DEPARTMENT OF MOLECULAR MEDICINE AND SURGERY
Karolinska Institutet, Stockholm, Sweden

COMPUTED TOMOGRAPHY BASED ASSESSMENT OF TREATMENT RESPONSE IN SOLID TUMORS

Chikako Suzuki

Stockholm 2012
For an idea that does not first seem insane,
there is no hope.
Albert Einstein

It always seems impossible until it’s done.
Nelson Mandela

To my mother Tomoko and my daughter Sakiko
ABSTRACT

A “substantial evidence” of effectiveness is required for new cancer treatment regimens to be approved. Objective tumor shrinkage/enlargement has been adopted as an indicator of drugs efficacy. The change of tumor size is assessed and quantified by various radiological techniques; most commonly computed tomography (CT). A high accuracy and reproducibility is, for obvious reasons, necessary in order to achieve a meaningful evaluation of such studies. For that purpose, the World Health Organization criteria (WHO-criteria) were launched in 1979 followed by the Response Evaluation Criteria In Solid Tumors (RECIST) in 2000 and the updated version RECIST 1.1 in 2009. There are, however, still several steps that may deteriorate consistencies.

The purpose of this thesis was to investigate causes that may affect inconsistency in evaluation procedure according to RECIST (study I and II) and to explore the percentage of tumor size change at the first follow-up CT as the potential new surrogate indicators for OS in patient with metastatic colorectal cancer (mCRC) (study III) and in patient with metastatic breast cancer (MBC) (study IV).

The number of discordant cases increased gradually when assessing fewer target lesions. Measuring fewer than four target lesions might cause discrepancies when more than five target lesions were present (study I). Interobserver variation using RECIST and WHO-criteria were moderate: 0.53 (95%CI 0.33 - 0.72) and 0.60 (0.39 – 0.80), respectively. Intraobserver variation using RECIST and WHO-criteria were substantial to perfect that ranged between 0.76 – 0.96 and 0.86 – 0.91, respectively (study II).

The initial change in tumor size 8 weeks after initiation of chemotherapy was prognostic for PFS: Hazard Ratio (HR) 2.21, 95%CI 1.97 – 2.49, and OS: HR2.01, 95%CI 1.75 – 2.31, in mCRC (study III). The initial change in tumor size also correlated with OS in MBC (study IV). A marked difference in OS between patients with or without new lesion was demonstrated in mCRC: HR 3.77, 95%CI 2.08 – 6.83 (study III) and in MBC: HR 4.29, 95%CI 2.44 – 7.53 (study IV).

In conclusion, the current tumor response evaluation criteria are associated with several subjective steps that may cause inconsistent results. The initial change in tumor size at the first follow-up CT may provide an alternative surrogate outcome. The findings obtained in this thesis may improve the development of future response evaluation criteria.

LIST OF PUBLICATIONS

I. DARKEH, M. H., SUZUKI, C. & TORKZAD, M. R. 2009. The minimum number of target lesions that need to be measured to be representative of the total number of target lesions (according to RECIST). *Br J Radiol*, 82, 681-6.


V. Appendix

CONTENTS

1 INTRODUCTION .............................................................................................................. 1
   1.1 Epidemiology and Economy of Cancer ................................................................. 2
   1.2 Evaluation of effectiveness of cancer therapy ......................................................... 3
      1.2.1 Clinical Trials ..................................................................................................... 3
      1.2.2 Biomarkers in Oncology ................................................................................... 4
      1.2.3 Survival as an Endpoint and it’s Surrogate Endpoints ................................. 5
   1.3 Radiological Imaging Based Response Evaluation .................................................. 6
      1.3.1 WHO-criteria and RECIST ............................................................................. 6
      1.3.2 Background of Threshold Setting .................................................................... 9
      1.3.3 Background of Target Lesion Number in RECIST (1.1) ................................. 10
      1.3.4 Revised RECIST guideline version 1.1 (RECIST 1.1) ................................. 11
   1.4 Tumor measurements beyond WHO-criteria and RECIST ................................. 13
      1.4.1 Waterfall analyses .......................................................................................... 13
      1.4.2 Functional Imaging Based Evaluation ............................................................. 13
   1.5 Emerging Image technology and cancer drugs ....................................................... 15

2 AIMS .............................................................................................................................. 17
   2.1 Study I .................................................................................................................... 17
   2.2 Study II ................................................................................................................... 17
   2.3 Study III ................................................................................................................ 17
   2.4 Study IV .................................................................................................................. 17

3 MATERIALS AND METHODS ................................................................................... 19
   3.1 Study Subjects ........................................................................................................ 19
   3.2 Definition of OR according TO RECIST and WHO-criteria ................................. 19
   3.3 Study I ................................................................................................................... 19
   3.4 Study II ................................................................................................................... 20
   3.5 Study III and IV ..................................................................................................... 21
      3.5.1 Definition of first change ................................................................................. 21
      3.5.2 Survival analyses ........................................................................................... 21
      3.5.3 Statistical computations .................................................................................. 24
      3.5.4 Patients in Study III ....................................................................................... 25
      3.5.5 Patients in Study IV ....................................................................................... 27

4 RESULTS ..................................................................................................................... 29
   4.1 Study I .................................................................................................................... 29
   4.2 Study II ................................................................................................................... 30
   4.3 Study III ................................................................................................................ 30
      4.3.1 Relationship between first change and OS in mCRC ................................. 32
      4.3.2 Relationship between first change and PFS in mCRC ......................... 35
      4.3.3 The 2nd follow-up examination ................................................................. 37
   4.4 Study IV .................................................................................................................. 39
      4.4.1 The extent of first change in MBC ............................................................... 39
      4.4.2 Relationship between first change and OS in MBC ................................. 40
      4.4.3 The Cox regression analysis of OS in MBC ............................................. 42

5 DISCUSSION .................................................................................................................. 43
   5.1 To measure or not to measure, that is the problem ............................................. 43
   5.2 Clinical implications, What matters most? ......................................................... 44
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA</td>
<td>accelerated approval</td>
</tr>
<tr>
<td>AIDS</td>
<td>acquired immune deficiency syndrome</td>
</tr>
<tr>
<td>BOR</td>
<td>best overall response</td>
</tr>
<tr>
<td>CAD</td>
<td>computer aided detection</td>
</tr>
<tr>
<td>CR</td>
<td>complete response</td>
</tr>
<tr>
<td>CRC</td>
<td>colorectal cancer</td>
</tr>
<tr>
<td>CRF</td>
<td>case record form</td>
</tr>
<tr>
<td>CT</td>
<td>computed tomography</td>
</tr>
<tr>
<td>DWI</td>
<td>diffusion weighted image</td>
</tr>
<tr>
<td>EGFR</td>
<td>the epidermal growth factor receptor</td>
</tr>
<tr>
<td>EORTC</td>
<td>The European Organization for Research and Treatment of Cancer</td>
</tr>
<tr>
<td>FDA</td>
<td>The U.S. Food and Drug Administration</td>
</tr>
<tr>
<td>18F-FDG</td>
<td>$[^{18}\text{F}]$-2-fluoro-2-deoxy-D-glucose</td>
</tr>
<tr>
<td>HER-2/neu</td>
<td>human epidermal growth factor 2</td>
</tr>
<tr>
<td>HR</td>
<td>hazard ratio</td>
</tr>
<tr>
<td>ICC</td>
<td>the intra-class correlation coefficient</td>
</tr>
<tr>
<td>KRAS</td>
<td>Kirsten rat sarcoma viral oncogene homolog</td>
</tr>
<tr>
<td>LD</td>
<td>the largest diameter</td>
</tr>
<tr>
<td>MBC</td>
<td>metastatic breast cancer</td>
</tr>
<tr>
<td>MRI</td>
<td>magnetic resonance imaging</td>
</tr>
<tr>
<td>mCRC</td>
<td>metastatic colorectal cancer</td>
</tr>
<tr>
<td>NCI</td>
<td>National Cancer Institute of United States</td>
</tr>
<tr>
<td>NCI Canada</td>
<td>National Cancer Institute of Canada</td>
</tr>
<tr>
<td>ORR</td>
<td>objective response rate</td>
</tr>
<tr>
<td>OS</td>
<td>overall survival</td>
</tr>
<tr>
<td>PD</td>
<td>progressive disease</td>
</tr>
<tr>
<td>PET</td>
<td>positron emission tomography</td>
</tr>
<tr>
<td>PFS</td>
<td>progression-free survival</td>
</tr>
<tr>
<td>PR</td>
<td>partial response</td>
</tr>
<tr>
<td>RECIST</td>
<td>Response Evaluation Criteria In Solid Tumors</td>
</tr>
<tr>
<td>ROI</td>
<td>regions of interest</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>------------------------------------</td>
</tr>
<tr>
<td>RR</td>
<td>response rate</td>
</tr>
<tr>
<td>SD</td>
<td>stable disease</td>
</tr>
<tr>
<td>SUV</td>
<td>standardized uptake value</td>
</tr>
<tr>
<td>TTP</td>
<td>time to progression</td>
</tr>
<tr>
<td>WHO</td>
<td>World Health Organization</td>
</tr>
</tbody>
</table>
1 INTRODUCTION

Clinical trials are mandatory in the evaluation of new tumor treatments. A commonly studied indicator of the effect of an instituted therapy is the change of size of the malignant lesion(s)\(^1\). Both tumor shrinkage (response) and growth (progress) are assumed to provide useful endpoints in clinical trials. The change of size of tumor burden before and after treatment is often assessed and quantified by various techniques; most commonly computed tomography (CT). A high accuracy and reproducibility are, for obvious reasons, necessary in order to achieve a meaningful evaluation of such studies.

In 1940’s to 1950’s, the radiological techniques were still primitive. Tumor size was mainly measured by oncologists placing a ruler or caliper over a deep-seated irregular mass through the muscle, fat and skin\(^2\). There was, obviously, an inevitable factor of human error in these measurements which could cause serious effects on the conclusion of a clinical trial. A simulation experiment in which balls of different sizes were measured through a layer of foam rubber demonstrated that tumor measurement error increased with the size of tumor and that variation between investigators was larger than the one within investigator. On the basis of simulation results, a 50% reduction in the product of perpendicular diameters of tumor is employed to keep an objective response rate (ORR) of 5 to 10% due to human error in tumor measurement\(^2\).

In 1960’s to 70’s, it became more common to obtain tumor size on “roentgenograms”, such as chest and abdominal X-ray\(^2\). In order to reduce the confusion caused by various methods used \textit{ad hoc} for therapy evaluation of solid tumors with radiological methods, the World Health Organization criteria (WHO-criteria) were launched in 1979\(^3\),\(^4\). As an extension and modification of these definitions, the Response Evaluation Criteria In Solid Tumors (RECIST) were published in 2000\(^5\). RECIST has become dominating\(^6\); an updated version was published as RECIST 1.1 in 2009\(^7\). By using these criteria, the therapy response is supposed to be assessed in a common way by measuring tumor size before, during and after the treatment. The effectiveness of cancer treatment is then categorized into four statuses: complete response (CR), partial response (PR), stable disease (SD) and progressive disease (PD). The first two categories are combined to define responders and the last two combined to define non-responders\(^1\).

Several studies have shown inconsistencies in different kinds of image based tumor measurements; either RECIST (uni-dimensional), WHO-criteria (bi-
dimensional), volume (three-dimensional) or standardized uptake value (SUV) obtained by positron emission tomography (PET), and discussed appropriate measurements for evaluation. Few have questioned other factors than the measurement itself causing inconsistent results of treatment response in clinical trials. Neither the stipulation of how many target lesions should be selected according to RECIST nor the threshold settings for categorization of response is supported by any scientific evidence. Few studies have evaluated the fewest number of target lesions being necessary for an adequate evaluation or the inter- and intraobserver variability when a reader is confronted with a series of radiological images having to select and measure target lesions from various organs. Several studies have questioned the ORR as a surrogate endpoint for overall survival (OS) since SD might also be due to the treatment effect, while few has questioned how and why certain thresholds in percent change in size of diameter of tumors have been defined for response classification. Instead, the value of emerging functional imaging techniques, such as PET-CT, has been increasingly discussed in terms of a potential evaluation tool beyond tumor size based on anatomical imaging.

The purpose of this thesis was to determine the minimum number of target lesions to be representative of full assessment of target lesions according to RECIST (study I), to examine the intra- and interobserver agreement using two different using two different response evaluation criteria (study II), to determine whether the change in tumor diameters at the first follow-up CT (first change) correlates with outcome in patients with metastatic colorectal cancer (mCRC) (study III) and in patients with metastatic breast cancer (MBC) (study IV).

1.1 EPIDEMIOLOGY AND ECONOMY OF CANCER

Cancer is now the leading cause of death in the world. According to the IARC report, the global burden of cancers is supposed double by 2020 and nearly triple by 2030. As many as 26 million new cancer diagnoses and up to 17 million deaths worldwide are supposed for 2030 compared with an estimated 12 million new cancer diagnoses and more than 7 million deaths in 2008. Almost two-thirds of deaths might occur in low- and middle-income countries. Many interventions for early detection and treatment might remain inaccessible, thus making cancer a leading cause of mortality for many people in developing countries. Better systemic therapy has contributed to improve prognosis in many cancers. Among patients with metastatic colorectal cancer, the median duration of survival was eight
months without chemotherapy\textsuperscript{34}. With fluorouracil, it increased to 12 months\textsuperscript{34}. The combination of fluorouracil, irinotecan and oxaliplatin extended the median survival to 21 months\textsuperscript{34}. Bevacizumab and cetuximab are likely to prolong median survival beyond 21 months\textsuperscript{34}. A concern when dealing with new systemic therapies is the cost of treatment. Estimated drug costs for a patient with colorectal cancer who receives chemotherapy for eight weeks are about $100 for fluorouracil, $10,000 for regimens containing irinotecan or oxaliplatin and $30,000 for regimens containing bevacizumab or cetuximab\textsuperscript{34}. In the United States approximately 56,000 patients need chemotherapy for colorectal cancer in a year, and the drug costs for an eight-week course of treatment will be approximately $666 million or $1.2 billion with the addition of monoclonal-antibody therapy\textsuperscript{34}. Obviously, these astronomical costs are a threat to the public health care systems. But why have the costs become so high? The cost for drug development is an important reason. In the United States the average cost of bringing a drug to the market increased from $318 million in 1991 to $802 million in 2003\textsuperscript{35}. Another possible answer is a lengthy process from initial laboratory and animal testing to phase I, II, III studies for drug approval which may take years\textsuperscript{36}.

1.2 EVALUATION OF EFFECTIVENESS OF CANCER THERAPY

1.2.1 Clinical Trials

Clinical trials are conducted to allow safety and efficacy data to be collected for various medical interventions; e.g. drugs, diagnostic methods and devices, treatment regimens. Clinical trials are mandatory for evaluation and implementation of new cancer treatments\textsuperscript{37, 38}. The whole procedure mainly consists of three to four phases. Each phase have a different purpose\textsuperscript{39}.

- **Phase I:** Screening for safety. To determine side effects and an appropriate dose for later study. Conducted in a small group of people (<100)

- **Phase II:** Efficacy assessment in people with a particular condition. Establish the testing protocol in phase III trials\textsuperscript{20, 40-42}. Primary endpoint: objective response rate (ORR). Conducted in a larger group of people (100-300)

- **Phase III:** Studies focuses on comparison of a new, experimental treatment with an existing standard treatment. To confirm findings of phase III trials\textsuperscript{41}. Primary endpoint: TTP/PFS, OS. Conducted in large groups of people (500-3000)
Phase IV: “Post-approval” studies to delineate additional information including the drug’s risks, benefits and optimal use.

The average probability that a drug candidate will successfully pass clinical phase I is in the range of 75%; the respective values for phase II and III are 50% and 65%. In total, the cumulative probability that a leading drug candidate will successfully proceed from the preclinical phase to approval is about 8%. This high failure rate, especially in phase III trials, may be due to inappropriate design of phase III trials based on the result of phase II trials. RR, which is the usual primary endpoint of phase II, has weak correlation with OS thus may indicate neither reliably drug activity nor subsequent phase III trial is warranted.

1.2.2 Biomarkers in Oncology

In oncology, there are three types of biomarkers which are of importance: prognostic, predictive and early response biomarkers. Prognostic biomarkers predict the likely course of disease, irrespective of treatment; for example, lymph node involvement predicts a poor outcome in the management of solid tumors, even though treatment can prolong survival of patients with or without evidence of nodal involvement. These biomarkers are valuable not only for identifying aggressive disease that requires immediate treatment, but for identifying indolent disease that can safely be left alone. Predictive biomarkers forecast the likely response to treatment; for example, KRAS mutation predicts lack of response to treatment with cetuximab and panitumumab in colon cancer, or hormone receptor status predicts response to endocrine therapies in breast cancer. Furthermore, some biomarkers, such as hormone receptor status and HER2/neu status in breast cancer, are both prognostic and predictive.

Early response biomarkers reveal treatment response or failure earlier than it is possible with conventional response assessment methods. Their use can spare patients from prolonged exposure to ineffective treatments and allow alternative therapies to be tried sooner, and also save money. If they can be served as surrogate endpoints, aiming at replacing a clinical endpoint with a faster and more-sensitive evaluation of the effect of experimental treatments, early response biomarkers can also hasten the progress of clinical trials.
1.2.3 Survival as an Endpoint and it’s Surrogate Endpoints

Overall survival (OS) is defined as the time from randomization to death from any cause, and has historically been regarded as the gold standard measure of clinical benefit 37-39, 48, 49. OS is favored due to its objectivity, clear indication of benefit, ease and reliability measurement. However, determination of OS requires a larger sample size, a prolonged follow-up period, and may be influenced by other therapies used after patient’s participating trial has ended; including crossover to the experimental arm 39. This has been of increasing concern because significant advances in the systemic treatment have altered the number and variety of available treatments. To avoid the influence by subsequent treatments and to shorten the time period that requires for response evaluation, time-to-event endpoints, such as progression-free survival (PFS) and the related endpoint time to progression (TTP) have been considered of surrogate endpoints for OS 37. PFS is defined as the time elapsed between randomization and tumor progression or death from any cause. TTP is defined as the time elapsed between randomization and disease progression.

To improve access to therapeutics for life-threatening diseases, namely acquired immune deficiency syndrome (AIDS), the U.S. Food and Drug Administration (FDA) established the accelerated approval (AA) besides regular approval in 1992. The use of AA was expanded to oncology in 1996 38, 39. AA is less stringent compared to the regular approval and it allows the sponsors to begin marketing relevant drugs on the basis of trials that indentify improvements in surrogate outcomes that are reasonably likely to predict clinical benefit, such as the ORR (= the number of complete and partial response) or PFS interval. These surrogate endpoints have been increasingly used as primary endpoints and are now the most frequently used in randomized clinical trials 39. However the degree of association between these surrogate endpoints and OS is sometimes inconsistent 24 19, 21, 23, 50, 51.
1.3 RADIOLOGICAL IMAGING BASED RESPONSE EVALUATION

1.3.1 WHO-criteria and RECIST

The development of cancer drugs can be traced back to the Manhattan Project (1945). In this project, a number of laboratories developed various kind of weapons of mass destruction such as an atomic bomb and alkylating agents; nitrogen mustard\textsuperscript{20}. Ever since Gilman et al described the possible value of the nitrogen mustards in the treatment of lymphoma, there has been a rapid and continuous increase in the number of investigations of cancer therapy carried out in the world\textsuperscript{52,53} (Figure 1). This has resulted in an enormous expansion of the cancer literature.

\textbf{Figure 1.} One of the primary example of demonstrating effect of cancer therapy by chest X-rays. A lymphoma patient with mediastinal and hilar lymphadnopathy was treated with nitrogen mustard\textsuperscript{53}. The arrow inserted between chests x-rays indicates that a treatment was given. According to the original article it is described, “sixteen days after the first injection (B) showed a reduction of the mediastinal mass to one-half of its former size”. There was, however, no statement how the mass was measured (reprinted from JAMA 1946\textsuperscript{53}, copyright 2012, by permission of American Medical Association)
However, these investigations were frequently reported in a way which made it difficult for investigators to compare their results with those of others. Therefore it became necessary to develop a “common language” to describe cancer treatment and to agree on internationally acceptable general principles for evaluating data\(^3\),\(^4\). In 1979, Miller et al. proposed uniform criteria that standardized the recording and reporting of response, recurrence, and disease-free interval and the grading of acute and sub-acute toxicity in the treatment of solid tumors\(^3\),\(^4\). These criteria (WHO-criteria) were based on bi-dimensional measurements (2D, Figure 2) since the tumor measurement was mainly obtained by palpation and it was impossible to measure tumor volume using available imaging technology at that time. These criteria were based on the assumption that the tumor shape is spherical\(^8\). They have received wide acceptance for reporting the results of cancer treatment.

WHO-criteria do not stipulate the minimum lesion size or the number of lesions to be selected in patients with multiple lesions. Nor do they consider the type of imaging modality that should be used. Application of the WHO criteria includes sources of variability and potential for overestimation\(^5\)-\(^6\).

As a consequence of the limitations and imaging technology developments, the RECIST guidelines were introduced in 2000 by the EORTC, the USNCI, and NCI Canada Clinical Trials Group\(^5\). The primary goal of RECIST was to try to unify the various modifications of WHO-criteria in order to allow for meaningful comparisons between studies. This included:

1) The need to maintain the four categories of responses: CR, PR, SD, and PD;
2) The need to maintain the same definition of PR so that favorable results of future therapies can be compared with WHO-criteria even though the measurements will be different; and
3) The need to modify the definition of PD.

There are five major differences between the RECIST and WHO-criteria

1) Adopting unidimensional (1D) measurements (Figure 2): this encourages measuring more lesions and minimizing labor\(^8\);
2) Stipulating the type of imaging to be used;
3) Defining the type of tumors that should or should not be chosen;
4) Specifying the number of tumor lesions for assessment; up to 5 lesions in the same organ and total 10 lesions in a patient and
5) Making the cut-off point for definition of PD larger.

![Image](image_url)

**Figure 2.** Measuring the size of a liver lesion according to RECIST: the longest diameter (A), and WHO-criteria: the product of A x B (the longest perpendicular diameter).

In spite of these differences, various studies demonstrated concordance between WHO criteria and RECIST for ORR. However, discrepancies for time to progression could be demonstrated since RECIST requires a larger increase of lesions and longer delay to detect disease progression than WHO-criteria does.

The evaluation process contains of four main steps: (i) selection of target lesions, (ii) measurement of target lesions, (iii) identification of new lesions or progression of non-target lesions and (iv) categorization of effect, best overall response (BOR) as PD, SD, PR or CR.

The number of patient who reaches CR or PR in study cohort is determined as RR and used for efficacy assessment mainly in phase II studies. PD is used to obtain PFS and TTP.
1.3.2 Background of Threshold Setting

Both tumor shrinkage (response) and tumor increase (disease progression) are useful endpoints in clinical trials. Since WHO-criteria were published in 1979, a decrease of >50% and an increase of >25% in the product of the largest and the perpendicular tumor diameters have been used as cut-off values in recording a BOR. There was no biological or radiological background to these original thresholds; instead, it was a way to minimize errors in reporting changes in size of spheres under foam rubber when the measurement of tumor was mainly obtained by palpation. Placing a ruler or caliper over a lump and attempting to estimate its size, there is clearly the inevitable factor of human error. On the basis of measurement of 12 simulated tumor mass models, the 50% reduction criterion was employed to keep the human error in ORR evaluation between 5 and 10% (Figure 3).

Figure 3. Frequency of correctly reported tumor changes (reprinted from Cancer 1976 (36), Moertel et al, copyright 2012, by permission of John Wiley and Sons)
When RECIST was determined, these cut-off values were modified after a mathematical conversion from bi- to uni-dimensional measurements in attempt to simplify tumor measurements. Additionally, the cut-off value for progression went from a 40% increase in volume using WHO-criteria to a 73% increase in volume using RECIST (Table 1). It is obvious that imaging techniques have evolved during the last decades and more accurate estimations can be obtained. Nevertheless, the choice of these thresholds is arbitrary without any biological background. However, thresholds for anatomical imaging based response evaluation criteria are difficult to abandon.

**Table 1.** Equivalent change in diameter, diameters and volume for spherical lesions

<table>
<thead>
<tr>
<th>Change in diameter</th>
<th>Change in diameters</th>
<th>Change in volume</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1D, RECIST)</td>
<td>(2D, WHO)</td>
<td>(3D)</td>
</tr>
<tr>
<td>Decrease (response)</td>
<td>30%</td>
<td>50%</td>
</tr>
<tr>
<td>Increase (progress)</td>
<td>12%</td>
<td>25%</td>
</tr>
<tr>
<td></td>
<td>20%</td>
<td>44%</td>
</tr>
</tbody>
</table>

Recent studies on the measurement variability of lung lesions on the CT reveal that an average relative difference of 10% in 1-D measurement. Thus a change less than 10% in 1-D can be a result of measurement error alone.

### 1.3.3 Background of Target Lesion Number in RECIST (1.1)

To solve one problem using WHO-criteria, RECIST stipulated the number of lesions to be recorded: all measurable lesions up to a maximum of five lesions per organ and 10 lesions in total. Again, there is no biological background rather than an effort to minimize time consumption as well as labor in the clinical trial. The number of lesions measured on response assessment was revised and reduced in to a maximum of two lesions per organ and 5 lesions in total in RECIST 1.1 after a simulation study.

It is, however, questionable if a few selected target lesions reflect the entire tumor burden. It is possible that the effect of the treatment will not be identical in all sites of metastases. It is sometimes experienced that during treatment some lesions shrink while other lesions grow. The variability in tumor response measurements is
substantially reduced as increasing numbers of lesions are measured. However, resources do not permit radiologists to evaluate every lesion.

1.3.4 Revised RECIST guideline version 1.1 (RECIST 1.1)

Since the RECIST was published in 2000, a number of questions and issues raised and led to the development of a revised version of RECIST, so called RECIST 1.1. Amongst these questions and issues were whether less than 10 lesions can be assessed without affecting the overall assigned response for patients or conclusion about activity; how to apply RECIST in randomized phase III trials where progression, not response, is the primary endpoint particularly if not all patients have measurable disease; whether or how to utilize newer imaging technologies such as $^{18}$F- 2-fluoro-2-deoxy-D-glucose ($^{18}$FDG)-PET, PET-CT and MRI; how to handle assessment of lymph nodes; whether response confirmation is truly needed; and the applicability of RECIST in trials of targeted drugs. The RECIST 1.1 touches on all these points.

Major changes in RECIST 1.1 include:

1. Number of lesions to be assessed: Based on evidence from numerous trial databases merged in to a data warehouse for analysis purposes, the number of target lesions is reduced from a maximum of 10 in RECIST to a maximum of 5 total and from 5 to 2 per one organ.

2. Assessment of pathological lymph nodes: Nodes with a short axis of $\geq 15$mm are considered measurable and assessable as target lesions. The short axis measurement should be included in the sum of lesions in calculation of tumor response. Nodes that shrink to $<10$mm short axis are considered normal.

3. Lytic bone lesions or mixed lytic-blastic lesions with identifiable soft tissue components that can be evaluated if the soft tissue component meets the definition of measurability. Totally blastic bone lesions are non-measurable.

4. Target lesions which are reported as “too small to measure” at follow-up should be assigned a default measurement of 5 mm if they are visible.

5. If isotropic reconstructions are performed on CT, measurements can be made on not only in trans-axial plane but also in a coronal or sagittal plane. Using the same plane of evaluation, the maximal diameter of each target lesion should always be measured at subsequent follow-up time points. Software tools that calculate the maximal diameter for a perimeter of tumor may be employed and may even reduce variability.
6. Confirmation of response is required for trials with response primary endpoint but is no longer required in randomized studies.

7. Disease progression is clarified in several aspects: in addition to the previous definition of progression in target disease of 20% increase in sum, a 5 mm absolute increase is required in order to guard against over calling PD when the total sum is very small. Furthermore, there is a stipulation what constitutes “unequivocal progression” of non-measurable/non-target disease.

8. The interpretation of $^{18}$F-FDG-PET scan assessment is included in a section on detection of new lesions.

9. CT scans of PET-CT: If the CT performed as part of a PET-CT is of identical diagnostic quality to a diagnostic CT with i.v. and oral contrast, then the CT portion of the PET-CT can be used for RECIST measurements.

10. A lesion identified on a follow-up study in an anatomical location that was not scanned at baseline is considered a new lesion.
1.4 TUMOR MEASUREMENTS BEYOND WHO-CRITERIA AND RECIST

1.4.1 Waterfall analyses

The use of tumor measurements extends beyond RECIST. Increasingly, investigators use measurements in waterfall analyses and evaluate responses, progressions and simply change in tumor size as a continuous variable rather than categorizing the changes in tumor size \(^{40, 67}\) (Figure 4). A patient whose tumor shrinks by 25% has a different outcome than one whose tumor increases by 10%, yet both are considered by RECIST as having SD. There is not much difference between 32% shrinkage and 27% shrinkage, yet the former patient is classified as a PR and the latter is not PR but SD. Moreover, the ORR as an endpoint especially in phase II study is questioned and considered as a problem for evaluating molecular target drugs such as bevacizumab, cetuximab and so on. These drugs may be active even if they do not lead to tumor regression \(^{20}\).

**Figure 4.** Example of Waterfall plots with low ORR (on the left) and with high ORR (on the right). The waterfall plots show the percent relative change in tumor size (y-axis) for each patient (x-axis). According to the distribution of plots, treatment effect can be estimated.

In this analysis, there are further advantages to treating size as a continuous variable \(^{69}\). It is easy to incorporate covariates into the analysis, to formulate a simulation model for better planning for a phase III trial \(^{69}\).

1.4.2 Functional Imaging Based Evaluation

There is no doubt that functional evaluation such as PET will provide a surrogate endpoint beyond tumor size for assessing the clinical efficacy of cancer drugs especially using new targeted agents. Though PET is costly, their availability is limited, and the sensitivity to detect changes in lesions less than 1 cm is uncertain \(^{70}\), several studies have shown that \(^{18}\)F-FDG-PET is of value for evaluating targeted drugs not only
for imatinib (Gleevec) treatment of GIST metastases but other drugs as well.\textsuperscript{26, 27, 71} Compared to one- and (RECIST) two-dimensional (WHO) measurements, SUV measurement has the highest reproducibility.\textsuperscript{15} Furthermore, the efficacy of \textsuperscript{18}F-FDG-PET has also shown promise for monitoring treatment response for earlier detection of responders (Figure 5). This may increase the possibility of shortening trial times.\textsuperscript{72, 73} Early identification of responders and non-responders is important in order to select patients who might not experience any therapeutic benefit, to consider alternative therapies, and to avoid toxicity; early identification may also help to minimize the expense of clinical trials.

Figure 5. CT/MRI may delay detection of response compared to PET A patient with multiple liver metastases from breast cancer treated with sunitinib (Sutent\textsuperscript{®}). \textsuperscript{18}F-FDG-PET-CT obtained 3 weeks after therapy showed decreased \textsuperscript{18}F-FDG-uptake indicating response. On the other hand, it took 6 month to be categorized as PR on CT and MRI. (The patient had an adverse reaction to contrast agent for CT and therefore the follow-up modality was changed to MRI.)

Until widely accepted guidelines were introduced, PET could not gain acceptance as a tool for functional evaluation of therapy. EORTC and NCI PET guidelines and more recently PET Response Criteria in Solid Tumors (PERCIST) have been stipulated and recommended how \textsuperscript{18}F-FDG-PET should be performed and translated for response evaluation.\textsuperscript{74-76} In these guidelines, the SUV is recognized as valuable in the evaluation of \textsuperscript{18}F-FDG uptake in PET. However, the SUV is depending on several factors such as uptake time, the patient’s body weight (which often changes during therapy), blood glucose levels, definition of Regions-Of-Interest (ROI), attenuation corrections, and the effects of metabolism by other tissues.
An advantage of PET/CT is that the CT-examination can be used for RECIST measurements if the CT exhibits the same diagnostic quality as a dedicated diagnostic CT (with i.v. and oral contrast medium)\textsuperscript{7}.

1.5 EMERGING IMAGE TECHNOLOGY AND CANCER DRUGS

In spite of remarkable progress in understanding molecular mechanisms in cancer cells, treatment response evaluation of solid tumors relies on changes in size assessed by imaging, mainly computed tomography (CT). Newer molecular targeted drugs have demonstrated an inherent limitation and unsuitability with “anatomical” tumor evaluation that only assesses the lesion size increase or decrease\textsuperscript{20,40,77,78}. The effect of these new target drugs demands the paradigm shift in evaluation.

Meanwhile, the widespread use of multi-detector computed tomography (MDCT) and other imaging innovations, such as PET-CT, MRI with functional capabilities such as dynamic contrast enhanced imaging and diffusion weighted image (DWI)\textsuperscript{79,80}, have led to concomitant need for modifications and up-date of RECIST (Figure 6).

![Figure 6](image)

**Figure 6.** Lymph node metastasis in a patient with sigmoid colon cancer.\textsuperscript{18}F-FDG-PET before (a) and after (b) treatment show only faint accumulation of \textsuperscript{18}F-FDG in a lymph node (white arrow). The T2-weighted image on MRI (c) shows heterogeneous higher intensity inside a node (arrow) that suggests pathological, possibly metastasis. DWI (b=500) highlights the node (black arrow) but it is unspecific without telling metastatic status. The primary lesion (arrowhead) in the sigmoid is also highlighted.
2 AIMS

To explore possible limitations with currently used image based tumor response criteria: WHO-criteria and RECIST, and to test if the tumor size change at the first response evaluation after initiation of chemotherapy correlates with outcome in patients with metastatic colorectal cancer and metastatic breast cancer.

Specific questions addressed for this purpose were the following:

2.1 STUDY I

RECIST does not specify the minimum number of target lesions. The aims of this study were whether the number of target lesions could be reduced, and whether the results still could be equally representative. If the number of target lesions could be reduced, how many target lesions would be needed to keep consistency.

2.2 STUDY II

What is the extent of inter- and intra-observer variations in RECIST and WHO-criteria based tumor response evaluation? What are the reasons for such variations?

2.3 STUDY III

Does the change in size of target lesions at the first follow-up CT after initiation of chemotherapy (1st change) correlate with PFS and OS in mCRC?

2.4 STUDY IV

Does the change in size of target lesions at the first follow-up CT after initiation of chemotherapy (1st change) correlate with OS in MBC?
3 MATERIALS AND METHODS

3.1 STUDY SUBJECTS

All studies were carried out retrospectively. Tumor size measurements and response definition were retrieved in the case recorded forms (CRFs) for study I, III and IV. In study II, tumor size measurement and BOR were reanalyzed independently by two board certified radiologists.

All studies were approved by the regional ethical committee.

3.2 DEFINITION OF OR ACCORDING TO RECIST AND WHO-CRITERIA

In study I, II, III and IV, response evaluation was conducted according to RECIST 1.0\textsuperscript{5}.

CR means disappearance of all signs of tumor; PR means at least a 30% decrease in the sum of the largest diameters of the target lesions; PD means at least a 20% increase in the sum of the largest diameters of the target lesions or the appearance of new lesions or unequivocal progression of non-target lesions; SD means neither PR nor PD.

In addition to the RECIST evaluation, the perpendicular longest diameter of target lesion was obtained to allow a bi-dimensional evaluation according to WHO-criteria was conducted in study II. Different from RECIST, a patient was classified as PD if the sum of the products of the largest diameter (LD) and perpendicular LD increased 25% and more in one or more target lesion.

Objective response rate (ORR) was defined as the percentage of the sum of CR and PR (thus those who were considered responders) in relation to all patients: (CR+PR)/all patients.

3.3 STUDY I

A total of 99 patients enrolled in 16 phase II clinical trial studies at Karolinska University Hospital Solna with at least two CT-examinations amenable to measurements according to RECIST and who had completed their studies were eligible.

Target lesions were referred as 1L, 2L … up to xL from the largest to the smallest. For 1L evaluation, only the largest; 1L, was chosen to assess the response. In the same fashion, for 2L evaluation, only the largest and the second largest; 1L+2L were chosen to assess the response. Thereafter, the discrepancy between full RECIST evaluation and
evaluation of fewer target lesions, in this study from seven down to one lesion, was analyzed.

3.4 STUDY II

Two board certified radiologists with experience in oncological imaging retrospectively re-evaluated 39 patients’ CT-examinations in nine phase II or III clinical trials for treatment of metastatic breast- and colorectal cancer according to RECIST and WHO-criteria. The radiologists independently selected and measured target lesions, assessed presence of new lesions or progression of non-target lesions and categorized treatment response. One radiologist repeated the procedure on two additional different occasions to examine intraobserver variation (Figure 7).

![Figure 7. Scheme of study II](Reprinted from Acta Oncologica 2010, copyright 2012, by permission of Informa Healthcare)

Kappa statistics was applied to examine inter- and intraobserver agreement. Kappa values above 0.81, 0.61 – 0.8, 0.41 - 0.60, 0.21 – 0.40, and 0-0.20 indicated almost perfect, substantial, moderate, fair and slight agreements, respectively. Wilcoxon matched pairs test was used to compare the number of selected target lesions by readers. Friedman ANOVA by ranks was used to compare the number of selected target lesions among three repeated evaluations. A value of p<0.05 was considered significant. Statistical analysis was carried out using StatXact 4 (CYTEL Software Corporation, Cambridge, MA, USA).
3.5 STUDY III AND IV

3.5.1 Definition of first change

The first change in tumor size (1st change) was calculated as the ratio of the sum of the LD of target lesions at the baseline study (baseline sum) and the sum of the LD of the same target lesions at the first follow-up CT (1st sum).

\[
1^{\text{st}} \text{ change} = \frac{[(1^{\text{st}} \text{ sum})-(\text{baseline sum})]}{(\text{baseline sum})}
\]

Likewise, the 2nd change was calculated by the ratio of the difference of the sum of the LDs at the first and second follow-up studies in the study III.

3.5.2 Survival analyses

PFS and OS are the time from randomization to the date of any progression or death. In study III, the primary endpoints were the correlations between 1st change and OS and PFS. In study IV, the primary endpoints were the correlation between 1st change and OS.

Patients alive or progression-free at the last follow-up were censored.

Chi-square test or Fisher’s exact test were applied to compare two independent groups. The Cox regression analysis was used to define prognostic variables and to estimate Hazard ratios (HRs) and 95% confidence limits (CL). The durations of PFS and OS were estimated by the Kaplan-Meier method. The log-rank test was used for comparison between groups.

3.5.2.1 Guarantee-time bias and the landmark method

The comparison of survival by response may be biased as patients need to live long enough to experience a response to treatment. Responders may survive longer than non-responders, not because of an effect of response on survival, but because response identifies patients with pretreatment characteristics that favor longer survival. Hence, the classification of patients as responders or non-responders is subject to “guarantee-time” bias, so that patients who die early before they have had a change of responding to treatment have a major negative impact upon the survival estimates of the group of patients classified as non-responders. The presence of guarantee-time bias can be eliminated by the “landmark method”. This essentially consists of setting the time origin for survival after a landmark period long enough for all responses to have been observed. \cite{1, 83, 84}

For example, Figure 8 shows the survival for responders versus non-responders without landmark method. Responders have a median survival which is more than twice as long...
as the median survival for the non-responders. The difference between the two curves is, however, due to the four non-responders who died early, before showing response or non-response to the treatment. The bias caused by the guarantee time is clearly visible in this example (Figure 8) ¹.

Figure 8. Evaluation of survival for responders vs. non-responders without landmark method.
(reprinted from Journal of Clinical Oncology 1983, Anderson et al ¹, by permission of American Society of Clinical Oncology)
The landmark time analysis in which survival from the landmark time is evaluated by response status is shown in Figure 9.

Figure 9. Evaluation of survival for responders vs. non-responders with landmark method.
(reprinted from Journal of Clinical Oncology 1983, Anderson et al, by permission of American Society of Clinical Oncology)

In this study, patients who died or relapsed prior to the landmark period have been excluded from the analysis. No effect of response is apparent when the guarantee-time (landmark time) bias has been removed (Figure 9).

Patients who survive the landmark period are classified as responders or non-responders, and their survival time after the end of the landmark period can be compared without bias. Using this method, either patients who die before the landmark time point are excluded from the analysis, or the landmark time is subtracted from PFS or OS if patients who live longer than the landmark time point. The landmark method was used to compensate for longer guarantee-time for responders and was applied in the PFS and OS analyses.
3.5.2.2 The intra-class correlation coefficient (ICC)

The intra-class correlation coefficient (ICC) measures the strength of agreement between repeated measurements by assessing the proportion of the between-patient variance (the square of the standard deviation) versus the total variance, comprising the sum of both the within and between variation. The ICC ranges from 0.00 to 1.00; with 1.00 representing perfect repeatability, hence high reliability. The ICC is the most commonly used method for assessing reliability with continuous data as well as categorical data. A value of at least 0.90 is often recommended if the measurements of concern are to be used for evaluating future patients for which therapeutic decisions are to be made.

To assess the reliability of 1st change values, an independent board certified radiologist retrospectively selected tumors and measured the baseline sum and the 1st sum in study III and IV. The radiologist had no information about the original radiological response assessments or patient outcomes.

3.5.3 Statistical computations

A value of $p<0.05$ was considered significant. Statistical computations were performed with StatXact 4 (CYTEL Software Corporation, Cambridge, MA, USA) and STATISTICA ver. 9 (StatSoft Inc., Tulsa, OK, USA).
3.5.4 Patients in Study III

The patients were enrolled in the multi-center randomized phase III Nordic VI trial (N=567) comparing irinotecan with either the Nordic bolus 5-fluorouracil (5-FU) and folinic acid schedule (FLIRI) or the de Gramont schedule (Lv5FU2-IRI) \(^8^6\). Both FLIRI and Lv5FU2-IRI treatments were repeated every 2 weeks until disease progression or unacceptable toxicity. CT scans of the chest and the abdomen were obtained at baseline within 3 weeks prior to randomization and repeated after every 4 cycles (every 8 weeks). Response evaluation was conducted according to RECIST 1.0.

According to the original study, there was no significant difference in PFS, median 9 months \((p=0.4)\), or OS, median 19 months \((p=0.4)\) in the two groups; however, ORR were more common using LVFU2-IRI compared to FLIRI (49% vs. 35%, \(p=0.001\)) \(^8^6\).

A total of 506 patients were eligible for the study. Patient characteristics are shown Table 2.
Table 2. Patient Characteristics in Nordic VI (n = 506)

<table>
<thead>
<tr>
<th>Characteristic</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>61</td>
</tr>
<tr>
<td>Median</td>
<td>62</td>
</tr>
<tr>
<td>Range</td>
<td>29 - 76</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>314 (62%)</td>
</tr>
<tr>
<td>Organ involved</td>
<td></td>
</tr>
<tr>
<td>liver</td>
<td>405 (40%)</td>
</tr>
<tr>
<td>lung</td>
<td>162 (16%)</td>
</tr>
<tr>
<td>lymph nodes</td>
<td>119 (12%)</td>
</tr>
<tr>
<td>others</td>
<td>336 (33%)</td>
</tr>
<tr>
<td>No. of organs involved</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>244 (48%)</td>
</tr>
<tr>
<td>2</td>
<td>162 (32%)</td>
</tr>
<tr>
<td>3+</td>
<td>100 (20%)</td>
</tr>
<tr>
<td>Overall Response / No. operated</td>
<td></td>
</tr>
<tr>
<td>CR</td>
<td>23 (5%)/5 (22%)</td>
</tr>
<tr>
<td>PR</td>
<td>208 (41%)/11 (5%)</td>
</tr>
<tr>
<td>SD</td>
<td>217 (43%)/10 (5%)</td>
</tr>
<tr>
<td>SD4&lt;sup&gt;a&lt;/sup&gt;</td>
<td>159/8</td>
</tr>
<tr>
<td>SD2&lt;sup&gt;a&lt;/sup&gt;</td>
<td>50/2</td>
</tr>
<tr>
<td>PD</td>
<td>58 (11%)/0 (0%)</td>
</tr>
<tr>
<td>Reason for PD at the 1&lt;sup&gt;st&lt;/sup&gt; follow-up (n=58)</td>
<td></td>
</tr>
<tr>
<td>Size only</td>
<td>24 (41%)</td>
</tr>
<tr>
<td>New, non-target</td>
<td>33 (57%)</td>
</tr>
<tr>
<td>Other clinical reasons</td>
<td>1 (2%)</td>
</tr>
<tr>
<td>Organ for new metastases, non-target progression</td>
<td></td>
</tr>
<tr>
<td>liver</td>
<td>23</td>
</tr>
<tr>
<td>lung</td>
<td>5</td>
</tr>
<tr>
<td>bone</td>
<td>3</td>
</tr>
<tr>
<td>others</td>
<td>6</td>
</tr>
</tbody>
</table>

<sup>a</sup> the numbers of SD4 and SD2 were given for those who underwent the 2<sup>nd</sup> follow-up after 4 months (n=432).

New, non-target: appearance of new lesion or progression of non-target lesion
3.5.5 Patients in Study IV

Totally, 287 patients with metastatic breast cancer were enrolled in a multi-center randomized phase III trial comparing a combination of epirubicin (Farmorubicin®) and paclitaxel (Taxol®) alone (ET) or in combination with capecitabine (Xeloda®, TEX) 87. Study subjects with morphologically confirmed loco-regional inoperable or disseminated carcinoma were randomized unless they had received treatment with an anthracycline, a taxane or 5-FU within one year before study entry. Previous endocrine treatment for advanced disease in patients with hormone receptor positive breast cancer was allowed. Patients with known brain metastases or other malignancies within the last five years were excluded.

In this trial, chemotherapy was repeated every 3rd week until disease progression or severe adverse effect. Baseline tumor measurement was obtained by CT, MRI or palpation 4 weeks prior to randomization. Tumor measurement was repeated after every 3 cycles (every 9 weeks). Response evaluation was conducted according to RECIST.

A total of 233 patients were eligible for the study. Cases with missing measurements in one of the two evaluations (n=26) or only non-measurable lesions (n=28) were excluded. Patient characteristics are shown in Table 3.
Table 3. Patient Characteristics in TEX (n = 233)

<table>
<thead>
<tr>
<th>Characteristic</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Age at entry</td>
<td>Mean</td>
<td>54</td>
</tr>
<tr>
<td></td>
<td>Range</td>
<td>29 - 74</td>
</tr>
<tr>
<td>No. of lesions/ patient</td>
<td>Mean</td>
<td>2.8</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Range</td>
<td>1 - 9</td>
</tr>
<tr>
<td>No. of target lesions evaluated</td>
<td>Liver</td>
<td>262 (39.9%)</td>
</tr>
<tr>
<td></td>
<td>Lymph node</td>
<td>235 (35.8%)</td>
</tr>
<tr>
<td></td>
<td>Lung</td>
<td>86 (13.1%)</td>
</tr>
<tr>
<td></td>
<td>Others</td>
<td>74 (11.3%)</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>657</td>
</tr>
<tr>
<td>Treatment arm</td>
<td>TEX</td>
<td>117</td>
</tr>
<tr>
<td></td>
<td>ET</td>
<td>116</td>
</tr>
<tr>
<td>Previous adjuvant treatment</td>
<td>Yes</td>
<td>116</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>117</td>
</tr>
</tbody>
</table>
4 RESULTS
4.1 STUDY I

The number of discordant results in response categories increased gradually from 0% for 4 lesions to 8% for one lesion assessments when all cases were considered. When only patients with \( \geq 5 \) measurable lesions (53 cases) were assessed, the discordant rate ranged from 0% for 4L to 15.1% (Table 4).

Table 4. Number of cases with discordant results between the measurement of fewer lesions and measurement of all target lesions

<table>
<thead>
<tr>
<th>Type of assessment</th>
<th>1L</th>
<th>2L</th>
<th>3L</th>
<th>4L, 5L, 6L, 7L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of “mistakes” (discordant cases) when all 99 patients are considered</td>
<td>8</td>
<td>6</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Percentage of discordant cases in all 99 cases</td>
<td>8%</td>
<td>6%</td>
<td>4%</td>
<td>0%</td>
</tr>
<tr>
<td>Percentage of discordant cases in patients with ( \geq 5 ) lesions (53 cases)</td>
<td>15.1%</td>
<td>11.3%</td>
<td>7.5%</td>
<td>0%</td>
</tr>
</tbody>
</table>

(Reprinted from The British Journal of Radiology 2009 \(^{88}\), copyright 2012, by permission of British Institute of Radiology)

The mean number of measurable lesions among those with discrepancies was 7.1. This was significantly higher than those demonstrating concordance when measuring fewer target lesions and all measurable lesions (4.1 lesions; \( p < 0.05 \)).

Figure 10 shows the percentage of discordances based on the available number of target lesions and the number of target lesions chosen.

Measuring less than four target lesions increased discrepancies in terms of BOR when more than five target lesions are present.

![Figure 10. Number of discordant cases based on available target lesions.](https://example.com/figure10.png)

(Reprinted from The British Journal of Radiology 2009 \(^{88}\), copyright 2012, by permission of British Institute of Radiology)
4.2 STUDY II

The number of CT-examinations followed for response evaluation was different between the readers because of judging the same patient as PD at different time points (Figure 11). The numbers of selected target lesions were also different between the readers (Figure 12).

Figure 11. Number of CT-examinations used for evaluations by reader A 1st and reader B. (Reprinted from Acta Oncologica 2010<sup>81</sup>, copyright 2012, by permission of Informa Healthcare)

![Figure 11: Number of CT-examinations](image1)

Figure 12. Number of target lesions selected by reader A and B. (Reprinted from Acta Oncologica 2010<sup>81</sup>, copyright 2012, by permission of Informa Healthcare)

![Figure 12: Number of target lesions](image2)
As shown in Figure 13, interobserver agreement was moderate and generally lower than the intraobserver agreement, which tended to be substantial to perfect. RR according to RECIST was 33% by reader A and 21% by reader B. RR according to bi-dimensional measurements simulating WHO-criteria was 33% by reader B and 23% by reader B.

Interobserver agreement in the detection of new lesion(s) and progression of non-target lesions was moderate, $\kappa = 0.50$ (0.28–0.73). Even among five patients whose target lesions were identical between the two readers, patients were judged differently mainly because inconsistency in interpretation of new lesions and progression of non-target lesions. In seven out of 39 patients, there was a clear discordance regarding evaluation of PD. Intraobserver agreements in the detection of new lesion(s) or progression of non-target lesions ranged from substantial to perfect.

![Figure 13. Non-weighted kappa coefficient value and corresponding 95% confidence interval for agreement. (Reprinted from Acta Oncologica 2010 81, copyright 2012, by permission of Informa Healthcare)](image)

It is concluded that radiological tumor response evaluation according to RECIST and bi-dimensional tumor measurement according to WHO-criteria is subjected to a considerable inter- and intraobserver variability.
4.3 STUDY III

The HRs for PFS and OS decreased with 1st change. A decrease between 10% to <30%, albeit RECIST does not regard this as a partial response, was a positive prognostic factor for PFS and OS. Patients who had new lesions or unequivocal progression of non-measurable lesions had a worse prognosis than those with only an increase in size of >20%.

4.3.1 Relationship between 1st change and OS in mCRC

The 1st change used as a continuous variable correlated strongly with OS. A decrease in size by 50%, or more, was the most positive prognostic factor for OS. In the absence of a new lesion or unequivocal progression of non-target lesions, even an increase of ≥20%, considered as PD by RECIST, was not statistically significantly associated with impaired OS vs. an increase of <10%. Appearance of a new lesion or progression of non-target lesion(s), irrespective of the change in diameters of the target lesions, was the most negative prognostic factor for OS (Table 5).
Table 5. Cox proportional hazards multiple regression model on overall survival in 506 patients with 327 deaths

<table>
<thead>
<tr>
<th>Covariate</th>
<th>No. of patients</th>
<th>p</th>
<th>HR</th>
<th>95% CL</th>
<th>OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuous*</td>
<td>506 (327)</td>
<td>&lt;0.001</td>
<td>2.01</td>
<td>1.75 – 2.31</td>
<td>-</td>
</tr>
<tr>
<td>Categorized</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>New, non-target</td>
<td>33 (29)</td>
<td>&lt;0.001</td>
<td>3.77</td>
<td>2.08 – 6.83</td>
<td>PD</td>
</tr>
<tr>
<td>Increase</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥20%</td>
<td>26 (22)</td>
<td>0.611</td>
<td>1.18</td>
<td>0.63 – 2.20</td>
<td>SD</td>
</tr>
<tr>
<td>≥10 - &lt;20%</td>
<td>15 (10)</td>
<td>0.722</td>
<td>1.15</td>
<td>0.53 – 2.49</td>
<td></td>
</tr>
<tr>
<td>&gt;0 - &lt;10%</td>
<td>25 (18)</td>
<td>-</td>
<td>1</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>0 – decrease &lt;10%</td>
<td>104 (69)</td>
<td>0.310</td>
<td>0.76</td>
<td>0.45 – 1.28</td>
<td></td>
</tr>
<tr>
<td>Decrease</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥10 - &lt;20%</td>
<td>95 (57)</td>
<td>0.031</td>
<td>0.56</td>
<td>0.33 – 0.95</td>
<td></td>
</tr>
<tr>
<td>≥20 - &lt;30%</td>
<td>57 (31)</td>
<td>0.002</td>
<td>0.39</td>
<td>0.22 – 0.70</td>
<td></td>
</tr>
<tr>
<td>≥30 - &lt;40%</td>
<td>68 (47)</td>
<td>0.046</td>
<td>0.57</td>
<td>0.33 – 0.99</td>
<td>PR</td>
</tr>
<tr>
<td>≥40 - &lt;50%</td>
<td>40 (25)</td>
<td>0.010</td>
<td>0.45</td>
<td>0.24 – 0.82</td>
<td></td>
</tr>
<tr>
<td>≥50%</td>
<td>43 (19)</td>
<td>&lt;0.001</td>
<td>0.25</td>
<td>0.13 – 0.49</td>
<td></td>
</tr>
</tbody>
</table>

* 1st change as a continuous valuable. Patients with new lesions or progression in non-target lesion at the 1st follow-up study converted into an increase of 1.0.

Number in parenthesis indicates the number of deaths.

HR: hazard ratio

95% CL: 95% confidence limits

OR: overall response if based upon the 1st change only

New, non-target: appearance of new lesion or progression of non-target lesion
Based on the results of the Cox regression model, patients were regrouped into four groups: (i) having new lesion/unequivocal non-target lesion progression, (ii) any increase to <10% decrease, (iii) ≥10% to <50% decrease or (iv) a decrease of ≥50%. The difference in OS was statistically significant among these four groups (landmark method, log-rank test, \( p<0.00001 \)) (Figure 14). The four RECIST categories could similarly discriminate OS (Figure 15).

**Figure 14.** Kaplan-Meier curves of overall survival according to first change values

**Figure 15.** Kaplan-Meier curves of overall survival according to RECIST
4.3.2 Relationship between 1st change and PFS in mCRC

The PFS increased as the 1st change value decreased. A patient whose tumor shrank more survived longer without progression. A decrease in size by ≥50% was again the most significant positive prognostic factor of PFS (Table 6).

Table 6. Cox proportional hazards multiple regression model on progression-free survival in 506 patients with 456 events

<table>
<thead>
<tr>
<th>Covariate</th>
<th>No. of patients</th>
<th>p</th>
<th>HR</th>
<th>95% CL</th>
<th>OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuous*</td>
<td>506 (456)</td>
<td>&lt;0.001</td>
<td>2.21</td>
<td>1.97 – 2.49</td>
<td></td>
</tr>
<tr>
<td>Categorized</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>New, non-target</td>
<td>33 (33)</td>
<td>&lt;0.001</td>
<td>12.21</td>
<td>6.75 – 22.09</td>
<td></td>
</tr>
<tr>
<td>Increase ≥ 20%</td>
<td>26 (25)</td>
<td>&lt;0.001</td>
<td>3.51</td>
<td>1.98 – 6.23</td>
<td></td>
</tr>
<tr>
<td>≥10% - &lt;20%</td>
<td>15 (14)</td>
<td>0.413</td>
<td>0.76</td>
<td>0.39 – 1.47</td>
<td></td>
</tr>
<tr>
<td>&gt;0 - &lt;10%</td>
<td>25 (24)</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>0 – decrease &lt;10%</td>
<td>104 (92)</td>
<td>0.167</td>
<td>0.73</td>
<td>0.46 – 1.14</td>
<td></td>
</tr>
<tr>
<td>Decrease ≥10 - &lt;20%</td>
<td>95 (82)</td>
<td>0.013</td>
<td>0.56</td>
<td>0.36 – 0.89</td>
<td></td>
</tr>
<tr>
<td>≥20 - &lt;30%</td>
<td>57 (51)</td>
<td>0.004</td>
<td>0.49</td>
<td>0.30 – 0.80</td>
<td></td>
</tr>
<tr>
<td>≥30 - &lt;40%</td>
<td>68 (66)</td>
<td>0.038</td>
<td>0.61</td>
<td>0.38 – 0.97</td>
<td></td>
</tr>
<tr>
<td>≥40 - &lt;50%</td>
<td>40 (34)</td>
<td>&lt;0.001</td>
<td>0.38</td>
<td>0.23 – 0.64</td>
<td></td>
</tr>
<tr>
<td>≥50%</td>
<td>43 (35)</td>
<td>&lt;0.001</td>
<td>0.35</td>
<td>0.21 – 0.59</td>
<td></td>
</tr>
</tbody>
</table>

* 1st change as a continuous valuable. Patients with new lesions or progression in non-target lesion at the 1st follow-up study converted into an increase of 1.0.
Number in parenthesis indicates the number of events.
HR: hazard ratio
95% CL: 95% confidence limits
OR: overall response if based upon 1st change only
New, non-target: appearance of new lesion or progression of non-target lesion
Patients were regrouped according to the 1\textsuperscript{st} change values: (i) having new lesion/non-target lesion progression, (ii) any increase to a <10\% decrease, (iii) ≥10\% to <50\% decrease or (iv) ≥50\% decrease. The difference in PFS among these four groups was statistically significant (landmark method, log-rank test, \(p<0.00001\)) (Figure 16). The same was seen for the RECIST categories (Figure 17).

\textbf{Figure 16.} Kaplan-Meier curves of progression-free survival according to first change values

\textbf{Figure 17.} Kaplan-Meier curves of progression-free survival according to RECIST
4.3.3 The 2nd follow-up examination

At the 1st evaluation, 283 patients were evaluated as SD. Of these, 43 patients (15%) had disease progression before 4 months (designated SD2).

Patients with SD4 had significantly longer OS compared to those with SD2 (median 497 vs. 358 days, log-rank test, \( p = 0.00742 \)). The mean 1st change values for SD2 and SD4 were -0.12 (SD 0.19) and -0.09 (SD 0.13), respectively, without any statistically significant difference (\( p=0.13 \)). However, the chance of being SD4 was significantly higher if a patient’s 1st change showed a decrease of ≥10% (\( p=0.00658 \)) (Figure 18).

![Waterfall plot of the first change. Patient with new lesion at the first follow-up was converted into 1.0 (an increase of 100 %).](image)

**Figure 18.** Waterfall plot of the first change. Patient with new lesion at the first follow-up was converted into 1.0 (an increase of 100 %).
Similar to the strong predictive capacity of PFS and OS by the 1st change from baseline to 2 months, the appearance of a new lesion or progression of non-target lesion(s) at the 2nd change was a single significant negative variable and a decrease by ≥50% was a single significant positive variable for both PFS and OS (Table 7).

**Table 7.** Prognostic variables by Cox proportional hazards multiple regression model on PFS and OS at the second time (n=432).

<table>
<thead>
<tr>
<th>Covariate (2nd change)</th>
<th>No. of patients*</th>
<th>p</th>
<th>HR</th>
<th>95% CL</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PFS</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>New, non-target</td>
<td>28 (28)</td>
<td>&lt; 0.001</td>
<td>17.85</td>
<td>9.78 – 32.58</td>
</tr>
<tr>
<td>Decrease ≥50%</td>
<td>41 (35)</td>
<td>0.008</td>
<td>0.50</td>
<td>0.30 – 0.84</td>
</tr>
<tr>
<td><strong>OS</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>New, non-target</td>
<td>28 (26)</td>
<td>0.014</td>
<td>2.01</td>
<td>1.15 – 3.52</td>
</tr>
<tr>
<td>Decrease ≥50%</td>
<td>41 (17)</td>
<td>&lt;0.001</td>
<td>0.25</td>
<td>0.13 – 0.47</td>
</tr>
</tbody>
</table>

The group which showed no change to up to a 10% increase at the second follow-up was referred to as a reference group (HR=1.0).

Other variables; from any increase to decrease <50%, were not significantly different from 1.0.

* Number in parenthesis indicates the number of events.

HR: hazard ratio

95% CL: 95% confidence limits

New, non-target: appearance of new lesion or progression of non-target lesion.
4.4 STUDY IV

An increase of the sum of metastatic lesions by more than 10% or the appearance of new lesions or progression of non-target lesions (increase >10%/new/non-target) predicted a shorter OS by univariable regression analysis (HR 4.29. 95% CI 2.44 – 7.53, p<0.001). After adjustment for previous adjuvant treatment and the treatment given within the frame of the randomized trial, OS was still significantly shorter in the group of patients with an increase of the sum of metastatic lesions by more than 10% or the appearance of new lesions or progression of non-target lesions (HR 4.58 95% CI 2.86 – 7.33, p<0.001).

4.4.1 The extent of 1st change in MBC

The median time from randomization to the first evaluation was 8.3 weeks (SD 1.6 weeks). The extent of the first change is shown in Figure 19. Twenty-three (10%) patients had a new lesion or progression of non-target lesions at this point; eight of these in combination with more than a 20% increase of target lesions; i.e. cut-off point for PD according to RECIST. No patient had solely more than a 20% increase of target lesions. Survival longer than the median OS (blue column in Figure 19) was seen both in patients who had at least a 30% decrease (PR for RECIST), and among those with a smaller decrease (10 – 30%), or no change (<10% increase or decrease) (SD for RECIST).

![Figure 19. Waterfall plot of the first change of tumor size.](image)
4.4.2 Relationship between 1st change and OS in MBC

Univariate regression analyses showed that the group of patients with more than a 10% increase in tumor size or appearance of new lesions or progression of non-target lesions at the first evaluation (increase >10%/new/non-target) had a significantly increased hazard of death (HR 4.29, 95% CI: 2.44 – 7.53, p<0.001). Groups with no change (±10%) and decrease in size by more than 10% had a similar outcome (Table 8). According to RECIST, no response (SD or PD) increased the risk for an unfavorable outcome (HR 1.97, 95% CI: 1.43 – 2.71, p<0.001, Table 8).

Table 8. Univariable Cox regression analyses on overall survival (OS) in 233 patients with 158 events

<table>
<thead>
<tr>
<th>Change of size at the first response evaluation</th>
<th>No. (%)</th>
<th>HR</th>
<th>95% CI</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Increase &gt;10%/new/non-target (a)</td>
<td>24 (10)</td>
<td>4.29</td>
<td>2.44 – 7.53</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>No change ± 10%</td>
<td>42 (18)</td>
<td>1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Decrease &gt;10%</td>
<td>167 (72)</td>
<td>0.86</td>
<td>-0.56 – 0.25</td>
<td>0.45</td>
</tr>
</tbody>
</table>

**Overall response**

| Responders (CR, PR)                          | 138 (59)| 1   | -            | -       |
| Non-responders (SD, PD)                      | 95 (41) | 1.97| 1.43 – 2.71  | < 0.001 |

(a) new/non-target: appearance of new lesion or progression of non-target lesion
There was a significant difference in OS between patients with an increase >10% /new/non-target and those without such findings (log-rank $\chi^2=20.4$, $p<0.001$; Figure 20). Corresponding differences were found between responders and non-responders according to RECIST (log-rank $\chi^2=15.8$, $p<0.001$, Figure 21).

**Figure 20.** Kaplan-Meier curves of overall survival according to the first change

**Figure 21.** Kaplan-Meier curves of overall survival according to RECIST
4.4.3 The Cox regression analysis of OS in MBC

Clinical data from the TEX trial indicated that previous adjuvant treatment reduced the chance to respond to the study treatment \(^8\). Adjusted for adjuvant treatment status and study treatment arm (ET vs. TEX) an increase >10%/new/non-target lesions was found to be an independent prognostic factor for OS (HR 4.58, 95% CI: 2.86 – 7.33, p<0.001, Table 9).

Table 9. Multivariate Cox regression model adjusting the first change of size to previous adjuvant treatment and study treatment arms in terms of overall survival (OS)

<table>
<thead>
<tr>
<th></th>
<th>HR</th>
<th>95% CI</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Increase &gt;10%/new/non-target (^{(a)})</td>
<td>4.58</td>
<td>2.86 – 7.33</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Previous adjuvant treatment (None = 1)</td>
<td>1.40</td>
<td>1.02 – 1.92</td>
<td>0.04</td>
</tr>
<tr>
<td>Study treatment (ET=1.00)</td>
<td>0.92</td>
<td>0.67 – 1.26</td>
<td>0.59</td>
</tr>
</tbody>
</table>

\(^{(a)}\) new/non-target: appearance of new lesion or progression of non-target lesion
5 DISCUSSION

5.1 TO MEASURE OR NOT TO MEASURE, THAT IS THE PROBLEM

In order to allow adequate evaluation of many various anti-neoplastic drugs examined in clinical trials, quantitative information is required. The increasing demand to objectively verify the treatment response of new cancer agents has resulted in an increased use of medical imaging, especially CT. WHO-criteria and RECIST (1.0 and 1.1) were set up in order to convert the analogous information provided by the radiological studies into the digital information in a strict and standardized way. The digital information, such as measurement values and changes before and after initiation of cancer agents appears to be very precise and objective, however the crucial, but subjective steps, when a lesion is selected and thereafter measured have surprisingly little been questioned.

In this thesis various sources of measurement variability in the process of tumor response evaluation has been revealed and described. These discordances can result in considerable discrepancies in response evaluation using both WHO-criteria and RECIST despite seemingly standardized conditions on how to carry out the evaluations. The possible negative effects of these findings on the reliability of clinical trials are obvious.

The significant variability between readers in the selection, in the measurement of the lesions as well as in tracking new lesions and in the attention of non-target lesions was shown in study II. To minimize diversity in selection of target lesions, reducing the number of target-lesions might be suggested as a possible solution. However as shown in study I as well as several previous retrospective studies, reducing the number has negative impact on consistency. This is because the contribution of a single lesion to the response evaluation increases when the number of lesions is reduced. It is also doubtful that only one or two arbitrary selected lesions in a patient with multiple metastases can represent the patient’s true response.

In our radiology department, one dedicated radiologist with experience in oncological imaging takes responsibility of response evaluation in each clinical trial. This one radiologist to one trial system ensures a more consistent evaluation, thus more reliable result. Independent central review (ICR) is a way to enable more objective, reproducible and independent evaluation of results compared to the response evaluation carried out in the daily workflow.
Moreover, considerable advances in medical imaging technology such as automated CT volumetric measurement and computer aided detection (CAD) may help to improve accuracy in the measurement and to reduce intra- and inter-observer inconsistencies\textsuperscript{12-14, 91-93}. Alternative measurement values, for example SUV based on the PET, has potentially higher reproducibility than tumor size measurement on CT\textsuperscript{15}. Even though all of those inconsistencies in measurement were minimized, it is still questionable how much the tumor measurement itself correlates with a patient’s prognosis. In studies II, III and IV, the main reason for PD was not size increase of target lesion but the appearance of new lesion and/or progression of non-target lesion. In addition, study III indicated a marked difference in prognosis between patients with or without a new lesion or unequivocal progression of non-target lesion(s) in patients with mCRC. To our knowledge, this has not been reported before, or incorporated in the simulation model\textsuperscript{42}. Since there was a complete overlap between patients with more than 20% increase and the development of new lesions and progression of non-target lesions in study IV, we were not able to find differences in survival between PD with and without development of new lesions and progression of non-target lesions in patients with MBC. Yet, the presence of a new lesion or progression of non-target lesions was highlighted as potentially more critical for survival as any change in size in study IV. On the basis of our studies, one may speculate that the appearance of new lesions or progression of non-target lesions could mean that the tumor disease is biologically more aggressive than when the progression is only due to an increase in the size of target lesions.

5.2 CLINICAL IMPLICATIONS, WHAT MATTERS MOST?

Few physicians and researchers have questioned the relevance of thresholds or categorization for response according to RECIST. Instead, the value of functional imaging for early response prediction has been the focus, with contradictory results\textsuperscript{25-27}. The relevance of early changes in size has not been well explored. The initial change in tumor size in mCRC (study III) and MBC (study IV) correlates with patient prognosis in terms of PFS (study III) and OS (study III and IV). This has previously never been shown besides for a modeling study to support end-of-phase II decision and design for phase III studies\textsuperscript{42}. On the basis of studies III and IV, the comparison of cytotoxic treatments in a trial can be achieved more rapidly by the 1st change approach than waiting for BOR using RECIST.
Furthermore, these studies question the currently adopted cut-off values to evaluate tumor response. The cut-off values used by RECIST between PD and SD on the one hand and between SD and PR on the other do not seem to be optimally chosen. The chosen cut-off values should be reappraised from a clinical standpoint 6 and not only from the extent of measurement errors 2, 5, 14, 15, 63, 64, 94. The BOR and ORR have been used as a surrogate outcome that is likely to predict clinical benefit for long time 39, 95. However, the value of the BOR and ORR as a surrogate indicator has been debated. Studies of different tumor types, particularly colorectal cancer (CRC) and MBC, have failed to show a strong correlation between a BOR and prolonged survival 19, 21, 23, 83. A survival benefit has been shown not only for patients who had a complete or partial response (CR+PR) but also for those who maintained SD for at least 4 months; the latter patients did not satisfy WHO-criteria requirements for PR 19. RECIST’s thresholds were also recently questioned in a mCRC trial including the epidermal growth factor receptor (EGFR) inhibitor cetuximab 22, 96 and in a MBC trial including the EGFR inhibitor bevacizumab 24. Even other surrogate outcomes, PFS or TTP could not demonstrate as a good surrogate for OS in a first-line randomized clinical trials that compared an anthracycline with a taxane 23 or in a phase III randomized trial that compared chemotherapy with or without bevacizumab 24. Why these surrogate outcomes do not correlate with OS? Besides thresholds setting, it should be questioned why dichotomization or categorization of response is necessary. Categorization often simplifies data analysis and interpretation. It is, however, obvious that once tumor size changes are dichotomized or categorized, e.g. responder and non-responder, it loses valuable information and statistical efficiency than continuous variable 40, 69. As shown study III, the first change as continuous variable might correlate better with OS.

It is crucial to find a surrogate outcome which correlates with OS. That may lead better planning for phase II and III trials, reduce the risk of failure of clinical trials, especially phase III trials and eventually lead reduction in drug developing costs 20. We hope the first change described in this thesis can be an effective surrogate for OS and help to deliver effective treatments to patients.
5.3 LIMITATIONS

There are several limitations in the studies that constitute this thesis. First, all studies were conducted retrospectively. Second, RECIST 1.1 was not used in any study since it is relatively new, introduced in 2009, and there are only about 15 published articles that adopted RECIST 1.1 so far. We cannot tell that issues addressed in study I and II are still valid in RECIST 1.1. However, the impact of initial tumor size change and appearance of new lesion/progression of non-target lesion to patient’s prognosis are of importance regardless which criteria is adopted.

Study I included a considerable number of patients while the mean number of measurable lesions per patient was limited. Larger studies with also more measurable lesions per patients are needed.

Study II is based on a restricted number of observations and on a heterogeneous patient group why it must be regarded as a pilot study. Ideally, it should be repeated in larger cohorts and at other institutions. Nevertheless and especially considering these limitations, the inconsistency of how lesions are selected, measured and interpreted between different readers remains an important finding.

Tumor measurement for WHO-criteria was actually “RECIST” based measurement, namely the perpendicular longest diameter was obtained for a lesion which was selected as a target lesion according to RECIST. This might cause underestimation of the inconsistency because there is no strict stipulation of number of target lesion or measurable lesion in WHO-criteria.

A limitation of studies III and IV is that only one time point defined by trial protocol for the follow-up study was analyzed (8 weeks for mCRC and MBC). We don’t know if this time point is optimal for measuring 1st change. Six weeks after initiation of chemotherapy with cetuximab for mCRC, 7 weeks after initiation of capecitabine or 5-FU for mCRC or 8 weeks after initiation of imatinib therapy for GIST have previously also been used.

In study III, a new lesion may represent an increase from an undetectable sub-cm to 1 cm lesion, and does not necessary indicate more aggressiveness than an increase from 1 cm to 1.2 cm. When a new lesion is reported in a trial, its size is not recorded in the CRFs. Without re-evaluating all imaging, we could not further explore this issue.

In study IV, the data also comes from a relatively restricted number of patients. The cut-off point in case of tumor shrinkage could not be determined because of lack of statistical power. Even though it could be possible, a majority of the patients who developed disease progression received second and later lines of treatments, thus the
relationship between the first tumor shrinkage and OS could be influenced by second and later lines of treatments.
5.4 FUTURE PERSPECTIVES

The findings of this thesis have revealed various limitations in the current tumor response evaluation criteria. Future studies have to be conducted in order to answer the following questions:

1. Is the first change applicable for molecular targeted anticancer agents like for cytotoxic anticancer agents?
   - The first change in metastatic colorectal cancer treated with the Nordic VII study including cetuximab (Erbitux®, anti-EGFR antibody) is planned.

2. Can the first change oriented tumor monitoring be beneficial for cancer care?
   - Enable better follow-up and care and ultimately prolong the overall survival?
   - Omit unnecessary CT or PET-CT examinations and economical burden?
   - Possibility to develop a better phase III clinical trial model based on the first change which providing better planning and minimizing the failure of phase III trials?

3. Does the tumor size really matter or not?
   - Is the change of tumor size before and after treatment really relevant for the prognosis?
   - Do measurements based on the Euclidian geometry reflect the biological behavior of a tumor?
   - Do the fractal dimension rather than Euclidian geometry serve a better indicator to evaluate tumors?

The Euclidian geometry is suitable to measure rather simple structures with smooth surface, e.g. spheres, which never be seen in tumors. Tumors have irregular and heterogeneous structures which can be measured by Fractal geometry rather than Euclidian geometry. Fractal geometry has been used to express the complex network of neurons, blood vessels and tumors 98-100.
6 CONCLUSIONS

1. WHO-criteria and RECIST contain a large source of random errors which may cause variability deteriorating the reliability of clinical trials.

2. The initial change in tumor size at the first follow-up CT examinations after initiation of treatment correlates with overall survival in mCRC and MBC.

3. The response evaluation must be more focused on detecting appearance of new lesion or the progression of non-target lesion. These indicates particularly poor prognosis. Anatomical imaging such as CT is still useful for response evaluation with wide accessibility. The current process of how to read-out response should be improved.
7 APPENDIX

7.1 ILLUSTRATIVE CHART FOR TUMOR EVALUATION

An illustrative chart for tumor response evaluation according to RECIST and WHO is shown.

Data are from a patient with multiple lung, liver, lymph node and bone metastasis. Extracted from RadioGraphics 2008[101], copyright 2012 by permission of The Radiological Society of North America (RSNA®).

<table>
<thead>
<tr>
<th>Baseline Study</th>
<th>Tumor Involvement</th>
<th>Presence of &quot;non-target&quot; lesion(s)</th>
<th>reason</th>
<th>Appearance of new lesion</th>
<th>Follow-up study</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (yy-mm-dd)</td>
<td>(1) Primary</td>
<td>Y N/U</td>
<td>&lt;10mm / &gt;5 / other</td>
<td>Yes PD</td>
<td>2 (yy-mm-dd)</td>
</tr>
<tr>
<td>2 LN</td>
<td>Y N/U</td>
<td>Y N/U</td>
<td>&lt;10mm / &gt;5 / other</td>
<td>No</td>
<td>3 (yy-mm-dd)</td>
</tr>
<tr>
<td>3 Lung</td>
<td>Y N/U</td>
<td>Y N/U</td>
<td>&lt;10mm / &gt;5 / other</td>
<td>Yes PD</td>
<td>4 (yy-mm-dd)</td>
</tr>
<tr>
<td>4 Liver</td>
<td>Y N/U</td>
<td>Y N/U</td>
<td>&lt;10mm / &gt;5 / other</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>5 Brain</td>
<td>Y N/U</td>
<td>Y N/U</td>
<td>&lt;10mm / &gt;5 / other</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>6 Other site</td>
<td>Y N/U</td>
<td>Y N/U</td>
<td>&lt;10mm / &gt;5 / other</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Bone</td>
<td>Y N/U</td>
<td>Malignancy related? Y / N / U</td>
<td></td>
<td>No PD</td>
<td></td>
</tr>
<tr>
<td>Pleural effusion</td>
<td>Y N/U</td>
<td>Malignancy related? Y / N / U</td>
<td></td>
<td>Yes PD</td>
<td></td>
</tr>
<tr>
<td>Ascoite</td>
<td>Y N/U</td>
<td>Malignancy related? Y / N / U</td>
<td></td>
<td>No PD</td>
<td></td>
</tr>
</tbody>
</table>

Other "truly" non-target lesions
Leptomeningeal / Inflammatory breast disease / Lymphangitis / Cystic / Necrotic / Lesion in the post-irradiated area / others ( )

<table>
<thead>
<tr>
<th>Target</th>
<th>Site *</th>
<th>comment</th>
<th>Size (mm)</th>
<th>Size (mm)</th>
<th>Size (mm)</th>
<th>Size (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(1) / (2) / (3) / (4) / (5) / (6)</td>
<td>S5</td>
<td>26.2 x 21.5</td>
<td>24.5 x 21.3</td>
<td>22.7 x 21.0</td>
<td>27.6 x 24.1</td>
</tr>
<tr>
<td>2</td>
<td>(1) / (2) / (3) / (4) / (5) / (6)</td>
<td>S4</td>
<td>22.3 x 17.7</td>
<td>19.9 x 15.3</td>
<td>16.5 x 14.3</td>
<td>23.1 x 20.3</td>
</tr>
<tr>
<td>3</td>
<td>(1) / (2) / (3) / (4) / (5) / (6)</td>
<td>S3</td>
<td>20.9 x 13.9</td>
<td>15.9 x 10.4</td>
<td>10.4 x 9.4</td>
<td>22.6 x 17.3</td>
</tr>
<tr>
<td>4</td>
<td>(1) / (2) / (3) / (4) / (5) / (6)</td>
<td>Rt. SIV</td>
<td>15.9 x 12.0</td>
<td>12.4 (9.8)</td>
<td>9.8 (8.4)</td>
<td>19.3 (15.8)</td>
</tr>
<tr>
<td>5</td>
<td>(1) / (2) / (3) / (4) / (5) / (6)</td>
<td>parRAo</td>
<td>14.5 x 13.4</td>
<td>12.4 (7.2)</td>
<td>10.6 (11.3)</td>
<td>17.3 (14.4)</td>
</tr>
<tr>
<td>6</td>
<td>(1) / (2) / (3) / (4) / (5) / (6)</td>
<td>Rt. SIV</td>
<td>12.1 x 11.5</td>
<td>6.8 x 12.2</td>
<td>5.2 (6.2)</td>
<td>15.6 (12.6)</td>
</tr>
<tr>
<td>7</td>
<td>(1) / (2) / (3) / (4) / (5) / (6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>(1) / (2) / (3) / (4) / (5) / (6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>(1) / (2) / (3) / (4) / (5) / (6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>(1) / (2) / (3) / (4) / (5) / (6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Sum (WHO) | 111.8 (1765.4) | 91.9 (1218.3) | 77.2 (1057.2) | 125.4 (1772.7) |

Evaluation (WHO)

<table>
<thead>
<tr>
<th>Ratio (%)</th>
<th>Category</th>
<th>SD (SO)</th>
<th>PR (PD)</th>
<th>PD (PD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>-18 (-31)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-31 (-41)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+62 (+115)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* each number under “Site” corresponds to the above-mentioned number of an organ, e.g., (2)=lymph node, (4)=liver

Comment: Hamartoma in the left lung
# values in parentheses indicate the longest perpendicular diameter of the target lesion for 2D (WHO) evaluation

**Base line evaluation**

A: Note the presence or absence of tumor involvement.

B: Start from the largest lesion.

C: Note in which window-setting the lesion is measured.

**Follow-up evaluation**

D: Before starting measurements, look for new lesions or unequivocal progression of non-target lesions that are critical to follow-up evaluation. The patient will then automatically be classified as PD regardless of other measurements.

E: Even if the target lesions become too small to obtain accurate measurement, it is important to keep on reporting them. *Or use default value of 5 mm (only valid for RECIST 1.1)*

F: In case of disease progression, the smallest sum of the longest diameters since the trial started should be referred rather than the baseline sum of the longest diameters.

G: Since lymph node metastasis often change size along their short axis, measurement of the longest diameter in the trans-axial plane may conceal change of size. *Lymph nodes should be measured in the short axis if RECIST 1.1 is used in the trial.*

In WHO-criteria, if the product of longest diameter and longest perpendicular diameter increased more than 25% in any lesion, the patient should be categorized as PD. On the other hand, in RECIST, progression in any lesion without an overall 20% increase in the sum of longest diameter is not considered progression.

As in this patient shown above, the lymph node metastasis increases more than 25% at the third follow-up. The patient should be classified as PD according to WHO-criteria. However, in RECIST, the sum of the longest diameters is more than 30% reduction of baseline sum and the patient is classified as PR.
8 ACKNOWLEDGEMENTS

I wish to express my sincere gratitude for the following persons whose help has been indispensable. Without you, I am sure, this would never come true.

My principal supervisor Lennart Blomqvist, who keeps on encouraging me and protecting me. Always paving the “Royal Road” for PhD candidates like me and guiding not to stray from the path. I cannot find any word to describe how I thank you! I have learnt a lot.

My co-supervisor Professor Hans Jacobsson, a brilliant person with abundant knowledge in many subjects and great sense of humor. The whole idea about this tumor response evaluation came from you. Without your support, I could not survive here.

My co-supervisor Hirofumi Fujii, one of the smartest people I have ever met. I am really lucky to have you as my supervisor ever since I was in Keio University in Japan. Professor Bengt Glimelius, you are so brilliant, needless to explain! Thank you very much for your kindly understanding to my primitive ideas of the first change and thoughtful comments, without them I were still be in the depths. I am looking forward to our future collaborations.

Professor Anders Sundin, a great educator as well as a brilliant scientist. Always keep on encouraging me and keep on trusting me. I will never forget your heartwarming words.

Thomas Hatschek, a great clinician and a scientist who keep on encouraging me from the beginning when I just came to Sweden. One of the key person who let me in this academic world. I am really lucky to meet you.

Maria Liljefors, my external mentor and one of my honorable role-model as a scientist, a woman and a mother. You rescued me, comforted me and encouraged me when I was in a terrible deep ditch.

Elisabeth Berg, the greatest statistician of all projects, I also do appreciate your patience to try to understand what I say in my bad Swedish or even worse in English.

Michael Torkzad, a distinguished colleague and a friend, I am looking forward to future collaborations.

Ann-Britt Wikstrom, excellent administrative of IMMK, always supporting me and guiding me when I was nearly drowning in a bunch of complex and long paper works.
Clementine Molin, always elegant, very smart and keep on encouraging me and telling me the importance of these projects for the development of cancer drugs.
The present head of Department of Diagnostic Radiology, Henry Lindholm and the former head of Department of Diagnostic Radiology, Lott Bergstrand for your kind arrangements and support.
Boel Söderén, IngMarie Meyer, Katarina Bodén, you are my sisters in Sweden!
Without your friendship I had been running back to Japan. Your jokes, a lot of laughs and encouragements keep me anchored here.
Veli Söderlund, Jan Bohlin, Haakon Viland, Magnus Tengvar, Erik Rollvén, Mikael Skorpil wonderful colleagues and friends. What a joy to work together with you!
Roberto, Yvonne, Manijeh, Mia and all fantastic MR staffs, thanks making me always happy!
Toya, Kirsi, Solweig for your great hospitality, it is always great pleasure to talk with you!
Noriko Kamata who is my “master Yoda” not only in my medical carrier in the radiology but in my life. Without you, I never be like me at present, yes, this is all your presents for me, thank you!
Fumikazu Sakai, Takashi Ushimi, both great radiologists and scientists. I appreciate your abundant knowledge that makes many radiology congresses fruitful.
All my friends both in Japan and in Sweden and all colleagues and staffs at the Department of Diagnostic Radiology, Karolinska University Hospital Solna and Department of Diagnostic Radiology Tokyo Metropolitan Cancer and Infectious Disease Center Komagome Hospital, Tokyo, Japan.
My father, Masayoshi Suzuki, he passed away in 2010. I am so sorry that I could not show you my thesis.
My mother, Tomoko Suzuki, who always pushes my back to move forward, always accept whatever I decide. You are the greatest mother and grandmother in the world.

And last but not least my beloved daughter, Sakiko. Saki had no idea how long she would stay in Sweden or how she would be like a “cosmopolitan” when she was taken here at the age of 5 by idiotic mother. Thank you for creating the cover jacket of this thesis. What a great privilege to be your mother!
9 REFERENCES


20. Ratain MJ, Eckhardt SG. Phase II studies of modern drugs directed against new targets: if you are fazed, too, then resist RECIST. *J Clin Oncol.* Nov 15 2004;22(22):4442-4445.


88. Darkeh MH, Suzuki C, Torkzad MR. The minimum number of target lesions that need to be measured to be representative of the total number of target lesions (according to RECIST). *Br J Radiol*. Aug 2009;82(980):681-686.


