IS highlighted that it is a robust, reproducible, quantitative and standardized immune assay, with a high prognostic performance, independent of all the prognostic markers currently used in clinical practice [42]. IS percentile values remained remarkably constant between formalin-fixed paraffin-embedded tissue blocks from the same patient. The correlation coefficients were  $R = 0.94$  and  $R = 0.97$  for CD8 and CD3, respectively. The concordance between results obtained with the selected blocks and the random blocks was 93% (95% CI 88–96%) [42]. The reproducibility of IS was evaluated on 13 slides per block for 10 patients and revealed excellent accuracy (95.7%), sensitivity (94.8%), specificity (100%) and an overall ROC area of 0.99 [42]. The technical variability of the method was evaluated with lot-to-lot reproducibility and IS assay precision measurements. Consecutive slides from three CCs were assessed for CD3+ and CD8+ T-cell densities using three different

antibody lots, three DAB revelation kit lots, two different benchmark auto-stainers, three different runs and three different operators. A concordance of 100% was observed between IS categories [42]. The analytical variability of the quantification by digital pathology was evaluated. Representative cases ( $n = 36$ ) with ISs ranging from 2.5th to 90th percentiles were re-analyzed by eight independent pathologists from different centers. Mean cell densities for CD3 and CD8 in each tumor region revealed a strong inter-observer reproducibility ( $r = 0.97$  for tumor;  $r = 0.97$  for invasive margin;  $p < 0.0001$ ) [22]. A full assessment of IS reproducibility was performed in two laboratories. Each laboratory had its own IS workflow, including staining, scanning and analysis. Non-consecutive cutting slides from the same tumor block were used to assess the IS of 100 representative cases. The inter-laboratory correlation for CD3+ and CD8+ cells was 0.94 ( $p < 0.001$ ), and the overall categorical IS concordance between the two centers was 93%. This also included biological variability of the tumor [42]. Moreover, the rare cases of discordance were all very close to the cut-point value of 25%, and it would be easy to re-test IS in such samples to correctly assign their score. Finally, the concordance from five independent IS quantifications using Cohen's Kappa statistics revealed an almost perfect concordance ( $K > 0.93$ ) between digital quantifications of IS [35].

#### 4.2. Non-Reproducibility of TIL on H&E Slides

A visual assessment of the density of TILs in tumor tissue stained with H&E was analyzed. H&E images from representative cases ( $n = 270$ ) from the international SITC cohort were assessed by 11 observers. Only 4% of cases were concordant between all observers, 8% of cases were concordant between 80% of observers and a total absence of concordance (50% discordance) was evident in 45% of the cases [22].

Concordance between all observers was obtained for only 8% of cases and was concordant with the digital IS for only 3% of cases. Discordant cases, with at least one evaluation different from others and different from IS, were found in 25% of cases. Strikingly, very discordant cases (the same H&E slide being evaluated as Low, Intermediate or High) were found in 72% of cases. The difference between IS quantification and TIL evaluation on H&E slides not only reflects the difficulty of such evaluation but also indicates that H&E staining of TILs is a crude and subjective semi-quantitative evaluation of undefined cell populations with possible opposite functions, such as CD4+ T cells with Th1 orientation vs. Th2 orientation vs. immune cells with regulatory functions (Treg cells), natural killer (NK) cells, NK-T cells, B-cells, subsets, innate lymphoid cells, cytotoxic CD8 T cells, or even round-shaped monocytes. This illustrates the complexity, subjectivity and discordance of TIL evaluation on H&E slides.

#### 4.3. Non-Concordance between Pathologists for T-Score Evaluation before and after Training

The semi-quantitative evaluation of 540 chromogenic (DAB) single-stain slides (CD3+ and CD8+) by 10 pathologists revealed major discordance between pathologists. Indeed, before training, evaluation showed 91% of non-concordant cases between pathologists. Furthermore, 26% of cases were very discordant (the same slides from the same patient being evaluated Low, Intermediate or High by different pathologists). These discrepancies were not improved after training with 12 representative reference CD3 and CD8 cases at the IS cut points (25th and 70th percentile).

A significant disagreement was observed between the semi-quantitative pathologist's T-score (into two (High or Low) or three categories (High, Intermediate, Low)) compared to the consensus digital pathology IS. Importantly, a high rate of disagreement was observed when comparing the pathologists' visual assessment with the reference IS, leading to misclassification of >96% cases, and this disagreement was even higher (100%) for the cases around the clinical cut point (of 25th percentile). The study revealed that the impact of training was heterogeneous between pathologists and that, overall, training did not improve the concordance between the visual assessment and IS. Changes in training

methods could be considered; however, this also illustrates the complexity, subjectivity and discordance of CD3+ and CD8+ evaluation by visual examination.

The lack of improvement in agreement between pathologists' evaluation and quantitative digital pathology, before and after training, is likely multifactorial. In fact, the size of a colon tumor is quite large, and a whole slide analysis revealed a heterogeneous pattern of CD3+ and CD8+ within different areas of the tumor. The total number of CD3+ cells on a given slide (CT + IM regions) is huge, with a mean of 88,000 CD3+ T-cell/slide, making visual evaluation very challenging. Furthermore, the mean density of these cells is higher at the invasive margin compared to the core of the tumor, rendering the overall visual evaluation difficult. In addition, these immune cells can be present at different densities within the tumor or the stroma and can be clustered or dispersed, even within the same tumor. CD3+, encompassing both CD8+ and CD4+ T-helper cells, and CD8+ cells also have different densities in different areas of the tumor, and the evaluation has to be performed twice for each of these markers on consecutive slides. Looking at the overall slide is tedious, and the semi-quantitative evaluation of so much heterogeneity is very complex and, in fact, very subjective. It is likely that poor concordance would also have been observed within pathological subgroups. The poor performance of pathologists' scoring even after training demonstrated that the novel tool of quantitative digital immune pathology is clearly much more appropriate for such evaluations. Even in an easier context of PD-L1 in non-small cell lung cancer (NSCLC), there is an impact of a pathologist's personality on the interobserver variability and diagnostic accuracy of immunostaining [43]. Furthermore, the subjective T-score showed low reproducibility across multiple pathologists with ONEST analysis, suggesting that the vast majority of pathologists will disagree about subjective evaluation of infiltrating T cells. Thus, based on previous data [22,35] and our actual results obtained herein, the digital consensus IS quantification shows a high level of reproducibility, with perfect software concordance. In contrast, the pathologists' visual subjective evaluation on H&E slides or the evaluation of CD3+- and CD8+-stained slides was not reliable enough for a precise therapeutic decision-making process. In conclusion, a reliable evaluation of CD3+ and CD8+ cells and of IS on a whole slide section shall not be a visual estimation but rather a real IS quantification using the dedicated reproducible software.

#### 4.4. Clinical Impact of Misclassification

The pre-existing immune contexture has an impact on the response to chemotherapy and immunotherapy treatments [16,20,44–55]. Multiple therapeutic approaches against cancer are ongoing [15,39,46,47,49,53,56–59], and for quantitative immune classification and the precision management of patients, biomarkers using quantitative pathology are becoming a necessity [42,60]. Misclassification of stage II and III CC patients by T-score semi-quantitative evaluation would result in inappropriate treatment decision making for many patients [24–26]. Similarly, another assay, Immunoscore-IC, which is also a quantitative and spatial evaluation of immune markers (CD8 and PD-L1), predicts response to immunotherapy and requires digital pathology [61,62].

For patients with stage II CC, many patients being misidentified as stage II CC at low clinical risk would, in fact, be at high risk based on IS. Such a situation would produce false expectations of recurrence for these patients who will not be monitored as closely as those at high risk of recurrence to detect signs of relapse earlier. These patients would not be appropriately considered as high-risk stage II patients and may be under-surveilled and under-treated. Similarly, misclassification of truly IS-High stage II CC patients as having tumors with low T-score by visual examination could result in patients recommended for adjuvant chemotherapy when their recurrence risk is low and exposing them to unnecessary toxicity.

For stage III CC, IS-Low cases being misclassified as high with visual T-score would be detrimental, as patients who may not get benefit from chemotherapy [26] or longer duration of adjuvant chemotherapy (6 months versus 3 months) [24] would not be identified as poor responders. These patients may be unnecessarily subjected to additional chemotherapy

and its associated long-term toxicity. Finally, stage III CC patients with IS-High could be misclassified as patients with poorly infiltrated tumors with visual evaluation (low T-score) and would not be identified as deriving significant benefit from a longer duration of adjuvant chemotherapy [24]. Such patients would be under-treated and subjected to an increased risk of relapse.

Based on previous treatment decision trees, IS would impact treatment decision for 44% of stage II patients and for 55% of stage III patients [63]. Based on visual T-scoring from a single pathologist after training, 70% of cases would be non-concordant with IS. Given an estimated incidence of 101,420 and 23,000 stage II and stage III CC patients per year, respectively, pathologists' visual evaluation of T-score would lead to 70,914 stage II, 16,100 stage III and more than 87,000 CC cases being misclassified and possibly receiving inappropriate patient care annually.

## 5. Conclusions

The very important difference between pathologists' T-score classification and the reproducible IS quantification highlights the importance of new tools for pathologists, namely quantitative digital pathology.

The potential negative impact of immune response misclassification due to pathologists' T-score may result in erroneous prognosis and risk evaluation for many CC patients. These results demonstrated that the IS assay helps to better stratify patients into reliable prognostic recurrence groups. We conclude that the standardized and robust IS assay outperforms the assessment of expert pathologists in the clinical setting for immune response and can, thus, provide most appropriate individualized therapeutic decisions for CC patients.

## 6. Patents

J.G. and B.M. have patents associated with immune prognostic biomarkers. J.G. is co-founder of HaliuDx, a Veracyte company. Immunoscore<sup>®</sup> is a registered trademark owned by the National Institute of Health and Medical Research (INSERM) and licensed to Veracyte.

**Supplementary Materials:** The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/cancers15164045/s1>, Figure S1: Detailed schematic representation of the experimental design; Figure S2: Representative images of CD3 and CD8 staining; Figure S3: Detailed concordance analysis between individual pathologist's T-score; Figure S4: ONEST plot showing T-score overall percent agreement between pathologists as a function of the number of observers; Figure S5: Representative images of CD3 staining from 5 cases, non-concordant patients (A–C), concordant patient with Low-IS (D) and concordant patient with High-IS (E); Figure S6: Target plot summarizing the study; Figure S7: Treatment and surveillance clinical decision-tree according to IS in Stage II (A) and Stage III (B) patients; Table S1: Cohort characteristics; Table S2: Cohen's Kappa statistical analysis highlighting agreements between pathologists' T-score and the reference IS for clinical subgroups of colon cancer patients (see Table 1 for details).

**Author Contributions:** Conceptualization, B.M. and J.G.; Methodology, J.W., R.A.A., T.T., Y.H., C.B., I.Z., B.M., S.D., W.-T.C., P.D., F.T., A.D.M., P.B., G.B., F.M., N.H., T.F., A.K., B.B., A.V., L.L., P.M., C.E.S., C.L., A.B., M.V.d.E., F.P. and A.L.; Software, B.M. and G.B.; Validation, J.W., R.A.A., T.T., Y.H., C.B., I.Z., S.D., W.-T.C., P.D., F.T., A.D.M., P.B., G.B., F.M., N.H., M.V.d.E., F.P., A.L. and J.G.; Formal analysis, J.W., R.A.A., T.T., Y.H., C.B., I.Z., B.M., S.D., W.-T.C., P.D., F.T., A.D.M., P.B., G.B., F.M., N.H., M.V.d.E., F.P., A.L. and J.G.; Investigation, J.G.; Writing—original draft, A.H., A.M. and J.G.; Supervision, J.G.; Project administration, J.G.; Funding acquisition, J.G. All authors have read and agreed to the published version of the manuscript.

**Funding:** The work was funded by grants from INSERM, LabEx Immuno-oncology (11LAXE61UPDE), Transcan ERAnet european project, Association pour la Recherche contre le Cancer (ARC RM22J20ARC03), Site de Recherche intégrée sur le Cancer (SIRIC), CAncer Research for PErsonalized Medicine (CARPEM, INCa-DGOS-Inserm-ITMO Cancer\_18006), La Ligue contre le Cancer (R19034DD), Assistance publique—Hôpitaux de Paris (AP-HP), Agence Nationale de la Recherche (ANR Grant TERMM

ANR-20-CE92-0001), Qatar National Research Fund (QNRF) grant number NPRP11S-0121-180351, Louis Jeantet Prize foundation. P.D. was supported by grant from the Ministry of Health, Czech Republic, MH CZ DRO-VFN 64165.

**Institutional Review Board Statement:** The study was approved by the ethics committees from each center. Ethical, legal and social implications were approved by an ethical review board of each center.

**Informed Consent Statement:** Informed consent was obtained from all subjects involved in the study.

**Data Availability Statement:** The materials described in the manuscript are freely available to use for non-commercial purposes. Detailed extracted data can be provided immediately following publication, upon request to the corresponding author. Proposals should be directed by email to the corresponding author J.G., at jerome.galon@crc.jussieu.fr.

**Acknowledgments:** This work was supported by grants from INSERM, LabEx Immuno-oncology (11LAXE61UPDE), Transcan ERAnet european project, Association pour la Recherche contre le Cancer (ARC RM22J20ARC03), Site de Recherche intégrée sur le Cancer (SIRIC), CAncer Research for PErsonalized Medicine (CARPEM), La Ligue contre le Cancer (R19034DD), Assistance publique–Hôpitaux de Paris (AP-HP), Agence Nationale de la Recherche (ANR Grant TERMM ANR-20-CE92-0001), Qatar National Research Fund (QNRF) grant number NPRP11S-0121-180351, Louis Jeantet Prize foundation.

**Conflicts of Interest:** J.G. and B.M. have patents associated with immune prognostic biomarkers. All other authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

## References

1. Galon, J.; Mlecnik, B.; Bindea, G.; Angell, H.K.; Berger, A.; Lagorce, C.; Lugli, A.; Zlobec, I.; Hartmann, A.; Bifulco, C.; et al. Towards the introduction of the ‘Immunoscore’ in the classification of malignant tumours. *J. Pathol.* **2014**, *232*, 199–209. [[CrossRef](#)]
2. Guinney, J.; Dienstmann, R.; Wang, X.; de Reynies, A.; Schlicker, A.; Soneson, C.; Marisa, L.; Roepman, P.; Nyamundanda, G.; Angelino, P.; et al. The consensus molecular subtypes of colorectal cancer. *Nat. Med.* **2015**, *21*, 1350–1356. [[CrossRef](#)]
3. Galon, J.; Costes, A.; Sanchez-Cabo, F.; Kirilovsky, A.; Mlecnik, B.; Lagorce-Pages, C.; Tosolini, M.; Camus, M.; Berger, A.; Wind, P.; et al. Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. *Science* **2006**, *313*, 1960–1964. [[CrossRef](#)] [[PubMed](#)]
4. Koelzer, V.H.; Dawson, H.; Andersson, E.; Karamitopoulou, E.; Masucci, G.V.; Lugli, A.; Zlobec, I. Active immunosurveillance in the tumor microenvironment of colorectal cancer is associated with low frequency tumor budding and improved outcome. *Transl. Res.* **2015**, *166*, 207–217. [[CrossRef](#)] [[PubMed](#)]
5. Laghi, L.; Bianchi, P.; Miranda, E.; Balladore, E.; Pacetti, V.; Grizzi, F.; Allavena, P.; Torri, V.; Repici, A.; Santoro, A.; et al. CD3+ cells at the invasive margin of deeply invading (pT3-T4) colorectal cancer and risk of post-surgical metastasis: A longitudinal study. *Lancet Oncol.* **2009**, *10*, 877–884. [[CrossRef](#)]
6. Lee, W.S.; Park, S.; Lee, W.Y.; Yun, S.H.; Chun, H.K. Clinical impact of tumor-infiltrating lymphocytes for survival in stage II colon cancer. *Cancer* **2010**, *116*, 5188–5199. [[CrossRef](#)] [[PubMed](#)]
7. Mlecnik, B.; Bindea, G.; Angell, H.K.; Maby, P.; Angelova, M.; Tougeron, D.; Church, S.E.; Lafontaine, L.; Fischer, M.; Fredriksen, T.; et al. Integrative Analyses of Colorectal Cancer Show Immunoscore Is a Stronger Predictor of Patient Survival than Microsatellite Instability. *Immunity* **2016**, *44*, 698–711. [[CrossRef](#)] [[PubMed](#)]
8. Mlecnik, B.; Tosolini, M.; Kirilovsky, A.; Berger, A.; Bindea, G.; Meatchi, T.; Bruneval, P.; Trajanoski, Z.; Fridman, W.H.; Pages, F.; et al. Histopathologic-based prognostic factors of colorectal cancers are associated with the state of the local immune reaction. *J. Clin. Oncol.* **2011**, *29*, 610–618. [[CrossRef](#)]
9. Noshu, K.; Baba, Y.; Tanaka, N.; Shima, K.; Hayashi, M.; Meyerhardt, J.A.; Giovannucci, E.; Dranoff, G.; Fuchs, C.S.; Ogino, S. Tumour-infiltrating T-cell subsets, molecular changes in colorectal cancer and prognosis: Cohort study and literature review. *J. Pathol.* **2010**, *222*, 350–366. [[CrossRef](#)]
10. Ogino, S.; Galon, J.; Fuchs, C.S.; Dranoff, G. Cancer immunology—Analysis of host and tumor factors for personalized medicine. *Nat. Rev. Clin. Oncol.* **2011**, *8*, 711–719. [[CrossRef](#)]
11. Ogino, S.; Noshu, K.; Irahara, N.; Meyerhardt, J.A.; Baba, Y.; Shima, K.; Glickman, J.N.; Ferrone, C.R.; Mino-Kenudson, M.; Tanaka, N.; et al. Lymphocytic reaction to colorectal cancer is associated with longer survival, independent of lymph node count, microsatellite instability, and CpG island methylator phenotype. *Clin. Cancer Res.* **2009**, *15*, 6412–6420. [[CrossRef](#)] [[PubMed](#)]
12. Pages, F.; Berger, A.; Camus, M.; Sanchez-Cabo, F.; Costes, A.; Molidor, R.; Mlecnik, B.; Kirilovsky, A.; Nilsson, M.; Damotte, D.; et al. Effector memory T cells, early metastasis, and survival in colorectal cancer. *N. Engl. J. Med.* **2005**, *353*, 2654–2666. [[CrossRef](#)] [[PubMed](#)]

13. Mlecnik, B.; Bindea, G.; Angell, H.K.; Sasso, M.S.; Obenauf, A.C.; Fredriksen, T.; Lafontaine, L.; Bilocq, A.M.; Kirilovsky, A.; Tosolini, M.; et al. Functional network pipeline reveals genetic determinants associated with in situ lymphocyte proliferation and survival of cancer patients. *Sci. Transl. Med.* **2014**, *6*, 228ra237. [[CrossRef](#)] [[PubMed](#)]
14. Pages, F.; Kirilovsky, A.; Mlecnik, B.; Asslaber, M.; Tosolini, M.; Bindea, G.; Lagorce, C.; Wind, P.; Marliot, F.; Bruneval, P.; et al. In situ cytotoxic and memory T cells predict outcome in patients with early-stage colorectal cancer. *J. Clin. Oncol.* **2009**, *27*, 5944–5951. [[CrossRef](#)]
15. Bruni, D.; Angell, H.K.; Galon, J. The immune contexture and Immunoscore in cancer prognosis and therapeutic efficacy. *Nat. Rev. Cancer* **2020**, *20*, 662–680. [[CrossRef](#)]
16. Bindea, G.; Mlecnik, B.; Angell, H.K.; Galon, J. The immune landscape of human tumors: Implications for cancer immunotherapy. *Oncoimmunology* **2014**, *3*, e27456. [[CrossRef](#)]
17. Bindea, G.; Mlecnik, B.; Fridman, W.H.; Galon, J. The prognostic impact of anti-cancer immune response: A novel classification of cancer patients. *Semin. Immunopathol.* **2011**, *33*, 335–340. [[CrossRef](#)]
18. Pages, F.; Galon, J.; Fridman, W.H. The essential role of the in situ immune reaction in human colorectal cancer. *J. Leukoc. Biol.* **2008**, *84*, 981–987. [[CrossRef](#)]
19. Angell, H.K.; Bruni, D.; Barrett, J.C.; Herbst, R.; Galon, J. The Immunoscore: Colon Cancer and Beyond. *Clin. Cancer Res.* **2020**, *26*, 332–339. [[CrossRef](#)]
20. Galon, J.; Bruni, D. Tumor Immunology and Tumor Evolution: Intertwined Histories. *Immunity* **2020**, *52*, 55–81. [[CrossRef](#)]
21. Kirilovsky, A.; Marliot, F.; El Sissy, C.; Haicheur, N.; Galon, J.; Pages, F. Rational bases for the use of the Immunoscore in routine clinical settings as a prognostic and predictive biomarker in cancer patients. *Int. Immunol.* **2016**, *28*, 373–382. [[CrossRef](#)] [[PubMed](#)]
22. Pages, F.; Mlecnik, B.; Marliot, F.; Bindea, G.; Ou, F.S.; Bifulco, C.; Lugli, A.; Zlobec, I.; Rau, T.T.; Berger, M.D.; et al. International validation of the consensus Immunoscore for the classification of colon cancer: A prognostic and accuracy study. *Lancet* **2018**, *391*, 2128–2139. [[CrossRef](#)] [[PubMed](#)]
23. Zhang, X.; Yang, J.; Du, L.; Zhou, Y.; Li, K. The prognostic value of Immunoscore in patients with cancer: A pooled analysis of 10,328 patients. *Int. J. Biol. Markers* **2020**, *35*, 1724600820927409. [[CrossRef](#)]
24. Pages, F.; Andre, T.; Taieb, J.; Vernerey, D.; Henriques, J.; Borg, C.; Marliot, F.; Ben Jannet, R.; Louvet, C.; Mineur, L.; et al. Prognostic and predictive value of the Immunoscore in stage III colon cancer patients treated with oxaliplatin in the prospective IDEA France PRODIGE-GERCOR cohort study. *Ann. Oncol.* **2020**, *31*, 921–929. [[CrossRef](#)]
25. Sinicrope, F.A.; Shi, Q.; Hermitte, F.; Zemla, T.J.; Mlecnik, B.; Benson, A.B.; Gill, S.; Goldberg, R.M.; Kahlenberg, M.S.; Nair, S.G.; et al. Contribution of Immunoscore and Molecular Features to Survival Prediction in Stage III Colon Cancer. *JNCI Cancer Spectr.* **2020**, *4*, pkaa023. [[CrossRef](#)]
26. Mlecnik, B.; Bifulco, C.; Bindea, G.; Marliot, F.; Lugli, A.; Lee, J.J.; Zlobec, I.; Rau, T.T.; Berger, M.D.; Nagtegaal, I.D.; et al. Multicenter International Society for Immunotherapy of Cancer Study of the Consensus Immunoscore for the Prediction of Survival and Response to Chemotherapy in Stage III Colon Cancer. *J. Clin. Oncol.* **2020**, *38*, 3638–3651. [[CrossRef](#)] [[PubMed](#)]
27. Imen, H.; Amira, H.; Fatma, K.; Raja, J.; Mariem, S.; Haithem, Z.; Ehsene, B.B.; Aschraf, C. Prognostic Value of Immunoscore in Colorectal Carcinomas. *Int. J. Surg. Pathol.* **2023**, *25*, 10668969231168357. [[CrossRef](#)]
28. Marliot, F.; Pages, F.; Galon, J. Usefulness and robustness of Immunoscore for personalized management of cancer patients. *Oncoimmunology* **2020**, *9*, 1832324. [[CrossRef](#)]
29. Mlecnik, B.; Lugli, A.; Bindea, G.; Marliot, F.; Bifulco, C.; Lee, J.J.; Zlobec, I.; Rau, T.T.; Berger, M.D.; Nagtegaal, I.D.; et al. Multicenter International Study of the Consensus Immunoscore for the Prediction of Relapse and Survival in Early-Stage Colon Cancer. *Cancers* **2023**, *15*, 418. [[CrossRef](#)]
30. Mlecnik, B.; Torigoe, T.; Bindea, G.; Popivanova, B.; Xu, M.; Fujita, T.; Hazama, S.; Suzuki, N.; Nagano, H.; Okuno, K.; et al. Clinical Performance of the Consensus Immunoscore in Colon Cancer in the Asian Population from the Multicenter International SITC Study. *Cancers* **2022**, *14*, 4346. [[CrossRef](#)]
31. Trabelsi, M.; Farah, F.; Zouari, B.; Jaafoura, M.H.; Kharrat, M. An Immunoscore System Based on CD3+ And CD8+ Infiltrating Lymphocytes Densities to Predict the Outcome of Patients with Colorectal Adenocarcinoma. *Onco Targets Ther.* **2019**, *12*, 8663–8673. [[CrossRef](#)]
32. Wang, F.; Lu, S.; Cao, D.; Qian, J.; Li, C.; Zhang, R.; Wang, F.; Wu, M.; Liu, Y.; Pan, Z.; et al. Prognostic and predictive value of Immunoscore and its correlation with ctDNA in stage II colorectal cancer. *Oncoimmunology* **2023**, *12*, 2161167. [[CrossRef](#)]
33. Argilés, G.; Tabernero, J.; Labianca, R.; Hochhauser, D.; Salazar, R.; Iveson, T.; Laurent-Puig, P.; Quirke, P.; Yoshino, T.; Taieb, J.; et al. Localised colon cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* **2020**, *31*, 1291–1305. [[CrossRef](#)]
34. Yoshino, T.; Argilés, G.; Oki, E.; Martinelli, E.; Taniguchi, H.; Arnold, D.; Mishima, S.; Li, Y.; Smruti, B.K.; Ahn, J.B.; et al. Pan-Asian adapted ESMO Clinical Practice Guidelines for the diagnosis treatment and follow-up of patients with localised colon cancer. *Ann. Oncol.* **2021**, *32*, 1496–1510. [[CrossRef](#)]
35. Boquet, I.; Kassambara, A.; Lui, A.; Tanner, A.; Latil, M.; Lovera, Y.; Arnoux, F.; Hermitte, F.; Galon, J.; Catteau, A. Comparison of Immune Response Assessment in Colon Cancer by Immunoscore (Automated Digital Pathology) and Pathologist Visual Scoring. *Cancers* **2022**, *14*, 1170. [[CrossRef](#)]
36. McHugh, M.L. Interrater reliability: The kappa statistic. *Biochem. Med.* **2012**, *22*, 276–282. [[CrossRef](#)]



37. Reisenbichler, E.S.; Han, G.; Bellizzi, A.; Bossuyt, V.; Brock, J.; Cole, K.; Fadare, O.; Hameed, O.; Hanley, K.; Harrison, B.T.; et al. Prospective multi-institutional evaluation of pathologist assessment of PD-L1 assays for patient selection in triple negative breast cancer. *Mod. Pathol.* **2020**, *33*, 1746–1752. [[CrossRef](#)]
38. Ascierto, P.A.; Capone, M.; Urba, W.J.; Bifulco, C.B.; Botti, G.; Lugli, A.; Marincola, F.M.; Ciliberto, G.; Galon, J.; Fox, B.A. The additional facet of immunoscore: Immunoprofiling as a possible predictive tool for cancer treatment. *J. Transl. Med.* **2013**, *11*, 54. [[CrossRef](#)] [[PubMed](#)]
39. Galon, J.; Bruni, D. Approaches to treat immune hot, altered and cold tumours with combination immunotherapies. *Nat. Rev. Drug Discov.* **2019**, *18*, 197–218. [[CrossRef](#)] [[PubMed](#)]
40. Galon, J.; Mlecnik, B.; Marliot, F.; Ou, F.S.; Bifulco, C.B.; Lugli, A.; Zlobec, I.; Rau, T.; Hartmann, A.; Masucci, G.; et al. Validation of the Immunoscore (IM) as a prognostic marker in stage I/II/III colon cancer: Results of a worldwide consortium-based analysis of 1336 patients. *J. Clin. Oncol.* **2016**, *34*, S3500. [[CrossRef](#)]
41. Fridman, W.H.; Dieu-Nosjean, M.C.; Pages, F.; Cremer, I.; Damotte, D.; Sautes-Fridman, C.; Galon, J. The immune microenvironment of human tumors: General significance and clinical impact. *Cancer Microenviron.* **2013**, *6*, 117–122. [[CrossRef](#)] [[PubMed](#)]
42. Marliot, F.; Chen, X.; Kirilovsky, A.; Sbarrato, T.; El Sissy, C.; Batista, L.; Van den Eynde, M.; Haicheur-Adjouri, N.; Anitei, M.G.; Musina, A.M.; et al. Analytical validation of the Immunoscore and its associated prognostic value in patients with colon cancer. *J. Immunother. Cancer* **2020**, *8*, e000272. [[CrossRef](#)] [[PubMed](#)]
43. Butter, R.; Hondelink, L.M.; van Elswijk, L.; Blaauwgeers, J.L.G.; Bloemena, E.; Britstra, R.; Bulkman, N.; van Gulik, A.L.; Monkhorst, K.; de Rooij, M.J.; et al. The impact of a pathologist's personality on the interobserver variability and diagnostic accuracy of predictive PD-L1 immunohistochemistry in lung cancer. *Lung Cancer* **2022**, *166*, 143–149. [[CrossRef](#)]
44. Aranda, F.; Vacchelli, E.; Eggermont, A.; Galon, J.; Fridman, W.H.; Zitvogel, L.; Kroemer, G.; Galluzzi, L. Trial Watch: Immunostimulatory monoclonal antibodies in cancer therapy. *Oncoimmunology* **2014**, *3*, e27297. [[CrossRef](#)]
45. Aranda, F.; Vacchelli, E.; Eggermont, A.; Galon, J.; Sautes-Fridman, C.; Tartour, E.; Zitvogel, L.; Kroemer, G.; Galluzzi, L. Trial Watch: Peptide vaccines in cancer therapy. *Oncoimmunology* **2013**, *2*, e26621. [[CrossRef](#)]
46. Buque, A.; Bloy, N.; Aranda, F.; Castoldi, F.; Eggermont, A.; Cremer, I.; Fridman, W.H.; Fucikova, J.; Galon, J.; Marabelle, A.; et al. Trial Watch: Immunomodulatory monoclonal antibodies for oncological indications. *Oncoimmunology* **2015**, *4*, e1008814. [[CrossRef](#)]
47. Galluzzi, L.; Vacchelli, E.; Fridman, W.H.; Galon, J.; Sautes-Fridman, C.; Tartour, E.; Zucman-Rossi, J.; Zitvogel, L.; Kroemer, G. Trial Watch: Monoclonal antibodies in cancer therapy. *Oncoimmunology* **2012**, *1*, 28–37. [[CrossRef](#)] [[PubMed](#)]
48. Galon, J.; Angell, H.K.; Bedognetti, D.; Marincola, F.M. The continuum of cancer immunosurveillance: Prognostic, predictive, and mechanistic signatures. *Immunity* **2013**, *39*, 11–26. [[CrossRef](#)] [[PubMed](#)]
49. Pol, J.; Bloy, N.; Buque, A.; Eggermont, A.; Cremer, I.; Sautes-Fridman, C.; Galon, J.; Tartour, E.; Zitvogel, L.; Kroemer, G.; et al. Trial Watch: Peptide-based anticancer vaccines. *Oncoimmunology* **2015**, *4*, e974411. [[CrossRef](#)]
50. Pol, J.; Bloy, N.; Obrist, F.; Eggermont, A.; Galon, J.; Cremer, I.; Erbs, P.; Limacher, J.M.; Preville, X.; Zitvogel, L.; et al. Trial Watch: Oncolytic viruses for cancer therapy. *Oncoimmunology* **2014**, *3*, e28694. [[CrossRef](#)]
51. Vacchelli, E.; Eggermont, A.; Fridman, W.H.; Galon, J.; Tartour, E.; Zitvogel, L.; Kroemer, G.; Galluzzi, L. Trial Watch: Adoptive cell transfer for anticancer immunotherapy. *Oncoimmunology* **2013**, *2*, e24238. [[CrossRef](#)]
52. Vacchelli, E.; Eggermont, A.; Fridman, W.H.; Galon, J.; Zitvogel, L.; Kroemer, G.; Galluzzi, L. Trial Watch: Immunostimulatory cytokines. *Oncoimmunology* **2013**, *2*, e24850. [[CrossRef](#)] [[PubMed](#)]
53. Vacchelli, E.; Eggermont, A.; Galon, J.; Sautes-Fridman, C.; Zitvogel, L.; Kroemer, G.; Galluzzi, L. Trial watch: Monoclonal antibodies in cancer therapy. *Oncoimmunology* **2013**, *2*, e22789. [[CrossRef](#)] [[PubMed](#)]
54. Vacchelli, E.; Eggermont, A.; Sautes-Fridman, C.; Galon, J.; Zitvogel, L.; Kroemer, G.; Galluzzi, L. Trial watch: Oncolytic viruses for cancer therapy. *Oncoimmunology* **2013**, *2*, e24612. [[CrossRef](#)] [[PubMed](#)]
55. Vacchelli, E.; Galluzzi, L.; Fridman, W.H.; Galon, J.; Sautes-Fridman, C.; Tartour, E.; Kroemer, G. Trial watch: Chemotherapy with immunogenic cell death inducers. *Oncoimmunology* **2012**, *1*, 179–188. [[CrossRef](#)]
56. Iribarren, K.; Bloy, N.; Buque, A.; Cremer, I.; Eggermont, A.; Fridman, W.H.; Fucikova, J.; Galon, J.; Spisek, R.; Zitvogel, L.; et al. Trial Watch: Immunostimulation with Toll-like receptor agonists in cancer therapy. *Oncoimmunology* **2016**, *5*, e1088631. [[CrossRef](#)]
57. Pol, J.; Buque, A.; Aranda, F.; Bloy, N.; Cremer, I.; Eggermont, A.; Erbs, P.; Fucikova, J.; Galon, J.; Limacher, J.M.; et al. Trial Watch-Oncolytic viruses and cancer therapy. *Oncoimmunology* **2016**, *5*, e1117740. [[CrossRef](#)]
58. Scholler, N.; Perbost, R.; Locke, F.L.; Jain, M.D.; Turcan, S.; Danan, C.; Chang, E.C.; Neelapu, S.S.; Miklos, D.B.; Jacobson, C.A.; et al. Tumor immune contexture is a determinant of anti-CD19 CAR T cell efficacy in large B cell lymphoma. *Nat. Med.* **2022**, *28*, 1872–1882. [[CrossRef](#)]
59. Vacchelli, E.; Senovilla, L.; Eggermont, A.; Fridman, W.H.; Galon, J.; Zitvogel, L.; Kroemer, G.; Galluzzi, L. Trial watch: Chemotherapy with immunogenic cell death inducers. *Oncoimmunology* **2013**, *2*, e23510. [[CrossRef](#)]
60. Marliot, F.; Lafontaine, L.; Galon, J. Immunoscore assay for the immune classification of solid tumors: Technical aspects, improvements and clinical perspectives. *Methods Enzymol.* **2020**, *636*, 109–128. [[CrossRef](#)]

61. Antoniotti, C.; Rossini, D.; Pietrantonio, F.; Catteau, A.; Salvatore, L.; Lonardi, S.; Boquet, I.; Tamberi, S.; Marmorino, F.; Moretto, R.; et al. Upfront FOLFOXIRI plus bevacizumab with or without atezolizumab in the treatment of patients with metastatic colorectal cancer (AtezoTRIBE): A multicentre, open-label, randomised, controlled, phase 2 trial. *Lancet Oncol.* **2022**, *23*, 876–887. [[CrossRef](#)] [[PubMed](#)]
62. Ghiringhelli, F.; Bibeau, F.; Greillier, L.; Fumet, J.D.; Ilie, A.; Monville, F.; Lauge, C.; Catteau, A.; Boquet, I.; Majdi, A.; et al. Immunoscore immune checkpoint using spatial quantitative analysis of CD8 and PD-L1 markers is predictive of the efficacy of anti-PD1/PD-L1 immunotherapy in non-small cell lung cancer. *EBioMedicine* **2023**, *92*, 104633. [[CrossRef](#)] [[PubMed](#)]
63. Pagès, F.; Taieb, J.; Laurent-Puig, P.; Galon, J. The consensus Immunoscore in phase 3 clinical trials; potential impact on patient management decisions. *Oncoimmunology* **2020**, *9*, 1812221. [[CrossRef](#)] [[PubMed](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.