Ann Arbor Stage vs. TMTV: time to prepare for change?

A/Prof. Judith Trotman,
MBChB, FRACP, FRCPA
Concord Hospital, Sydney

Disclosures: nil
Disclaimer: Haematologist not a PET Physician

concord haematology & CLINICAL RESEARCH UNIT

ALLG AUSTRALASIAN LEUKAEMIA & LYMPHOMA GROUP

The Ann Arbor classification

Carbone P, Cancer Research 1971

- Named after Ann Arbor, Michigan, where the Committee on Hodgkin's Disease Staging Classification met in 1971.

[CANCER RESEARCH 31, 1860–1861, November 1971]

Report of the Committee on Hodgkin's Disease Staging Classification

Paul P. Carbone (Chairman), Henry S. Kaplan, Karl Musshoff, David W. Smithers, and Maurice Tubiana

National Cancer Institute, Bethesda, Maryland 20014 [P. P. C.]; Stanford University, Stanford, California 94305 [H. S. K.]; Roentgen-Radium-Abteilung, Freiburg, Germany [K. M.]; Royal Marsden Hospital, London, England [D. W. S.]; and Institut Gustave Roussy, Villejuif, France [M. T.]

Since the Rye classification for staging was produced in 1965, the significance of 2 important observations with major impact on staging has been appreciated. First, extralymphatic disease, if localized and related to adjacent lymph node disease, does not adversely affect the survival of patients. Patients with localized extralymphatic disease do as well as comparable patients of the same stage without extralymphatic spread. Secondly, laparotomy with splenectomy has been introduced as a method of obtaining more information on disease extent in the abdominal region. Thus, it has become necessary to reconsider the Rye classification and to recommend a modified scheme.

Staging has 2 aims. The first is to facilitate communication and exchange information. This can be done only at the expense of a loss of some information, as it is necessary to condense in one number a considerable amount of data. Furthermore, intercomparison demands that all the staging procedures performed should be as similar as possible in each center to avoid bias in staging and interpretation of the therapeutic results. The second aim is to provide guidance of prognosis and to assist in therapeutic decisions. This latter aim is best achieved when the greatest amount of information is collected for each patient. It has been recognized that a single staging procedure cannot achieve these 2 purposes. For

tests of urine and blood, and the initial biopsy results. Clinical evidence of liver involvement must include an enlarged liver and at least an abnormal serum alkaline phosphatase value, 2 different liver function test abnormalities, or an abnormal liver scan and 1 abnormal liver function test. Either palpable enlargement of the spleen confirmed by radiographic or radioisotopic studies or an isotopic scan of the spleen showing marked filling defects will be acceptable as clinical evidence of spleen involvement.

Pathological Staging (PS)

The Committee recognizes the wide diversity in the kinds and amounts of surgical removal of tissue to improve the accuracy of clinical staging at different institutions. To increase the amount of data reported and to allow for more precise comparisons, we recommend the use of a simultaneously recorded PS staging in all patients. The PS classification is to be subscripted by symbols indicating the tissue sampled and the results of histopathological examination by + when positive for Hodgkin's disease or – when negative. The abbreviations recommended are as follows:

Staging has two aims:

Carbone, 1971

1. to facilitate communication and exchange information.

- This can be done only at the expense of a loss of some information, as it is necessary to condense in one number a considerable amount of data.

- Inter-comparison demands that all the staging procedures performed should be <u>as similar as possible in each center to avoid bias</u> in staging and interpretation of the therapeutic results.

STANDARDISATION

The second aim of staging:

Carbone, 1971

2. guidance of prognosis and assist in therapeutic decisions.

“This latter aim is best achieved when the greatest amount of information is collected for each patient.”

“a single staging procedure cannot achieve these two purposes.”

Inherent to staging is a tension between being succinct vs. comprehensive; between being a lumper or splitter.

Conventional AA staging:

- On the face of it simplified to Stage I-IV A or B,
- But even in 1971 it was complicated!

$$\mathbf{CS\ IIA_3\ PS\ III_{S+N+H-M-}}$$

Clinical Stage IIA,3 lymph node regions involved,
Pathological Stage III with:
- spleen positive,
- abdominal lymph node positive,
- liver biopsy negative,
- bone marrow biopsy negative.

$$\mathbf{CS\ IVB_{L\,H}\ PS\ IV_{H+M-}}$$

Clinical Stage IVB with gross evidence of lung and liver involvement, and
Pathological Stage IV due to positive liver biopsy.
Bone marrow biopsy negative.

Cotswold modification

Lister et al, JCO 1989

- Structure of the classification maintained.
- CT included
- 'X' to designate bulky disease (>10 cm)
- Introduced CRu to accommodate persistent radiological abnormalities

Criteria for response assessment

Cheson 1999

- Defined abnormal node as >1.0cm in short axis.
- Noted value of bilateral BMBx in staging

- Anatomic definitions of response:
  -CR: normal node size after treatment of ≤1.5 cm in the longest diam by CT, or
  - nodes 1.1-1.5 cm should decrease to <1.0 cm to be considered normal, or by >75% SPD.
  - CRu: for >75% ↓SPD
  - PR: >50% ↓ in SPD (six dominant nodes / masses).

International Working Group

Cheson, Juweid 2007

- Introduced PET for HL and aggressive NHL
- Mediastinal blood pool used as a reference for PET status
- Abolished CRu

Lugano Classification & Imaging Working Group Consensus

Cheson, Barrington 2014

- Defined PET-CT as the imaging modality for ALL FDG-avid lymphomas,
- BM Biopsy not necessary in HL or DLBCL
- B symptoms only assigned for HL
- Bulk: 10cm (or 1/3rd mediastinal diameter) for HL
  6 for FL?,
  6-10cm for DLBCL?
- Splenomegaly: 13cm as best fit

Recognised the limitations of AA system

Cheson, Barrington 2014

"*the increased use of systemic and multimodality approaches has made Ann Arbor stage less relevant in directing the choice of therapy*".

- Recommended modification:

- Limited (stages I-II, non-bulky) or Advanced
- Stage II bulky - Limited or Advanced - determined by histology and prognostic factors.
- E not relevant to advanced stage

PET-CT

- More sensitive at detecting disease esp. extra-nodal disease.
- 10-30% stage migration - mostly up-staging
- *Quantitative imaging parameters for assessing disease burden and response should be explored as potential prognosticators. The standardisation of methods is mandatory ...*

  Barrington JCO 2014
- SUVmax (the semi-quantitative measure of maximal FDG uptake in a single lesion amongst many) is inadequate.
- Several investigators have focussed on the development of TMTV: the sum of the 3D measurements of lesions with FDG uptake - a measure of the viable fraction of tumours and microenvironment.

- In combining the metabolic & anatomical features could we potentially have "*a single staging investigation*"?

Methodology issues with TMTV

- What threshold SUV to measure disease?
  - Absolute SUV?
  - Per-lesion threshold of 41% SUV_{max}?
  - Per-patient adapted threshold based on SUV_{max} liver?
  - Other more sophisticated methods

  Comparative studies needed to determine best method for each lymphoma
- If we underestimate the volume of low-avidity lymphomas/lesions is this of relevance if uniformly done?
- Different scanner resolutions
- Different standardised uptake periods - are we achieving uniformity in practice?
- Different software algorithms of different vendors – source of bias in determination and reproducibility of TMTV. Different shapes of VOI drawing.
- How operator dependent / subjective is TMTV?
- How reproducible?

41% SUVmax threshold

- Sums the metabolic volume of each lesion, derived from calculating all voxels ≥41% of the lesion's SUVmax.
- Best approximation for volume determination according to phantom studies, and DLBCL/HL cohort. Meignan, Eur J Nucl Med Mol Imaging 2014
- Good reproducibility
- Recommended by the EANM.
- Available in a range of clinical imaging software

- Volume of the part of a lesion with the lowest SUV max might be underestimated, hence areas with heterogeneous uptake should be split into separate ROIs (Subjectivity of different operators?)
- When lesion has low FDG uptake overall the volume can be overestimated if background activity is erroneously included.

Flat SUV cut-off: including all voxels with SUV >2.5

- Arbitrarily determined cut-off.
- The simplest determination and widely available
- Less time-consuming.
- Limitations in heterogeneous uptake diseases like FL?
- May overestimate MTV esp. when background around the tumour has high activity leading to inclusion of background voxels in the TMTV equation (freq in lesions in BM, spleen and liver)
- Limited by lack of reproducibility of SUV values: variability in SUV max on diff PET scanners, PET acquisition protocol and reconstruction methods

Adaptive thresholds

- Now easier to apply in practice with more sophisticated software.

Hodgkin Lymphoma

We have already realised the limited value of AA Stage in HL, with

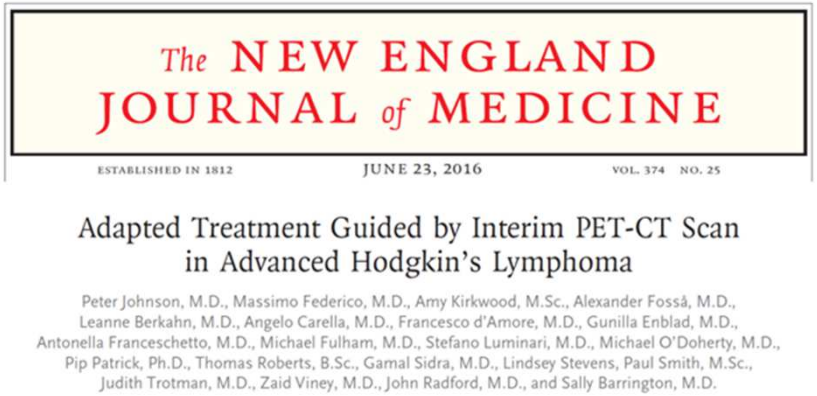
- different study group/trial definitions of ES Good vs. Poor risk, &

EORTC	GHSG (HD10,11,12)	ECOG/NCIC (HD6) bulky disease excluded
• **Large mediastinal lymphadenopathy** • **ESR ≥50 without B sy** • **ESR ≥30 with B sy** • **Age>50** • **≥4 LN sites involved**	• Large mediastinal lymphadenopathy (MMR>0.33 at T5-T6) • ESR ≥50 without B sy • ESR ≥30 with B sy • ≥3 LN areas involved	• Age ≥40 • ESR ≥50 • Mixed cellularity or lymphocyte depleted • ≥3 sites of disease

Hodgkin Lymphoma

We have already realised AA Stage has limited value in HL, with

- different study group/trial definitions of ES Good vs. Poor risk, &
- by including patients with IIB, or IIA with risk factors, in trials for AS



The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 JUNE 23, 2016 VOL. 374 NO. 25

Adapted Treatment Guided by Interim PET-CT Scan in Advanced Hodgkin's Lymphoma

Peter Johnson, M.D., Massimo Federico, M.D., Amy Kirkwood, M.Sc., Alexander Fossá, M.D., Leanne Berkahn, M.D., Angelo Carella, M.D., Francesco d'Amore, M.D., Gunilla Enblad, M.D., Antonella Franceschetto, M.D., Michael Fulham, M.D., Stefano Luminari, M.D., Michael O'Doherty, M.D., Pip Patrick, Ph.D., Thomas Roberts, B.Sc., Gamal Sidra, M.D., Lindsey Stevens, Paul Smith, M.Sc., Judith Trotman, M.D., Zaid Viney, M.D., John Radford, M.D., and Sally Barrington, M.D.

Advanced stage was defined as an Ann Arbor stage of IIB to IV, or stage IIA with adverse features: bulky disease (>33% of the transthoracic diameter or >10 cm elsewhere) or at least three involved sites.

TMTV calculation in HL:

Kanoun, PLOS one, Oct 2015

Compared the influence of both software tool and TMTV calculation method on prognostic stratification in HL

Methods

- 59 patients retrospectively included. Median f/u 39 months.
- Four sets of baseline TMTV calculation with free Beth Israel (BI) software:
  - an **absolute** threshold selecting voxels with SUV >2.5 ($TMTV_{2.5}$)
  - a **per-lesion threshold** of 41% SUV_{max} ($TMTV_{41}$),
  - a **per-patient adapted** threshold based on >125% & >140% SUV_{max} liver. ($TMTV_{125}$ & $TMTV_{140}$)
- $TMTV_{41}$ was also determined with commercial software for comparison
- ROC curves used to determine the optimal predictive threshold for each TMTV.

Results

- Excellent correlation between $TMTV_{41}$ determined with BI and commercial software ($r = 0.96$, $p<0.0001$).

Newly diagnosed HL,

Kanoun, PLOS one, Oct 2015

	TMTV41	TMTV2.5	TMTV125	TMTV140
Median	160	210	183	143
Optimal Threshold	313	432	450	330
AUC of ROC	0.70	0.68	0.68	0.68
4yr PFS (%) Low vs. high TMTV	83 vs. 42 p = 0.006	83 vs. 41 p = 0.003	85 vs. 40 p < 0.001	83 vs. 42 p = 0.004

Conclusion

- Baseline MTVs significantly influenced by the choice of method used for determination of volume …
- But no significant differences in terms of prognosis

How reproducible will these parameters in different HL populations?

Other HL studies

- Retrospective in Stage **I-II** HL MTV delineated on PET using ≥SUV2.5
- 127 patients: 66 received 6 ABVD only, 61 received CMT
- TMTV >198ml identified as best cut-off on ROC
- High MTV independently prognostic for PFS (p = 0.008) and OS (p = 0.007)

Song, Cancer Science 2013

- Retrospective in 30 patients with newly diagnosed & relapsed disease
- Baseline TMTV not predictive, but interim/baseline TMTV was.
- Hypermetabolic tumor foci were segmented with a software application, RT_Image, with a semi-automatic delineation of FDG uptake which did not follow EANM guidelines.

Tseng, Radiat Oncol 2012

DLBCL: The IPI (APLES)

Three factors which are surrogates for tumour volume,

- AA stage I-IV
- LDH: reflects cellular turnover – bulk of malignant cells and proliferation.
- ≥2 extra-nodal sites: more easily identified on PET

plus

- Age >60: a continuous unchangeable variable split into a dichotomous one
- Performance status: subjective definition of ECOG 2/3, and poorly measured
  - baseline before becoming symptomatic of lymphoma?
  - at time of presentation?
  - after pre-phase prednisone?

The more discriminating 8 factor NCCN-IPI splits age into >40-60, >60-75, >75 and LDH normalised ratio into 1-3 , >3, and identifies bone marrow, CNS, liver/GI tract, or lung involvement as the only relevant extra-nodal involvement

Zhou, Blood 2014

TMTV in DLBCL

	Song Ann Hem 2011	Mikhaeel EJNMMI 2016	Cottereau Clin Canc Res 2016	Sasanelli EJNMMI 2014	Adams Eur J Haem 2015
Patients	169	147	81	114	73
Med TMTV (cm^3)	198	595	320	315	445 (used as cut-off)
ROC cut-off	220	396	300	550	No ROC analysis
Stage	II-III	All stages	80% III-IV	82% III-IV	All stages II-IV 62%
SUVmax threshold	2.5	2.5	41%	41%	40% Spherical ROI only

TMTV in DLBCL

	Song	Mikhaeel	Cottereau	Sasanelli	Adams
PFS	90 vs. 66% 3yrs p= 0.001	92% vs. 48%. 3yrs p <0.001 30% 5yr PFS if MTV>400 and iPET DS 4-5 after 2R-CHOP	75% vs. 42% 5yrs p=0.0023 Plus significant molecular factors…	77% vs. 60% , 3yrs p=0.04 Only significant factor on MVA	p=0.059 NS for PFS. Not on MVA vs. NCCN IPI
OS	93 vs. 58% 3yrs p= 0.001		78% vs. 46% 5yrs p=0.0047 Plus significant molecular factors…	87 vs. 60% 3yrs p=0.003 Only significant factor on MVA	p = 0.037 Not on MVA vs. NCCN IPI

DLBCL

- All retrospective analyses, with methodologic variations, but same highly significant predictive of survival (excl. Adams with no ROC analysis).
- Optimal cut-off depends on both pop^{n} characteristics (age, stage, treatment) and methodology
- Validation studies comparing methods needed.
- Refinement in a larger series incl. all AA stages uniformly treated. Doubtful such a series exists?

- Can TMTV be prognostic for different cohorts of DLBCL:
  GCB, non-GCB, Double expressers, Double hits?

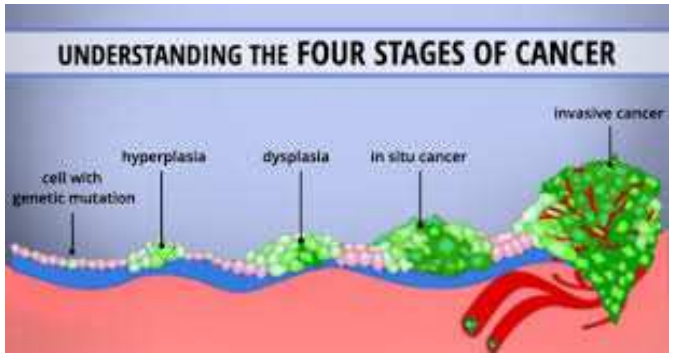
  Yes, Cottereau, Clin Cancer Res 2016
- Will it have the same predictive value in R-ACVBP and R-DA-EPOCH as in R-CHOP populations?

Follicular Lymphoma

a very heterogeneous lymphoma

- Patients focussed on "*I have Stage IV*"

Dr Google



- Physician focussed on:
  - multiple FLIPI & FLIPI2 prognostic factors:
    B_2M, LDH, Stage, Hb, LODLIN, age, # nodal sites
  - Treatment criteria (GELF / BNLI)
  - patient age and comorbidities …

  which affect timing, choice and outcome of therapy.

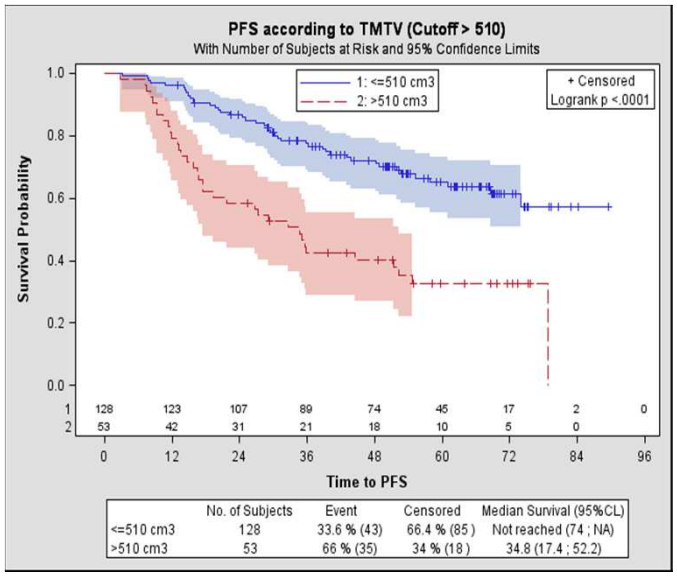
HTB Follicular Lymphoma

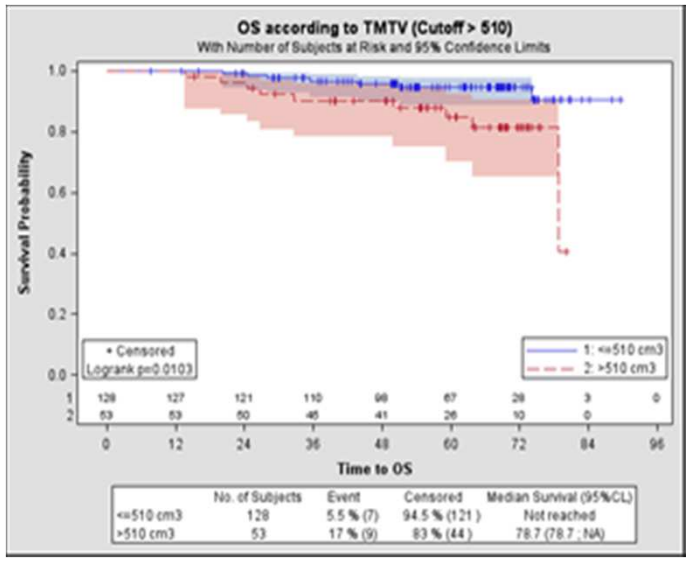
Meignan et al, JCO online 22 August

- Retrospective, but centrally reviewed baseline PETs for 185 patients in: PRIMA/FOLL05/PET Folliculaire prospective studies
- 41% SUVmax method
- Median TMTV 297ml (Q1-Q3: 135-567)
- Optimal cutoff on ROC/X-tile analysis of 510ml: 29% patients

  5-yr PFS 33% vs. 65%, (HR 2.90, *p* <0.0001)

  5-yr OS 85% vs. 95%, (HR 3.45, *p* =0.010)





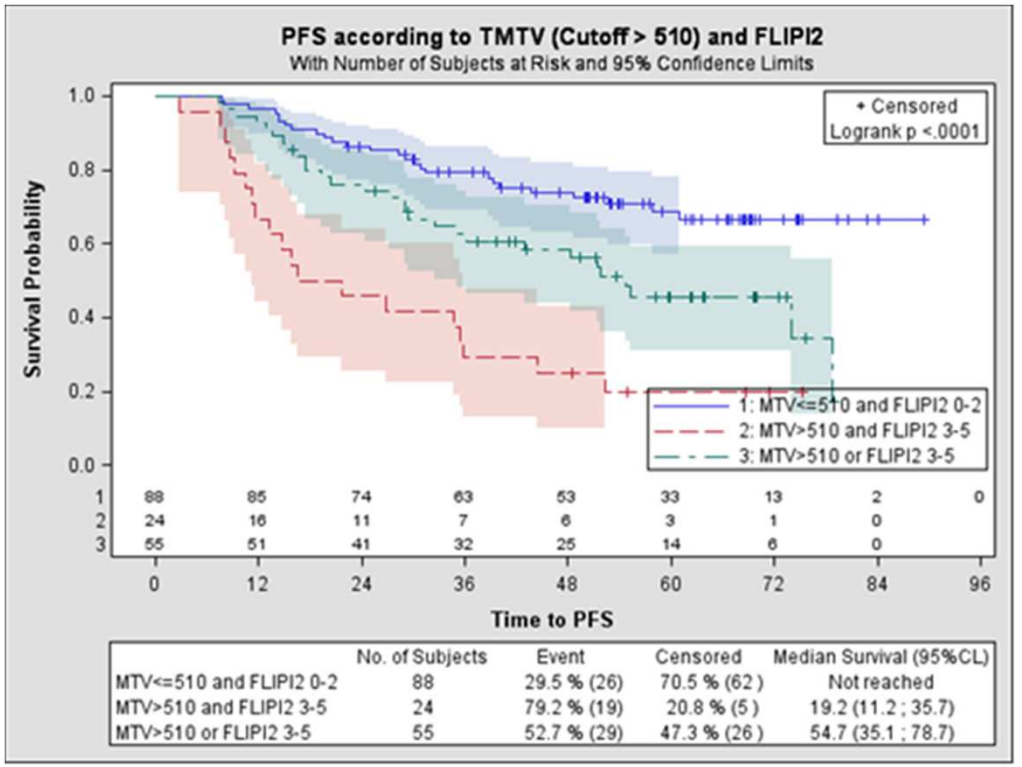
HTB Follicular Lymphoma

Meignan et al, JCO online 22 August

MVA, only TMTV (HR 2.3, *p* =0.002)
& FLIPI2 (HR 2.2, *p* =0.002) independent predictors of PFS.

5-year PFS

- Low TMTV & low FLIPI2
  69%
- High TMTV or int-high FLIPI2
  46% (HR 2.1, *p*=0.007)
- High TMTV & int-high FLIPI2
  20% (HR 5.0, p <0.0001)



Peripheral T Cell Lymphoma

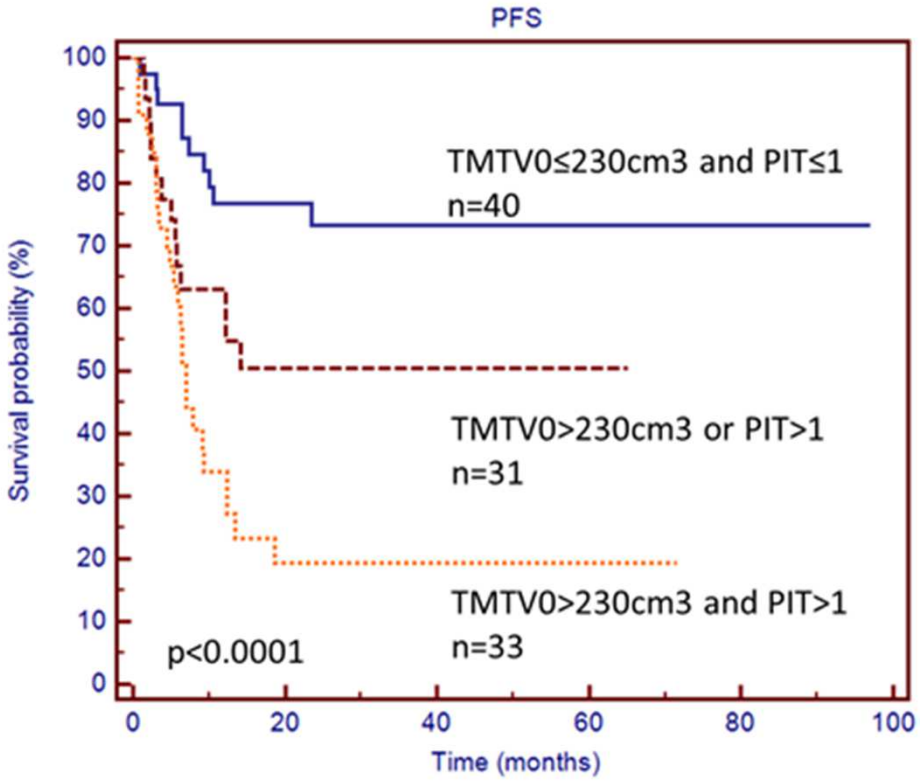
Cottereau, Annals of Oncology 2016

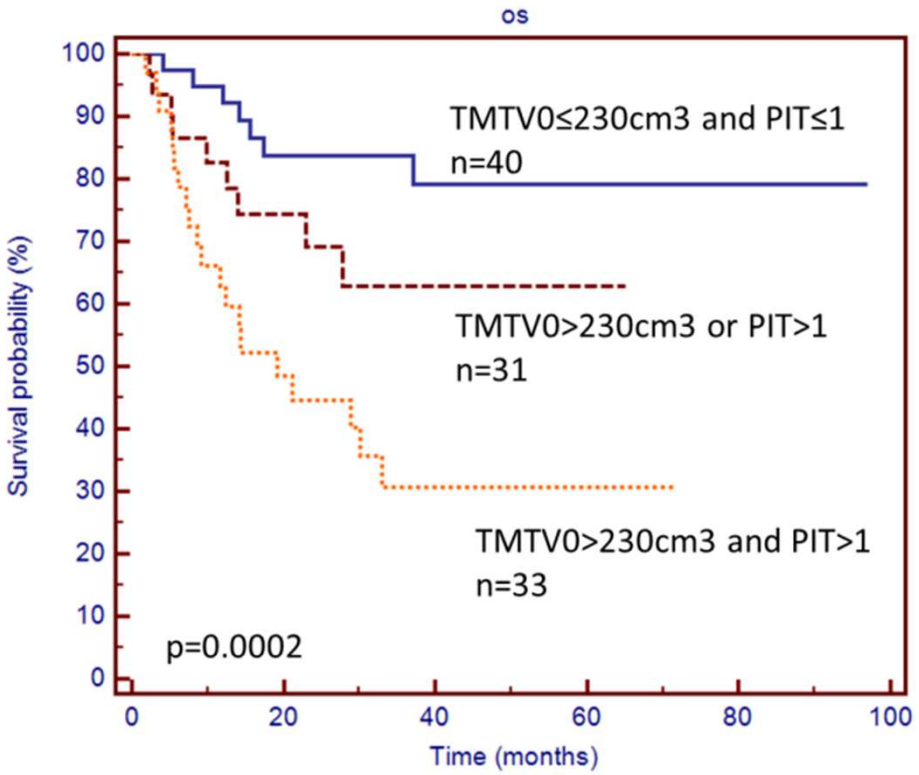
- Retrospective: 108 PTCL patients treated with anthracycline
- TMTV 41% max SUV threshold method
- Med TMTV 224 cm^3 (5-3824)
- Best cut-off 230 cm^3

- 2yr PFS 26% vs. 71%, p <0.0001, HR 4.0
- 2yr OS 50% vs. 80%, p =0.0005, HR 3.1.

- MVA - TMTV the only significant independent factor for both PFS & OS. (PIT significant for OS only, p=0.05)

TMTV combined with PIT





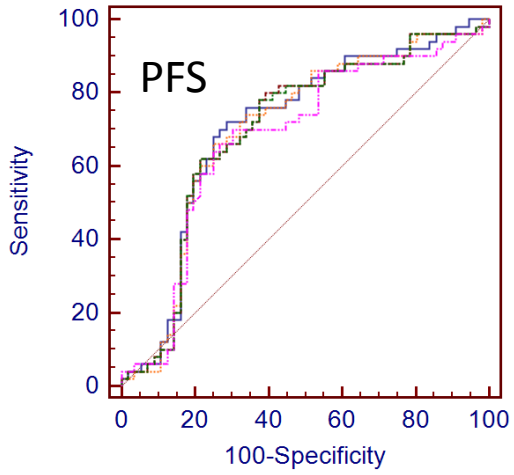
2y PFS 73% vs 50% vs 19%

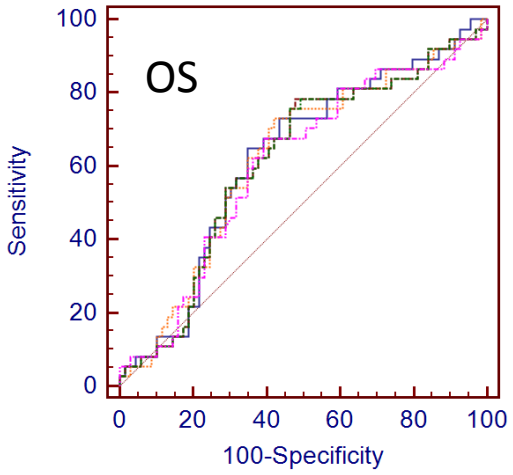
2y OS 81% vs 68% vs 43%

Peripheral T Cell Lymphoma

Cottereau, Menton 2016, J Nucl Med In press

- Same series, comparing different TMTV methods:
  - fixed 41% SUVmax threshold
  - and four adaptive thresholding methods.
- On PFS & OS ROC curves of 41% and 3 adaptive methods, there was no significant differences for outcome prediction.
- Advantage of such relative methods is minimisation of errors linked to the use of different devices at different centres.





ENNK/TCL

- AA staging of little merit.
- 80 patients IE/IIE disease
- Cut-off SUV2.5
- ROC established 35cm^3
- On MVA:
  High TMTV (HR 4.2 p=0.002 for PFS and HR 4.1 p=0.003 for OS) and up-front RT were independent prognostic factors.

Song, Leuk Research 2013

Similar results by Kim, Eur J Nucl Med Mol Imaging 2013

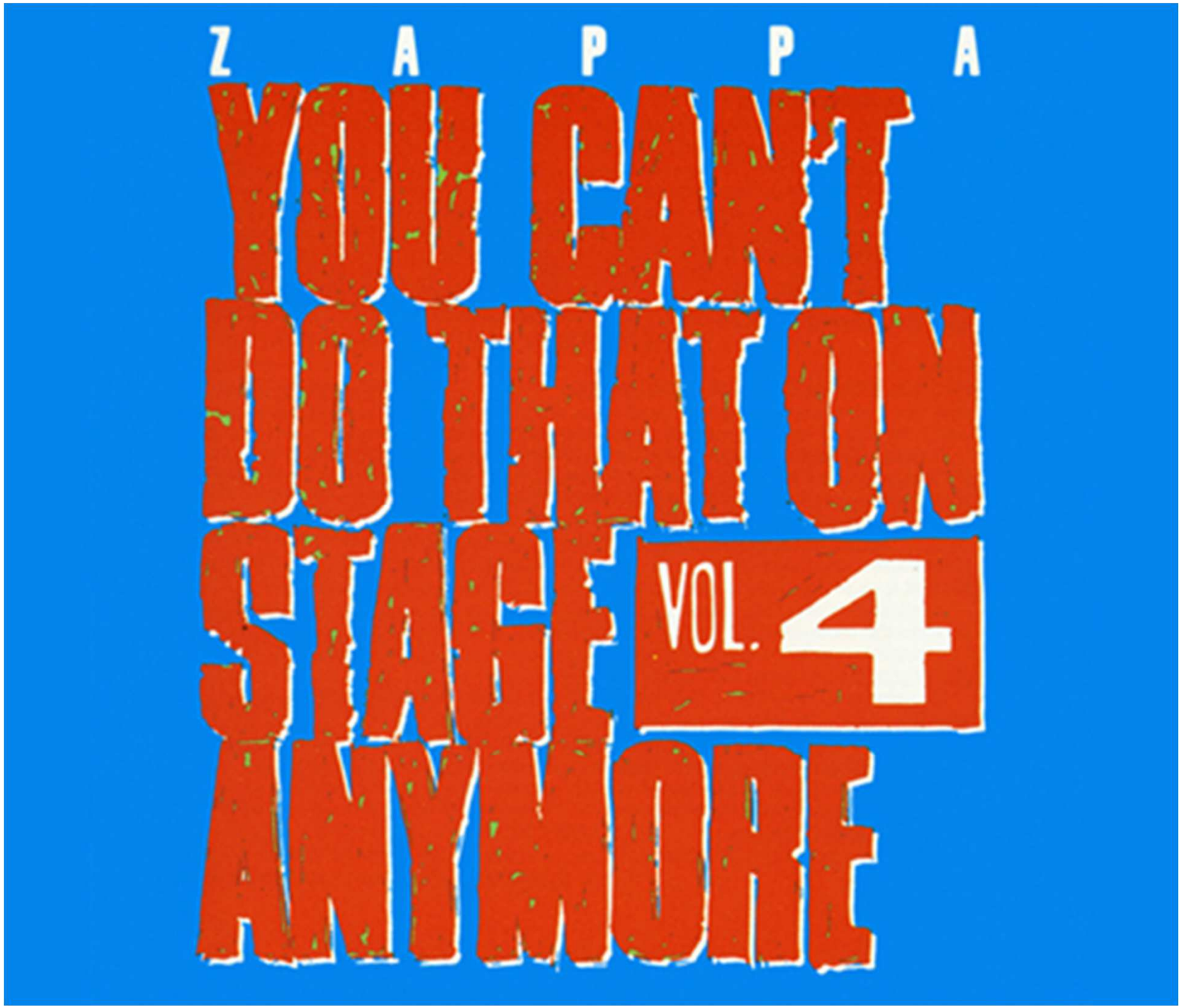
Staging has two aims:

Carbone P, Cancer Research 1971

1. *"to facilitate communication and exchange information.*

- *Done at the expense of a loss of some information, as it is necessary to condense.*
- *Inter-comparison demands all staging procedures performed should be as similar as possible in each center to avoid bias"*

2. "to provide guidance of prognosis and assist in therapeutic decisions".


ZAPPA
YOU CAN'T DO THAT ON STAGE ANYMORE
VOL. 4

Next staging consensus … in 2020?

Time to prepare for further change because

- PET-CT is central to staging of FDG-avid lymphoma
- 2-D determination of anatomic bulk of limited use
- Separation between local & advanced, nodal & extra-nodal disease less relevant with systemic therapies.
- Every FDG-avid lymphoma has its own multi-factor prognostic index with surrogates for TMTV: IPS, IPI, FLIPI(2), MIPI, PIT…

Next staging consensus:

Time to prepare for further change because…

- Retrospective studies suggest TMTV gives an accurate estimation of tumor burden for prognosis, but
- We need standardisation to prospectively identify reliably reproducible TMTV cut-offs for prognostication in clinical practice,
- and as a platform for study of baseline PET-adapted therapy.

- We need a better imaging measure of tumour burden to incorporate with the growing genetic data about our lymphomas

- We need something easier to convey prognosis and rationale for tailored therapy to our patients!

TMTV within current and future studies

Exploratory analysis within large of uniformly treated cohorts:

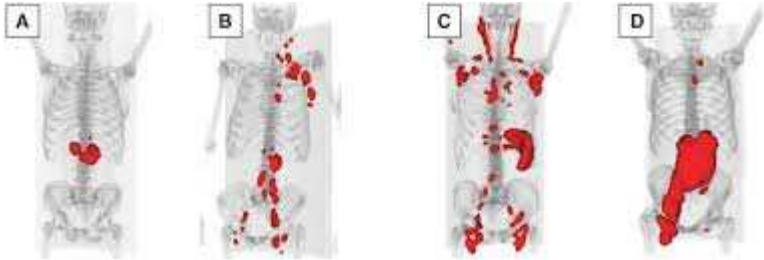
- in the GOYA (DLBCL) and GALLIUM (n = 600 FL) studies
- ??

Then prospective observational studies before studying a stratified / randomised therapeutic approach.

1971

2014

2020-



Shift from qualitative to quantitative assessment provided we can find