ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΤΜΗΜΑ ΙΑΤΡΙΚΗΣ
ΠΜΣ «Μεθοδολογία Βιοϊατρικής Έρευνας, Βιοστατιστική και Κλινική Βιοπληροφορική

Έτος 2018-2019

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Assessment of the reporting quality of studies investigating the prognostic accuracy of immune markers, CD8 and Foxp3, in the prediction of response to cancer immunotherapy published from 2000-2019 using the STARD-TRIPOD statements

Εκτίμηση της ποιότητας αναφοράς μελετών που διερευνούν την προγνωστική ακρίβεια ανοσολογικών δεικτών, CD8 και Foxp3, στην πρόβλεψη ανταπόκρισης στην ανοσοθεραπεία του καρκίνου που δημοσιεύθηκαν κατά την περίοδο 2000-2019 χρησιμοποιώντας τις οδηγίες STARD-TRIPOD

Τριμελής επιτροπή:
Δοξάνη Χρυσούλα (επιβλέπουσα)
Στεφανίδης Ιωάννης
Ζιντζαράς Ηλίας

Νικόλαος Πασχαλίδης
(ΑΜ 222)
2019

Ευχαριστίες.

Θα ήθελα να ευχαριστήσω τα μέλη της τριμελούς επιτροπής κα Δοξάνη, κ. Ζιντζαρά και τον κ. Στεφανίδη για τις προτάσεις και οδηγίες για την ολοκλήρωση της εργασίας αυτής.

Ένα μεγάλο ευχαριστώ στην Έλενα Κριτσέλη για την καθοδήγηση και τις διορθώσεις.

Δεν μπορώ να μην ευχαριστώ τους γονείς μου, Γιώργο και Εύη, για την στήριξη όλα αυτά τα χρόνια και ειδικά τη φετινή, δύσκολη, χρονιά.

Επίσης θα ήθελα να ευχαριστήσω την γυναίκα μου Έλενα και την κόρη μου Ναταλία για την υπομονή, την στήριξη και την αγάπη τους καθώς επίσης και την πεθερά μου Λέλα και τον πεθερό μου Κώστα για τη βοήθεια και στη στήριξη ειδικά τα σαββατοκύριακα που ήμουν στη Λάρισα.

*Τέλος, θα ήθελα να αφιερώσω την εργασία αυτή στην εκλιπούσα καθηγήτρια του εργαστηρίου που πραγματοποιώ την μεταδιδακτορική μου έρευνα στο Ι.ΙΒ.Ε.Α.Α., και καλή μου φίλη, **Βίλη Πανουτσακοπούλου**, που πέθανε τον Νοέμβριο του 2018 από λευχαιμία. Μακάρι να μπορούσα να μοιραστώ τις εμπειρίες και τις γνώσεις που πήρα από το μεταπτυχιακό αυτό μαζί της.*

Για τη Βίλη.

Abstract (*English*)

Introduction: Despite the proven potential of immunotherapy (IT) in increasing overall survival rates in cancer patients, low response rates, side effects and high costs are still big hurdles to deal with. Reliable predictive biomarkers for IT (and especially immune-checkpoint inhibition, ICI, therapy), other than PD-L1 and tumor mutational burden (TMB) are needed. Less widely investigated predictive biomarkers are tumor-infiltrating lymphocytes (TILs) like $CD8^+$ and $Foxp3^+$ T cells. Inadequate reporting of clinical studies for the development of prognostic tools impedes identification and reproducibility and restrains applicability of results and standardization of methods. This study sought to evaluate the reporting quality of studies that investigate predictive ability of baseline, pre-treatment, CD8 and Foxp3 levels in cancer immunotherapy responses.

Aims: To identify studies with predictive claims on response to IT based on baseline phenotyping of CD8 and FoxP3, record study characteristics and asses their reporting quality using STARD and TRIPOD statements.

Methods: A systematic review was conducted according to PRISMA guidelines. 201 studies were retrieved from MEDLINE and screened for eligibility. 27 studies were assessed for reporting quality using STARD statement and 6 out of 27 with STARD and TRIPOD statements.

Results: A lack of consensus in the use of methods, reagents and analysis to make predictive claims on the use of these markers was evident. High quality reporting studies were stronger in reporting measures of prognostic accuracy. All the studies that were scored with TRIPOD scored high using the STARD statement.

Conclusions: In order for baseline measurements of markers CD8 and Foxp3 to be more efficiently evaluated as potential predictive markers for IT, the reporting quality of subsequent studies has to be improved. This can be aided by STARD and TRIPOD guidelines. This report also serves this task by aiding in the dissemination of these guidelines.

Keywords: *Immunotherapy, CD8, Foxp3, prediction, STARD, TRIPOD*

Abstract (*Greek*)

Εισαγωγή: Παρά την αποδεδειγμένη δυναμική της ανοσοθεραπείας στην αύξηση του συνολικού ποσοστού επιβίωσης ασθενών με καρκίνο, τα χαμηλά ποσοστά ανταπόκρισης, οι ανεπιθύμητες ενέργειες και το υψηλό κόστος παραμένουν ακόμη μεγάλα εμπόδια. Απαιτούνται αξιόπιστοι προβλεπτικοί βιοδείκτες για την ανοσοθεραπεία (και ειδικότερα για τις θεραπείες αναστολής σημείου ελέγχου) πέρα από τον δείκτη PD-L1 και το TMB. Λιγότερο μελετημένοι προβλεπτικοί βιοδείκτες είναι τα Τ λεμφοκύτταρα $CD8^+$ και $Foxp3^+$. Τα ανεπαρκή στοιχεία στις αναφορές κλινικών μελετών για την ανάπτυξη προβλεπτικών εργαλείων εμποδίζουν την εφαρμοσιμότητα των αποτελεσμάτων και την τυποποίηση των μεθόδων. Η μελέτη αυτή αξιολόγησε την ποιότητα αναφοράς των μελετών που διερευνούν την προβλεπτική αξία των προ-θεραπείας επιπέδων των δεικτών CD8 και Foxp3 στην ανοσοθεραπεία στον καρκίνο.

Στόχοι: Η αναγνώριση των μελετών που ερευνούν την προβλεπτική αξία των προ-θεραπείας μετρήσεων των δεικτών CD8 και FoxP3, η καταγραφή των χαρακτηριστικών των μελετών αυτών και η αξιολόγηση της ποιότητας αναφορών χρησιμοποιώντας τις οδηγίες STARD και TRIPOD.

Μέθοδοι: Πραγματοποιήθηκε συστηματική ανασκόπηση και 201 υποψήφιες μελέτες ανακτήθηκαν από την MEDLINE και εξετάστηκαν για επιλεξιμότητα. 27 μελέτες αξιολογήθηκαν για την ποιότητα αναφοράς χρησιμοποιώντας τις οδηγίες STARD. 6 μελέτες αξιολογήθηκαν με τις οδηγίες STARD και TRIPOD.

Αποτελέσματα: Οι μελέτες που επεξεργάσθηκαν παρουσίασαν ετερογένεια στη χρήση μεθόδων, αντιδραστηρίων και ανάλυσης. Οι μελέτες με υψηλό σκορ ποιότητας αναφοράς ήταν ισχυρότερες όσον αφορά τις αναφορές στην ακρίβεια των προγνωστικών μετρήσεων. Οι μελέτες που βαθμολογήθηκαν με τη λίστα TRIPOD σημείωσαν υψηλή βαθμολογία στη λίστα STARD.

Συμπεράσματα: Προκειμένου οι μετρήσεις προ-θεραπείας των δεικτών CD8 και Foxp3 να αξιολογηθούν πιο αποτελεσματικά ως δυνητικοί δείκτες πρόβλεψης στην ανοσοθεραπεία πρέπει να βελτιωθεί η ποιότητα αναφοράς των σχετικών μελετών. Η βελτίωση αυτή μπορεί να γίνει υπό την καθοδήγηση των οδηγιών STARD και TRIPOD. Η παρούσα μελέτη εξυπηρετεί και αυτό το καθήκον καθώς βοηθάει στη διάδοση αυτών των κατευθυντήριων γραμμών.

Λέξεις κλειδιά: *Ανοσοθεραπεία, CD8, Foxp3, πρόβλεψη, STARD, TRIPOD*

Table of **C**ontents

1. Introduction

1.1 Cancer immunotherapy

In recent years, the development of new therapies for cancer has made rapid progress. Traditional therapies are considered surgery, chemotherapy and radiotherapy. Radiotherapy and chemotherapy involve direct killing of tumor cells (interference with cell division, DNA damage, and intercalation, synthesis-blocking). However, these therapies come with a high toxicity burden and some treated patients exhibit certain patterns of resistance. Immunotherapy (IT) represents the fourth generation of therapeutics against cancer.

The immune system is highly capable of specifically destroying tumors with minimal toxicity to normal tissue while maintaining long-term memory, preventing future recurrence of the disease. The great advances in cancer immunology in recent years have provided the knowledge and techniques to develop innovative immunotherapeutic approaches. Although by definition immunotherapy consists of many ways in which the immune system of a patient can be modulated in order detect and eliminate tumor cells (with the use of vaccines , recombinant cytokines and preformed monoclonal antibodies) its the discerning of mechanisms that inhibit anti-tumor immunity that has led to the development of immune-checkpoint inhibitors (ICI) that generated a paradigm shift in cancer treatment[1]. In reality, this type of therapy works by removing immune system's "brakes". The different forms of IT and a simplified graphic of anti-tumor immunity are displayed in **Figure 1.**

A

Table 1. Different forms of immunotherapy used to manage cancer.

Form of immunotherapy	Example(s)	Tumor type(s) where used
Cytokines	IL2, interferon-α	Melanoma, renal
Vaccines	Sipuleucel-T	Prostate
	BCG[a]	Noninvasive bladder
Antibodies	Rituximab	B-cell lymphoma
	Trastuzumab*	Breast, stomach
Adoptive cell transfer	Tisagenlecleucel	Acute lymphoblastic leukemia
ICIs	Ipilimumab, pembrolizumab, nivolumab	Melanoma, lung, kidney, urothelial, Merkel cell carcinoma, Hodgkin lymphoma

[a] BCG, Bacillus Calmette-Guerin. Acts in part via an immune mechanism, i.e., via antibody-dependent cell-mediated cytotoxicity.

B

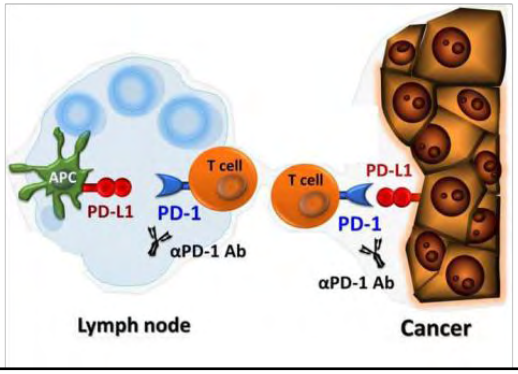

Figure 1. ***(A)*** *Different forms of immunotherapy with examples and applied tumor type (adopted from[2]) **(B)** Simplified graphic on anti-tumor immunity and key targets (PD-1/PD-L1) in immunotherapy (adopted from[3]). The anti-tumor immune response briefly involves cancer-cell antigen presented on an antigen presenting cell which then leads to T cell activation, response and proliferation followed by T cell migration to the tumor site and immune-mediated tumor-cell death (mainly mediated by $CD8^+$ cytolytic T cells This activation and response cascade is regulated by a balance of stimulatory and inhibitory signals called immune checkpoints, which control the magnitude of response.*

1.2 Emerging biomarkers in cancer immunotherapy

Despite the proven potential of immunotherapy in increasing overall survival rates in cancer patients, low response rates, side effects and high costs are big hurdles to deal with. Reliable predictive biomarkers for IT (and especially ICI therapies) are needed to manage cancer patients more efficiently in the long term. It is not the scope of this report to extensively review all different categories of biomarkers as this is a rapidly expanding landscape and has been extensively reviewed recently elsewhere[4]. The most thoroughly investigated predictive biomarkers for IT are PD-L1 (approved for first-line pembrolizumab monotherapy in lung cancer)[5], microsatellite instability/defective mismatch repair (MSI/dMMR), and tumor mutational burden (TMB). MSI/dMMR concerns mutations in genes that are involved in correcting mistakes of DNA replication (approved for the clinic irrespective for any tumor) whereas TMB measures the quantity of mutations found in a tumor (shown to predict response to several different forms of immunotherapy, across multiple cancer types[6]).

Less widely investigated predictive biomarkers for IT are tumor-infiltrating lymphocytes (TILs) like $CD8^+$ and $Foxp3^+$ T cells (**Figure 2A**). $CD8^+$ T cells are key players in immune-mediated tumor-cell death as upon activation they release granzymes and other lytic enzymes at the tumor site. CD8 is an extracellular protein (co-receptor to T cell Receptor, TCR) found mainly in these cells. It is generally considered that their accumulation is associated with favorable prognosis and positive predictive value for IT, however there is still a lack of consensus on the methods to optimize and standardize determination of these cells[2].

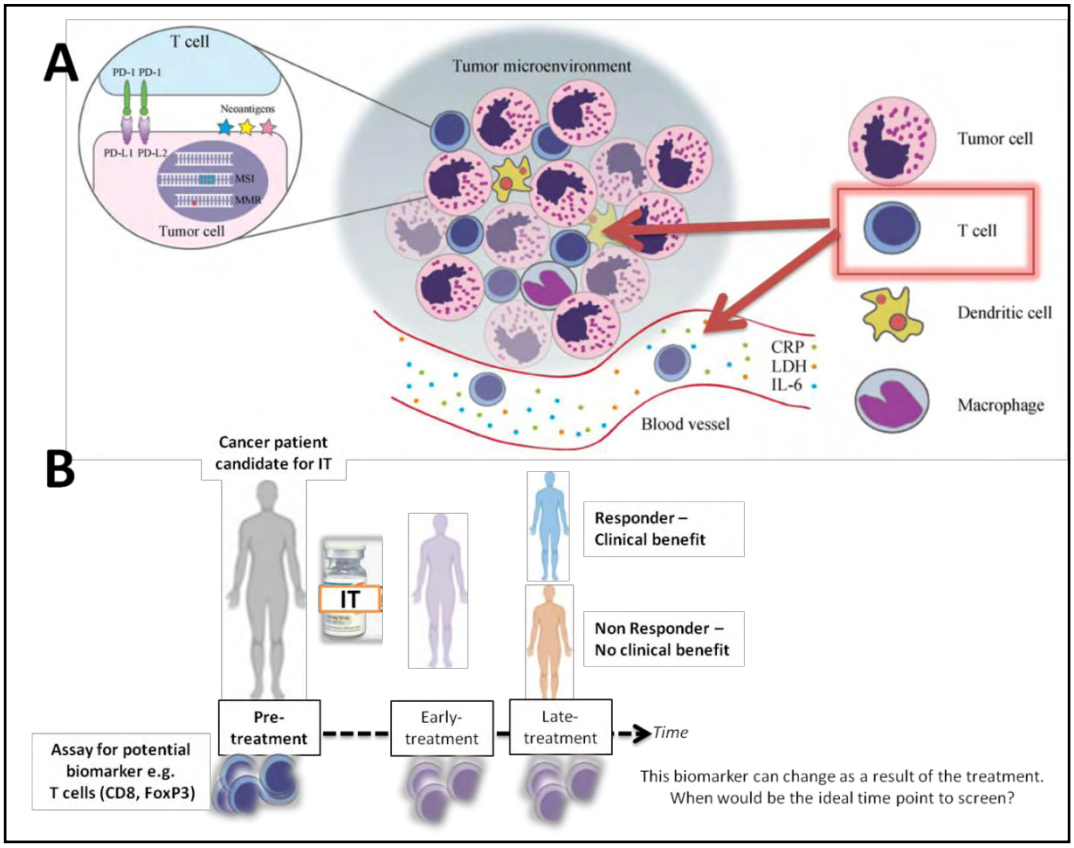

Figure 2. (A) *Biomarker diversity can be broken in different categories identified by type (soluble, cellular and genomic) as well as source (serum, peripheral blood, tumor infiltrating and tumor-originating). These include soluble factors in blood (CRP, LDH, IL-6), cells (like T cells-red box- macrophages, dendritic cells in the TME or blood) and MSI/MMR/TMB at the genomic level. Especially for T cells identification is based on the detection of specific surface/intracellular markers (CD8 and Foxp3) that are usually expressed exclusively by these populations (adopted from[7]).* ***(B)*** *Possible crucial determinants for standardizing predictions to be made based on the measurement of T cell markers are the assay used and the timing of sampling.*

Foxp3$^+$ regulatory (Treg) cells are specialized lymphocytes that halt exacerbated immune responses[8]. However, Tregs can also infiltrate malignant tumors and suppress beneficial anti-tumor immunity[9]. Their accumulation in the tumor microenvironment (TME) is associated with poor prognosis in various types of cancer, including colorectal cancer and melanoma[10]. However, recent studies have challenged this paradigm by showing that FoxP3$^+$ T cells exhibit heterogeneous phenotypes and, in some cohorts, are associated with favorable prognosis (reviewed elsewhere[11]). Foxp3 is a transcription factor (nuclear protein) that controls most of Tregs key phenotypic characteristics related to their function[12]. Despite its wide use as a Treg specific marker, in humans, activated T cells also exhibit patterns of transient FoxP3 expression which makes their detection difficult and sometimes impossible without the use of surrogate markers (like CD4, CD25, CD127 and others). Recent studies have shown that nivolumab reduces Treg suppressive effects by decreasing Foxp3 expression[13]. Thus, the timing (prior or after therapy) of assessing the phenotype of these cells could be critical **(Figure 2B).**

1.3 The importance of accurate reporting – STARD and TRIPOD guidelines

It is widely accepted that inadequate reporting of clinical studies can impede the identification and reproducibility of the study and restrain applicability of results and standardization of methods. Many guidelines have been developed (by an international group of methodologists, statisticians, reviewers and editors) to improve this. For diagnostic/prognostic studies, in particular, the Standards for Reporting of Diagnostic Accuracy (STARD) and the Transparent Reporting of a multivariable prediction model of Individual Prognosis Or Diagnosis (TRIPOD) have been proposed (http://www.equator-network.org/reporting-guidelines/stard/, http://www.equator-network.org/reporting-guidelines/tripod-statement/)[14,15]. The purpose of these guidelines is not to assess the actual quality of research performed, but to guide to increase the transparency and clarity of reporting. This will ultimately facilitate better access, interpretation and applicability of the suggested methods and results.

The updated STARD checklist (launched in 2003 and updated in 2015) contains 25 items that should be clearly reported to make the study manuscript fully informative. A full list of these items with a brief description can be found in the **Appendix** section (Table A2, *p. 17).* In brief, these items enquire about whether specific details in introduction, methods, results and discussion like information regarding participants, index test, reference standards, statistical analysis and relevant results are reported.

The TRIPOD statement is more focused for reporting of studies that propose the development, validation of a prediction model (for diagnostic or prognostic purposes).It consists of a checklist of 22 items that interrogate the reporting of characteristics like source of data, participants, predictors, model development and performance. Clearly, there is some overlapping with the STARD statement, however the TRIPOD statement allows for a deeper investigation of reporting focused around development and validation cohorts, necessary for predictive model design. A full list of these items in the TRIPOD statement with brief descriptions can be found in the **Appendix** section (Table A3, *p. 18).*

1.3 Aims of this report

The main aims of this report were: **(*1*)** to identify studies that report predictive claims on response to immunotherapy based on baseline phenotyping data on markers CD8 and Foxp3, **(*2*)** to extract data from these studies concerning basic, methodological, technical characteristics and results and **(*3*)** to asses reporting quality of the identified studies using STARD and TRIPOD statements.

2. Methods

2.1 Study identification and selection

A systematic search for clinical studies was performed, based on PRISMA guidelines, in database MEDLINE, using PubMed. The aim was to identify clinical studies published from January 1999 to July 2019 that investigated the association of phenotyping of immune markers CD8 and/or FoxP3 in the prediction of response to cancer immunotherapy. The search used the following terms: ((“predictive” or “prediction”) and (“immunotherapy”) and (“tumor” or “cancer”) and (“CD8” or “Cd8” or “Foxp3” or “FOXP3” or “FoxP3” or “foxp3” or “Tregs”)). The references of the retrieved articles were also screened for potential additional eligible studies. A list with all retrieved articles from PubMed prior to screening can be also found in a pdf file in the link provided in the **Appendix** section (see *bottom of p.15*).

From the retrieved studies, a selection step was performed with the following criteria: (1) they contained patients diagnosed with cancer (any type) to be treated as part of a trial/study with immunotherapy (any type), (2) reported baseline (pre-treatment, before immunotherapy) measurements of markers CD8 and/or FoxP3 levels (protein and/or mRNA), (3) published in English language, (4) they were human studies (and not animal model studies) and (6) they were full reports (and not conference abstracts, presentations short reports, or commentaries).

2.2 Data abstraction

All selected studies were critically evaluated using a variety of study settings, technical details of phenotyping assays as well as statistical analyses and models used. A detailed list of all the information abstracted from the studies can be found in the **Appendix** section, Table A1**,** *p.16*.

2.3 Assessment of study reporting quality using reporting tools STARD and TRIPOD

To assess the quality of reporting of the selected studies the STARD guidelines (***Appendix****, Table A2 p. 17*). Undoubtedly, all items in this list are important to fully assess the quality of reporting of diagnostic/prognostic studies, however for the purposes of this report the focus was mainly on items related to methodology, results and analysis. Thus, from the STARD checklist, items 1,2,3,4 and 26, 27, 28, 30 were excluded and items 5 to 25 were included. According to the principles of the STARD guidelines, these items were investigated on the basis of whether they were fully and clearly reported and not whether they were actually carried out during the study. Using a scoring approach previously described[16], items were scored as ‘1’ if they were reported in adequate detail to allow reader to judge that the definition/description had been met or ‘0’ if the item was absent or poorly described, a scoring approach which has been previously described[16].
In addition to STARD we also used the TRIPOD tool (***Appendix****, Table A3, p. 18*) for the studies that clearly stated the use of a model that was developed using discovery/training/developing and validation/testing cohorts. For this purpose, and in a similar way as above, the focus was on methodological, results and analysis descriptions so items 6a to 20 were included in the analysis. As above, items were scored as ‘1’ if they were clearly reported in detail and as ‘0’ if absent or unclear. Finally, studies were qualified as high quality and low quality. For STARD assessment studies were qualified as high when score was ≥12 and lower quality when score was <12. For TRIPOD assessment studies were qualified as high when score was ≥16,5 and lower quality when score was <16,5. The choice of quality score cut-offs was the median of the overall quality scores of the studies. The overall quality score was calculated by summing the score of reported items.

2.4 Statistical analysis of STARD scores

To express the association between proportions of reporting an item in the STARD checklist across the high/low grouping (based on the median STARD score of all the studies), the Fisher’s exact test was used, *P* values <0.05 indicated statistical significance.

3. Results

3.1 Eligible Studies

A search for studies was conducted using PubMed and identified 165 articles that met the search criteria (*see Methods*). Additionally, 36 articles were identified from references of retrieved articles and reviews. All together, 201 articles were screened for eligibility. A flow diagram of retrieved and excluded articles with specification reasons is shown in **Figure 3**. In total, 27 unique articles remained for complete full-text evaluation and assessment of reporting quality using STARD and TRIPOD statements. The full list of the 27 articles can be found in the **Appendix** (*see List of selected studies, p. 15*).

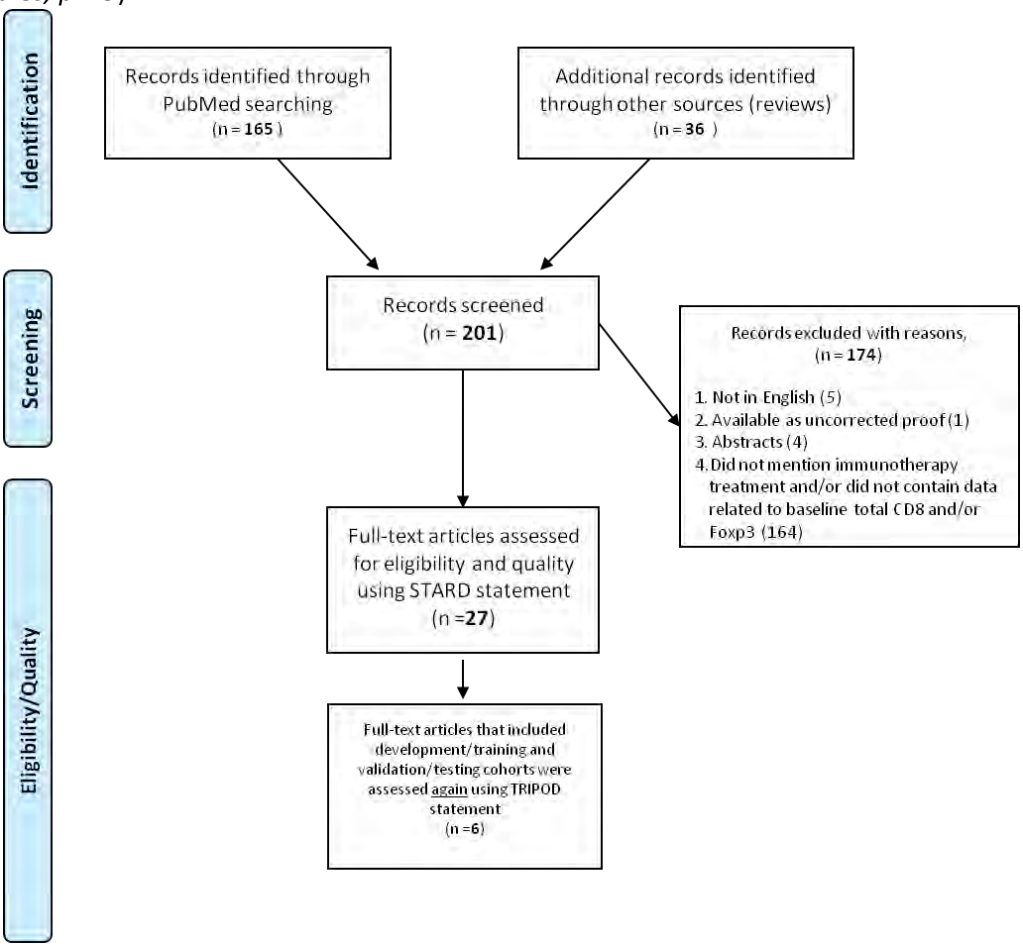

Figure 3. *A Flow diagram of identified and eligible studies. A full list of selected studies can be found in the* ***Appendix,*** *List of selected studies, p. 15).*

3.2 Study characteristics

All eligible studies were subjected to screening and data extraction. Briefly, data extracted included study main characteristics (First author, date, Journal, country, setting etc), study design (retrospective or prospective, sample size, type of cancer, type of immunotherapy etc) as well as methodological details (markers assessed, type of assay, antibody clones, blind assessment etc), type of analysis performed (Univariate, multivariate) and information related to the predictive claim of the study. A full list with of the extracted characteristics with a brief description can be found in the **Appendix** (**Table A1**, *p.16*).

An excel file with all the extracted data from the 27 identified studies can be found in this link:
https://drive.google.com/drive/folders/1C0xSkKvUPd_pU4cJRg6ZEqR0NR_NOVvm?usp=sharing

Percentage analysis of basic characteristics of the studies revealed that 18.5% of the studies (5 studies) were published in *AACR Clinical Cancer Research* followed by *Cancer Immunology and Immunotherapy* (11.1%, 3 studies) (**Figure 4A**). Interestingly, 50% of the journals were STARD endorsers and 2 out of 16 journals explicitly endorsed TRIPOD guidelines. In addition, 51.8% of the identified studies were published in STARD endorsing journals **(Figure 4A).**

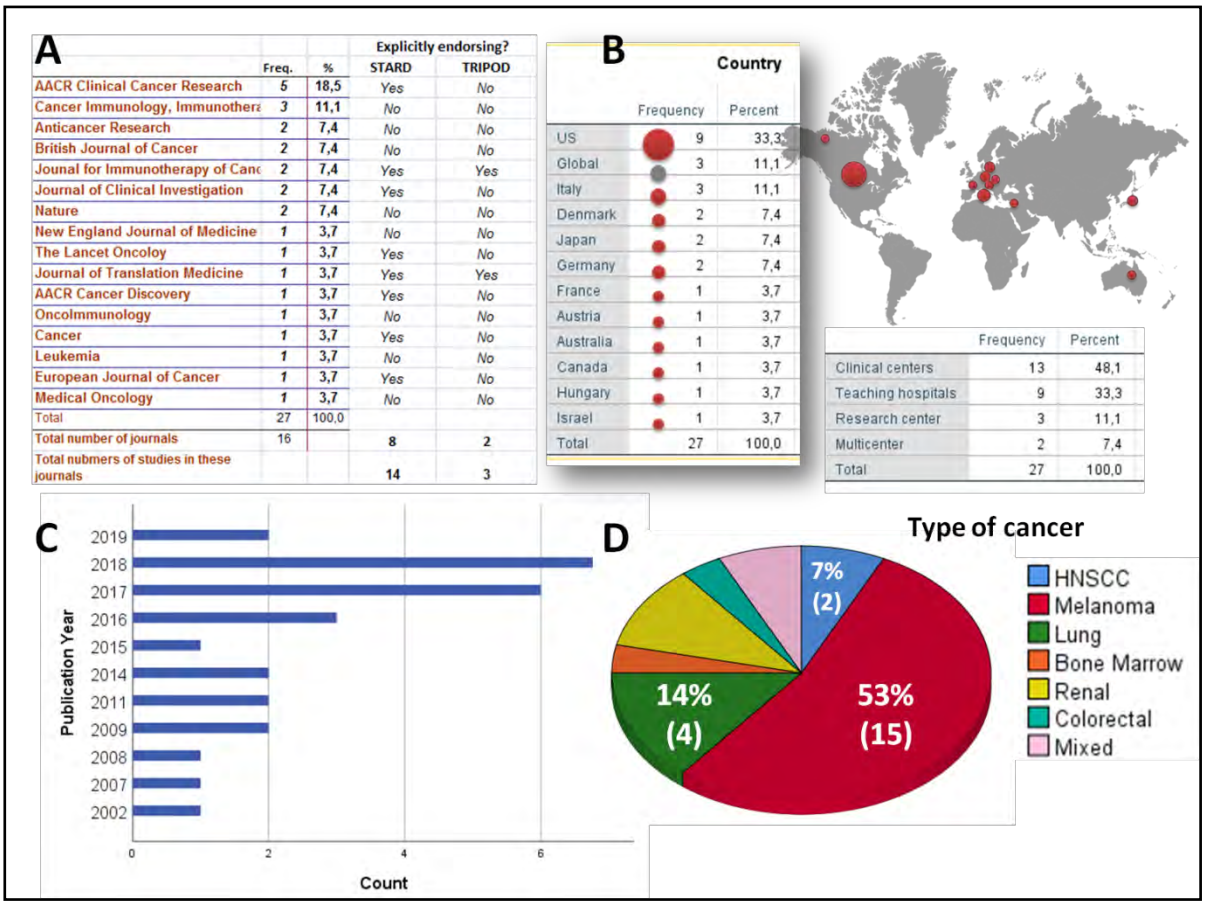

Figure 4. *Percentage analysis of retrieved studies regarding* ***(A)*** *Journal and endorsement of reporting guidelines,* ***(B)*** *Population,* ***(C)*** *Year Published and* ***(D)*** *type of cancer under treatment*.

A big majority of identified studies were performed in the US (33.3%) in clinical centers. Most publications were dated from 2016 onwards and in half of the studies the type of cancer under treatment was melanoma (53%) followed by Lung (14%) and Head and Neck cancer (7%) (**Figure 4B-D**).

Regarding the reported information on basic characteristics of patients, the mean of the all the median ages was 58.8 years, the mean of total study size was 99.2 (range 7-401) and the mean number of samples (events) with baseline marker data included in subsequent analysis was 57 (range 7-175) (**Figure 5A**). Concerning therapy, 26% of studies reported as prior therapy chemotherapy and the majority of studies involved administration of anti-CTLA4 (Ipilimumab) followed by anti-PD1 (nivolumab) and anti-PD1/PDL1 (**Figure 5B,C**). It is interesting to note here that 19% of studies did not report prior therapy and 26% of studies reported the existence of prior therapy but did not specify the type. The majority of the studies had a prospective design, whereas 8 studies had a retrospective design (**Figure 5D**).

The features that very often dictate the quality of study output, but most importantly overall transparency and reproducibility are methodological and technical details. Concerning the markers screened in the identified studies, 48% studied CD8, 44% CD8 and FoxP3 whereas 7% (2 studies) screened for FoxP3 alone (**Figure 6A**). In addition, 70% of studies used additional markers either as surrogate markers or as parallel/reference tests (**Figure 6B**). Among these, the most prevalent was PD-L1 (in 9 studies). One study used also TMB in the analysis of predictors. Concerning technical details, percentage analysis revealed that most of the studies used immunohistochemistry (IHC) techniques and reported using sections from 'tumor biopsies' (17 studies), 'tumor' (9 studies), 'tumor and stroma' (7 studies) or 'tumor center' (1 study) (**Figure 6C**). A smaller fraction of studies used blood (7 studies) as the biological material under investigation. Finally, the majority of studies had at least one screening technique that involved the use of an antibody for the detection of CD8

and Foxp3. The spectrum of antibodies used can be seen in **Figure 6D**. Many studies did not report the specific clone that was used. The clone SP16 (anti-CD8) was used in 4 studies. clone (anti-CD8) (**Figure 6D**).

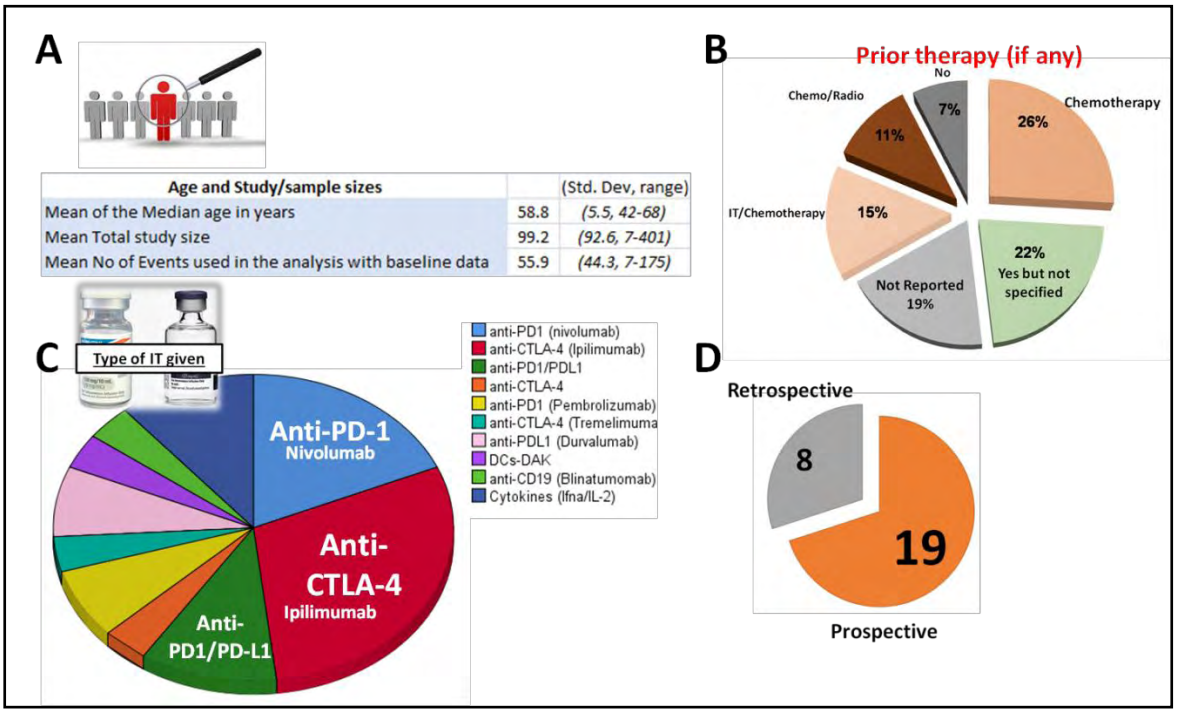

Figure 5. Descriptive statistics and *percentage analysis on retrieved studies regarding **(A)** Patient age and study sizes, **(B)** prior therapy, **(C)** type of Immunotherapy (IT) and **(D)** type of study design.*

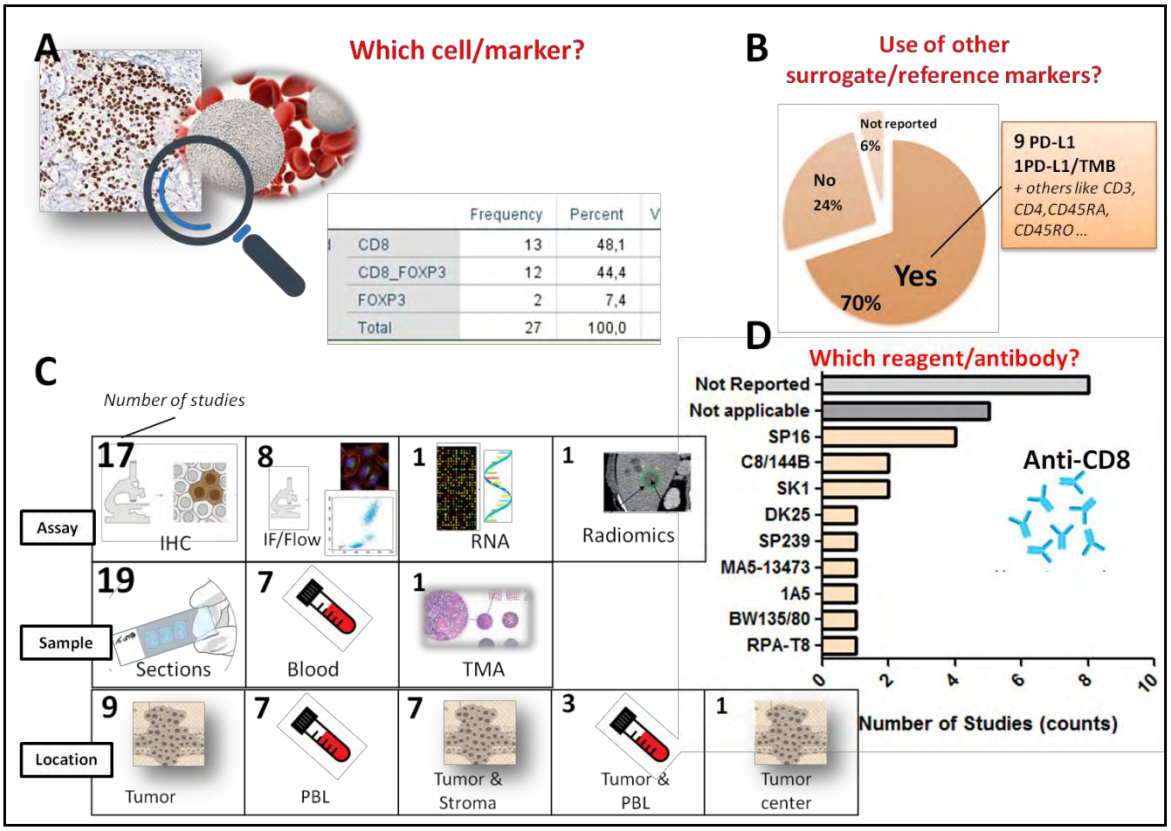

Figure 6. *Percentage analysis on retrieved studies regarding **(A)** which marker was measured, **(B)** whether surrogate and/or reference markers were used **(C)** methodological details such as the type of assay, sample and location used and **(D)** the clones of antibodies used, if applicable.*

Next, information related to methodological aspects of these studies such as the outcome index, assessment, model development/validation and blindness of assessment (*Extracted data with codes C16, C28, C43 and C29* **Appendix Table A1**, *p.20*) was extracted. Most of the studies (66%) reported the use of the Response Evaluation Criteria in Solid Tumors (RECIST) or RECIST 1.1[17] (**Figure 7A**). Assessment of index test was performed by investigator(s)/pathologist(s) in 13 studies and investigator and computer (automated- algorithm based system) in 8 studies whereas 3 studies report the exclusive use of computer based assessment method and 3 studies do not specify how assessment was done **(Figure 7B**). Only 6 studies report the design and use of a predictive model using training/discovery/development and test/validation cohorts and only these studies were further analyzed for reporting quality using the TRIPOD statement **(Figure 7C**). Finally, percentage analysis on the blindness of the assessment of the index test in relation to the therapy outcome revealed that 55.5% of studies (15 studies) did not report such information whereas 12 did (10 reported 'yes' and 2 reported 'no') **(Figure 7D**).

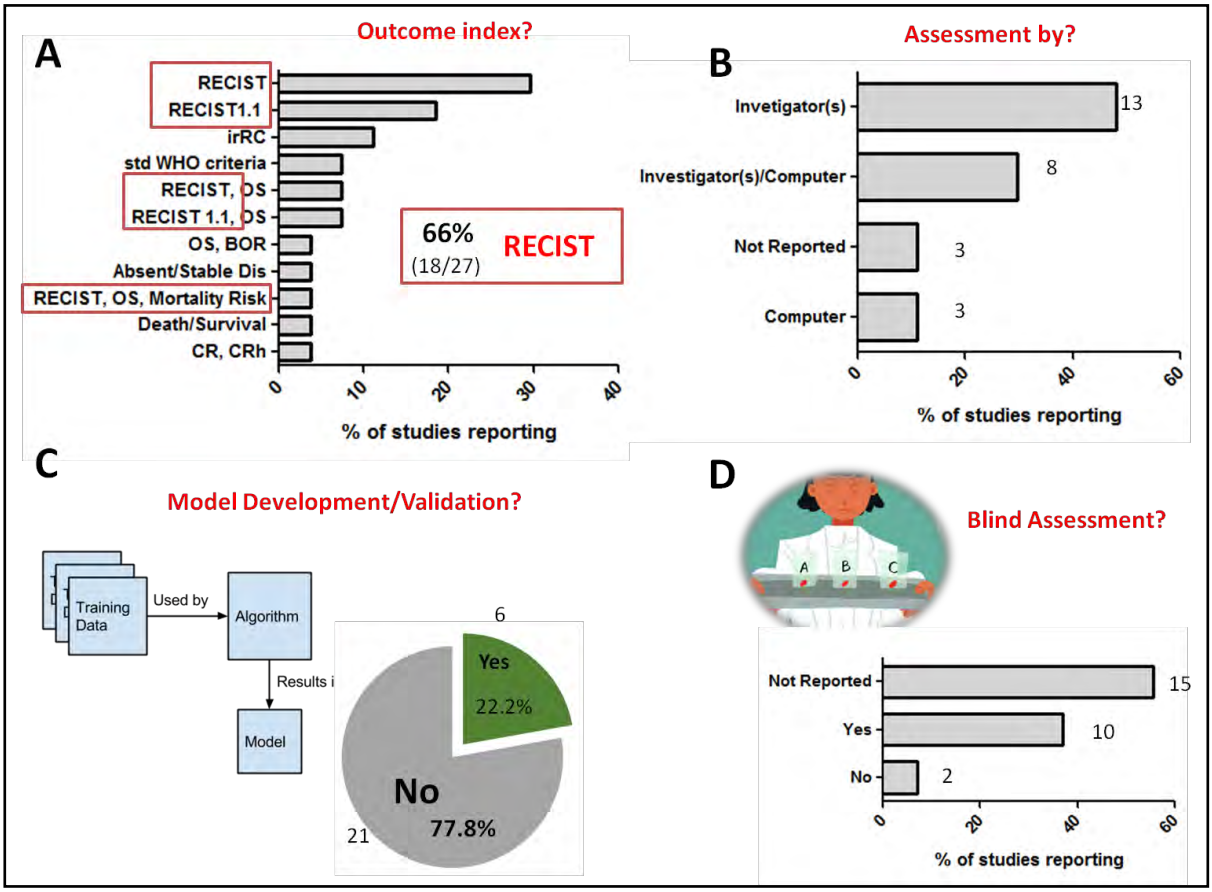

Figure 7. *Percentage analysis of retrieved studies regarding* ***(A)*** *outcome index,* ***(B)*** *assessment method,* ***(C)*** *whether the study included training/validation cohort and* ***(D)*** *whether assessment was performed in a blind fashion.*

Concerning the reporting of results in these studies, the analysis of responders and non-responders throughout the cohort of the identified studies reveals that the mean percentage of non-responders is much higher than non-responders, however (despite the small sample size) this trend changes if these results are displayed based on patient stratification according to type of IT given **(Figure 8A)**. Nearly half of the studies (44%, 12 studies) report positive association (meaning that more/positive cells are found in responders) of these markers (any) with clinical benefit whereas 11 studies report no association and 4 studies report negative or poor association (either positive or negative) **(Figure 8B)**. Overall, 46.4% of the studies report that baseline measurements of these markers can be used to predict responses **(Figure 8C)**. Finally, only a few studies explicitly specified a cutoff metric such as a percentage of cells in blood/tissue or a specific number of cells per unit area. Overall, a wide spectrum of cutoffs (numbers/methods) for discrimination of patients based on these markers (either CD8 or Foxp3) were used **(Figure 8D)**.

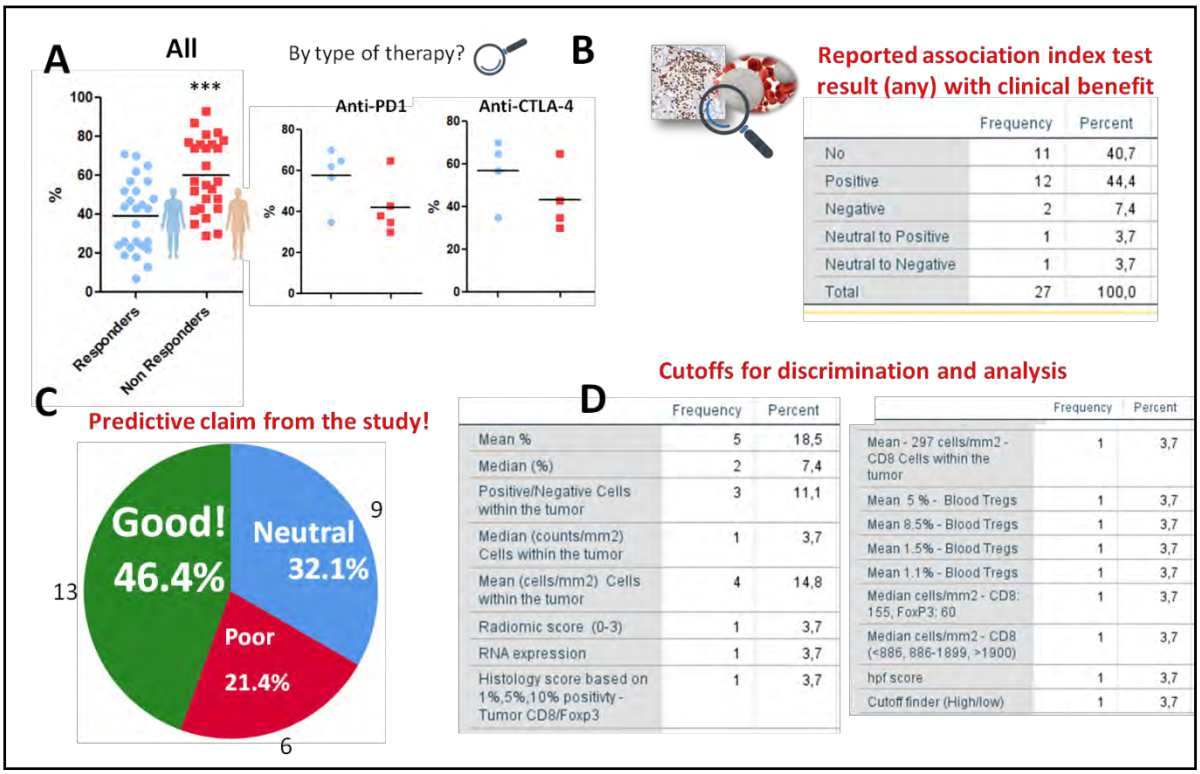

Figure 8. *Percentage analysis of retrieved studies regarding* ***(A)*** *response to IT by any type (all) or stratified by anti-PD1/anti-CTLA-4 studies,*** p<0.001, Mann-Whitney test,* ***(B)*** *association with clinical benefit based on marker examined,* ***(C)*** *predictive claim from the study and* ***(D)*** *the wide range of cutoffs and ways of discriminating patients based on the measurements of these markers.*

3.3 Assessment of reporting quality using STARD and TRIPOD statements

3.3.1 Assessment of reporting and scoring using STARD and TRIPOD

In parallel with data the reporting quality of identified studies was assessed and all studies were scored using the STARD statement (*see Methods, section 2.3*). As mentioned in the *Introduction,* (*section 1.3)* from the 25 items of the STARD checklist the focus was on items of methodology and analysis whereas items related to title, abstract, hypothesis and funding were excluded from the scoring scheme (*see highlighted items in yellow in* ***Appendix, Table A2****)*. However, it has to be noted that most of the studies did adequately report on the aforementioned features. The full table that contains the marks for every identified study can be found here in an excel file in this link:*https://drive.google.com/drive/folders/1C0xSkKvUPd_pU4cJRg6ZEgR0NR_NOVvm?usp=sharing*

The percentages of reporting items of STARD can be seen in **Figure 9A.** The percentage of reporting was very high in the methods section (study design and participants). Then moving into the analysis and results section, percentage of reporting gradually decreased and reached 0% in item 18 which relates to sample size and how it was determined. None of the studies reported how sample size was determined. Very low percentage (almost 0%) can be seen in items 22 (which is about "*time interval and any clinical interventions between index test and reference standard*" and item 23 which is about "*Cross tabulation of the index test results (or their distribution) by the results of the reference standard"*. Finally, all studies do not report adverse events from performing index test since in this particular group this was mainly blood collection and tumor biopsies.

Further assessment and marking was done in studies that include the reporting of a training/validation design for a prediction model in response in immunotherapy. From the 27 identified studies only 6 studies reported such design; these were studies with id *5, 6, 11, 12, 17 and 24*. In a similar way with the STARD scoring scheme the percentage of reporting for each item of the TRIPOD checklist can be seen in **Figure 9B.**

Reporting percentage was 0 for items 8, 10e, 14b and 17. Item 8 relates to study size, item 10e relates to *"model updating (e.g., recalibration) arising from the validation",* item 14b is about reporting on "*the unadjusted association between each candidate predictor and outcome*" and item 17 is about reporting on "results from any model updating (i.e., model specification, model performance)". Regarding item 8, it was expected since these studies were also scored using the STARD guidelines. Regarding items 10e and 1, clearly these issues do not apply to current studies as yet. Perhaps in future follow-ups, if more data are available these issues could be reported. Regarding item 14b, the unadjusted association would enhance transparency in reporting in these studies and could have been included perhaps as supplementary material. Reporting percentages were 100% in items 6a, 12, 13c and 18 to 20. These items relate to the reporting of a clear outcome to be predicted (*item 6a*) as well as reporting of comparisons of training and test cohorts (*items 12, 13c*) to identify study limitations and future use of the model (*items 18 to 20*). Finally, the studies that were marked with TRIPOD were found to be on the high range of STARD scoring **(Figure 9C)**.

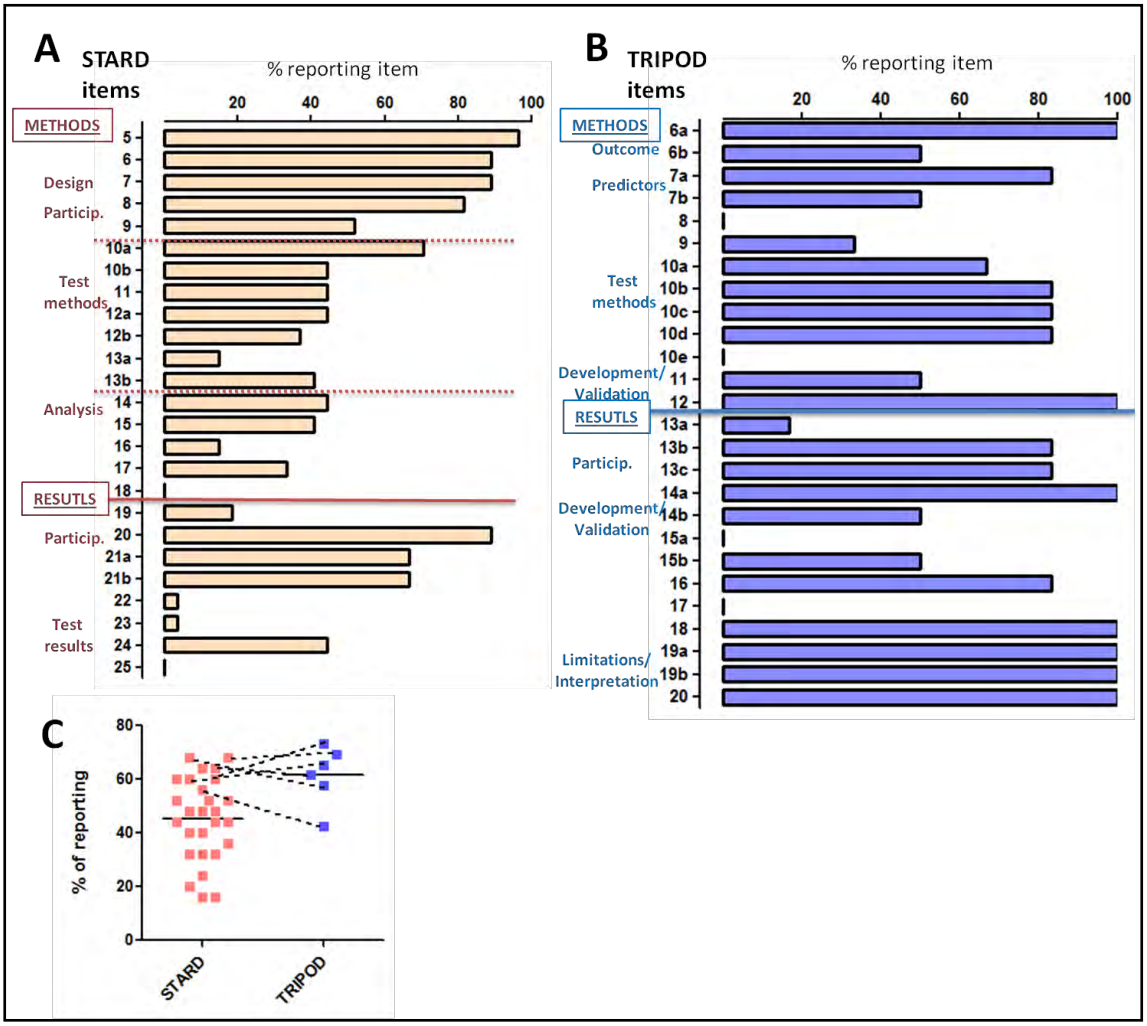

Figure 9. ***(A)*** *Bar graph showing percentage of reporting items (5 to 25) of the STARD 2015 checklist for all the 27 identified studies.* ***(B)*** *Bar graph showing percentage of reporting items (6 to 20) of the TRIPOD checklist for 6 studies (reporting development/validation cohorts for prediction model) with id 5, 6, 11, 12, 17 and 24 (see list in Appendix p.15).* ***(C)*** *Scatter plot showing the percentage of reporting for all items using STARD (n=27) and TRIPOD (n=6) guidelines. Dashed lines connect same studies assessed with both reporting tools.*

3.3.2 Effect of study quality

The median score from the above marking schemes was calculated. The median score for STARD from all studies was 12. This score was used to dichotomize the group to high/low scoring, so 15 studies (*id: 1,2,3,4,5,6,8,11,12,15,16,17,18 and 27*) were classified as high quality and 12 studies (*id: 7,9,10,13,14,19,20,21,23,24,25 and 26*) were classified as low quality. This dichotomization was used

to investigate for associations between proportions of an item across the two groups of articles as previously described[16]. Similarly, the same rationale was applied to calculate median scores and dichotomize and classify the TRIPOD assessed studies. The median score from all studies was 16.5 resulting in 3 studies (*id: 5, 6 and 17*) to be classified as high quality score and 3 studies (*id: 11, 12 and 27*) to be classified as low quality score. However, the small sample size did not allow for correlation statistics to be performed in this case. Interestingly, the percentage of studies published in journals endorsing STARD was higher in high quality score studies (58% vs 51% in low high score), but this correlation was not found to be significant (*chi-square test, not shown*).

Next, the quality of reporting in high scoring (score <12) versus low scoring (score ≥12) was compared item-by-item and presented in **Table 1**. Significant differences were seen in 5 items ($p<0.05$) (3 items in methods: reference standard, rationale for choosing and measures of diagnostic (prognostic in this case) accuracy and 2 items in results: distributions of severity of diseases in those with and without the target condition. In all these items, high scoring articles showed better performance.

	STARD items	Overall % of reporting item n=27	% of reporting item		*P-value
			Low score articles (< 12)	High score articles (≥ 12)	
5	*Whether data collection was planned before the index test and reference standard were performed (prospective study) or after (retrospective study)*	96.3	90.9	100	0.407
6	*Eligibility criteria*	88.9	81.8	93.8	0.549
7	*On what basis potentially eligible participants were identified (such as symptoms, results from previous tests, inclusion in registry)*	88.9	81.8	93.8	0.549
8	*Where and when potentially eligible participants were identified (setting, location and dates)*	81.5	72.7	87.5	0.370
9	*Whether participants formed a consecutive, random or convenience series*	51.9	27.4	68.8	0.054
10a	*Index test, in sufficient detail to allow replication*	70.4	72.7	68.8	1,000
10b	***Reference standard, in sufficient detail to allow replication***	44.4	18.2	62.5	**0.047**
11	***Rationale for choosing the reference standard (if alternatives exist)***	44.4	18.2	62.5	**0.047**
12a	*Definition of and rationale for test positivity cut-offs or result categories of the index test, distinguishing pre-specified from exploratory*	44.4	45.5	43.8	1,000
12b	*Definition of and rationale for test positivity cut-offs or result categories of the reference standard, distinguishing pre-specified from exploratory*	37.0	18.2	50.0	0.124
13a	*Whether clinical information and reference standard results were available to the performers/readers of the index test*	14.8	9.1	18.8	0.624
13b	*Whether clinical information and index test results were available to the assessors of the reference standard*	40.7	45.5	37.5	0.710
14	***Methods for estimating or comparing measures of diagnostic accuracy***	44.4	18.2	62.5	**0.047**
15	*How indeterminate index test or reference standard results were handled*	40.7	27.3	50.0	0.427
16	*How missing data on the index test and reference standard were handled*	14.8	0.0	25.0	0.123
17	*Any analyses of variability in diagnostic accuracy, distinguishing pre-specified from exploratory*	33.3	18.2	43.8	0.231
18	*Intended sample size and how it was determined*	0.0	0.0	0.0	-
19	*Flow of participants, using a diagram*	18.5	0.00	31.3	0.06
20	*Baseline demographic and clinical characteristics of participants*	88.9	81.8	93.8	0.549
21a	***Distribution of severity of disease in those with the target condition***	66.7	36.4	87.5	**0.011**
21b	***Distribution of alternative diagnoses in those without the target condition***	66.7	36.4	87.5	**0.011**
22	*Time interval and any clinical interventions between index test and reference standard*	3.7	0.00	6.3	1.000
23	*Cross tabulation of the index test results (or their distribution) by the results of the reference standard*	3.7	0.00	6.3	1.000
24	*Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)*	44.4	27.3	56.3	0.239
25	*Any adverse events from performing the index test or the reference standard*	0.0	0.0	0.0	-

Table 1. Proportion of reporting of the items in the STARD statement, overall, and in a total of 27 identified studies involving predictive claim of response to immunotherapy based on phenotyping for CD8 and/or Foxp3 markers. * P values from Fisher's exact test (association between proportions of reporting an item across the two groups of articles).

4. Discussion

The global cost of cancer care is constantly rising. Advanced immunotherapy (IT) approaches such as the Immune checkpoint inhibitors (ICIs) do have the prospect of eliminating cancer efficiently and reduce the costs of cancer patient management in the long term. However, despite clinical success, cancer immunotherapy remains ineffective in a large proportion of patients. Importantly, the lack of biomarkers that could predict clinical responses is a major drawback to therapeutic decisions and is accompanied by high expense. Identifying good biomarkers as predictors is not and will not be an easy task. A predictive signature of multiple aspects of the immunological (and non-immunological profile, such as genomic instability) characteristics of the patient rather than individual markers will most likely prove to be useful for the clinic. Nevertheless, the need to assess individual predictors by design, performance and reporting remains very important.

This report identified studies that assessed baseline (prior to IT) measurements of immune markers CD8 and/or Foxp3 as predictors of response to IT and investigated the reporting quality of the identified studies. The evaluation of reporting quality was aided by the STARD and TRIPOD guidelines which have a reported impact on improving reporting transparency in diagnostic/prognostic studies over the past ten years[14]. This was performed in parallel with data extraction from these studies which can assist in a better evaluation of the reporting quality in detail and feed in to a future meta-analysis on this subject.

The focus of this report was on baseline measurements of these immune markers rather than early or late-treatment assessments. Predictions based on baseline could be equally important as later assessments since it has been suggested that, ICI IT approaches such as anti-CTLA-4 (ipilimumab) and anti-PD-1 (nivolumab) could partially act via modulating Foxp3 levels and Treg cell function[13]. As a consequence, CD8 levels could also be affected by IT making the interpretation of these results more complex. From the identified studies, some contained comparison of CD8 and/or Foxp3 changes after treatment (compared to baseline), and this also could be a useful marker that has to be thoroughly investigated[18]. Other types of treatments such as radiation could increase the frequency of proliferating $CD8^+$ cells, therefore such details should always be reported in similar studies[19].

From the data extraction of the different characteristics of the studies many interesting points can be highlighted. First, some studies had a very low number of participants which makes the predicted output poor in terms of statistical power. To some extent, this is expectable since many of the identified studies had a retrospective design. However, it was striking to note that none of the studies, especially the ones with prospective design reported sample size estimation even as a discussion point to describe the ideal situation needed to create enough statistical power. In general, the effects of study power were not specified and adequately reported. Second, many identified studies did not have as a primary objective to develop a model to predict responses based on baseline measurements of CD8 and/or FoxP3, however they have used one or both markers to make assumptions and hypotheses around this issue and, as secondary endpoint, this comparison was included to support the primary endpoint.

The assessment with reporting guidelines also revealed reporting weaknesses. Many studies did not report whether assessment of the index test was blinded or not (*item 8 in TRIPOD, item 13b in STARD*). As described in the relevant explanation and elaboration documents[15] this is not about whether blinding is desirable or undesirable, but rather that readers of the study report need information about blinding for the index test and the reference standard to be able to interpret the study findings. The introduction of reporting and/or measurement bias cannot be excluded. Moreover, low reporting was observed in the item that describes cross-tabulation of index test and/or to a reference standard. Again, this could be due to context as many of the identified studies do not relate to a reference standard. However, better reporting of relating index test (CD8/Foxp3 phenotyping results with high/low densities) and the future clinical events (response or no response to therapy) could have been reported to allow for a more comprehensive understanding of the data.

This is particularly useful in such cases where there is no 'gold' standard and evaluation of a prognostic test can rely on relative risks, event rates and other correlation statistics [20].

Another very important point especially for diagnostic/prognostic studies is the use of reference standards and predictive models. The majority of the identified studies lack reporting on using generally accepted emerging reference standards PD-L1 and TMB. None of these markers can be considered as 'gold standard', thus it is somehow justifiable that many reports do not use them or consider other markers. A recent study in melanoma patients showed that 6–41% of patients negative for PD-L1 do respond to anti-PD-1 therapy, while half of the patients positive for PD-L1 positive do not respond[21]. Another layer that fuels this heterogeneity is the fact that discordances in PD-L1 measurements have been observed between biopsies and surgical resections[22]. Furthermore, the use of different antibodies and cut-off criteria can also account for these discrepancies[22]. To a similar extent, in the current report, many different antibodies were reported to measure CD8 expression but few studies also did not provide the clone that was used. In addition, some of the identified articles were published long before these reference markers and relevant ICI IT emerged.

Regarding more technical details, most studies used tumor tissue slides by immunohistochemistry as this is the 'gold' standard to assess tumor immune infiltrates because it allows for exact quantification of different parameters such as density, type and localization of infiltrating immune cells [23-25]. A small fraction of studies used blood as the biological material under investigation. This is due to either context (for leukemia) or due to the notion that peripheral immune profiles can also be used to predict anti-tumor responses and to a certain extent responses to immunotherapy[26].

Another key finding on this report was revealed through reporting of surrogate markers and the use of immune models rather than individual markers to associate with clinical responses. Recent studies have shown that high percentage of Tregs in primary cutaneous and metastatic melanomas was associated with reduced overall survival[27]. Similar studies however did not support these findings and such correlations[27]. It is believed that the ratio $CD8^+$ T to Treg cells in the TME will better predict favorable outcome[27]. None of the identified studies in this report have reported measurement of a similar baseline ratio of cytolytic T to regulatory T cells in their design.

Finally, a very interesting and important aspect of the analysis of this report was that high quality reporting studies reported more strongly on measures of diagnostic (prognostic) accuracy like sensitivity, specificity, positive and negative predictive values.

Biomarkers for predicting response to immunotherapy are paving the way for personalized treatment for patients with diverse cancer types. However, standardization of the available biomarker assays is an urgent requirement. In general, studies that do not meet quality standards cannot be accepted by Evidence based level grading and therefore do not meet quality standards for being considered in future review panels of practice guidelines. This report focused on identifying and assessing for reporting quality, studies that tested baseline measurements of immune markers as predictors of immunotherapy response. This report also serves to disseminate STARD and TRIPOD guidelines to improve reporting quality of relevant studies. Assessment of the quality of reporting of these data will lead to future evaluation of identification of novel biomarkers that will reliably predict responses to immunotherapy, allowing therefore patient stratification. There is a high demand for studies that will present panels with prediction ability that will be then validated in a separate cohort of patients before establishment as a diagnostic/prognostic tool. Sophisticated multi-disciplinary approaches that are needed to develop these diagnostic/prognostic tools will be more quickly produced if clinical data reporting becomes more transparent and clear.

5. References

1. Couzin-Frankel J. Breakthrough of the year 2013. Cancer immunotherapy. *Science.* Dec 20 2013;342(6165):1432-1433.
2. Duffy MJ, Crown J. Biomarkers for Predicting Response to Immunotherapy with Immune Checkpoint Inhibitors in Cancer Patients. *Clinical chemistry.* Jul 17 2019.
3. Hamanishi J, Mandai M, Matsumura N, Abiko K, Baba T, Konishi I. PD-1/PD-L1 blockade in cancer treatment: perspectives and issues. *International journal of clinical oncology.* Jun 2016;21(3):462-473.
4. Havel JJ, Chowell D, Chan TA. The evolving landscape of biomarkers for checkpoint inhibitor immunotherapy. *Nature reviews. Cancer.* Mar 2019;19(3):133-150.
5. Kourie HR, Awada G, Awada AH. Learning from the "tsunami" of immune checkpoint inhibitors in 2015. *Critical reviews in oncology/hematology.* May 2016;101:213-220.
6. Samstein RM, Lee CH, Shoushtari AN, et al. Tumor mutational load predicts survival after immunotherapy across multiple cancer types. *Nature genetics.* Feb 2019;51(2):202-206.
7. Zhang M, Yang J, Hua W, Li Z, Xu Z, Qian Q. Monitoring checkpoint inhibitors: predictive biomarkers in immunotherapy. *Frontiers of medicine.* Feb 2019;13(1):32-44.
8. Kim JM, Rasmussen JP, Rudensky AY. Regulatory T cells prevent catastrophic autoimmunity throughout the lifespan of mice. *Nature immunology.* Feb 2007;8(2):191-197.
9. Saito T, Nishikawa H, Wada H, et al. Two FOXP3(+)CD4(+) T cell subpopulations distinctly control the prognosis of colorectal cancers. *Nature medicine.* Jun 2016;22(6):679-684.
10. Chaudhary B, Elkord E. Regulatory T Cells in the Tumor Microenvironment and Cancer Progression: Role and Therapeutic Targeting. *Vaccines.* Aug 6 2016;4(3).
11. deLeeuw RJ, Kost SE, Kakal JA, Nelson BH. The prognostic value of FoxP3+ tumor-infiltrating lymphocytes in cancer: a critical review of the literature. *Clinical cancer research : an official journal of the American Association for Cancer Research.* Jun 1 2012;18(11):3022-3029.
12. Fontenot JD, Rasmussen JP, Williams LM, Dooley JL, Farr AG, Rudensky AY. Regulatory T cell lineage specification by the forkhead transcription factor foxp3. *Immunity.* Mar 2005;22(3):329-341.
13. Wang C, Thudium KB, Han M, et al. In vitro characterization of the anti-PD-1 antibody nivolumab, BMS-936558, and in vivo toxicology in non-human primates. *Cancer immunology research.* Sep 2014;2(9):846-856.
14. Cohen JF, Korevaar DA, Altman DG, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ open.* Nov 14 2016;6(11):e012799.
15. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Annals of internal medicine.* Jan 6 2015;162(1):W1-73.
16. Zintzaras E, Papathanasiou AA, Ziogas DC, Voulgarelis M. The reporting quality of studies investigating the diagnostic accuracy of anti-CCP antibody in rheumatoid arthritis and its impact on diagnostic estimates. *BMC musculoskeletal disorders.* Jun 25 2012;13:113.
17. Schwartz LH, Litiere S, de Vries E, et al. RECIST 1.1-Update and clarification: From the RECIST committee. *European journal of cancer.* Jul 2016;62:132-137.
18. Chen PL, Roh W, Reuben A, et al. Analysis of Immune Signatures in Longitudinal Tumor Samples Yields Insight into Biomarkers of Response and Mechanisms of Resistance to Immune Checkpoint Blockade. *Cancer discovery.* Aug 2016;6(8):827-837.
19. Kamphorst AO, Pillai RN, Yang S, et al. Proliferation of PD-1+ CD8 T cells in peripheral blood after PD-1-targeted therapy in lung cancer patients. *Proceedings of the National Academy of Sciences of the United States of America.* May 9 2017;114(19):4993-4998.
20. Rutjes AW, Reitsma JB, Coomarasamy A, Khan KS, Bossuyt PM. Evaluation of diagnostic tests when there is no gold standard. A review of methods. *Health technology assessment.* Dec 2007;11(50):iii, ix-51.
21. Patel SP, Kurzrock R. PD-L1 Expression as a Predictive Biomarker in Cancer Immunotherapy. *Molecular cancer therapeutics.* Apr 2015;14(4):847-856.
22. Fumet JD, Richard C, Ledys F, et al. Prognostic and predictive role of CD8 and PD-L1 determination in lung tumor tissue of patients under anti-PD-1 therapy. *British journal of cancer.* Oct 2018;119(8):950-960.
23. Becht E, de Reynies A, Giraldo NA, et al. Immune and Stromal Classification of Colorectal Cancer Is Associated with Molecular Subtypes and Relevant for Precision Immunotherapy. *Clinical cancer research : an official journal of the American Association for Cancer Research.* Aug 15 2016;22(16):4057-4066.
24. Fridman WH, Zitvogel L, Sautes-Fridman C, Kroemer G. The immune contexture in cancer prognosis and treatment. *Nature reviews. Clinical oncology.* Dec 2017;14(12):717-734.
25. Tumeh PC, Harview CL, Yearley JH, et al. PD-1 blockade induces responses by inhibiting adaptive immune resistance. *Nature.* Nov 27 2014;515(7528):568-571.
26. Gnjatic S, Bronte V, Brunet LR, et al. Identifying baseline immune-related biomarkers to predict clinical outcome of immunotherapy. *Journal for immunotherapy of cancer.* 2017;5:44.
27. Jacobs JF, Nierkens S, Figdor CG, de Vries IJ, Adema GJ. Regulatory T cells in melanoma: the final hurdle towards effective immunotherapy? *The Lancet. Oncology.* Jan 2012;13(1):e32-42.

6. Appendix

List of selected studies. *Reference id 1-27.*

1. Haddad R, Concha-Benavente F, Blumenschein G, Jr., et al. Nivolumab treatment beyond RECIST-defined progression in recurrent or metastatic squamous cell carcinoma of the head and neck in CheckMate 141: A subgroup analysis of a randomized phase 3 clinical trial. Cancer. Sep 15 2019;125(18):3208-3218.
2. Hamid O, Schmidt H, Nissan A, et al. A prospective phase II trial exploring the association between tumor microenvironment biomarkers and clinical activity of ipilimumab in advanced melanoma. Journal of translational medicine. Nov 28 2011;9:204.
3. Hanna GJ, Lizotte P, Cavanaugh M, et al. Frameshift events predict anti-PD-1/L1 response in head and neck cancer. JCI insight. Feb 22 2018;3(4).
4. Chen PL, Roh W, Reuben A, et al. Analysis of Immune Signatures in Longitudinal Tumor Samples Yields Insight into Biomarkers of Response and Mechanisms of Resistance to Immune Checkpoint Blockade. Cancer discovery. Aug 2016;6(8):827-837.
5. Tumeh PC, Harview CL, Yearley JH, et al. PD-1 blockade induces responses by inhibiting adaptive immune resistance. Nature. Nov 27 2014;515(7528):568-571.
6. Sun R, Limkin EJ, Vakalopoulou M, et al. A radiomics approach to assess tumour-infiltrating CD8 cells and response to anti-PD-1 or anti-PD-L1 immunotherapy: an imaging biomarker, retrospective multicohort study. The Lancet. Oncology. Sep 2018;19(9):1180-1191.
7. Ji RR, Chasalow SD, Wang L, et al. An immune-active tumor microenvironment favors clinical response to ipilimumab. Cancer immunology, immunotherapy : CII. Jul 2012;61(7):1019-1031.
8. Herbst RS, Soria JC, Kowanetz M, et al. Predictive correlates of response to the anti-PD-L1 antibody MPDL3280A in cancer patients. Nature. Nov 27 2014;515(7528):563-567.
9. Ribas A, Comin-Anduix B, Economou JS, et al. Intratumoral immune cell infiltrates, FoxP3, and indoleamine 2,3-dioxygenase in patients with melanoma undergoing CTLA4 blockade. Clinical cancer research : an official journal of the American Association for Cancer Research. Jan 1 2009;15(1):390-399.
10. Madonna G, Ballesteros-Merino C, Feng Z, et al. PD-L1 expression with immune-infiltrate evaluation and outcome prediction in melanoma patients treated with ipilimumab. Oncoimmunology. 2018;7(12):e1405206.
11. Fumet JD, Richard C, Ledys F, et al. Prognostic and predictive role of CD8 and PD-L1 determination in lung tumor tissue of patients under anti-PD-1 therapy. British journal of cancer. Oct 2018;119(8):950-960.
12. Althammer S, Tan TH, Spitzmuller A, et al. Automated image analysis of NSCLC biopsies to predict response to anti-PD-L1 therapy. Journal for immunotherapy of cancer. May 6 2019;7(1):121.
13. Wada J, Yamasaki A, Nagai S, et al. Regulatory T-cells are possible effect prediction markers of immunotherapy for cancer patients. Anticancer research. Jul-Aug 2008;28(4C):2401-2408.
14. Duell J, Dittrich M, Bedke T, et al. Frequency of regulatory T cells determines the outcome of the T-cell-engaging antibody blinatumomab in patients with B-precursor ALL. Leukemia. Oct 2017;31(10):2181-2190.
15. Mazzaschi G, Madeddu D, Falco A, et al. Low PD-1 Expression in Cytotoxic CD8(+) Tumor-Infiltrating Lymphocytes Confers an Immune-Privileged Tissue Microenvironment in NSCLC with a Prognostic and Predictive Value. Clinical cancer research : an official journal of the American Association for Cancer Research. Jan 15 2018;24(2):407-419.
16. Vilain RE, Menzies AM, Wilmott JS, et al. Dynamic Changes in PD-L1 Expression and Immune Infiltrates Early During Treatment Predict Response to PD-1 Blockade in Melanoma. Clinical cancer research : an official journal of the American Association for Cancer Research. Sep 1 2017;23(17):5024-5033.
17. Martens A, Wistuba-Hamprecht K, Geukes Foppen M, et al. Baseline Peripheral Blood Biomarkers Associated with Clinical Outcome of Advanced Melanoma Patients Treated with Ipilimumab. Clinical cancer research : an official journal of the American Association for Cancer Research. Jun 15 2016;22(12):2908-2918.
18. Jamal R, Lapointe R, Cocolakis E, et al. Peripheral and local predictive immune signatures identified in a phase II trial of ipilimumab with carboplatin/paclitaxel in unresectable stage III or stage IV melanoma. Journal for immunotherapy of cancer. Nov 21 2017;5(1):83.
19. Wistuba-Hamprecht K, Martens A, Heubach F, et al. Peripheral CD8 effector-memory type 1 T-cells correlate with outcome in ipilimumab-treated stage IV melanoma patients. European journal of cancer. Mar 2017;73:61-70.
20. Romano S, Simeone E, D'Angelillo A, et al. FKBP51s signature in peripheral blood mononuclear cells of melanoma patients as a possible predictive factor for immunotherapy. Cancer immunology, immunotherapy : CII. Sep 2017;66(9):1143-1151.
21. Balatoni T, Mohos A, Papp E, et al. Tumor-infiltrating immune cells as potential biomarkers predicting response to treatment and survival in patients with metastatic melanoma receiving ipilimumab therapy. Cancer immunology, immunotherapy : CII. Jan 2018;67(1):141-151.
22. Uryvaev A, Passhak M, Hershkovits D, Sabo E, Bar-Sela G. The role of tumor-infiltrating lymphocytes (TILs) as a predictive biomarker of response to anti-PD1 therapy in patients with metastatic non-small cell lung cancer or metastatic melanoma. Medical oncology. Jan 31 2018;35(3):25.
23. Kobayashi M, Suzuki K, Yashi M, Yuzawa M, Takayashiki N, Morita T. Tumor infiltrating dendritic cells predict treatment response to immmunotherapy in patients with metastatic renal cell carcinoma. Anticancer research. Mar-Apr 2007;27(2):1137-1141.
24. Daud AI, Loo K, Pauli ML, et al. Tumor immune profiling predicts response to anti-PD-1 therapy in human melanoma. The Journal of clinical investigation. Sep 1 2016;126(9):3447-3452.
25. Jensen HK, Donskov F, Nordsmark M, Marcussen N, von der Maase H. Increased intratumoral FOXP3-positive regulatory immune cells during interleukin-2 treatment in metastatic renal cell carcinoma. Clinical cancer research : an official journal of the American Association for Cancer Research. Feb 1 2009;15(3):1052-1058.
26. Donskov F, Bennedsgaard KM, Von Der Maase H, et al. Intratumoural and peripheral blood lymphocyte subsets in patients with metastatic renal cell carcinoma undergoing interleukin-2 based immunotherapy: association to objective response and survival. British journal of cancer. Jul 15 2002;87(2):194-201.
27. Le DT, Uram JN, Wang H, et al. PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. The New England journal of medicine. Jun 25 2015;372(26):2509-2520.

A pdf file that contains all the 165 articles retrieved prior to screening can be found here:
https://drive.google.com/drive/folders/1C0xSkKvUPd_pU4cJRg6ZEgR0NR_NOVvm?usp=sharing

Table A1. *Characteristics of information extracted (C1-C55) from the 27 selected studies with a brief description.*

Code	Characteristic	Brief description
C1	First author	*Name of first author*
C2	Year	*Publication year*
C3	Journal	*Publication name*
C4	Study type	*Type of study (retrospective, prospective, RCT etc)*
C5	Setting	*Where the study was conducted (e.g. clinical center)*
C6	Population	*Participant population*
C7	Tumor type	*Type(s) of cancer in participants*
C8	Stage/Grade	*Stage/ Grade of tumor in participants*
C9	IT type	*Type of Immunotherapy (ICI, cytokines etc)*
C10	Description of prior therapy (if any)	*Description of prior or concurrent therapy (if any other than surgery, e.g. chemotherapy)*
C11	Study size	*Total number of participants*
C12	Median age (range)	*Median age and range of participants*
C12	Events	*Number of patients with baseline data of CD8 and/or Foxp3 markers*
C13	no. of R	*Number of responders to IT based on the predefined criteria (outcome)*
C14	no. of NR	*Number of non-responders to IT based on the predefined criteria (outcome)*
C15	Follow-up period (from treatment) months	*Follow-up period during or after therapy that assessment occurred*
C16	Outcome index	*Assessment of outcome based on specified criteria (e.g. RECIST, OS, BOR etc)*
C17	Marker (CD8/FoxP3)	*Which marker was measured*
C18	Assay	*Type of assay to measure CD8 and or FoxP3*
C19	Reagent used (clones if Ab)	*Which antibody clone was used in applicable*
C20	Use of multiple markers?	*Whether there was in parallel marking of additional surrogate/reference markers*
C21	Other markers	*If C20 was a yes, specify markers*
C22	Index test Cutoff point (if used)	*Cut-off values or methodology of the index test*
C23	Reference to PD-L1/TMB	*Whether the study used emerging reference markers to immunotherapy response like PD-L1/TMB*
C24	Ref std Cutoff point (if used)	*Cut-off values or methodology of the reference test (if applicable)*
C25	Immune model used	*Whether a ratio of markers indicative of immune model was used (e.g. CD8:CD4, CD8:CD3 etc)*
C26	Counting Location	*Which location in the tissue was used to measure*
C28	TMA/Sections/Blood	*Whether the study used tissue sections, blood or tissue micro arrays as sampling material*
	Index test assessment by	*How was the assessment done (investigator(s) pathologist, computer etc)*
C29	Blinding	*Whether index/test results were blinded to outcome assessment*
C30	Predictive claim from study	*Predictive claim for the specific immune markers (Neutral, poor, good)*
C31	Statistical sig difference between Foxp3/CD8 between Resp and non Resp	*Whether there was statistical significance difference in any type of measurement by any type of statistical test between responders and non responders*
C32	Association with clinical benefit	*The type of association between index test result and clinical benefit (of any type) (neg, pos etc)*
C33	Reported analysis method	*Type of analysis used (univariate, multivariate)*
C34	Univariate p-value	*Univariate analysis p-value*
C35	Univariate HR	*Univariate analysis Hazard Ratio*
C36	Univariate CI	*Univariate analysis confidence intervals*
C37	Multivar correction to clinicopath	*Multivariate analysis correction to clinicopathological characteristics*
C38	Multivar p-value	*Multivariate analysis confidence p-value*
C39	Multivar CI	*Multivariate analysis confidence intervals*
C40	AUC	*Area under the Curve*
C41	AUC CI	*Area under the Curve Confidence intervals*
C42	AUC p-value	*Area under the Curve Confidence p-value*
C43	Predictive model and validation	*Whether the study used in the design development/discover and validation/test cohorts (Y/N)*
C44	Sensitivity	*Sensitivity*
C45	Specificity	*Specificity*
C46	PPV	*Positive predictive Value*
C47	NPV	*Negative predictive Value*
C48	DOI	*Digital Object Identifier System*
C50	JIF (2019)	*Journal Impact factor - 2019*
C51	Citations	*Article citations based on latest metrics - 2019*
C52	STARD score	*STARD score from this report*
C53	STARD score hi/lo	*STARD score dichotomization to high/low*
C54	TRIPOD score	*TRIPOD score from this report*
C55	TRIPOD score hi/lo	*TRIPOD score dichotomization to high/low*

Table A2. *STARD 2015 checklist. Yellow boxes denote items used in analysis and subsequent marking.*

Section & Topic	No	Item
TITLE OR ABSTRACT		
	1	*Identification as a study of diagnostic accuracy using at least one measure of accuracy (such as sensitivity, specificity, predictive values, or AUC)*
ABSTRACT		
	2	*Structured summary of study design, methods, results, and conclusions (for specific guidance, see STARD for Abstracts)*
INTRODUCTION		
	3	*Scientific and clinical background, including the intended use and clinical role of the index test*
	4	*Study objectives and hypotheses*
METHODS		
Study design	**5**	Whether data collection was planned before the index test and reference standard were performed (prospective study) or after (retrospective study)
Participants	**6**	Eligibility criteria
	7	On what basis potentially eligible participants were identified (such as symptoms, results from previous tests, inclusion in registry)
	8	Where and when potentially eligible participants were identified (setting, location and dates)
	9	Whether participants formed a consecutive, random or convenience series
Test methods	**10a**	Index test, in sufficient detail to allow replication
	10b	Reference standard, in sufficient detail to allow replication
	11	Rationale for choosing the reference standard (if alternatives exist)
	12a	Definition of and rationale for test positivity cut-offs or result categories of the index test, distinguishing pre-specified from exploratory
	12b	Definition of and rationale for test positivity cut-offs or result categories of the reference standard, distinguishing pre-specified from exploratory
	13a	Whether clinical information and reference standard results were available to the performers/readers of the index test
	13b	Whether clinical information and index test results were available to the assessors of the reference standard
Analysis	**14**	Methods for estimating or comparing measures of diagnostic accuracy
	15	How indeterminate index test or reference standard results were handled
	16	How missing data on the index test and reference standard were handled
	17	Any analyses of variability in diagnostic accuracy, distinguishing pre-specified from exploratory
	18	Intended sample size and how it was determined
RESULTS		
Participants	**19**	Flow of participants, using a diagram
	20	Baseline demographic and clinical characteristics of participants
	21a	Distribution of severity of disease in those with the target condition
	21b	Distribution of alternative diagnoses in those without the target condition
	22	Time interval and any clinical interventions between index test and reference standard
Test results	**23**	Cross tabulation of the index test results (or their distribution) by the results of the reference standard
	24	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)
	25	Any adverse events from performing the index test or the reference standard
DISCUSSION		
	26	Study limitations, including sources of potential bias, statistical uncertainty, and generalisability
	27	Implications for practice, including the intended use and clinical role of the index test
OTHER INFORMATION		
	28	Registration number and name of registry
	29	Where the full study protocol can be accessed
	30	Sources of funding and other support; role of funders

Table A3. *TRIPOD checklist. Yellow boxes denote items used in analysis and subsequent marking.*

Section/Topic			Checklist Item
Title	1	D;V	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.
Abstract	2	D;V	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.
Background and objectives	3a	D;V	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.
	3b	D;V	Specify the objectives, including whether the study describes the development or validation of the model or both.
Source of data	4a	D;V	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation datasets, if applicable.
	4b	D;V	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.
Participants	5a	D;V	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centers.
	5b	D;V	Describe eligibility criteria for participants.
	5c	D;V	Give details of treatments received, if relevant.
Outcome	6a	D;V	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.
	6b	D;V	Report any actions to blind assessment of the outcome to be predicted.
Predictors	7a	D;V	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.
	7b	D;V	Report any actions to blind assessment of predictors for the outcome and other predictors.
Sample size	8	D;V	Explain how the study size was arrived at.
Missing data	9	D;V	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.
Statistical analysis methods	10a	D	Describe how predictors were handled in the analyses.
	10b	D	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.
	10c	V	For validation, describe how the predictions were calculated.
	10d	D;V	Specify all measures used to assess model performance and, if relevant, to compare multiple models.
	10e	V	Describe any model updating (e.g., recalibration) arising from the validation, if done.
Risk groups	11	D;V	Provide details on how risk groups were created, if done.
Development vs. validation	12	V	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.
Participants	13a	D;V	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.
	13b	D;V	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.
	13c	V	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).
Model development	14a	D	Specify the number of participants and outcome events in each analysis.
	14b	D	If done, report the unadjusted association between each candidate predictor and outcome.
Model specification	15a	D	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).
	15b	D	Explain how to the use the prediction model.
Model performance	16	D;V	Report performance measures (with CIs) for the prediction model.
Model-updating	17	V	If done, report the results from any model updating (i.e., model specification, model performance).
Limitations	18	D;V	Discuss any limitations of the study (such as non representative sample, few events per predictor, missing data).
Interpretation	19a	V	For validation, discuss the results with reference to performance in the development data, and any other validation data.
	19b	D;V	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.
Implications	20	D;V	Discuss the potential clinical use of the model and implications for future research.
Supplementary information	21	D;V	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and datasets.
Funding	22	D;V	Give the source of funding and the role of the funders for the present study.

*Items relevant only to the development of a prediction model are denoted by D, items relating solely to a validation of a prediction model are denoted by V, and items relating to both are denoted D;V. The TRIPOD Checklist was used in conjunction with the TRIPOD Explanation and Elaboration document.