Development of Methods and Strategies for Optimisation of X-ray Examinations

Jonny Hansson

Department of Radiation Physics
Institute of Clinical Sciences
Sahlgrenska Academy, University of Gothenburg

UNIVERSITY OF GOTHENBURG

Gothenburg 2019

Development of methods and strategies for optimisation of x-ray examinations © Jonny Hansson 2019
jonny.hansson@gu.se/jonny.hansson@vgregion.se

ISBN 978-91-7833-684-5 (PRINT)
ISBN 978-91-7833-685-2 (PDF)
http://hdl.handle.net/2077/61691

Printed in Gothenburg, Sweden 2019
Printed by BrandFactory

Es ist manchmal ganz nützlich, kräftige Muskeln zu besitzen

Baron Munchhausen, after rescuing himself and his horse from a swamp by lifting himself and the horse out by his pigtail.

Erich Kästner, *Des Freiherrn von Münchhausen wunderbare Reisen und Abenteuer zu Wasser und zu Lande*

Development of Methods and Strategies for Optimisation of X-ray Examinations

Jonny Hansson

Department of Radiation Physics, Institute of Clinical Sciences
Sahlgrenska Academy, University of Gothenburg
Gothenburg, Sweden

ABSTRACT

The overall aim of the work presented in this thesis was to develop methods and strategies for the optimisation process prescribed by legal authorities for medical X-ray imaging. This overall aim was divided into four detailed aims: 1) to analyse and describe the conditions for the optimisation of a given projectional X-ray examination in a digital environment, 2) to develop an overall strategy for the optimisation work in a radiology department, 3) to develop and implement a suitable method for statistical analysis of visual grading characteristics (VGC) data, and, 4) to evaluate the characteristics of the new statistical method by comparison with receiver operating characteristics (ROC) statistical methodology and by simulations.

The four aims are coupled to the five papers presented in this thesis. In Paper I, the conditions for the optimisation of a given projectional X-ray examination in a digital environment are analysed and a proposed optimisation strategy, based on the analysis, is described. In Paper II an overall strategy for the prioritisation of the optimisation work in a radiology department is presented. Paper III describes the development of a suitable method for statistical analysis of VGC data, which is implemented in the software VGC Analyzer. In Papers IV and V, the characteristics of the new statistical method are thoroughly evaluated by comparison with ROC statistical methodology and by simulations.

The strategies developed helped clarify the prerequisites in the process of optimising medical X-ray imaging and were shown to be useful in clinical applications. However, the objective of optimising the radiation protection in medical use of radiation is not fully clarified in legal requirements, and needs further discussion. The development of resampling methods for statistical analysis of VGC data, implemented in VGC Analyzer, provides a method that is easy to apply in clinical optimisation projects where visual grading is judged to be the appropriate evaluation method.

Keywords: optimisation, visual grading, VGC Analyzer

ISBN 978-91-7833-684-5 (PRINT)
ISBN 978-91-7833-685-2 (PDF) http://hdl.handle.net/2077/61691

SAMMANFATTNING PÅ SVENSKA

Användandet av joniserande strålning har bidragit stort till sjukvårdens utveckling under de senaste 120 åren. Den risk för skador som följer av att människor utsätts för strålning gör dock att användandet måste ske med stor försiktighet. Myndigheter är också mycket tydliga med att den som utsätter patienter för joniserande strålning måste göra detta på ett kontrollerat och optimerat sätt, även om strålning i sjukvården används med ett gott syfte. Motiverat av dessa myndighetskrav på kvalitetssäkring och det medicinska behovet av ständiga förbättringar lägger sjukvården stora resurser på att optimera sin strålningsanvändning, dvs. balansera nytta mot risk. Det är därför av stor vikt att optimering görs på ett så effektivt och högkvalitativt sätt som möjligt. Syftet med denna avhandling har varit att bidra till förbättring inom detta område.

I första delen av denna avhandling har ett förslag till strategi för att genomföra en systematisk optimering av en undersökningsmetod tagits fram, liksom en praktisk metod för att prioritera i vilken ordning olika undersökningsmetoder ska optimeras. I utvärdering av den förslagna optimeringsstrategin framkom ett behov av ett statistiskt verktyg för att testa den statistiska säkerheten i en uppmätt skillnad mellan två jämförda undersökningsmetoder. Målsättningen för andra delen av avhandlingsarbetet blev därför att utveckla ett sådant verktyg och att utvärdera hur väl det fungerar för sitt syfte. Det statistiska verktyget som består av programvaran, VGC Analyzer, kan skatta den statistiska osäkerheten i en värderingsstudie av bildkvalitet, en s.k. visual grading-studie. Skattningen av osäkerheten görs genom återanvändning av insamlade data, bootstrapping och permutation, som simulerar den verkliga fördelningen och möjliggör att inga antaganden behövs om hur granskarna har tolkat den skala på vilken bilkvalitetsbedömningen är gjord. Utvärderingen av VGC Analyzer visar att den ger korrekt analys för studier som är utförda med god statistisk grund. För studier med begränsade data minskar korrektheten i analysen.

Den föreslagna strategin för att genomföra en optimering av en undersökningsmetod och det praktiska sättet för att prioritera i vilken ordning underökningar ska optimeras kan förhoppningsvis bidra till att optimeringsarbetet effektiviseras och kan genomföras med högre kvalitet. Framtagandet av det statistiska verktyget, VGC Analyzer, är ett bidrag till att på ett enkelt sätt utvärdera den statistiska säkerheten i en uppmätt skillnad mellan två jämförda undersökningsmetoder.

LIST OF PAPERS

This thesis is based on the following papers, referred to in the text by their Roman numerals.

I. M Båth, M Håkansson, J Hansson and L G Månsson. A conceptual optimisation strategy for radiography in a digital environment. Radiat. Prot. Dosimetry 114, 230-235, 2005.

II. J Hansson, P Sund, P Jonasson, L G Månsson and M Båth. A practical approach to prioritise among optimisation tasks in x-ray imaging: introducing the 4-bit concept. Radiat. Prot. Dosimetry 139, 393-399, 2010.

III. M Båth and J Hansson. VGC Analyzer: A software for statistical analysis of fully crossed multiple-reader multiple-case visual grading characteristics studies. Radiat. Prot. Dosimetry 169, 46-53, 2016.

IV. J Hansson, L G Månsson and M Båth. The validity of using ROC software for analysing visual grading characteristics data: an investigation based on the novel software VGC Analyzer. Radiat. Prot. Dosimetry 169, 54-59, 2016.

V. J Hansson, L G Månsson och M Båth. Evaluation of resampling methods for analysis of visual grading data by comparison with state-of-the-art ROC methodology and analysis of simulated data. Submitted.

Papers I – IV are reprinted by permission of Oxford University Press.

OTHER RELATED PUBLICATIONS NOT INCLUDED IN THIS THESIS

1. J Hansson, M Båth, M Håkansson, H Grundin, E Bjurklint, P Orvestad, A Kjellström, H Boström, M Jönsson, K Jonsson and L G Månsson. An optimisation strategy in a digital environment applied to neonatal chest imaging. Radiat. Prot. Dosimetry 114, 278-285, 2005.
2. A Carlander, J Hansson, J Söderberg, K Steneryd and M Båth. Clinical evaluation of a dual-side readout technique computed radiography system in chest radiography of premature neonates. Acta Radiol. 49, 468-474, 2008.
3. S Zachrisson, J Hansson, Å Cederblad, K Geterud and M Båth. Optimisation of tube voltage for conventional urography using a Gd_2O_2S:Tb flat panel detector. Radiat. Prot. Dosimetry 139, 86-91, 2010.
4. A Carlander, J Hansson, J Söderberg, K Steneryd and M Båth. The effect of radiation dose reduction on clinical image quality in chest radiography of premature neonates using a dual-side readout technique computed radiography system. Radiat. Prot. Dosimetry 139, 275-80, 2010.
5. J Hansson, S Eriksson, A Thilander-Klang and M Båth. Comparison of three methods for determining CT dose profile: presenting the tritium method. Radiat. Prot. Dosimetry 139, 434-438, 2010.
6. A Thilander-Klang, K Ledenius, J Hansson, P Sund and M Båth. Evaluation of subjective assessment of the low-contrast visibility in constancy control of computed tomography. Radiat Prot Dosimetry 139, 449-54, 2010.
7. H Precht, J Hansson, C Outzen, P Hogg and A Tingberg. Radiographers' perspectives on Visual Grading Analysis as a scientific method to evaluate image quality. Radiography 25, 14-18, 2019.

CONTENT

ABBREVIATIONS

ALARA	As low as reasonably achievable
AUC	Area under the curve
CI	Confidence interval
CT	Computed tomography
DBM	Dorfman, Berbaum and Metz
DRL	Diagnostic reference level
ICRP	International Commission on Radiological Protection
ICS	Image criteria score
IOC	Intraoperative cholangiography
FOM	Figure of merit
FPF	False positive fraction
K_{air}	Air kerma
KAP	Kerma-area product
Kerma	Kinetic energy released per unit mass
MRMC	Multiple-reader multiple-case
ROC	Receiver operating characteristics
TPF	True positive fraction
VGA	Visual grading analysis
VGC	Visual grading characteristics
VGR	Visual grading regression

1 INTRODUCTION

The use of X-rays in medical diagnostics has been important since the first X-ray image of Frau Röntgen's left hand, just after her husband's discovery of X-rays in 1895. However, there has been a constant need to improve the image quality and increase the information content in X-ray images. Also, the risks of excessive exposure to ionising radiation became evident after a few years. Early improvement efforts were focused on technical developments to achieve better X-ray output and longer life-times of the X-ray tubes, more sensitive detector materials to allow shorter exposure times, and the development of medical applications for new patient groups[1]. Furthermore, the global epidemic of tuberculosis in the middle of the 20th century led to an urgent need for more time- and cost-effective X-ray equipment to meet the enormous diagnostic need.

In 1947 Birkelo et al.[2] presented a comparison of existing X-ray techniques for detection of tuberculosis. The aim of their study was to ascertain that newly developed equipment for the effective examination of a large number of patients could deliver images with a quality as good as the gold standard equipment. However, their ground-breaking findings were that radiologists were not united in their diagnostic conclusions (i.e., consensus was not a reliable measure of the truth), and that improvement of the statistical methods used to analyse the results was required. After these discoveries, extensive efforts were started among American radiologists to develop methods of measuring how many lesions were missed by an observer (also referred to as reader) and to identify the underlying reasons for lesions being missed by an observer[3].

The terms "underreading" and "overreading" were used by Birkelo et al., but the more general terms, "sensitivity" and "specificity" came into use after their introduction by Lusted in 1960[4]. His introduction of the statistical-decision-theory approach to the analysis of observer response data led to the observers in a study not only stating whether pathology was present or not, but also to state the confidence with which they made that decision. This new approach led to the use of receiver operating characteristics (ROC) for the presentation of the observer's response on the images. It was, however, not until 1979 that Swets et al.[5] presented a study in which the ROC approach was used in the assessment of clinical images[3].

A great number of studies has followed this first clinical ROC study, contributing to more validated examination techniques. The ROC

methodology has also been further developed to better describe the statistical properties of a study. Statistical methods for different study set-ups have also been improved. Nevertheless, the basis for any ROC study is that the observer's assessments are compared to a known truth. The advantage of the ROC method is that it can be used to measure an absolute result in a clinically relevant situation, and if the detection of the abnormality of search is critical for the outcome of an examination, the study has high validity. However, in many situations it is not easy to establish the true condition, resulting in time-consuming studies. There is also a risk that the need for truth will reduce the clinical validity, as the selection of examinations that can be studied is limited to images in which information is provided of the existence or nonexistence of pathology.

To overcome the obstacle of having to produce a truth for the ROC study a parallel method of image quality evaluation has been developed. The concept of visual grading has evolved from the established tradition among radiologists of validating a clinical image by assessing the visibility of known structures that should be visible in a good-quality clinical image. The idea is that by defining standardised structures to be assessed, the observer's rating of the visibility of the structures is a good measure of the clinical value of that image. Results from visual grading studies have also been shown to agree with ROC studies performed in parallel[6-8]. The methodological aspects of visual grading have been improved during recent decades, although there is considerable potential for further development of e.g. statistical methods in visual grading.

The optimisation of an X-ray examination method, with the goal to establish the most favourable examination method in terms of a high-quality medical outcome at a low cost/risk, is a complex process that must be performed in a structured way to ensure a reliable result. Therefore, as the evaluation of the quality of clinical images by human observers is the final step in the assessment of the optimal technique for a specific diagnostic task, this step must be preceded by a structural evaluation process in order to ensure that the methods being compared are assessed on equal terms. Evaluation studies of new imaging techniques with improved image quality, albeit with higher radiation dose, or conversely, studies of new imaging technique using a lower radiation dose, albeit with an assessed lower image quality, may be difficult to interpret.

During the technical transformation from analogue imaging to digital imaging in projectional X-ray imaging during the last 30 years, many of the examination settings, adapted for analogue technique, remained in the digital environment for a long time after the transformation. This was partly due to the lack of strategies for optimisation of the new imaging techniques and the

assessment of their excellence. Furthermore, several hundreds of different X-ray examinations are offered in a radiology department. To go through with optimisation of all these examination types, it is required to establish a prioritisation order based on the optimisation criteria set by expert organisations and legal regulations.

Medical diagnostics provides essential information in patient management and is an integrated part of patient care that cannot be separated from the outcome of other activities in this management. Therefore, the information obtained from medical imaging should be used in the most optimal way to ensure the greatest benefit to the patient in his or her medical care. However, the process of medical imaging in itself is complex, and must be optimised. The overall aim of the research described in this thesis was to develop methods and strategies for the optimisation of medical X-ray imaging. Attention was specifically directed to developing methods for the process of optimisation, including improved visual grading methodology. Some of the most critical steps in the optimisation process were identified and elucidated with the purpose of improving this process.

In the next two chapters, a more thorough overview of the subjects studied in this thesis is presented. The overview is also aimed to provide background information to the aims of the research described in this thesis. Therefore, the specific aims are presented after the overview chapters.

2 OPTIMISING THE USE OF IONISING RADIATION IN MEDICAL IMAGING

The International Commission on Radiological Protection (ICRP) is a central resource for knowledge and guidance in the field of radiation protection. Being an independent international organisation, ICRP is free to exclusively focus on issues that "advance for the public benefit the science of radiological protection"[9]. Recommendations and guidance regarding suitable approaches and good practice in the use of ionising radiation are distributed through the frequent compiling of emerging scientific research. The first recommendations of the ICRP regarding the use of radiation in medicine were published in 1928[10].

The constant need for quality improvement in medical diagnostics is driving the development of suitable quality evaluation methods. In the use of ionising radiation, where the radiation exposure presents a risk to both patients and staff, the choice of evaluation procedure is also driven by the compromise between image quality and radiation risk. This is motivated by ethical and legal demands intended to ensure that the use of ionising radiation is justified, i.e. that the benefit is greater than the potential harm. Birkelo et al.[2] used the term *effectiveness* to describe the quality measure studied. According to Fryback and Thornbury[11], *efficacy* became more frequent in the 1970s in describing clinical quality of a procedure. According to them, the efficacy of a process is closely related to the cost-effectiveness of the process, meaning that efficacy is a more relative description (the probability of benefit) than effectiveness which describes the process in more absolute terms (the performance).

According to the Oxford English Dictionary, *efficacy* is defined as the "Power or capacity to produce effects"[12], whereas *optimisation* is defined as "The action or process of making the best of something"[12]. From the radiation protection point of view, the ICRP uses the word *optimisation* to describe the process used to minimise radiation exposure. However, based on the strong connection to legal requirements on the use of ionising radiation in medicine, and the thereof extensive resources used for quality assurance in the medical use of radiation, *optimisation* has become the general description of a quality improvement process in which radiation exposure is a risk factor to be weighted with the benefit of the exposure to increase *efficacy*. This difference in the interpretation of an optimisation process, based on the two different points of view, can be a source of misunderstanding in a discussion on the design or goal of an optimisation project.

In general, the purpose of radiological protection is to minimise levels of detrimental exposures from ionising radiation. The statement from the ICRP of the principle that "all doses be kept as low as readily achievable" in publication 9[13] had the objective to express an overall recommendation of the optimisation of radiation protection. In later publications the principle is expressed "as low as reasonably achievable" (ALARA). However, the principle has been problematic for medical services to adopt as the use of radiation in medicine does not only have negative consequences for the patient, but is also used as a means of treatment or diagnostics. The ICRP has therefore been working on the clarification of the ALARA principle, for example by introducing diagnostic reference levels[14], dynamically adapted for specified imaging procedures, and suggesting the use of cost-benefit analysis as to facilitate optimisations[15]. The special application of radiation protection in medicine was addressed in ICRP 60[16] in 1991, and a later publication in 1996, ICRP 73[14], was aimed more specifically at medical users. In these publications the ICRP states the first two principles of radiation protection, i.e. the justification requirement to do more good than harm and that "all reasonable steps should be taken to adjust the protection so as to maximise the net benefit"[14]. The deliberate use of radiation in medicine is, however, addressed as a separate problem, where difficulties in making a "quantitative balance between loss of diagnostic information and reduction of dose to the patient"[14] are identified. However, the only help given by the ICRP is that the method of reducing the dose to a level where the image quality criterion is just fulfilled "is not the best method of optimising protection"[14] as it assumes a fixed limit at which image quality changes from acceptable to unacceptable. In ICRP 105 published in 2007[17], the ICRP continues the discussion on the deliberate exposure of patients, and states that the exposure "cannot be reduced indefinitely without prejudicing the intended outcome". The objective of the ICRP, to express an overall recommendation on the principles of radiation protection, seems to restrict its ability to provide more helpful recommendations on the use of radiation in medicine. Criticism of this constraint and suggestions for new approaches have recently been expressed[18-21], as discussed further in Section 2.3.

2.1 Regulation of optimisation

The recommendations of the ICRP are a fundamental source for the more regulated directives issued by legal authorities. In the current EU Directive 2013/59/Euratom[22], the Council of the European Union has described a system of radiation protection that member states must employ as legal requirements, in general terms. The regulations shall include "a system of radiation protection based on the principles of justification, optimisation and

dose limitation". The special perspective of radiation protection in medical exposure is clarified here as the directive emphasises that the optimisation of "medical exposure shall apply to the magnitude of individual doses and be consistent with the medical purpose of the exposure". This principle can be applied in terms of equivalent doses (the radiation type weighted organ dose) and effective dose (the tissue weighted sum of the mean organ equivalent doses), where appropriate. Member states shall ensure that exposures with a medical purpose are "kept as low as reasonably achievable consistent with obtaining the required medical information, taking into account economic and societal factors". According to this directive, the management of a medical exposure activity is obliged to ensure that every exposure of patients is performed according to the regulations, via national regulations in each member state.

The current EU Directive was implemented in Swedish law in 2018 through the Radiation Protection Act[23]. This act states that in medical exposure, each method used must be justified in general, and also that the specific use of radiation must be justified in each individual case. In any activity including human exposure, radiation protection must be optimised with the goal to reduce

1. the likelihood of exposure,
2. the number of individuals exposed and
3. the magnitude of the individual doses.

The Radiation Protection Act was followed by an ordinance from The Swedish Government[24] and several regulations from the Swedish Radiation Safety Authority. According to the regulation concerning medical exposure[25], the goal of exposure optimisation is to adapt the extent of a diagnostic procedure and the radiation dose to the exposed individual so that the required diagnostic information is obtained with a radiation dose that is as low as reasonably achievable. In the case of the examination or treatment of a pregnant patient, the radiation dose to the foetus must be considered in the planning and execution of the examination, to ensure that the radiation dose to the foetus is as low as reasonably achievable. It should, however, be noted, that in this context the difference in the goals of optimisation in the protection of a patient who is examined voluntarily and the foetus, which has no choice in the matter, is very small. Obtaining the required diagnostic information can be interpreted here as a fixed level, where the aim of optimisation is to reach a predefined level of information with a radiation dose that is as low as reasonably achievable. The declaration in ICRP 105, that "an optimisation of radiation protection in medical exposure does not necessarily imply a reduction of the

radiation dose to the patient"[17], is not referred to in the regulation concerning medical exposure itself[25], but only in the guidance to the regulation[26]; the achievement of the examination or treatment result that is intended is of utmost importance in medical care.

2.2 Justification of the use of radiation in medical imaging

According to the ICRP, the proper use of radiation in medicine is justified on a general level to do more good than harm to society[17]. The appropriate use of the deliberate exposure of patients that cannot be reduced indefinitely is thus a fundamental starting point in the process of optimising examination routines. The objective of using radiation as a tool to obtain information in diagnostic radiology should therefore focus on making as much use as possible of the radiation that is used, whereas radiation that is not needed to obtain the requested information should be reduced to a minimum. The need for optimisation is expressed more specifically in the second and third levels of justification, i.e. the justification of a procedure with a specified objective to improve the diagnosis or treatment for a group of patients with certain problems and, the justification of an individual patient to fit into this group of patients.

It is necessary to assess in each individual case whether the examination of a patient may do more good than harm. A frequently performed examination with limited medical impact on most of the examined individuals, e.g. mammography screening, is only justified if the examination is performed with a limited radiation dose to the individual. However, a lifesaving vascular treatment or preparation for cancer therapy is justified at a higher dose[17]. Therefore, a general guide to reasonable exposure levels for the collection of examination procedures is an important basis in the optimisation process at a radiology department. A crucial task in the justification process is thus to decide if the individual patient fits into a standardised request group, in the referral process, associated with a specific examination routine.

The problematic compromise between risk and benefit in medical care, specifically in diagnostic radiology, has been thoroughly described by the European Society of Radiology in Brochure IV (*Risk Management in Radiology in Europe*, 2004)[27], where risk factors affecting the outcome of visiting a radiology department are reviewed. The report is mainly focused on the direct risks associated with poor quality in a radiological examination, where the risk of false positive and false negative reading is especially highlighted. Radiology departments are recommended to perform regular audit

programmes to review the quality of the service provided. The radiation exposure is mainly identified as a risk factor when performing inadequate examinations, which have little value in patient management and therefore should be avoided for justification reasons.

2.3 The objective of optimisation

With a well-grounded justification of a medical examination at hand, the objective of an optimisation process is to maximise the net benefit to each patient with a practical application ranging from simple common sense to advanced research studies. To ensure both fast throughput and high validity in the optimisation process a compromise between the time dedicated for optimisation and the validity in the final result is required. A method of prioritising optimisation tasks in a radiology department, focused on maximising the reduction of the radiation risk, has been suggested by Månsson et al.[28]. This study has contributed to the discussion on the appropriateness of focusing only on dose reduction in the prioritisation of optimisation tasks and was part of the motivation behind the work described in this thesis.

The pedagogical difficulties in conveying the traditionally used radiation protection nomenclature to those in clinical practice have also been discussed during recent years. Malone and Zölzer[29] suggested making use of a more pragmatic ethical basis, based on the general principles of ethics in medicine (i.e. the Hippocratic Oath). They claim that "for the most part, scholarship in medical ethics does not attend to the problems in radiation protection". Rather, such problems are dealt with through the strict regulations of radiation protection in a separate system with "exceptional independence, which allowed it unique access to management and resources". However, this independence has led to a poor recognition in the medical world where the assumption is often that the problems associated with using radiation have been solved, and as long as examinations are performed within the diagnostic reference levels, the level of exposure to the patient is safe. Should practitioners discuss the ethical problems of using radiation in the same way as other ethical dilemmas, the authors' conclusion is that it would be "advantageous to frame ethical dilemmas in radiology in terms of these values, rather than relying solely on the established principles of justification, optimisation and dose limitations."

Malone and Zölzer refer to four basic principles for ethical decision making, first suggested by Beauchamp and Childress[30]:

- Respect for autonomy (of the individual)

- Non-maleficence (do no harm)
- Beneficence (do good)
- Justice (be fair)

Malone and Zölzer add two further principles that are more specific for ethical decision-making in the radiological context:

- Prudence (keep in mind possible long-term risks of actions)
- Honesty (share knowledge with those concerned truthfully).

Regarding the radiation protection principle of optimisation, the transition to the ethical compromise between non-maleficence and beneficence is easily understood. From the utilitarian's point of view the best action is the one that produces the best well-being. The fundamental issue of the ICRP principle of justification is that "no practice involving exposure to radiation should be adopted unless it produces sufficient benefit"[14] or, in other words, "Any decision that alters the radiation exposure situation should do more good than harm"[31]. However, the more well-known ALARA principle, may lead to the interpretation that the entire focus of optimisation is that the exposure should be reduced to a level that is "as low as reasonably achievable".

The management of radiation protection in diagnostic radiology is thoroughly discussed by Moores in a series of publications[18-21]. In the first publication, a cost-risk-benefit analysis, based on calculations using published values of prevalence, sensitivity and specificity, is carried out, showing that the total number of false positive and false negative outcomes in X-ray examinations is surprisingly high. Compared to the number of patients that are likely to suffer from the stochastic induction of cancer caused by the examinations, the number of patients that will suffer from the incorrect outcome of the examinations is probably a factor of 1000 higher. Based on this relation, Moores notes that from examinations performed in economic Level 1 countries worldwide (roughly 2.4×10^9 per year) that result in an incorrect outcome (approximated to 1.2×10^8), the number of induced cancers can be estimated to more than 10 000 per year. According to the three levels of justification given by the ICRP, an examination that results in an incorrect outcome is not justified, and can therefore not be deemed to fulfil the basic principles of radiation protection. Moores states that, although optimisation of radiation protection is mostly focused on the reduction of radiation exposure, true optimisation in diagnostic radiology "cannot be assessed or verified without knowledge of diagnostic performance".

Moores continues his analysis with an ethical review of radiation protection optimisation in diagnostic radiology[19]. He claims that, from the knowledge of incorrect outcomes from diagnostic examinations, it is unethical not to include the diagnostic risk in the optimisation process in order to improve the diagnostic outcome. He invites the ICRP to broaden their view on recommendations for radiation protection in diagnostic radiology, and medical societies are likewise invited to develop methods for continuous assessment of the diagnostic outcome of examinations and measurements of improved outcomes by the introduction of new methods.

In the third publication in his series[20], Moores analyses the nature of decision-making in the context of radiation protection in diagnostic radiology. His finding in this study is that decisions to deliberately expose patients in diagnostic radiology should be taken based on an as well-founded balance between risk factors as possible. According to Moores, the risk resulting from radiation is only one factor of many associated with patient care, and the separate handling of this risk, e.g. by the introduction of diagnostic reference levels (DRLs), tends to separate the subject from others. The use of DRLs thus tends to represent a public health initiative rather than the ethical basis for patient protection, as stipulated in the Hippocratic Oath, i.e. to do more good than harm.

Moores continues the discussion on diagnostic risk in the fourth publication in the series[21] by proposing the use of “cost-risk-benefit” in the evaluation of optimised use of radiation in diagnostic radiology. The process of cost-benefit was proposed in the 1972 ICRP Publication 22[15] for establishing the optimum levels of radiation protection. Although the 1983 ICRP Publication 37[32] gave examples to how a medical optimisation process can lead to a decision of increased exposure with the gain of increased benefit, the process was not recommended for use in medicine. The argument was that reduction of exposure is often difficult as it can have negative effect on the intended result and that the full effect in a cost-benefit process is complicated to foresee. The suggestion from Moores is now that, by adding all known effects from alterations of medical use of radiation, the net “cost-risk-benefit” can be summed as a measure of the alteration effect. The transformation of methodology from e.g. conventional X-ray to computed tomography (CT), that has been ongoing since the 1970s, is an example for which this method would be suited. Another example, proposed by Moores, is the employment of referral guidelines where the effect of reduced non-justified examinations can be compared with the cost for the introduction of the system. By the help of modern information systems, more reliable calculations of cost-risk-benefit

should be possible today, compared to the 1980s, when the uncertainties were judged too high.

Although arguments for delimiting the area for an optimisation process are often motivated, the overall benefit of a diagnostic procedure also must be evaluated. A good example of a case in which a broader perspective has been used in the optimisation process, is the system of quality assurance applied in mammography screening programmes used in many countries[19]. Another recently published example, is a report on the justification of diagnostic X-ray use during surgery, evaluated by the Swedish Agency for Health Technology Assessment and Assessment of Social Services[33], where the routine use of intraoperative cholangiography (IOC) in cholecystectomy (surgical removal of the gallbladder) was compared to selective use, decided during surgery. A meta-analysis of published reports showed that surgically inflicted injuries to the bile ducts could be reduced by 30% when IOC was used routinely, compared to only selectively during surgery. The cost of a quality-adjusted life year was approximated to 30 000 EUR, and the reduced incidence of severe inflicted injuries was approximated to a factor of 10 higher than the incidence of cancer caused by radiation from the X-ray examination. This provides an example of a study in which the resulting health of patients are measured after a specific procedure, and where the radiation risk is one of the input factors determining the outcome. A further optimisation process with the intention of increasing the benefit of the use of X-rays in the procedure may have increased the positive effect of the outcome. However, as the strategy of the study was to collect data from several reports for meta-analysis, this combination was not possible in this case. Therefore, it is suggested that a follow-up study would be to perform an optimisation study with the intention of increasing the net benefit to the patients undergoing the procedure in question.

In summary, the deliberate use of ionising radiation in medicine requires a special perspective on the optimisation of radiation protection. During a large technological improvement, such as the transition from analogue to digital imaging in projectional radiography, optimisation strategies must be adapted for this new environment. Therefore, the development of improved strategies for the performance and the prioritisation order of optimisation projects was an important part of the work described in this thesis.

3 VISUAL GRADING IN OPTIMISATION OF X-RAY IMAGING

If the purpose of a diagnostic procedure is to provide useful information in the investigation of a medical problem, the method used to evaluate how well this procedure performs should be the method with the highest validity for the overall purpose of the medical investigation. However, as a diagnostic procedure is only one link in a chain of events aimed at achieving the overall purpose, a measurement of the outcome for a group of patients, although very important, would only depend to a small degree on the quality of the diagnostic procedure. Nevertheless, the value of the diagnostic procedure, in itself as well as in the chain, must be evaluated with methods that are relevant to the diagnostic task.

The process of defining measurable variables thought to describe a phenomenon is called operationalisation. In this process the reliability and the validity of the measurable variables are evaluated to identify the variable that best describes the phenomenon. The reliability describes the precision of the measurement, a high reliability requiring small stochastic errors. The validity indicates of how well the variables describe the phenomenon, a high validity requiring small systematic errors[34].

Physical quantities, such as the modulation transfer function, signal-to-noise ratio, noise power spectrum or detective quantum efficiency, can be measured with the intention of evaluating the efficacy in transforming information related to a biological condition in a patient (e.g. electronic density in different tissues and thus the attenuation of X-ray radiation) to a visual display (e.g. X-ray image) with high reliability. Thanks to their high reliability, physical measurements are very suitable for quality assurance purpose and can be valuable in direct comparisons between imaging techniques. Furthermore, well-performed measurements of physical quantities can be of great help in understanding the characteristics of a system, and can often be important in the refined adjustment of parameters in the initial optimisation process[35-38]. However, these types of quantities are often limited in that they describe only a particular part of the imaging chain, and it has been shown to be difficult to achieve high validity in the description of the ability to detect diseases with this type of methods[39]. For the purpose of identifying the imaging system that provides the best clinical value, methods with high validity for the overall result of a complex procedure are often more appropriate.

Methods of evaluating image quality intended to cover the complete imaging chain in radiology, from exposure to interpretation, are generally described as image perception studies. The imaging objects used in these studies can vary from standardised physical phantoms (psychophysical studies), to human-like (anthropomorphic) phantoms, to clinical images of healthy individuals or patient volunteers. With the right conditions, these methods can be assumed to have high validity for their purpose (increasing with similarity between the imaging objects in the study and the intended patient group). Correspondingly, the reliability of the measurements is assumed to be relatively low and a large number of cases are required to achieve high power[34]. The reliability will decrease as the variation in the imaging objects increases or as the difference in the interpretation of the images between observers (human or machine) increases.

3.1 Image perception studies

Image perception studies are divided into two main kinds.

1. Observer performance studies, where the ability of the system – including the observer – to detect abnormality is measured
2. Visual grading studies, where the ability of the system to visualise defined anatomical structures is graded by an observer

The difference between the two methods is that an observer performance study requires knowledge of the true state of the studied objects as it measures the performance of the system, whereas in a visual grading study, the rating from the observer is adopted as the outcome of the system.

3.1.1 observer performance studies

The fundamental task in medical diagnostics is to distinguish healthy individuals from diseased ones (or normal from abnormal). The performance of a medical imaging system is therefore preferably evaluated by measuring the accuracy with which the system can fulfil this task. Motivated by the shortcomings identified by Birkelo et al.[2], ROC has become the leading method for observer performance studies in medical imaging where the ability of an observer or a system to distinguish normal from abnormal is measured, including characterisation of the confidence with which the decision is made[40].

The proportion of correct answers is an intuitively appropriate measure to use to characterise the performance of a system. However, this measure would depend strongly on the prevalence of the disease (abnormal). For example, the most effective way of obtaining a high score in a study with a low prevalence of disease (say one out of hundred, i.e. 1%) would be to state all cases as "normal", resulting in a correct score of 99%. Considering the sensitivity (relative number of correctly identified abnormal cases) and the specificity (relative number of correctly identified negative cases) separately, will provide a prevalence-independent measure of the accuracy of the system. However, determining the sensitivity and specificity with no rating of the degree of confidence by the observer, will not only be limited by the lack of rating information in the result, but the result will also depend on the observer's choice of confidence level. If, for example, the reported advantage of one of the systems being compared is high sensitivity, while the other has high specificity, the measured difference between the two systems can arise from "threshold effects" depending on the observer's prioritisation between identifying the abnormal and rejecting the normal[41]. The observer's confidence rating can be measured by collecting the answers on an ordinal decision scale (e.g. from certainly normal to certainly abnormal). By pairwise registration of the ratings collected on each threshold level for the compared groups, operating points of the accumulated sensitivity (in ROC normally denoted the true positive fraction (TPF)) and specificity (in ROC normally denoted 1-false positive fraction (FPF)) can be created. The operating points can be connected to form the so-called ROC curve, as illustrated in Figure 1. The more distant the ROC curve is from the diagonal line, the better the system distinguishes abnormal from normal. The fundamental figure of merit (FOM) in ROC analysis is the area under the ROC curve (AUC). The AUC is also a transformation of the ratings collected on an ordinal scale to a rank invariant FOM on an interval scale, more appropriate for the statistical uncertainty testing of the result[42].

The ROC curve can be created either by a trapezoidal rule between the operating points or by parametric estimation of the "most probable" curve from (0,0) to (1,1) formed by the measured operating points. As the thresholds are collected on an ordinal axis, the distances between the rating steps are unknown, and the thresholds can take any position on the decision axis as long as the order is maintained. An interval decision axis can therefore be obtained by any appropriate parametric curve fitting. For simplicity normal distributions of the two underlying probability distributions are often assumed[43] (binormal distribution). In binormal curve fitting, the sample points are fitted to two normal distribution curves on an interval axis by adjusting the position of the sampled thresholds. The two normal distribution curves can then be used to

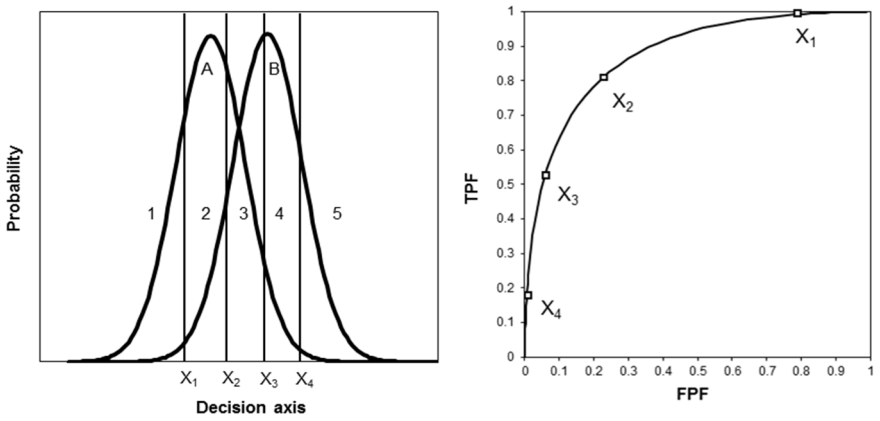


Figure 1. Left: Probability distributions (A=normal and B=abnormal) of a detection task showing 4 levels of decision thresholds, X_1-X_4. Values of $X < X_1$ correspond to the first rating category (1), $X_1 \leq X < X_2$ to the second (2), etc. Right: The resulting ROC curve, giving the true positive fraction (TPF) as a function of the false positive fraction (FPF). The four operating points corresponding to the four decision thresholds (left) are given by the four boxes (right). The smooth binormal curve (right) is created by adjustment of the position of X_1-X_4 on the ordinal decision axis (left) so that two normal distributed curves can be created, based on the operating points, on an interval axis.

create an infinite number of operating points forming a "binormal curve" in the ROC diagram. It is, however, important to note that transformation from the ordinal to the interval axis is performed while maintaining the positions of the operating points in the ROC diagram. In practice, the binormal curve is formed through conversion of the obtained operating points from the study into plots on "normal-deviate" axes, where the operating points can form a straight line by maximum-likelihood estimation[44]. Experience has shown that the method is unreliable when the datasets are too small, or the decision scale has too few steps. Although alternative distributions have been evaluated[45], the binormal distribution is still the most commonly used curve fitting method[46].

Early statistical methods used to test the significance of the result of an ROC study were in practice limited to the estimation of either case variation or reader variation[46], but in 1992, Dorfman, Berbaum and Metz (DBM) presented a method of estimating the variation in the AUC by the generation of pseudo-value AUCs from jackknifing of the original reader data on both readers and cases[42]. In jackknifing, the uncertainty in the sampled data can be simulated by removing one sample at a time and recalculate a pseudo-FOM. The number of pseudo-FOMs that can be obtained is equal to the number of samples in the original data set. After transforming resampled ratings on an ordinal scale to pseudo-AUC values on an interval scale, classical analysis of

variance could be used to calculate the uncertainty in the original AUC. The DBM multiple-reader multiple-case (MRMC) approach is now the benchmark in statistical testing of ROC studies[3]. The use of resampling methods has been further developed by introducing the more general bootstrap method to the DBM MRMC approach[47]. (The bootstrap method is more thoroughly described in Section 5.3.)

3.1.2 Visual grading studies

Performing an ROC study may be cumbersome and time consuming, due to the request for a known truth. As an alternative, methods for evaluation of image quality by visual grading of defined structures in clinical images have been developed. Visual grading is founded on two traditions for evaluating the diagnostic value of clinical images. The first is the clinical routine of reviewing the visibility of defined structures that should be visible in a diagnostically valid image. This is designed to ensure that the image presented to the observer is of adequate quality to visualise the medical problems in question. The second is the traditional method (also used by Birkelo et al.[2]) to evaluate different imaging conditions by simply asking a group of observers which image is preferable. The risk of misinterpretation between the observers can be minimised by developing questionnaires with the aim to focus on specific details[43]. The method has a low performance threshold as clinical images can be used, and the observers are relatively easily motivated as the method makes use of the skills of an experienced radiologist in the sense that the clinical value of the studied structures has an impact on the validity of the study. Therefore, the validity of a visual grading study is dependent on a strong connection between the visibility of the evaluated structures in the image and the demand on image quality on the clinical image. The reliability of a visual grading study also depends on the observers' consensus regarding their interpretation of the visual grading task, i.e. what should be evaluated and how. The validity and reliability of such a study should benefit from the use of image quality criteria in the form of standardised clinical landmarks and levels of the reproduction, such as those described by the European guidelines on quality criteria[48-50].

The rating scale in a visual grading study can be either dichotomous with the evaluation alternatives fulfilled or not fulfilled, leading to an image criteria score (ICS, the proportion of fulfilled image criteria in the images evaluated), or a more complex scale where the observers rate their assessment of the visibility of the evaluated structures, usually referred to as visual grading analysis (VGA). In ICS, the statistical handling of the result is straightforward as the central limit theorem states that for large samples the sample mean will be normally distributed[51]. However, a disadvantage of ICS is the limited rating in the observer's judgement (fulfilled/not fulfilled). In VGA, the rating

scale is increased from two to three or more steps. This allows the observers to rate their opinion. However, the ordinal structure of the scale prevents the assumption of normal distribution of the mean, and the statistical handling of a VGA study is therefore more complex. Furthermore, observers may interpret the scale steps differently, which complicates the handling of data in studies with multiple observers since ratings from different observers cannot be directly compared.

3.2 Challenges in visual grading

Studies on the correlation between ROC and visual grading have shown both low[52-54] and high correlation[6-8] between the methods. However, as these studies compared the methods in different diagnostic situations, the divergence in the results does not necessarily indicate that the results are contradicting. Arguments have been voiced against the visual grading method for having low scientific value and amount to a beauty contest[55] while other researchers are of the opinion that this argument is a simplification and an underestimation of the ability of radiologists to recognise the image quality required for the reproduction of anatomy in order to make a diagnosis[43]. Similar contradictory arguments can be found in the comparison between observer evaluation and physical image quality measurements, where both non-correlation[56] and correlations have been reported[57, 58]. The finding of no correlation between two measurement methods can be explained by the different abilities of the methods to measure specific parameters[52], or differences in the study design where different parameters have been measured with the methods compared[54]. In a review from 2008, Tapiovaara concluded that: “the various image quality evaluation tasks in an X-ray department are best done by different methods” and that, “which of the image quality evaluation methods should be used is clearly dependent on the purpose of the image quality evaluation task and the resources that can be used to accomplish it”[59].

The problems associated with comparing different methods for image quality evaluation highlights the importance of operationalisation in the planning of an optimisation study. As previously described, this process is strongly connected to the reliability and validity of the results from a study[34]. The choice of physical measurements due to their high reliability has been criticised because of the risk of low validity[39]. Kundel emphasises that: “Diagnostic information can only be defined in the light of real clinical problems because it is not an absolute”, and that: “the highest quality image is the one that enables the observer to most accurately report diagnostically relevant structures and features”[60]. His arguments strongly underline the need for a thorough operationalisation process in the planning of an image quality evaluation study,

and that in a highly complex environment, such as that in diagnostic imaging, there is no gold standard that will always give the overall answer in a comparison between imaging methods. The quality of the operationalisation process in the planning stage determines the value of the final result of the study in each optimisation project.

The argument against visual grading for image quality evaluation based on the method being too subjective has been addressed in comprehensive studies supported by the European Commission where guidelines on quality criteria for diagnostic radiographic images were presented[48-50]. The development of strict image quality criteria was basically intended for clinical audits of radiological departments but has also been shown to be suitable for image quality optimisation purposes[39, 61]. Based on clear definition of the criteria to be evaluated, the objectives of the guidelines are: "to provide the basis for accurate radiological interpretation of the image". Hopefully, a stricter use of image quality criteria will reduce the subjective influence of the observers' opinions in the rating of the evaluated images.

Another argument against the use of visual grading in the evaluation of medical images is the lack of well-founded statistical methods of dealing with the data that result from a visual grading experiment. Ratings collected by observers in a visual grading study are typically collected on an ordinal scale where the order of the rating steps is defined, but the scaling distance between the steps is not (e.g. low, medium, high). It is thus unwise to use methods where a specific distribution is presumed (i.e. parametric). The basic non-parametric method for statistical analysis of two compared groups is the Mann-Whitney U test where the ratings collected for the groups are ranked on a scale and the rank order sum for each group is calculated to provide a measure of the difference between the groups. The Kruskal-Wallis test is an extension of the Mann-Whitney U test that is valid for the comparison of more than two compared groups. If the samples in the compared groups are dependent (matched/paired samples), analysis should preferably be performed using the Wilcoxon signed-rank test.[62] These "basic" methods are standard statistical tools, available in statistical softwares and commonly used by general statisticians. However, these methods were developed in the middle of the last century and were intended for calculation by hand or simple calculators. Furthermore, regarding analysis of data from image perception studies, they suffer from limitations in the handling of the variation between multiple-readers (although jackknifing has been suggested as a method to perform this analysis[42]). Another problem is the finite rating scale that is often used in image perception studies and that will lead to a relatively high number of ratings being collected on the same level (ties). Ties are problematic in the

Mann-Whitney U test. The usefulnes of these tests in evaluating visual grading data is therefore limited. However, efforts have been made to overcome this obstacle by the developments of new methods especially dedicated for use in the statistical analysis of visual grading studies[43, 63]. One of these methods is visual grading characteristics (VGC) analysis[43], discussed in the next section.

3.3 Visual grading characteristics

Inspired by the approach in ROC analysis of transforming the ratings collected on an ordinal scale to an AUC on an interval scale without confounding the data, Båth and Månsson presented a method of producing a corresponding AUC from visual grading data, i.e. VGC[43]. In VGC analysis, rating data (e.g. image quality ratings) for two conditions (a reference condition and a test condition) are compared by producing a VGC curve, as shown in Figure 2, similar to the way in which rating data for normal and abnormal cases in ROC analysis are used to create an ROC curve. In analogy with the ROC methodology, the operating points in VGC are defined by assigning the threshold levels used on the rating scale to an ICS on each threshold level. The ICS in VGC corresponds to the dichotomous ICS in the primary image criteria scoring method, although it is extended to multiple ICS levels by sampling the accumulated ratings collected on each threshold level. Hence, the VGC curve is a plot of the proportion of ratings above a certain threshold for the test condition against the same proportion for the reference condition, as the threshold is changed. The VGC curve is thus independent of the labeling of the scale steps, and the ordinal nature of the data is therefore not confounded.

The separation between the two rating distributions can then be characterised by the area under the VGC curve (AUC_{VGC}) ($0 \leq AUC_{VGC} \leq 1$). A low value of AUC_{VGC} (<0.5) indicates that the reference condition is in general rated higher, whereas a high AUC_{VGC} (>0.5) indicates that the test condition is rated higher. An AUC_{VGC} of 0.5 indicates that the image quality for the two conditions is rated the same, on average. If multiple observers are used in the study, the total AUC_{VGC} is calculated by averaging the individual AUC_{VGC} from the participating observers. Since each observer's interpretation of the rating scale steps does not affect the determined individual AUC_{VGC}, the total AUC_{VGC} is unaffected by the observers' different interpretations of the scale steps.

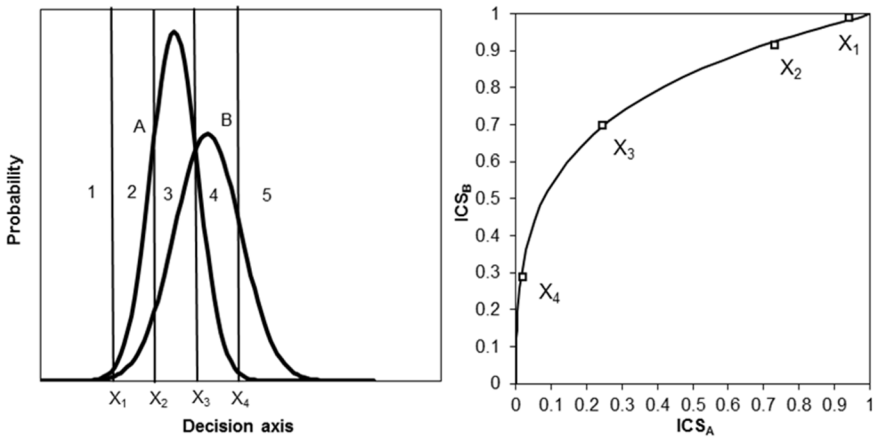


Figure 2. The visual grading characteristics (VGC) probability distributions for imaging condition A and B (left). Image criteria scores (ICS) are pairwise registered at the operating points X_1-X_4 and plotted in a diagram to form a VGC curve (right).

Båth and Månsson suggested the use of existing ROC methods for the statistical analysis of VGC data, i.e. determining the statistical uncertainty of the obtained FOM, AUC_{VGC}. However, important differences in the approach to the analysis of the collected data from the two types of studies make the suggested solution questionable. First, ROC studies are almost exclusively based on independent normal and abnormal data sets and, to the best of the author's knowledge, independence between the two data sets is a basic assumption in the statistical analysis used in contemporary ROC methodology. However, dependency between the two sets of rating data for the two conditions compared is common in VGC studies, e.g. data resulting from one group of patients examined with two types of equipment. Second, there is a fundamental difference between the properties of an ROC study and a VGC study, when evaluating two imaging conditions, in that in ROC the statistical analysis is focused on the uncertainty in the difference between the two ROC curves originating from the two conditions[64], whereas in VGC the analysis is focused on the uncertainty in the single VGC curve originating from the two conditions.

In summary, VGC was developed as a result of the need to transform ordinal data to an AUC_{VGC} on an interval scale. This transformation can be accomplished without using any parametric assumptions or confounding of data due to the variation in observer interpretation of the ordinal rating scale. However, for VGC to become a useful method in the analysis of visual grading data, it is necessary to develop statistical methods with the ability to estimate the uncertainty in a calculated AUC_{VGC} using no parametric assumptions.

4 AIMS

The overall aim of the work presented in this thesis was to develop methods and strategies for the optimisation process prescribed for medical X-ray imaging. Specifically, methods of conducting and prioritising the optimisations of examinations, including improved visual grading methods, were investigated. The four specific aims of this work were:

1. to analyse and describe the conditions for the optimisation of a given projectional X-ray examination in a digital environment,

2. to develop an overall strategy for the optimisation work in a radiology department,

3. to develop and implement a suitable method for statistical analysis of VGC data, and

4. to evaluate the characteristics of the new statistical method by comparison with ROC statistical methodology and by simulations.

Aims 1-3 are directly connected to the studies presented in Papers I-III, respectively, whereas aim 4 is connected to the studies presented in Papers IV and V.

5 FULFILMENT OF THESIS AIMS

In this chapter, the papers included in this thesis are summarised in relation to the aims of this work. In connection with the presentation of Paper III, a thorough description of the resampling methods used in the developed method for statistical analysis of VGC data is given.

5.1 Paper I

A Conceptual Optimisation Strategy for Radiography in a Digital Environment

In Paper I, the effect of the technical transformation from analogue to digital radiography on the optimisation of projectional X-ray imaging was analysed. The paper focuses on describing an optimisation strategy that takes full advantage of the fundamental differences between digital systems and screen/film systems.

During the final decade of the previous century, projectional X-ray imaging in diagnostic radiology went through radical technical developments which led to the change from analogue to digital image registration, communication, visualisation and archiving. This led to a need for the optimisation of examination parameters adapted to the new technology. In parallel with this technical revolution, demands on the management of radiation for medical use were expressed, first by the recommendations issued by the ICRP in ICRP 73[14] and then by the stricter legislation in the European Medical Exposure Directive[65]. Furthermore, research at the time showed the limited validity of basing optimisation on traditional signal-to-noise ratio measurements[52, 56, 66-74], which previously had been common. This combination of a new technical era, new recommendations and directives, and a paradigm shift in the view on image quality measurements, provided the motivation for the development of a conceptual optimisation strategy in a digital environment, presented in Paper I. The proposed strategy was summarised in three main parts:

a) Include the anatomical background when evaluating image quality.
b) Perform all comparisons at a constant effective dose.
c) Make full advantage of the digital system for separation of the image collection step from the image display step in the imaging chain.

5.1.1 Include the anatomical background when evaluating image quality

The traditional Rose model[75], describing the inverse relationship between the size of an object and the contrast needed for its detection in images with white noise background, has been constituting the foundation for many optimisation studies[76-78]. However, in the early 2000s, many studies showed the limitation of basing optimisation of projectional X-ray imaging on the Rose model[52, 56, 66-74]. Burgess demonstrated that mammographic images containing anatomical background did not follow the Rose model and that larger objects in fact required higher contrast for their detection[79]. Other studies showed that, when performing optimisation studies on clinical images, the result of a detection task is often more dependent on the anatomical background than on the quantum noise in the image[67, 71-74, 80, 81]. From these insights, it was concluded that, to ensure high validity, an optimisation strategy should contain the recommendation that the appropriate anatomical background should be included in all stages of the optimisation process.

The method recommended as suitable, in Paper I, for optimisation of clinical imaging procedures was visual grading. ROC studies are often specific for the type of signal/background combination studied and a generalised measure of clinical image quality is difficult to obtain. Therefore, four arguments to use visual grading were listed: 1) the validity is assumed to be high with the use of clinically relevant criteria, 2) agreements have been shown with both ROC-based methods and with calculations of physical image quality, 3) visual grading studies are relatively easy to conduct, and 4) a visual grading study can be performed with moderate time consumption.

5.1.2 Perform all comparisons at a constant effective dose

In the analogue imaging environment, the reference exposure level for an examination was the reference giving the optimum grey level for the combination of intensifying screen and photographic film that resulted from the examination exposure, e.g. the air kerma (K_{air}) in the imaging plane (kerma, kinetic energy released per unit mass). Optimisation projects succeeded in both increasing image quality and decreasing exposure[82]. However, restricted by the limitations in the analogue technique, other optimisation projects, with the aim of either increasing the image quality or decreasing the patient exposure with maintained grey level of the film, reported that the aim was reached at the expense of the counteracting parameter (image quality vs patient exposure)[83, 84]. In the digital environment, the exposure of the imaging detector can be set within a large dynamic range and is no longer a technical limitation in the

optimisation process. This increases the freedom to vary technical parameters which may be beneficial in the design of the optimisation process.

A more relevant parameter for risk estimation is the kerma-area product (KAP) that is a measure of the total amount of radiation exposed to the patient. However, neither K_{air} in the imaging plane or KAP are proportional to risk estimating quantities such as effective dose when beam quality is altered[85]. Therefore, to preserve the freedom to alter parameters in the optimisation, the choice of risk reference parameter should be the parameter with the highest validity for the risk estimation. Thus, in Paper I, the effective dose, or an analogue relevant measure of radiation risk, was recommended to be kept constant during beam quality optimisation. By keeping the relevant risk parameter on a reference level during the optimisation of the image collection, the necessary dose level for the examination can instead be determined in a later stage when the image collection and the image processing are optimised.

(Note that the presentation of dosimetric quantities in Paper I is unclear concerning the use of K_{air}. K_{air} is used for denotation of both air kerma in the imaging plane and incident air kerma to the patient.)

5.1.3 Make full advantage of the digital system

In the process of optimising the exposure settings for a specific type of examination, any change in the settings will result in a change in the dynamic distribution of the signal detected in the imaging detector. When analogue film technique was used, this change in the detected signal (grey level) led to a corresponding change in the image signal displayed or a change in contrast in the resulting X-ray film. This will not be the case in the digital environment. Any change in signal distribution in the detection stage can be compensated for in the display stage, either by simple windowing, or by more advanced software-driven adjustments of the dynamic signal level. The visualisation of object contrast is therefore only limited by the signal-to-noise ratio in the region of interest.

Since the image display stage in theory is separated from the image collection stage for a digital radiographic system, it was in Paper I argued that an optimisation task can with some validity be treated as a procedure with three independent steps. These steps can then be optimised one at a time, with the suggested order:

1) Determine the optimal setting of adjustable technique factors in the image collection stage (tube voltage, filtration, grid,

etc.) while keeping the effective dose constant. (Maximise information/risk ratio; image collection)
2) Determine the optimal setting of adjustable image presentation parameters (edge enhancement, contrast amplification, etc.). (Maximise information/risk ratio; image display).
3) Determine the optimal amount of radiation to use. (Optimise information/risk ratio).

It could be argued that complete independence between these three steps cannot be guaranteed. As an example, we can consider the situation where the optimum beam quality for a specific examination is to be determined. Any alteration in the energy distribution of the incident X-ray beam will lead to a change in the detected signal distribution, due to variation in the object contrast. This contrast variation must be compensated for, either by pre-setting of the windowing or by free windowing by the observer. Once the optimal technical parameters in the collection stage have been determined, the optimal image presentation parameters can be determined in the next optimisation step. These two steps will lead to a maximised information/risk ratio, enabling the final step to be carried out, i.e. the determination of the absolute exposure level for an optimised information/risk ratio. Furthermore, it can be argued that the probability of reversed effects, i.e. that previous steps must be re-optimised, would be reduced by performing the optimisation procedure in the suggested order, and that the reversed effects would be minimised by using the initial settings as close to the optimised settings as possible. Therefore, the parameters that will be optimised in a later stage (image presentation and exposure level) should be pre-optimised based on the knowledge at hand, so that each setting is evaluated as fairly as possible.

The strategy proposed in Paper I was applied in a separate study with the purpose of optimisation of neonatal chest imaging to find the optimum tube voltage for the examination in computed radiography[86]. The study was designed to take full advantage of the benefits of digital imaging, for example by comparing the tube voltages at constant effective dose. A phantom study using a living rabbit under anaesthesia was first conducted. Images were collected at tube voltages ranging from 40 to 90 kVp, where the change in collection dynamic range was compensated for by the display windowing. As this study was performed during the semi-digital era when the images were displayed on printed film, signal levels in defined anatomical regions, measured on a monitor, were set to a standardised level before printing. The reproduction of four structures (central vessels, peripheral vessels, the carina and the thoracic vertebrae) was rated by 10 radiologists. The reproduction of

the central and peripheral vessels was found to be relatively independent of tube voltage. However, the carina was better reproduced at higher tube voltages whereas the thoracic vertebrae were better reproduced at lower tube voltages. Based on the greater importance of the reproduction of the carina it was decided that 90 kVp was the optimal tube voltage for neonatal chest imaging. To validate the results of the phantom study, a follow-up study was conducted in which chest images of neonates collected at the tube voltage regularly used at Sahlgrenska University Hospital (70 kVp) were compared with images collected at 90 kVp. The follow-up study confirmed the results of the phantom study, namely that the reproduction of the carina was better at 90 than at 70 kVp.

The application of the new optimisation strategy by practical application to neonatal chest imaging showed that the strategy is effective in the performance of an optimisation project in a completely digital environment. However, an overall strategy will be required to determine the order in which different types of examinations at radiological departments should be optimised. This was the motivation for aim II of the work presented in this thesis. Furthermore, experience from the investigation of the optimisation strategy described above showed that the use of visual grading in optimisation projects requires improved methods for reliable statistical analysis of the examination conditions being evaluated. In combination with the suggestion in Paper I, to primarily use visual grading in optimisation of clinical X-ray imaging, this experience was the inspiration for aims III and IV described in this thesis.

5.2 Paper II

A Practical Approach to Prioritise Among Optimisation Tasks in X-ray Imaging: Introducing the 4-bit Concept

The legal requirement to optimise all medical procedures employing ionising radiation means that the hospitals must not only develop routines for optimising radiological examinations, but also determine the order in which these examinations should be optimised. Bearing in mind, the hundreds of different kinds of X-ray examinations performed at radiology departments, and the limited resources available, it will be difficult to prioritise their optimisation order. Therefore, the study presented in Paper II focused on developing a method that could be used to determine the order in which radiological examinations should be optimised.

The European Commission prescribes the content of an optimisation process in relation to medical exposure in the Medical Exposure Directive from

1997[65]. The directive states that all doses "shall be kept as low as reasonably achievable consistent with obtaining the required diagnostic information". A reasonable interpretation of the directive is that the assurance of fulfilment of the medical purpose of a justified examination overrides the need to decrease the radiation dose. This interpretation, that the primary focus in the optimisation process is the diagnostic information, is further supported in the directive by the statement that the process shall include the selection of equipment, the consistent production of adequate diagnostic information or therapeutic outcome as well as the practical aspects, quality assurance including quality control and the assessment and evaluation of patient doses or administered activities, taking into account economic and social factors. Thus, it can be argued that quality assurance problems are of greater importance than dose issues when prioritising the order of optimisation of different radiological examinations.

Although the demand for justification of all radiological examinations cannot be questioned, different examinations have different impacts on patient health, and the consequence of an inadequately performed examination may vary. The number of patients undergoing a certain examination is also an obvious factor in the optimisation process. Thus, both the consequence for the individual patient of an inadequately performed examination and the frequency of the examination should be taken into account in the prioritisation of optimisations.

According to the ALARA principle, examinations performed with unnecessary high doses to the patients should be optimised before those in which radiation doses to the patients are considered reasonable. However, following the argumentation above that quality problems connected to a medical X-ray procedure are of greater importance than reducing dose when prioritising optimisation tasks, a reduction of radiation dose should only be considered when the issues regarding the diagnostic outcome (image quality and impact of the examination) are judged to be equal.

There may be special dose considerations among the examinations than can be considered to involve unnecessary high doses. For example, many countries have adopted the concept of DRLs for certain examinations[14, 65]. These examinations are typically associated with high collective doses. Examinations with these concerns should therefore be of greater priority than others if all other issues are judged equal.

The above arguments were used as the basis for developing a method for the prioritisation of optimisation tasks. Using the method, the following four questions are applied to each type of examination:

i. Is the present image quality unacceptable? (*Cf. "Poor quality?" in Figure 3.)*

ii. Is the examination of particular importance? (*Cf. "Important examination?" in Figure 3.)*

iii. Is the radiation dose suspiciously high? *(Cf. "Suspiciously high dose?" in Figure 3.)*

iv. Are there special dose level concerns, e.g. diagnostic reference levels? (*Cf. "Dose considerations?" in Figure 3.)*

Arguing that the questions are asked in decreasing order of importance and that a given issue is more important than all the following issues combined, it can be shown that the resulting flow chart, determining the order in which the examinations should be optimised, can be described by a 4-bit binary number. In this way, each type of examination is assigned a number from 0 to 15; a higher number indicating higher priority. The flow-chart illustrating the prioritisation procedure is shown in Figure 3.

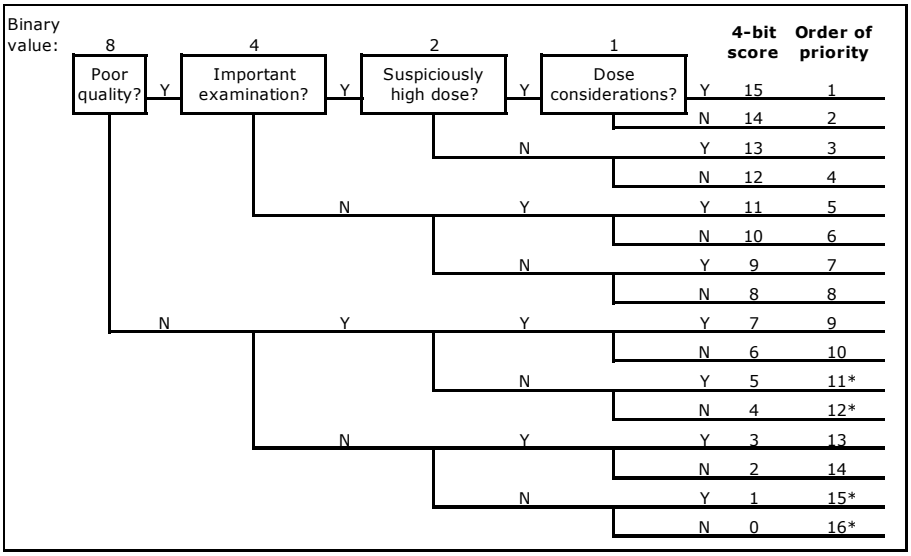


*Figure 3. Question flow chart proposed in Paper II to prioritise optimisation tasks. Calculation by the use of 4-bit scores enables the order of priority to easily be generated in an MS Excel® chart. E.g. a binary result of (1010) will lead to a 4-bit score of 10 and the order of priority 6. Order of priorities marked * are examination types that are judged non-problematic and hence need no further consideration in the optimisation process.*

The proposed method of prioritisation was applied to the examinations carried out at a general radiology department at a university hospital, with eight X-ray rooms including two CT rooms at the time of the study (2009). Supporting information was obtained from various sources: a list of the frequency of all examinations performed during one year in each examination room, and extracted from the hospital radiological information system; documentation from equipment quality control; and the results of diagnostic standard dose measurements. A group consisting of a radiologist, a radiographer and a medical physicist, all with good knowledge of the activities at the department, was asked to score examinations with poor quality (Question i) and/or were of particular importance (Question ii), taking into account the frequency of each examination. Examinations with noticeably high dose levels were identified by the medical physicist (Question iii), from equipment quality control reports and standard dose measurements by comparison with other similar equipment and DRLs. Examinations associated with DRL were identified as examinations with special dose level concerns (Question iv). Finally, the score for each examination was determined and the examinations were ranked in order of increasing score, score 15 indicating the highest priority, 1. The summarised prioritisation list for the tested radiology department is given in Table 1.

Table 1. Summary of scores from the evaluation of examinations performed at the radiology department. Examination types appear more than once if they are performed in more than one examination room

15	Chest (erect), Lumbar spine
14	Thoracic spine
13	Pelvis, Pelvis, Hip, IVU[†], Lumbar spine, Chest (erect)
12	Knee joint, Knee joint, Pelvimetry, Thoracic spine, Venogram, KUB[††]
8	Sacro-iliac-joints, Sacrum and Coccyx, Shoulder/acromio-clavicular-joint, Scapula, Humerus, Elbow, Wrist, Hand, Fingers, Femur, Tibia and fibula, Ankle joint, Foot, Scoliosis, Long-leg
7	CT Brain
5	11 examination types
4	44 examination types
1	7 examination types
0	102 examination types

[†]: Intra venous urogram, [††]: Kidney, ureters and bladder

After two one-hour meetings in the scoring group, an action plan was established regarding the priority of the optimisation of the examinations. Examples of measures listed on the action plan were; technical service of equipment, revised methods, harmonisation with other examination rooms,

training of staff, adjustment in image processing, investigation of optimal technical parameter settings, and exchange of examination room. In total, 16% of the types of examinations performed at the department were judged to be in need of optimisation. When establishing the action list, not only the order priority was considered, but also practical aspects, such as envisaged complexity of an optimisation task, and future plans for investments in new equipment at the department.

To summarise, the method proposed to score the examinations at a radiological department is efficient, and the order of priority for the optimisation of examinations takes into account both medical outcome and potential risk to the patient.

5.3 Paper III

VGC Analyzer: A Software for Statistical Analysis of Fully Crossed Multiple-Reader Multiple-Case Visual Grading Characteristics Studies

Paper III describes the development and implementation (in a dedicated software) of a method for statistical analysis of VGC data. The purpose was to develop a method adapted for the data used in VGC, i.e. taking into account the dependence of paired data in the statistical analysis. The software, VGC Analyzer, determines the area under the VGC curve and its uncertainty (CI and p-value) using non-parametric resampling techniques.

5.3.1 Introduction

As pointed out above, one finding arising from the practical use[86] of the optimisation strategy presented in Paper I, was the need for improved methods for reliable statistical analysis of the data concerning the examination conditions evaluated in a visual grading study. Visual grading is a practical choice for the comparison of different conditions, and is useful because of its adaptability to clinical situations. As described in Section 3.2, the results obtained with this method have been shown to agree well with results from detection studies as well as with advanced calculations of physical image quality. However, as pointed out by Geijer et al. in 2001, statistical methods adapted for the handling of ordinal data in a visual grading study were not fully developed at the time[87]. The attempt to solve this problem, presented by Båth and Månsson in 2007[43], was the introduction of VGC analysis with conversion of the ratings collected on an ordinal scale to a FOM, AUC_{VGC}, on an interval scale. However, this first presentation of VGC did not contain a method to calculate the uncertainty in the presented AUC_{VGC}. As described in Chapter 3.3, the similarity between VGC and ROC was initially an inspiration

to suggest the usage of available methods for statistical analysis of ROC studies also for statistical analysis of VGC data. However, due to the important difference between the methods, it was decided that a dedicated method was needed to perform statistical analysis of VGC data. This inspired the third aim of this work. By renewed inspiration from the development of ROC statistics, where the use of resampling for uncertainty estimation had been introduced[42], attention was directed towards dedicated non-parametric analysis by resampling for the estimation of the uncertainty in VGC data.

5.3.2 Estimation of uncertainty by data resampling

In sampling studies, where the uncertainty in the sampled result cannot be calculated using parametric assumptions, e.g. a normal distribution, a method has been developed to reuse the sampled data, i.e. resampling. The bootstrap technique was introduced by Efron in 1977[88] as a generalisation of the previously used jackknifing method, introduced by Quenouille in 1947[89] and further developed by Tukey in 1958[90]. These stochastic methods have provided researchers with improved tools for the analysis of data when their probability distributions are unknown. In bootstrapping, the collected data are reused by stochastically picking one element at a time (with replacement) from the sample, to construct a new, resampled, data set. The nominal number of data sets that can be constructed is n^n, where n is the number of samples in the original data set. However, the number of unique resampled data sets that can be obtained will be reduced because the order of the resampled data is irrelevant. Also, in image perception studies the number of rating scale steps used is limited, and hence collected rating values can appear more than once. Assuming that the original sample is a good representative of the population, bootstrapping creates a simulated distribution giving the information required for statistical evaluation of the study, with no need for assumptions regarding the underlying distribution[91-94]. The distribution of resampled values can then be used to estimate the uncertainty in the original data, for example, by the confidence interval (CI).

5.3.3 The etymology of bootstrapping

To pull oneself up by one's bootstraps is an idiom describing a (physically) impossible task with no help but your own. It was used in the USA from the first half of the 19th century (Workingman's Advocate 1834: *"It is conjectured that Mr. Murphee will now be enabled to hand himself over the Cumberland river or a barn yard fence by the straps of his boots.")*[95]. A similar idiom was coined in Europe in the 18th century by Rudolf Erich Raspe in Baron Munchausen's Narrative of his Marvellous Travels and Campaigns in Russia[96]. In one episode, Munchhausen saves himself from a swamp by

pulling himself and his horse out using his own pigtail[97]. The real Baron Munchhausen had participated on the side of the Russians in the war against the Ottoman Empire from 1735-1739, and was later well-known in the German aristocracy for telling tall tales about his adventures. The stories by Raspe were later translated and expanded by several writers[98] and in the USA the two idioms of bootstrapping and pulling one's hair seem to have become mixed (the author's own speculation). Therefore, although no actual episode of "bootstrapping" can be found in the original editions, the idiom of bootstrapping has in some parts of the world been attributed to Munchhausen. Efron and Tibshirani, for example, comment on the origin of the name for this method in An Introduction to the Bootstrap[99]: *"The Baron had fallen to the bottom of a deep lake. Just when it looked like all was lost, he thought to pick himself up by his own bootstraps"*.

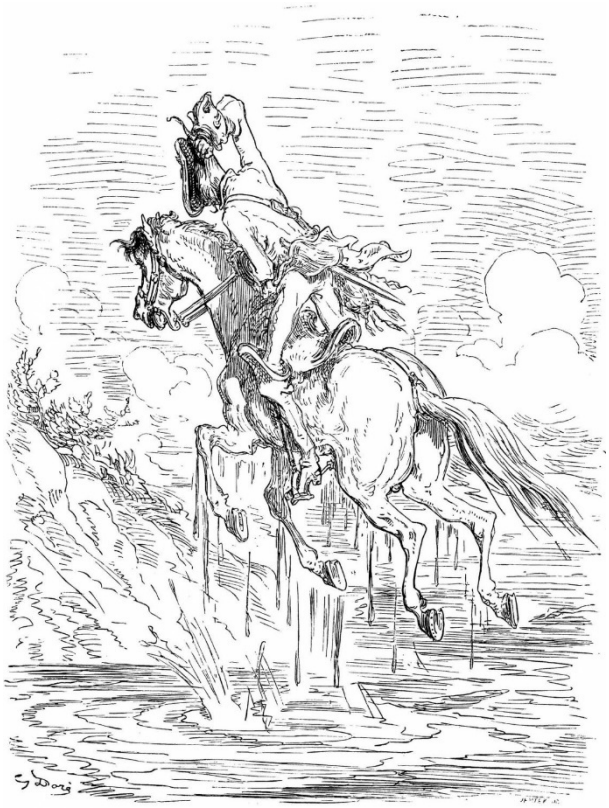


Figure 4. Baron Munchhausen rescuing himself and his horse from sinking in a swamp by pulling on his pigtail. Illustration by Gustave Doré, Wikimedia[100].

In the acknowledgments in his first publication on the bootstrap method[88] Efron thanks all his friends who had suggested alternative names for the new method. His personal favourite among the alternatives was the Shotgun,

"which, to paraphrase Tukey, can blow the head of any problem if the statistician can stand the resulting mess".

The development of resampling of data by jackknifing and bootstrapping has had a major impact in the field of statistics during the last decades, as an alternative to traditional algebraic derivations. The method has become a useful tool in statistical analysis where the distribution of the data cannot be predicted, increasing in parallel to the availability of computer capacity.

5.3.4 The use of bootstrapping to estimate the uncertainty in a sample

In general terms, resampling by bootstrapping is based on the assumption that the true probability distribution of an unknown parameter can be estimated by the distribution that will be generated from resampling the original data a large number of times.

In analogy with Efron and Tibshirani[99], let $P \rightarrow \mathbf{x} = (x_1, x_2 \ldots, x_n)$ indicate a sample **x** drawn from the unknown probability distribution *P*, where the sample elements $(x_1, x_2 \ldots, x_n)$ are all independent and identically distributed. The general distribution of *P* is a consequence of the complex mixture of affecting factors, whereas, in a specific study, the collected samples will be a point estimate of *P,* here denoted $\hat{P}$. Hence, $\mathbf{x} = (x_1, x_2 \ldots, x_n)$ is the discrete distribution of the point estimate, $\hat{P}$, where x_i, $i=1, 2, \ldots, n$, all have the probability *1/n.* From **x** we can compute a statistic of interest s(**x**), e.g. sample mean.

$\mathbf{x}^* = (x_1^*, x_2^* \ldots, x_n^*)$ is a bootstrap sample randomly collected from $\hat{P}$ where the star symbol indicates that $\mathbf{x}^*$ is a resampled version of **x**, i.e. $\hat{P} \rightarrow \mathbf{x}^* = (x_2^*, x_2^* \ldots, x_n^*)$, where the resampled elements can can be collected several times. $\hat{P} \rightarrow \mathbf{x}_\mathrm{B}^* = \left(\mathbf{x}^{*1}, \mathbf{x}^{*2} \ldots, \mathbf{x}^{*\mathrm{B}}\right)$ is the full sample of B bootstrap samples from $\hat{P}$, where the size of B is unlimited. For each bootstrap sample a statistic of interest, s($\mathbf{x}^*$), corresponding to a bootstrap replication of s(**x**), can be computed. Thus, we can write $\hat{P} \rightarrow \mathrm{s}(\mathbf{x}_\mathrm{B}^*) = \left(s(\mathbf{x}^{*1}), s(\mathbf{x}^{*2}) \ldots, s(\mathbf{x}^{*\mathrm{B}})\right)$, where the distribution of $\mathrm{s}(\mathbf{x}_\mathrm{B}^*)$ is interpreted as a simulated distribution of the real distribution of s(**x**) from repeated samples of **x**.

5.3.5 The use of bootstrapping to estimate the CI of a VGC study

When performing a VGC study, the ratings of the images from the conditions to be compared are collected separately in two lists. The two conditions are

then compared by comparing the image quality ratings, and VGC analysis is used to calculate the area under the VGC curve, AUC_{VGC}, which acts as the FOM. In analogy with the description above, AUC_{VGC} is the statistic of interest, s(**x**), of the study. Resampling of the observer's ratings by bootstrapping will form new ICSs and result in a bootstrap replication of AUC_{VGC} which in analogy with the description above can be denoted AUC^*_{VGC}, i.e. s($\mathbf{x}^*$). A full bootstrap sample of the collected ratings, resulting in new bootstrap replications of the statistics of interest, AUC_{VGC}, can accordingly be written $\hat{P} \rightarrow \mathbf{AUC}^*_{VGC,B} = \left(AUC^{*1}_{VGC}, AUC^{*2}_{VGC} \ldots AUC^{*B}_{VGC}\right)$, where the characteristic single FOM, AUC_{VGC}, from the study is expanded to give a series of values, thereby enabling the characteristics of the original AUC_{VGC} value to be estimated. In this way, the variation in $\mathbf{AUC}^*_{VGC,B}$ can be used to determine a simulated non-parametric measure of the uncertainty, e.g. the CI of the AUC_{VGC}.

In a single-reader situation, the bootstrap process is a straight-forward process of resampling of the ratings, resulting in repeated bootstrap replications of AUC_{VGC}. However, in a multiple-reader study with r observers, where a description of a random reader situation is often required, a generalisation function must be included in the bootstrap. In the method presented in Paper III, cases are treated as in the single-reader process for each bootstrap session. These cases are then used for all observers selected in a bootstrap of observers, resulting in a bootstrap replication of the AUC_{VGC} for each bootstrapped observer j (j =1, 2, … ,r), denoted $AUC^*_{VGC,j}$. Calculation of the mean value of all $AUC^*_{VGC,j}$ completes each AUC^*_{VGC}. For a fixed-reader study, the bootstrapped cases are reused for all observers, who are all included in each bootstrap session. If the data sets of compared systems are correlated (paired), e.g. a study where all the study objects (patients) are examined under both conditions, the data must be handled as being correlated. The correlation between the compared conditions is maintained throughout the bootstrapping by copying the case order in each bootstrap session for the reference condition to the test condition (pairwise resampling).

Referring to the description above, the distribution of all AUC^*_{VGC} is a simulation of the unknown distribution of the measured AUC_{VGC} and can be used to estimate the significance (parametric or non-parametric) in a detected difference between the conditions. The non-parametric CI of the measured AUC_{VGC} is calculated from the bootstrap data as the levels of pre-defined $\mathbf{AUC}^*_{VGC,B}$ percentile boundary conditions, feasible for use in hypothesis testing. For example, if AUC=0.5 is not included in the asymmetric 95% CI of the $\mathbf{AUC}^*_{VGC,B}$ distribution, bounded by the 2.5% and 97.5% percentiles, the

separation of the test condition from the reference condition is statistically significant at the level of 0.05.

5.3.6 The use of permutation to estimate the p-value of a VGC study

The p-value is an alternative measure to CI in the statistical evaluation of the collected ratings, and is defined as the probability of obtaining the detected score (or a more extreme value) if the null hypothesis, H_0, were true, i.e. there is no difference between the compared conditions. A non-parametric method of calculating the p-value is to use permutation, where collected ratings of images from the conditions being compared are merged into one collection of ratings. If H_0 were true, a repeated random separation of the ratings into two compared fabricated conditions would indicate the probability of obtaining the detected difference by chance.

In VGC analysis, the p-value is the probability of obtaining a score at least as separated from AUC=0.5 as that obtained, given that H_0 is true. In VGC Analyzer, the permuted p-value ($\boldsymbol{p}_{PERM}$), is calculated using a permutation test in which ratings collected from the compared conditions are regarded as if they originated from the same distribution, i.e. H_0 is true (in this case meaning that the real value of the AUC is 0.5). Ratings are randomly selected (without replacement) from the combined data to create a resampled data set for two pseudo-conditions A and B, and the AUC_{VGC} for the comparison between these two pseudo-conditions is determined. By resampling the ratings from the total collection of the compared conditions and randomly regrouping them, i.e. permuting them, new VGC studies can be replicated, where statistics of interest, $AUC^{\bullet}_{VGC}$, can be calculated, as illustrated by
$\hat{P} \rightarrow \mathbf{AUC}^{\bullet}_{VGC,N} = \left(AUC^{\bullet 1}_{VGC}, AUC^{\bullet 2}_{VGC} \ldots . AUC^{\bullet N}_{VGC}\right)$, where the bullet symbol indicates that $AUC^{\bullet}_{VGC}$is a permuted version of AUC_{VGC}.

The permuted distribution of $\mathbf{AUC}^{\bullet}_{VGC,N}$ will reveal the probability of obtaining the actually measured difference between the conditions (AUC_{VGC}) or more extreme, including the probability of obtaining the opposite outcome, i.e. double sided. Assuming that $\mathbf{AUC}^{\bullet}_{VGC,N}$ has a symmetric distribution around 0.5, the double-sided p-value is given by:
$\boldsymbol{p}_{PERM} = (\#[|\, AUC^{\bullet}_{VGC} - 0.5| \geq |\, AUC_{VGC} - 0.5|])/N$, i.e. the number of permutations performed where the resulting $AUC^{\bullet}_{VGC}$ is equal to or more distant from the originally measured AUC_{VGC} divided by the number of permutations performed (N). The effect of observer variability on the p-value is, in the random-reader situation, added to the permutation by bootstrapping which of the observers to include in each permutation sequence, in the same

way as in the bootstrapping for CI. If the original data are uncorrelated (unpaired), the permutation is performed over all cases. If the original data are correlated (paired), the permutation randomly selects which rating in each pair of ratings from the test and reference conditions that should be assigned to pseudo-condition A, and which should be assigned to pseudo-condition B.

5.3.7 Description of VGC Analyzer

The resampling methods described above for the estimation of CI and the p-value from a VGC study were used in the development of VGC Analyzer, written in Interactive Data Language (Research Systems, Inc., Boulder, CO, USA). The software performs a statistical analysis of the rating data from a fully crossed MRMC VGC study, in which multiple observers (readers) and multiple cases are used, and all observers assess all cases. The software determines AUC_{VGC} averaged over the observers and applies non-parametric resampling methods for the statistical tests: bootstrapping to determine the CI of the AUC_{VGC}, and permutation to determine the p-value for testing the null hypothesis that the two compared systems are equal (H_0: $AUC_{VGC} = 0.5$).

The input to VGC Analyzer is given as a plain text file shown and described in Figure 5. The input file is preferably created using a spreadsheet software, such as MS Excel®, and saved as a text file. All input data are collected in the first columns of each line, except for the rating data from the observers, which are collected in multiple columns (one column per observer). An example input file, modifiable according to the data in a specific study, is included in the delivery of VGC Analyzer. As long as the ratings are integer based, they can belong to an arbitrarily chosen ordinal scale. VGC Analyzer automatically identifies the categories actually used by the observers, and the results of the analysis depends only on the number of ratings in each category, not on the labelling of the scale steps (the ratings collected). The study must be fully crossed (all observers rate all cases) and no data may be missing from the input file. If cases from the compared conditions are paired (for example, if the collected data result from the same group of patients examined under both conditions), the cases collected in the case columns must have the same order for both conditions. Thus, in the analysis of paired data, the number of cases in the compared conditions must be equal.

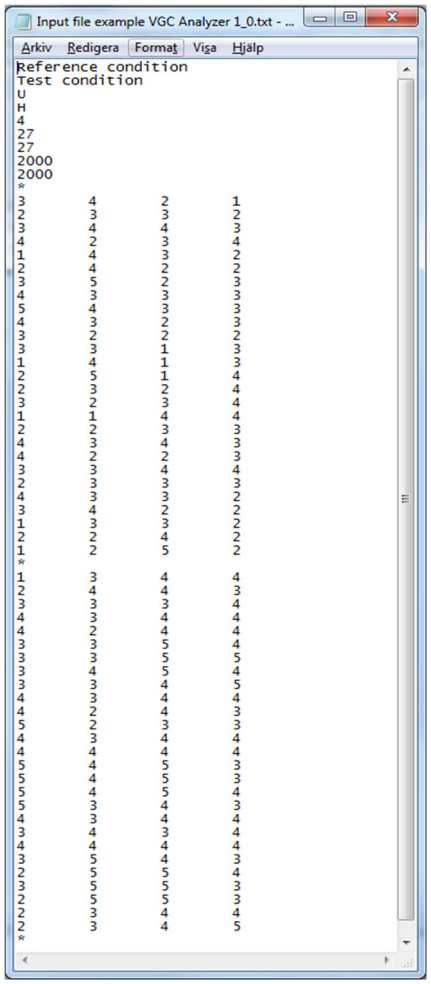
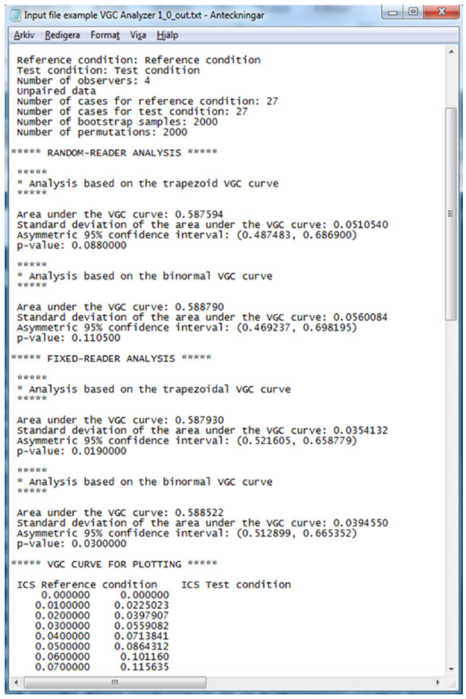


Figure 5. Example of an input file to VGC Analyzer, describing a study in which four observers rated 27 cases each for 2 compared conditions. The data are unpaired and higher ratings correspond to higher image quality.
Line 1: *Name of reference condition.*
Line 2: *Name of test condition.*
Line 3: *'U' for unpaired data (or 'P' for paired).*
Line 4: *'H' for high ratings corresponding to high image quality (or 'L' for low).*
Line 5: *Number of observers.*
Line 6: *Number of cases in the reference condition.*
Line 7: *Number of cases in the test condition.*
Line 8: *Number of bootstrap samples used to determine the distribution of* AUC_{VGC}.
Line 9: *Number of permutation samples used to determine the p-value.*
Line 10: *Delimiter.*
Lines 11–37: *The ratings for the reference condition. Each observer is separated by a tab, and each case is collected on a new line. The case order must be the same for all observers.*
Line 38: *Delimiter.*
Lines 39–67: *The ratings for the test condition.*
Line 68: *Delimiter.*

The results of the statistical analysis by VGC Analyzer are provided as a plain text file, as shown in Figure 6. In addition, the binormal VGC curve determined from the pooled data is presented in a graphical window in Figure 7, providing instant visualisation of the results of the VGC analysis.



Input file example VGC Analyzer 1_0_out.txt - Anteckningar

Arkiv Redigera Format Visa Hjälp

```
Reference condition: Reference condition
Test condition: Test condition
Number of observers: 4
Unpaired data
Number of cases for reference condition: 27
Number of cases for test condition: 27
Number of bootstrap samples: 2000
Number of permutations: 2000

***** RANDOM-READER ANALYSIS *****

*****
* Analysis based on the trapezoid VGC curve
*****

Area under the VGC curve: 0.587594
Standard deviation of the area under the VGC curve: 0.0510540
Asymmetric 95% confidence interval: (0.487483, 0.686900)
p-value: 0.0880000

*****
* Analysis based on the binormal VGC curve
*****

Area under the VGC curve: 0.588790
Standard deviation of the area under the VGC curve: 0.0560084
Asymmetric 95% confidence interval: (0.469237, 0.698195)
p-value: 0.110500

***** FIXED-READER ANALYSIS *****

*****
* Analysis based on the trapezoidal VGC curve
*****

Area under the VGC curve: 0.587930
Standard deviation of the area under the VGC curve: 0.0354132
Asymmetric 95% confidence interval: (0.521605, 0.658779)
p-value: 0.0190000

*****
* Analysis based on the binormal VGC curve
*****

Area under the VGC curve: 0.588522
Standard deviation of the area under the VGC curve: 0.0394550
Asymmetric 95% confidence interval: (0.512899, 0.665352)
p-value: 0.0300000

***** VGC CURVE FOR PLOTTING *****

ICS Reference condition    ICS Test condition
    0.000000     0.000000
    0.0100000    0.0225023
    0.0200000    0.0397907
    0.0300000    0.0559082
    0.0400000    0.0713841
    0.0500000    0.0864312
    0.0600000     0.101160
    0.0700000     0.115635
```

Figure 6. Excerpt from an output file from VGC Analyzer. For each type of reader analysis (fixed and random) and for each type of curve (trapezoid and binormal), the following measures are given: the bootstrap-averaged AUC_{VGC}, the standard deviation of the AUC_{VGC}, the non-parametric 95% CI of the AUC_{VGC}, and the p-value for H_0: $AUC_{VGC} = 0.5$. For plotting purposes, data describing the binormal VGC curve obtained by pooling the data from all readers are given at the end of the file (the input data for the illustration shown in Figure 7).

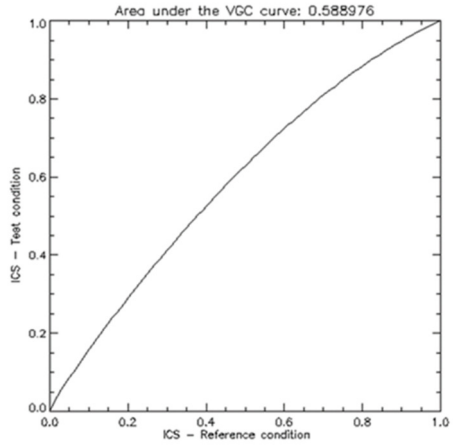


Figure 7. Example of the binormal VGC curve presented in a graphical window as output after the analysis. In order to maximise the probability of a successful binormal fit, the data from all readers are pooled. Note that this curve is only for presentation purposes: the statistical analysis is based on non-pooled data.

Based on the VGC data points resulting from the ratings collected, the value of AUC_{VGC} is determined using both the trapezoidal rule and a binormal fit, as described for ROC in Section 3.1.1. VGC Analyzer determines the AUC_{VGC} for each observer, and the FOM is obtained by averaging AUC_{VGC} values across the observers. The same resampled cases are used for all the observers included in each resampled data set. If there is a correlation between the cases for the two conditions in the input data (paired data), the correlation is stated

by the user in the set-up so that the resampling procedure takes this into account by pairwise resampling of the cases under the two conditions. The analysis is performed for both the fixed-reader situation (results applicable only to the actual observers in the study) and the random-reader situation (results applicable to the population of observers).

5.4 Paper IV

The Validity of Using ROC Software for Analysing Visual Grading Characteristics Data: An Investigation Based on the Novel Software VGC Analyzer

The purpose of the study presented in Paper IV was to investigate the validity of using single-reader-adapted ROC software, where the rating data from multiple-readers are pooled, for the analysis of VGC data. VGC data (actual ratings) from four published VGC studies on the optimisation of X-ray examinations, originally analysed using ROCFIT (C E Metz, University of Chicago, Chicago, IL, USA), were reanalysed using VGC Analyzer, and the outcomes (the mean and 95% CI of the AUC_{VGC} and the p-value) were compared.

ROC-dedicated software for the analysis of VGC data has been used in a number of previous optimisation studies[101-104]. ROCFIT has been appreciated for the unique feature to provide an estimation of the uncertainty in a single AUC value. However, as ROCFIT has no ability but to pool the observers, the validity of the analysis of multiple-reader studies using ROCFIT can be assumed to be limited. The studies reanalysed in Paper IV included both paired and non-paired data and the data were reanalysed for pooled readers, the fixed-reader and the random-reader situations.

In the original studies, all rating data had been handled as non-paired due to the use of ROCFIT. Furthermore, the CIs and the p-values had been determined parametrically, based on the standard deviation (SD) of the AUC provided by ROCFIT; the CIs had been determined as +1.96 SD around the AUC_{VGC}, and the p-values had been determined using the z-test, since the number of degrees of freedom were not provided by ROCFIT.

In the original studies, the AUC_{VGC} had been determined by binormal curve fitting. In the re-analyses, the AUC_{VGC} was determined by both the binormal curve fitting and the trapezoidal rule, for evaluation purposes.

The results for the mean AUC_{VGC} showed excellent agreement between the methods when the analysis was performed in the same way (pooled readers, binormal). Using trapezoid curve fitting, the AUC_{VGC} calculated by VGC Analyzer was less distant from 0.5. When including the reader variation, the resulting AUCs, for non-paired data, were more distant from 0.5, and wider CIs were obtained with VGC Analyzer than previously reported (see Figure 8). For paired data, the previously reported CIs were similar or even wider (see Figure 9). Similar observations were made for the p-values. These results indicate that the use of pooling in single-reader-adapted ROC software such as ROCFIT to analyse non-paired VGC data may lead to an increased risk of committing Type I errors, i.e. overestimating the probability of significant difference, especially in the random-reader situation. On the other hand, the use of analysis methods adapted for ROC data in the analysis of paired VGC data may lead to an increased risk of committing Type II errors, i.e. underestimating the probability of significant difference, especially in the fixed-reader situation.

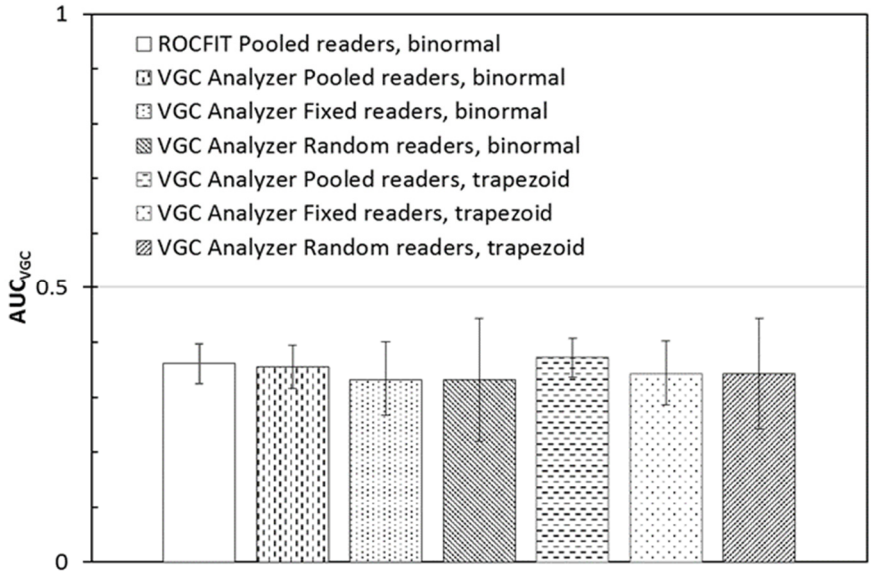


Figure 8. AUC_{VGC} and CI from Zachrisson et al.[104] comparing two tube voltages for urography (55 kV vs 73 kV), using five observers, three criteria, and 31 cases per condition (non-paired data). On the 55 kV images, the effective dose was reduced to 32%, compared to the 73 kV images, by software simulation[105]. In the original study, observers and criteria were pooled and ROCFIT was used to determine the AUC_{VGC} and the CI. In the reanalysis, using VGC Analyzer, criteria were pooled but the observers were treated either as fixed or random.

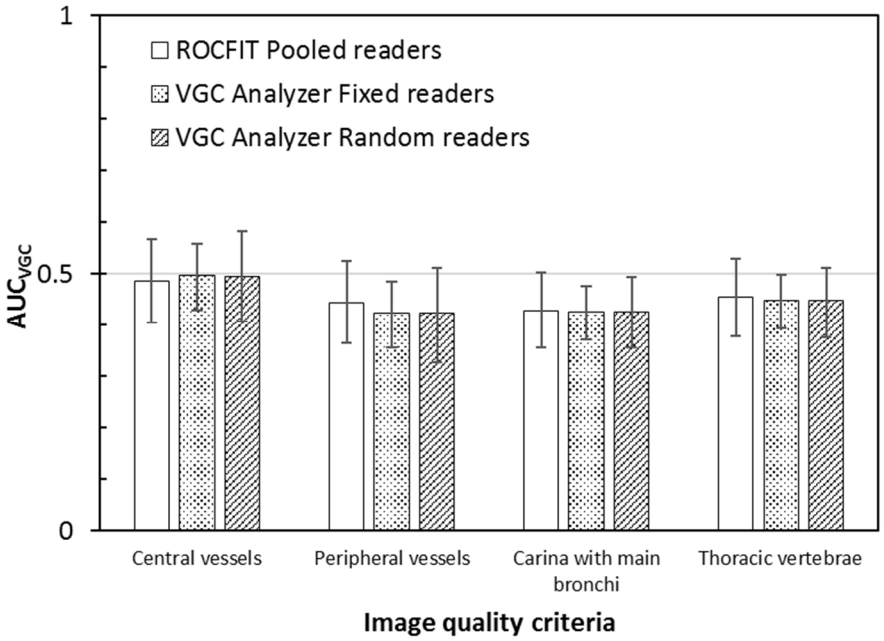


Figure 9. AUC_{VGC} (binormal curve fitting) and CI from Carlander et al.[102] comparing two dose levels for neonatal chest imaging (80% vs 100%), using five observers, four criteria, and 24 cases per condition (paired data). In the original study, observers were pooled and ROCFIT was used to determine the AUC_{VGC} and the CI for each criterion. Hence, the paired-data effect could not be taken into account. In the reanalysis using VGC Analyzer, analysis was performed on pooled observers, but also the paired-data effect was taken into account in the analysis where observers were treated either as fixed or random.

5.5 Paper V

Evaluation of Resampling Methods for Analysis of Visual Grading Data by Comparison with State-of-the-art ROC methodology and Analysis of Simulated Data

The purpose of the study presented in Paper V was to evaluate the use of resampling statistical methods for the analysis of visual grading data – implemented in VGC Analyzer. Reanalysed results from previously performed visual grading studies were compared with the results calculated by a gold standard in ROC analysis, OR-DBM MRMC (The Medical Image Perception Laboratory, The University of Iowa, Iowa City, Iowa, USA)[106], and by analysis of simulated visual grading data where the true distribution was assumed to be known.

In the study described in Paper IV, VGC Analyzer was compared with now outdated ROC methodology. Therefore, one purpose of the study presented in Paper V was to compare the performance of VGC Analyzer with state-of-the-art ROC software on visual grading data where the ROC methodology can be assumed to provide valid results, i.e. with correct handling of multiple-readers, but also where it can be assumed to provide invalid results, i.e. lacking the correct handling of paired data adapted for visual grading. The validity of an evaluation based on a comparison between two methods where access to the exact truth is limited. Therefore, the second purpose of the study was to extend the evaluation of the method by performing a simulation study in which the truth is known. By creating a large number of simulated visual grading studies and performing statistical analysis on each study, the distribution of the results can be compared with the expected frequency of rejected null hypotheses in the simulated studies.

The comparison between VGC Analyzer and OR-DBM MRMC was based on rating data from two of the studies reanalysed in Paper IV. In Zachrisson et al.[104] the images for the two compared conditions were acquired from two different patient groups (non-paired data) whereas in Carlander et al.[102] the images for both conditions were acquired from the same group of patients (paired data). In the reanalysis, VGC Analyzer was configured for analysis of either non-paired or paired data, depending on the type of data, whereas OR-DBM MRMC, for methodological reasons, treats the data as non-paired (it can therefore be assumed that OR-DBM MRMC overestimates the CI of the AUC from the paired data study[102]). The uncertainty of the AUC_{VGC} was determined both for the actual observers (fixed readers) and for the population of observers (random readers). The AUC_{VGC} was determined from curve fitting by the trapezoidal rule, both for OR-DBM MRMC and VGC Analyzer. The OR-DBM MRMC software was set to use the Jackknifing resampling technique, since the bootstrapping technique was not available in the used version, 2.51.

Specially designed simulation software was used to evaluate the validity of the method through the analysis of simulated studies. The same resampling method was used in the simulations as in VGC Analyzer, but the input and output routines were modified to suit the simulation procedures. The study was performed by analysing the results from a large number of simulated VGC studies, in which ratings were produced by random sampling from pre-defined distributions. For each simulated VGC study, the simulation software was used to determine the CI of the AUC_{VGC} and the p-value.

The resampling methods used in VGC Analyzer were evaluated by simulating VGC studies with the null hypothesis, H_0, set to true (i.e., the probability

distributions for the two conditions were equal). A similar assessment strategy has been used previously for ROC analysis by Roe and Metz[107], where simulated stochastically distributed ROC data ("null case studies") were used to test the DBM MRMC analysis method. The correctness of the method for the calculation of CI was evaluated by testing whether the resampling using bootstrapping "erroneously" indicated a significant difference between the two conditions (i.e. AUC_{VGC}=0.5 was not included in the 95% CI) in the intended H_0 rejection rate (5%) of studies (thereby performing Type I errors at an α-level of 0.05). Correspondingly, the correctness of the method for the calculation of p-value was evaluated by testing whether the resampling using permutation "erroneously" indicated a significant difference between the two conditions (i.e. $p<0.05$) in the intended H_0 rejection rate (5%) of studies.

A random sequence was applied to produce artificial visual grading ratings from two simulated conditions, defined by various properties of the simulated observers' probability distributions. The properties altered were the shape (uniform, normal, or wedged), the number of scale steps, and the statistical variation between cases as well as within and between observers. Different visual grading studies were simulated by varying the number of cases and observers, all with the prerequisite that the compared conditions were equal in terms of probability distributions. Multiple-reader situations were simulated by introducing a randomised "dummy-observer" rating, indicating the general impression among the observers for each case. The rating for each observer was then randomised from a normal distribution, centered around the "dummy-observer" rating for the given case, and with an observer-specific standard deviation, constant for all cases. The observer-specific standard deviation was randomised from another normal distribution with zero mean and standard deviation σ_{obs}. The ratings of the "dummy-observer" were not included in the simulated result. Paired data studies were generated by introducing a correlation between each pair of cases in the simulations of ratings for the two conditions.

For each combination of property values, 100 000 studies were simulated, separated into groups of 10 000 to determine the standard error. The relative number of results from the resampling analysis that indicated a statistically significant difference (by CI and p-value) was recorded as the H_0 rejection rate. In the resampling analysis, the number of bootstraps was varied from 200-20 000. The standard settings in the simulations were 2000 bootstraps and 2000 permutations.

The comparison with OR-DBM MRMC showed good agreement when analysing non-paired data for both fixed-reader and random-reader settings for

the calculated values of AUC_{VGC} and CI (see Figure 10). For paired-data analysis, VGC Analyzer showed significantly lower CIs than OR-DBM MRMC (see Figure 11), indicating the necessity of using adequately adapted methods for paired-data analysis. This effect was also illustrated by the simulation study, where the H_0 rejection rate decreased to 0.1% when paired data were treated as non-paired (see Figure 12).

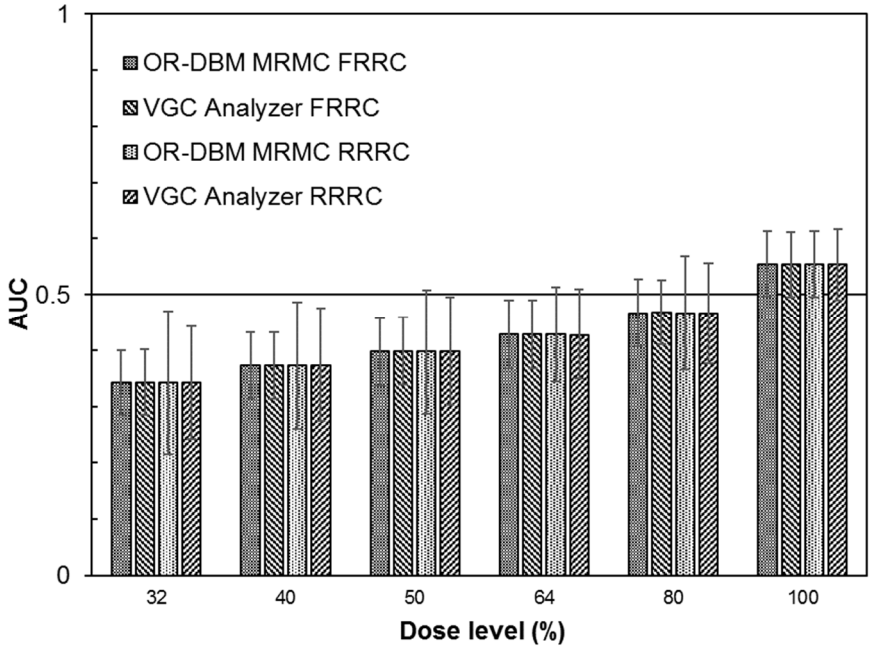


Figure 10. AUC_{VGC} and CI based on data from Zachrisson et al.[104] comparing two tube voltages for urography [55 kV at different dose levels vs 73 kV at 100% dose], using five observers, three criteria, and 31 cases per condition (non-paired data). In the analysis using OR-DBM MRMC and VGC Analyzer, criteria were pooled but the observers were treated either as fixed or random (FR or RR) and the cases were treated as random (RC). Error bars are the 95% CIs given by the analysis software.

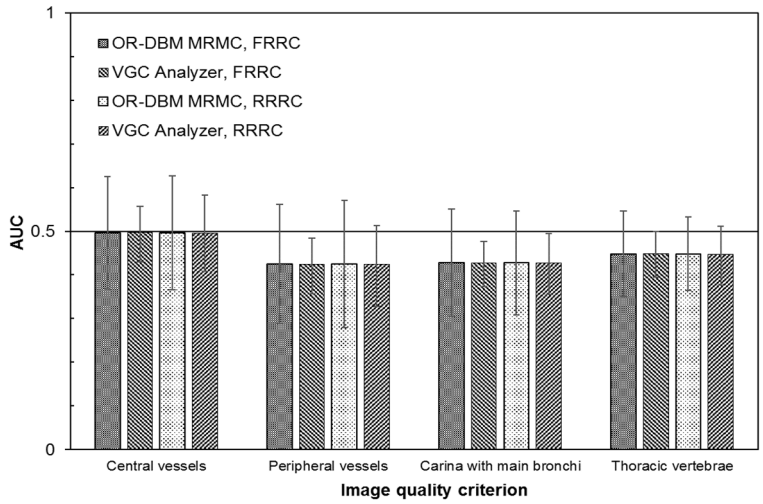


Figure 11. AUC_{VGC} and CI based on data from Carlander et al.[102] comparing two dose levels for neonatal chest imaging (80% vs 100%), using five observers, four criteria, and 24 cases per condition (paired data). In the analysis using VGC Analyzer, the paired-data effect was taken into account, whereas in the analysis using OR-DBM MRMC, ROC-data were assumed, and the effect of paired-data could not be taken into account in the analysis. Observers were treated either as fixed or random (FR or RR) and the cases were treated as random (RC). Error bars are the 95% CIs given by the analysis software.

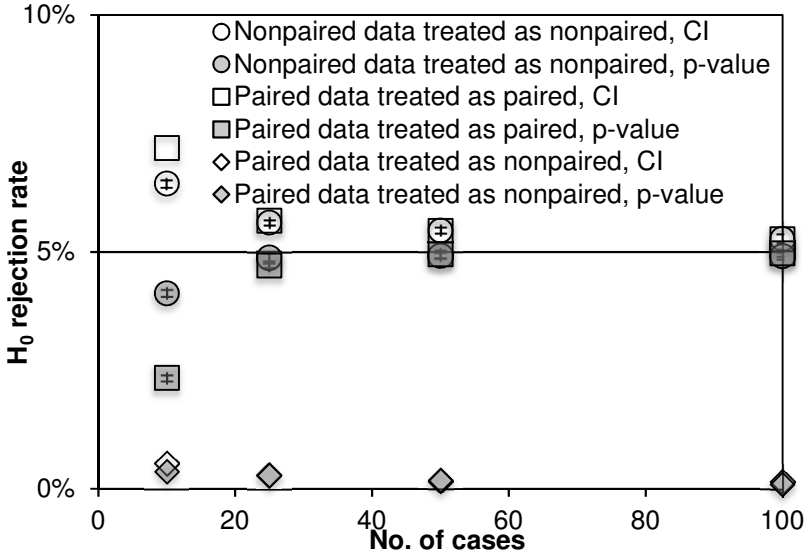


Figure 12. Simulated data after 2000 bootstraps (CI) or 2000 permutations (p-value) on 10 000 VGC studies with varying number of cases in a single-reader setting. Simulated non-paired data and paired data treated as non-paired or paired. The H_0 rejection rate obtained from the CI or p-value and AUC was calculated with trapezoidal curve fitting for normal distributions of ratings on a 5-step scale (mean value=3, SD=1). Error bars indicate the standard error from 10 consecutive simulation sessions.

The discrepancies between VGC Analyzer and OR-DBM MRMC observed in the reanalysis of real data were confirmed in the simulation study. In general, VGC Analyzer showed good accuracy for simulated studies with stable statistical basis, insusceptible to assessment scale and distribution of ratings. On the other hand, for simulated studies with unstable statistics (low numbers of cases or observers as well as large data variation) the accuracy in the H_0 rejection rate decreased, as partly illustrated in Figure 13.

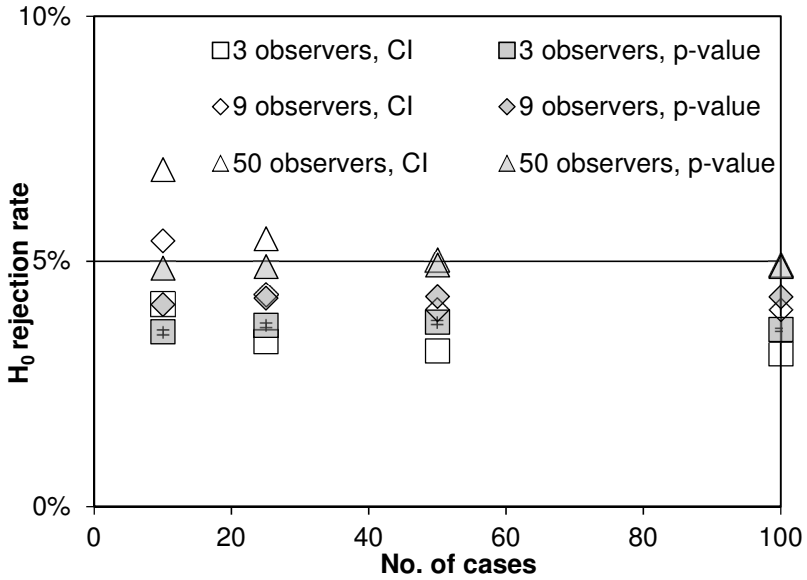


Figure 13. H_0 rejection rate obtained from the CI or p-value in studies where observers were generalised and treated as random, for a wide range of number of observers. The "dummy-observer" ratings were obtained from a normal distribution with mean value=3 and SD=1 on a 5-step rating scale. A value of σ_{obs}=1 was used to create the multiple-reader situation. Simulated data after 2000 bootstraps (CI) or 2000 permutations (p-value) on 10 000 VGC studies with varying number of cases in multiple-readers settings. Non-paired data were treated as non-paired, normal distribution of ratings, AUC calculated with trapezoid curve fitting. Error bars indicate the standard error from 10 consecutive simulation sessions.

To summarise, the study presented in Paper V showed that VGC Analyzer can be used to accurately perform the statistical analysis of a VGC study, although the resampling technique used makes the method sensitive to small data sets. However, the non-parametric resampling method makes the analysis insusceptible to the assessment scale and the distribution of ratings. The adaption to handle paired data included in VGC Analyzer makes it more suitable for these types of studies, than software intended for the analysis of ROC data, where the handling of paired data is not adapted for visual grading studies. The effect of incorrect handling of paired data was verified in the simulation study.

6 DISCUSSION

The purpose of the aims stipulated in this thesis was to contribute to improvements in the process for the optimisation of X-ray examinations in medical care. The fulfilment of these aims can contribute in the development of this process. However, the research presented here only contributes to some areas in the complex process of optimisation, and the complexity of this process makes it difficult to obtain a good overview of how the process is best pursued in the clinical environment. In this chapter the experiences from the work presented in this thesis are discussed in relation to other research performed, related to the topic of optimisation of X-ray imaging.

According to the EU current Directive and national legislation, all use of ionising radiation in medicine shall be optimised. As discussed in Chapter 2, it is the author's opinion that the objective of optimisation is to balance the goal to achieve a good medical outcome with the goal to ensure high radiation safety. This compromise is a particular task for the medical physicists that requires careful reflection in the work with quality development. The argumentation that beneficial factors as well as diagnostic risk factors should be included in the planning of optimisation efforts in the use of ionising radiation in diagnostic radiology has been thoroughly discussed by Moores in a recent series of publications[18-21]. The ICRP has addressed the problem from solely focusing on exposure reduction in medicine[17]. However, constrained by maintaining the ALARA principle, no practical solutions to solve the problem has been suggested. Also, in directives and legislations from authorities, it is unclear if the objective of optimisation of radiation protection is to reduce the risk and detriment from radiation or if the objective is to maximise the efficacy of radiation use in medicine.

For the effective optimisation of an X-ray procedure, good knowledge of the conditions under which the procedure is performed is crucial. An adequate justification process will therefore be of great guidance in the identification of procedures that are in special need of optimisation. Furthermore, quality assurance programmes (including equipment control and monitoring patient exposure) and incident reports are other sources of information that can be used for the identification of optimisation needs. Information obtained from interviews with staff or from training exercises can also be added to these systematic data. All the available information on the examinations or treatment procedures used in a radiology department can form the basis for determining the order of priority for the optimisation of radiological procedures.

A systematic prioritisation process is presented in Paper II, the 4-bit concept, in which the priority is graded on a scale from 1 to 16. This prioritisation strategy is based on the principle that the need for improved medical outcome of a procedure should be prioritised over the desire to reduce exposure. The proposed strategy is also supported by the argumentation from Moores[18-21], where the need for assessed medical outcome is addressed. It is by no means a perfect strategy in this sense, but can be used to simply and quickly identify the procedures that are in the greatest need of optimisation. As the purpose of this strategy is to prioritise the hundreds of combinations of examination or treatment rooms and procedure types performed in a radiology department, it is important that it is relatively easy to collect basic relevant and reliable data on which decisions can be based. Once a prioritisation process is initiated, and optimisation is underway, more information on the need for improvements will become available. The collection of information on the need for optimisation will facilitate the prioritisation of examinations requiring optimisation, in turn leading to optimisation of these examinations, thus forming a continuous loop. The progress of the optimisation loop should be yearly monitored and audited.

Technical developments after the publication of Paper II have enabled better data collection regarding patient exposure from individual examinations, i.e. continuous recording of equipment output from each examination. This development enables improved possibilities to provide useful data as the basis for the prioritisation of optimisation task. Almén and Båth[108] have suggested a conceptual framework where the appropriate exposure from introduced examination types are assessed in four steps, including a feedback step where the actual outcome from performed examinations are compared with expected levels established in previous steps. This framework could be implemented using modern technology with the purpose to support optimisation, e.g. by automatic calculation of DRLs.

Based on the results of prioritisation, decisions must be taken regarding the method that should be used to optimise the procedures with the highest priorities. Extensive image perception evaluation will not always be necessary to determine the optimal examination technique. In some cases, change in the examination set-up, adjustments of technical parameters for image collection, improved logistics of the referral process or staff training will be sufficient to achieve the desired improvement. In other situations, more thorough investigation will be needed, aiming to find the best patient benefit from a procedure in a wider medical care perspective than the specific exposure event.

In the case of examination types where needs for optimisation of image collection and presentation have been identified as being of high priority, a

suitable optimisation procedure must be chosen, as well as an appropriate evaluation method to determine the best technique. A proposed optimisation procedure for projectional X-ray imaging is presented in Paper I where the strategy is summarised in the following way:

a) include the anatomical background in image quality optimisation,
b) perform all comparisons at a constant effective dose, and
c) make full advantage of the digital system for separation of the image collection step from the image display step in the imaging chain.

The first and last steps are probably not controversial, as several studies have shown the necessity of performing image quality evaluations in a situation as close to the clinical one as possible. Since the publication of Paper I in 2005 the use of digital data has evolved such that processing of image data is a natural part of the process before display. The second step may, however, be somewhat controversial. It was argued in Paper I that a constant risk level is the best reference when comparing image quality resulting from different imaging techniques. This is also valid for specific evaluation situations. However, other research groups have demonstrated successful optimisation procedures using different approaches. For example, Wiltz et al. showed that it was possible to reduce the patient absorbed dose in urography by lowering the tube voltage while maintaining image quality[109], and Smedby et al. quantified the potential for dose reduction by allowing variation of the exposure level in a visual grading regression (VGR) study[110]. VGR is a method using ordinal regression tools in standard statistical software for the analysis of visual grading data, introduced by Smedby and Fredriksson[63]. In VGR, the effect of adjusting multiple factors affecting the diagnostic outcome can be analysed to obtain an indication of the optimal setting for a specific diagnostic method, with the option of defining individual scaling and distribution for each factor. As described in Section 5.1, the optimisation strategy suggested in Paper I is performed in separated stages, where image collection and image display are optimised in the first two stages, and the absolute exposure level is optimised in the last stage. This is in contrast to VGR, where the three stages of optimisation can be performed simultaneously.

As pointed out above, the optimisation process in medical imaging using ionising radiation is a compromise between information content (image quality) and the risk (radiation dose). When designing an optimisation procedure, it is important to determine whether higher image quality (with the risk of higher dose) or lower dose (with the risk of poorer image quality) is

most important. One important difference between these two quantities is that radiation dose is measured on an interval scale, and the result of changing the exposure setting between measurements is relatively easy to predict, whereas image quality in observer performance studies is rated on an ordinal axis, where the distance between ratings is unknown, and hence the effect of change in exposure setting is more difficult to predict. Also, the uncertainty in the radiation dose should be relatively well known and independent of the uncertainties in the observer performance study. This is not the case for the measured difference in image quality between two examination conditions, where the uncertainty in the result is not known beforehand, and is strongly dependent on the variation between the observers. The recommendation in the optimisation strategy described in this thesis is therefore to use patient exposure as the reference quantity when determining the best setting that provides optimal image quality. In Paper I, the effective dose, or an analogue relevant measure of radiation risk was suggested as the reference risk quantity, to be kept constant when comparing different examination techniques. The reason for suggesting this was to emphasise the need to use a risk estimate that is as relevant as possible for the optimisation process. If a parameter that is evaluated in the process, e.g. the tube-voltage level, affects the quantity used as the reference, e.g. skin dose, in such a way that the risk, e.g. the stochastic risk, does not remain constant, the validity of the result of the study will decrease. If, however, in the example above, the deterministic skin effect is the main risk factor in the optimisation, the skin dose is obviously the best reference quantity. Therefore, the choice of reference risk quantity should always be based on the specific situation in the planned optimisation process.

The most valid risk parameter in medical diagnostic use of ionising radiation, is the quantity describing the risk to the patient, either the stochastic risk, e.g. the effective dose, or the deterministic risk, e.g. the equivalent skin dose. The concept of effective dose using tissue weighting factors, that represent the proportion of stochastic risk was proposed by Jacobi[111]. Although primarily intended for use in the radiological protection of workers and the public, the concept was adopted as a dose restriction quantity by the ICRP in 1977[112] and was further recommended for use in medical exposure in the 1990 recommendations[113]. It has been suggested that the effective dose should be used carefully, especially in situations where the local exposure distribution to specific organs is uncertain (e.g. CT, fluoroscopy and intra-oral radiography)[114]. However, software tools have been developed to facilitate the calculation and increase the accuracy in the estimation of the effective dose[115-117]. Furthermore, dosimetric studies have shown that the use of more simplified risk estimates, such as KAP or the energy imparted, would be generally less valid[85]. Based on the overall conclusion presented in a review

of relevant studies[118], the recommendation in Paper I was to primarily use the effective dose as the reference risk estimate in an optimisation strategy when comparing imaging conditions. The appropriate use of effective dose as an estimate of risk from X-ray imaging has, however, been further discussed. Brenner and Huda[119] have criticised the ambition to combine genetic and cancer risk into one quantity, without including the variation of risk by age. They therefore suggest a further development of the concept by introducing the quantity *effective risk* where focus is set on cancer risk estimation, including sensitivity variation by age.

Once it has been decided how optimisation should be accomplished the most appropriate evaluation method must be decided. If human observers and clinical images are involved, image perception studies are preferable. If it is possible to perform a detection study with high validity for the purpose of optimisation and a known truth for the detection task is available, the best method would be some form of ROC study. If, however, it is difficult to establish a detection task with high clinical validity, or there is no well-founded truth, a visual grading study is preferable. Another reason for using a visual grading study is that it is less time consuming than an ROC study as images to be evaluated can be collected directly from the clinical production, whereas an ROC study often requires some form of preparation for image collection. Nevertheless, in order to ensure high validity in a visual grading study, it is important to carefully define the image quality criteria that are to be assessed in the evaluation. Strict definition of the structures to be rated and of their rating scale will reduce the risk of "preference-bias" and will increase the reliability of the study.

Regardless of the choice of evaluation method in an optimisation process, an effective image viewer with integrated recording of the ratings is desirable. ViewDEX (Viewer for Digital Evaluation of X-ray images)[120-122] is free-ware especially developed for this purpose. Both the viewing properties and registration properties of the ratings can be edited by the user, and the system can be adapted for both ROC and visual grading studies.

To the author's knowledge, two dedicated approaches have been established for the statistical evaluation of visual grading studies, VGC and VGR. VGC, with the statistical analysis by VGC Analyzer, was developed for the statistical analysis of the difference between two imaging settings. If multiple settings are to be included in the study, testing of the significance of differences between the settings can be carried out pairwise. VGC Analyzer is easy to run via a plain text file interface, and the properties of the method are well documented from the evaluations presented in this thesis. Being a non-

parametric method for statistical analysis, the resampling methods used in VGC Analyzer have an advantage in that the results are not affected by any assumptions regarding the underlying data distribution. However, a disadvantage is that non-parametric methods cannot easily be used to handle more complex data with multiple dependencies[62].

VGR, on the other hand, is a suggested method where ordinal regression is used to analyse multiple parameters with varying data scaling, including ordinal scaling. If an optimisation study is planned, VGR has been shown to be an effective method for identifying the optimal settings if the evaluation of multiple parameters is of interest. Methods for ordinal regression analysis are based on the assumption that the relationship between the ordinal distributions being compared can be fitted to a pre-defined model. The "proportional odds model" described by McCullagh[123] was the first model suggested and has become the most popular[124]. The resulting 'logit' value is assigned a proportional dependence on the relationship between the compared conditions through a logarithmic transformation of the odds ratio between two compared conditions. The method combines the relations between the assigned ratings over all levels of outcomes without dichotomisation of the data. Also, the proportionality of the odds ratios from each comparison between conditions will allow them to be combined by addition[125]. By maximising the outcome of the covariate effect sum, the proportional odds model enables a straightforward means of optimising the evaluated method. Despite being characterised by high parameter flexibility and correct handling of the ordinal rating scale, it has been pointed out in the literature that the stringent assumption of proportional odds on which this model is based, is not automatically valid for all ordinal response variables[124].

The FOM resulting from a visual grading study evaluated using VGC is a single rank-invariant AUC_{VGC} value, which can be determined without any parametric assumptions regarding the underlying rating distribution. The uncertainty of the AUC_{VGC} is in VGC Analyzer determined using non-parametric resampling. This non-parametric treatment enables a complete conservation of the rank-invariant nature of VGC, requiring no assumptions concerning the underlying probability distribution in ratings or resulting resampled data. The insusceptibility to limited expansion of the assessment scale in VGC Analyzer is interpreted as being the consequence of the rank-invariant nature of VGC. A non-parametric approach was paramount in the development of a method for statistical analysis of VGC data. The advantage of this approach is the insusceptibility of VGC Analyzer to the distribution of the analysed data. A possible disadvantage of the approach could be the sensitivity to small data sets. In this situation, the resampled distribution of

statistics of interest (AUC_{VGC}) is limited by the restricted number of unique events possible. This effect was identified in Paper V, where the analysis of simulated studies with low statistical basis decreased in accuracy.

7 CONCLUSIONS

The overall aim of the work presented in this thesis was to develop methods and strategies for the optimisation process prescribed for medical X-ray imaging. Specifically, methods of conducting and prioritising the optimisations of examinations, including improved visual grading methods, were investigated.

In Paper I, the conditions for the optimisation of a given projectional X-ray examination in a digital environment are analysed and a proposed optimisation strategy, based on the analysis, is described. In Paper II an overall strategy for the prioritisation of the optimisation work in a radiology department is presented. Paper III describes the development of a suitable method for statistical analysis of VGC data, which is implemented in the software VGC Analyzer. In Papers IV and V, the characteristics of the new statistical method are thoroughly evaluated by comparison with ROC statistical methodology and by simulations.

The strategies developed helped clarify the prerequisites in the process of optimising medical X-ray imaging and were shown to be useful in clinical applications. However, the objective of optimising the radiation protection in medical use of radiation is not fully clarified in legal requirements, and needs further discussion.

The development of resampling methods for statistical analysis of VGC data, implemented in VGC Analyzer, provides a method that is easy to apply in clinical optimisation projects where visual grading is judged to be the appropriate evaluation method. The results from VGC Analyzer are comparable to those obtained by state-of-the-art ROC methods in situations where the latter are valid for analysis of visual grading data. Furthermore, the simulation study showed that VGC Analyzer performs the statistical analysis of VGC data with high accuracy when the statistical basis of the study is stable. In studies with small data sets, the resampling methods used in VGC Analyzer are limited by the restricted number of unique events possible in the resampling of the data. Therefore, the accuracy decreases in studies where the statistical basis is weak.

ACKNOWLEDGEMENTS

I wish to express my deep and sincere gratitude to all those who have helped me in the work resulting in this thesis.

Professor Magnus Båth, my supervisor and valued colleague, to whom I owe the greatest gratitude. Your enthusiasm and extensive professional knowledge have been vital in the fulfilment of this mission.

Dr Lars Gunnar Månsson, my co-supervisor and boss who has provided professional support through my whole career. Your endurance is amazing!

Markus Håkansson, **Pernilla Jonasson** and **Dr Patrik Sund**, my co-authors in Papers I and II. Our inspiring discussions on the optimisation strategy had a considerable influence on the realisation of this thesis.

Dr Jakobina Grétarsdóttir and **Dr Barbro Vikhoff Baaz**, my two bosses who supported me in completing this project. It wouldn't have been possible without you.

Dr Angelica Svalkvist, thank you for reading the manuscript and for giving constructive comments that improved the final quality of this thesis.

All my colleagues on the X-ray physics corridor, who covered for me during this past year when I was often absent from my clinical duties.

My fellow PhD students, who were so generous in welcoming me to the group. Your youth and encouragement were stimulating! However, all good things must come to an end.

All my friends and workmates at the Department of Medical Physics and Biomedical Engineering and the Department of Radiation Physics. Your friendship and positive attitude ensure a warm workplace.

Dr Helen Sheppard, for her professional help in the revision of the English in this thesis.

Finally, my beloved **Helena** and my extra daughters **Nanna**, **Moa** and **Maja**. Your understanding during the past ten years has been the best support I could ever have asked for.

.

REFERENCES

1. Simon M., "Stray remarks on the history of medical radiology in Sweden," *Acta Radiol.* **7**(1-6), 476-490 (1926).
2. Birkelo C.C., Chamberlain W.E. et al., "Tuberculosis case finding; a comparison of the effectiveness of various roentgenographic and photofluorographic methods," *J. Am. Med. Assoc.* **133**(6), 359-366 (1947).
3. Kundel H.L., "History of research in medical image perception", *J. Am. Coll. Radiol.* **3**(6), 402-408 (2006).
4. Lusted L.B., "Logical analysis in roentgen diagnosis*," Radiology* **74**,178-193 (1960).
5. Swets J.A., Pickett R.M., Whitehead S.F., Getty D.J., Schnur J.A., Swets J.B. et al., "Assessment of diagnostic technologies," *Science* **205**(4408), 753-759 (1979).
6. Sund P., Herrman, C., Tingberg, A., Kheddache, S., Månsson, L.G., Almén, A. and Mattsson, S., "Comparison of two methods for evaluating image quality of chest radiographs," *Proc. SPIE* **3981**, 251-257 (2000).
7. Tingberg A., Herrmann C., Almén A., Besjakov J., Mattsson S., Sund P. et al., "Comparison of two methods for evaluation of the image quality of lumbar spine radiographs," *Radiat. Prot. Dosimetry* **90**(1), 165-168 (2000).
8. Redlich U., Hoeschen C. and Doehring W., "Assessment and optimisation of the image quality of chest-radiography systems," *Radiat. Prot. Dosimetry* **114**(1-3), 264-268 (2005).
9. International Commission on Radiological Protection, http://www.icrp.org/page.asp?id=3, (updated 18 Oct 2019).
10. International Commission on Radiological Protection, *1928 International recommendations for X-ray and radium protection,* P.A. Norstedt & Söner, Stockholm (1928).
11. Fryback D.G. and Thornbury J.R., "The efficacy of diagnostic imaging," *Med. Decis. Mak.* **11**(2), 88-94 (1991).
12. Oxford University Press, Oxford English Dictionary, https://www.oed.com (updated 28 Aug 2019).
13. International Commission on Radiological Protection. *ICRP Publication 9: Recommendations of the International Commission on Radiological Protection,* Pergamon Press, Oxford (1966).

14. International Commission on Radiological Protection, *ICRP Publication 22: Implications of Commission recommendations that doses be kept as low as readily achievable*, Pergamon Press, Oxford (1973).
15. International Commission on Radiological Protection. *ICRP Publication 60: 1990 Recommendations of the International Commission on Radiological Protection*, Ann. ICRP **21**(1-3), Pergamon Press, Oxford (1991).
16. International Commission on Radiological Protection, *ICRP Publication 73: Radiological protection and safety in medicine*, Ann. ICRP **26**(2), Pergamon Press, Oxford, (1996).
17. International Commission on Radiological Protection, *ICRP Publication 105: Radiation protection in medicine*, Ann. ICRP **37**(6), 1-63, Pergamon Press, Oxford, (2007).
18. Moores B.M., "Cost-risk-benefit analysis in diagnostic radiology: A theoretical and economic basis for radiation protection of the patient," *Radiat. Prot. Dosimetry* **169**(1-4), 2-10 (2016).
19. Moores B.M., "A review of the fundamental principles of radiation protection when applied to the patient in diagnostic radiology," *Radiat. Prot. Dosimetry* **175**(1), 1-9 (2017).
20. Moores B.M., "The psychology of decision making and its relevance to radiation protection of the patient in medicine," *Radiat. Prot. Dosimetry* **178**(3), 245-253 (2018).
21. Moores B.M., "Cost-risk-benefit analysis in diagnostic radiology with special reference to the application of referral guidelines," *Radiat. Prot. Dosimetry,* doi:10.1093/rpd/ncz054 (2019).
22. The Council of the European Union, *Council Directive 2013/59/Euratom of 5 December 2013, laying down basic safety standards for protection against the dangers arising from exposure to ionising radiation*. Official Journal of the European Union, No. L 013/1-17.1.2014 (2014).
23. The Government Office of Sweden, *Strålskyddslag (2018:396),* SFS 2018:396 (2018).
24. The Government Office of Sweden, *Strålskyddsförordning (2018:506),* SFS 2018:506 (2018).
25. Swedish Radiation Safety Authority, *SSMFS 2018:5 Strålsäkerhetsmyndighetens föreskrifter och allmänna råd om medicinska exponeringar* (2018).

26. Swedish Radiation Safety Authority, *Vägledning med bakgrund och motiv till Strålsäkerhetsmyndighetens föreskrifter (SSMFS 2018:5) och allmänna råd om medicinska exponeringar* (2018).
27. European Society of Radiology, *Risk management in radiology in Europe*, ESR/EAR Office, Vienna, Austria (2004).
28. Månsson L.G., Båth M. and Mattsson S., "Priorities in optimisation of medical X-ray imaging – a contribution to the debate," *Radiat. Prot. Dosimetry* **114**(1-3), 298-302 (2005).
29. Malone J. and Zölzer F., "Pragmatic ethical basis for radiation protection in diagnostic radiology," *Br. J. Radiol.* **89**(1059), 20150713 (2016).
30. Beauchamp T.L. and Childress J.F., *Principles of biomedical ethics,* Oxford University Press, New York (1979).
31. International Commission on Radiological Protection, *ICRP publication 103: The 2007 Recommendations of the International Commission on Radiological Protection*, Ann. ICRP **37**(2-4), 1-332, Pergamon Press, Oxford (2007).
32. International Commission on Radiological Protection, *ICRP publication 37: Cost-benefit analysis in the optimization of radiation protection*, Ann. ICRP **10**(2-3), Pergamon Press, Oxford (1983).
33. Swedish Agency for Health Technology Assessment and Assessment of Social Services, *Intraoperativ kolangiografi vid kolecystektomi. En systematisk översikt och utvärdering av medicinska, hälsoekonomiska, sociala och etiska aspekter*. Stockholm: Statens beredning för medicinsk och social utvärdering (SBU), SBU-rapport nr 292, ISBN 978-91-88437-34-1 (2018).
34. Båth M., "Evaluating imaging systems: Practical applications," *Radiat. Prot. Dosimetry* **139**(1-3), 26-36 (2010).
35. Sandborg M. and Carlsson G.A., "Influence of X-ray-energy spectrum, contrasting detail and detector on the signal-to-noise ratio (SNR) and detective quantum efficiency (DQE) in projection radiography," *Phys. Med. Biol.* **37**(6), 1245-1263 (1992).
36. Sandborg M., Dance D.R., Carlsson G.A. and Persliden J., "Selection of anti-scatter grids for different imaging tasks – the advantage of low atomic-number cover and interspace materials," *Br. J. Radiol.* **66**(792), 1151-1163 (1993).
37. Båth M., Håkansson M. and Månsson L.G., "Determination of the two-dimensional detective quantum efficiency of a computed radiography system," *Med. Phys.* **30**(12), 3172-3182 (2003).

38. Båth M., Sund P. and Månsson L.G., "Evaluation of the imaging properties of two generations of a CCD-based system for digital chest radiography," Med. Phys. **29**(10), 2286-2297 (2002).
39. Månsson L.G., "Methods for the evaluation of image quality: A review," *Radiat. Prot. Dosimetry* **90**(1-2), 89-99 (2000).
40. Swets J.A. and Picket R.M., *Evaluation of diagnostic systems: Methods from signal detection theory*, Academic Press, New York (1982).
41. Metz C.E., "ROC methodology in radiologic imaging," *Invest. Radiol.* **21**(9), 720-733 (1986).
42. Dorfman D.D., Berbaum K.S. and Metz C.E., "Receiver operating characteristic rating analysis. Generalization to the population of readers and patients with the jackknife method," *Invest. Radiol.* **27**(9), 723-731 (1992).
43. Båth M. and Månsson L.G., "Visual grading characteristics (VGC) analysis: A non-parametric rank-invariant statistical method for image quality evaluation," *Br. J. Radiol.* **80**(951), 169-176 (2007).
44. Dorfman D.D. and Alf E., "Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals – rating-method data," *J. Math. Psychol.* **6**(3), 487-496 (1969).
45. Hanley J.A., "The use of the 'binormal' model for parametric ROC analysis of quantitative diagnostic tests," *Stat. Med.* **15**(14), 1575-1585 (1996).
46. Metz C.E., "ROC analysis in medical imaging: A tutorial review of the literature," *Radiol. Phys. Technol.* **1**(1), 2-12 (2008).
47. Dorfman D.D., Berbaum K.S. and Lenth R.V., "Multireader, multicase receiver operating characteristic methodology – a bootstrap analysis," *Acad. Radiol.* **2**(7), 626-633 (1995).
48. European Commission, *EUR 16260 – European guidelines on quality criteria for diagnostic radiographic images*, Office for official publications of the European Communities, Luxembourg (1996).
49. European Commission, *EUR 16261 – European guidelines on quality criteria for diagnostic radiographic images in paediatrics*, Office for official publications of the European Communities, Luxembourg (1996).
50. European Commission, *EUR 16262 – European guidelines on quality criteria for computed tomography*, Office for official publications of the European Communities, Luxembourg (1996).

51. Altman D.G., *Practical statistics for medical research*, Chapman and Hall, London (1991).
52. Tingberg A., Båth M., Håkansson M., Medin J., Sandborg M., Alm-Carlsson G. et al., "Comparison of two methods for evaluation of image quality of lumbar spine radiographs," *Proc. SPIE* **5372**, 251-262 (2004).
53. Tingberg A., Båth M., Håkansson M., Medin J., Besjakov J., Sandborg M. et al., "Evaluation of image quality of lumbar spine images: A comparison between FFE and VGA," *Radiat. Prot. Dosimetry* **114**(1-3), 53-61 (2005).
54. Zanca F., Van Ongeval C., Claus F., Jacobs J., Oyen R. and Bosmans H., "Comparison of visual grading and free-response ROC analyses for assessment of image-processing algorithms in digital mammography," *Br. J. Radiol.* **85**(1020), 1233-1241 (2012).
55. Chakraborty D.P., "Problems with the differential receiver operating characteristics (DROC) method," *Proc. SPIE* **5372**, 138-143 (2004).
56. Sund P., Båth M., Kheddache S. and Månsson L.G., "Comparison of visual grading analysis and determination of detective quantum efficiency for evaluating system performance in digital chest radiography," *Eur. Radiol.* **14**(1), 48-58 (2004).
57. Moore C.S., Wood T.J., Beavis A.W. and Saunderson J.R., "Correlation of the clinical and physical image quality in chest radiography for average adults with a computed radiography imaging system," *Br. J. Radiol.* **86**(1027) (2013).
58. Sandborg M., Tingberg A., Ullman G., Dance D.R. and Carlsson G.A., "Comparison of clinical and physical measures of image quality in chest and pelvis computed radiography at different tube voltages," *Med. Phys.* **33**(11), 4169-4175 (2006).
59. Tapiovaara M.J., "Review of relationships between physical measurements and user evaluation of image quality," *Radiat. Prot. Dosimetry* **129**(1-3), 244-248 (2008).
60. Kundel H.L., "Images, image quality and observer performance: New horizons in radiology lecture," *Radiology* **132**(2), 265-271 (1979).
61. Vaño E., Guibelalde E., Morillo A., Alvarez-Pedrosa C.S. and Fernandez J.M., "Evaluation of the European image quality criteria for chest examinations," *Br. J. Radiol.* **68**(816), 1349-1355 (1995).
62. Machin D., Campbell M.J. and Walters S.J., *Medical statistics: A textbook for the health sciences.* 4th ed. Wiley, Chichester, (2007).

63. Smedby Ö. and Fredrikson M., “Visual grading regression: Analysing data from visual grading experiments with regression models,” *Br. J. Radiol.* **83**(993), 767-775 (2010).
64. Hanley J.A., McNeil B.J., “A method of comparing the areas under receiver operating characteristic curves derived from the same cases,” *Radiology* **148**(3), 839-843 (1983).
65. The Council of the European Union, *Council Directive 97/43/Euratom on health protection of individuals against the dangers of ionising radiation in relation to medical exposure*, Official Journal of the European Communities, No. L180/22-27.9.07.97 (1997).
66. Bochud F.O., Valley J.F., Verdun F.R., Hessler C. and Schnyder P., “Estimation of the noisy component of anatomical backgrounds,” *Med. Phys.* **26**(7), 1365-1370 (1999).
67. Samei E., Flynn M.J. and Eyler W.R., “Detection of subtle lung nodules: Relative influence of quantum and anatomic noise on chest radiographs,” *Radiology* **213**(3), 727-734 (1999).
68. Samei E., Flynn, M.J. and Eyler, W.R., *Effects of anatomical structure on signal detection. In: Handbook of Medical Imaging, Volume 1: Physics and Psychophysics.* Van Metter R.L., Beutel J. and Kundel H.L. editors, Bellingham: SPIE, 655-682 (2000).
69. Huda W., Ogden K.M., Scalzetti E.M., Dudley E.F. and Dance D.R., “How do radiographic techniques affect mass lesion detection performance in digital mammography?” *Proc. SPIE* **5372**, 372-382 (2004).
70. Keelan B.W., Topfer K., Yorkston J., Sehnert W.J. and Ellinwood J.S., “Relative impact of detector noise and anatomical structure on lung nodule detection,” *Proc. SPIE* **5372**, 230-241 (2004).
71. Båth M., Håkansson M., Börjesson S., Kheddache S., Grahn A., Ruschin M. et al., ”Nodule detection in digital chest radiography: Introduction to the RADIUS chest trial,” *Radiat. Prot. Dosimetry* **114**(1-3), 85-91 (2005).
72. Båth M., Håkansson M., Börjesson S., Kheddache S., Grahn A., Bochud F.O. et al., “Nodule detection in digital chest radiography: Part of image background acting as pure noise,” *Radiat. Prot. Dosimetry* **114**(1-3), 102-108 (2005).
73. Håkansson M., Båth M., Börjesson S., Kheddache S., Johnsson Å.A. and Månsson L.G., “Nodule detection in digital chest radiography: Effect of system noise,” *Radiat. Prot. Dosimetry* **114**(1-3), 97-101 (2005).

74. Håkansson M., Båth M., Börjesson S., Kheddache S., Grahn A., Ruschin M. et al., ”Nodule detection in digital chest radiography: Summary of the RADIUS chest trial,” *Radiat. Prot. Dosimetry* **114**(1-3), 114-120 (2005).
75. Rose A., “A unified approach to the performance of photographic film, television pickup tubes, and the human eye,” *J. Soc. Motion Pict. Eng*. **47**(4), 273-294 (1946).
76. Verdun F.R., Meuli R.A., Bucher G., Noël A., Stines J., Schnyder P. et al., “Dose and image quality characterisation of CT units,” *Radiat. Prot. Dosimetry* **90**(1-2), 193-196 (2000).
77. Thilander-Klang A., Ledenius K., Hansson J., Sund P. and Båth M., “Evaluation of subjective assessment of the low-contrast visibility in constancy control of computed tomography,” *Radiat. Prot. Dosimetry* **139**(1-3), 449-454 (2010).
78. Konst B., Weedon-Fekjær H. and Båth M., “Image quality and radiation dose in planar imaging – Image quality figure of merits from the CDRAD phantom,” *J. Appl. Clin. Med. Phys.* **20**(7), 151-159 (2019).
79. Burgess A., Jacobson F. and Judy P., “Mass discrimination in mammography: Experiments using hybrid images,” *Acad. Radiol.* **10**(11), 1247-1256 (2003).
80. Håkansson M., Båth M., Börjesson S., Kheddache S., Flinck A., Ullman G. et al., ”Nodule detection in digital chest radiography: Effect of nodule location,” *Radiat. Prot. Dosimetry* **114**(1-3), 92-96 (2005).
81. Båth M., Håkansson M., Börjesson S., Hoeschen C., Tischenko O., Kheddache S. et al., ”Nodule detection in digital chest radiography: Effect of anatomical noise,” *Radiat. Prot. Dosimetry* **114**(1-3), 109-113 (2005).
82. Hintzer R.A., Matthies H.J., Neiman H.L., Rogers L.F. and Lin P.P., “Comparison of xeroradiograph and a rare earth intensifying screen film system (Kodak HIN-R) for mammography,” *Invest. Radiol.* **11**(5), 386 (1976).
83. Pärtan G., Partik B., Mayrhofer R., Pichler L., Urban M., Gindl K. et al., “Feasibility of 0.3 mm Cu additional beam filtration for digital gastrointestinal fluororadiography,” *Radiat. Prot. Dosimetry* **90**(1-2), 217-220 (2000).
84. Staniszewska M.A., Biegański T., Midel A. and Barańska D., “Filters for dose reduction in conventional X-ray examinations of children,” *Radiat. Prot. Dosimetry* **90**(1-2) 127-133 (2000).

85. Månsson L.G., Wallström E. and Mattsson S., “Relations between effective dose, effective dose-equivalent, area-kerma product, and energy imparted in chest radiography,” *Radiat. Prot. Dosimetry* **49**(4), 421-431 (1993).
86. Hansson J., Båth M., Håkansson M., Grundin H., Bjurklint E., Orvestad P. et al., ”An optimisation strategy in a digital environment applied to neonatal chest imaging,” *Radiat. Prot. Dosimetry* **114**(1-3), 278-285 (2005).
87. Geijer H., Beckman K., Jonsson B., Andersson T. and Persliden J., “Digital radiography of scoliosis with a scanning method: Initial evaluation,” *Radiology* **218**(2), 402-410 (2001).
88. Efron B., “Bootstrap methods: Another look at the jackknife,” Ann. Stat. **7**, 1-26 (1979).
89. Quenouille M.H., “Approximate tests of correlation in time-series,” *J. R. Stat. Soc. B* **11**(1), 68-84 (1949).
90. Tukey J., “Bias and confidence in not-quite large samples (Abstracts of Papers),” *Ann. Math. Stat.* **29**(2), 614 (1958).
91. Efron B. and Tibshirani R., “Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy,” *Stat. Sci.* **1**(1), 54-75 (1986).
92. Efron B., “Better bootstrap confidence intervals,” *J. Am. Stat. Assoc.* **82**(397), 171-185 (1987).
93. Efron B., “Bootstrap confidence intervals: Good or bad?” *Psychol. Bull.* **104**(2), 293-296 (1988).
94. Efron B., “Second thoughts on the bootstrap” *Stat. Sci.* **18**(2), 135-140 (2003).
95. Wikipedia, Bootstrapping: Wikipedia, Available from: https://en.wikipedia.org/wiki/Bootstrapping#Etymology (2019).
96. Wikipedia, Baron Munchausen: Wikipedia, Available from: https://en.wikipedia.org/wiki/Baron_Munchausen (2019).
97. Raspe R.E., *Münchhausens underbara äventyr*, Almqvist & Wiksell Förlag AB, Stockholm (1979).
98. Kästner E., *Des Freiherrn von Münchhausen wunderbare Reisen und Abenteuer zu Wasser und zu Lande*, Atrium Verlag, Zürich (2001).
99. Efron B. and Tibshirani R.J., *An Introduction to the Bootstrap*, Chapman & Hall/CRC, Boca Raton, FL (1993).
100. Wikimedia, *Gustave Doré – Baron von Münchhausen,* Wikimedia, Available from: https://commons.wikimedia.org/wiki/File:Gustave_Doré_-_Baron_von_Münchhausen_-_037.jpg (2019)

101. Carlander A., Hansson J., Söderberg J., Steneryd K. and Båth M., "Clinical evaluation of a dual-side readout technique computed radiography system in chest radiography of premature neonates," *Acta Radiol.* **49**(4), 468-474 (2008).
102. Carlander A., Hansson J., Söderberg J., Steneryd K. and Båth M., "The effect of radiation dose reduction on clinical image quality in chest radiography of premature neonates using a dual-side readout technique computed radiography system," *Radiat. Prot. Dosimetry* **139**(1-3), 275-280 (2010).
103. Larsson L., Båth M., Engman E.L. and Månsson L.G., "Harmonisation of the appearance of digital radiographs from different vendors by means of common external image processing," *Radiat. Prot. Dosimetry* **139**(1-3), 92-97 (2010).
104. Zachrisson S., Hansson J., Cederblad Å., Geterud K. and Båth M., "Optimisation of tube voltage for conventional urography using a Gd2O2S:Tb flat panel detector," *Radiat. Prot. Dosimetry* **139**(1-3), 86-91 (2010).
105. Båth M., Håkansson M., Tingberg A. and Månsson L.G., "Method of simulating dose reduction for digital radiographic systems," *Radiat. Prot. Dosimetry* **114**(1-3), 253-259 (2005).
106. Schartz K., *Medical Image Perception Laboratory,* http://perception.radiology.uiowa.edu/: The University of Iowa, IA, USA (2019).
107. Roe C.A. and Metz C. E., "Dorfman-Berbaum-Metz method for statistical analysis of multireader, multimodality receiver operating characteristic data: Validation with computer simulation," *Acad. Radiol.* **4**(4), 298-303 (1997).
108. Almén A. and Båth M. "A conceptual framework for managing radiation dose to patients in diagnostic radiology using reference dose levels," *Radiat. Prot. Dosimetry* **169**(1-4), 17-23 (2016).
109. Wiltz H.J., Petersen U. and Axelsson B., "Reduction of absorbed dose in storage phosphor urography by significant lowering of tube voltage and adjustment of image display parameters," *Acta Radiol.* **46**(4), 391-395 (2005).
110. Smedby Ö., Fredrikson M., De Geer J., Borgen L. and Sandborg M., "Quantifying the potential for dose reduction with visual grading regression," *Br. J. Radiol.* **86**(1021), 31197714 (2013).
111. Jacobi W., "The concept of the effective dose – a proposal for the combination of organ doses. *Radiat. Environ. Biophys*. **12**(2), 101-109 (1975).

112. International Commission on Radiological Protection, *ICRP Publication 26: Recommendations of the International Commission on Radiological Protection*, Ann. ICRP **1**(3), Pergamon Press, Oxford (1977).
113. International Commission on Radiological Protection, *Proceedings of the Second International Symposium on the System of Radiological Protection*, Ann. ICRP **44**(S), 1-2 (2015).
114. Carlsson G.A., Dance D.R., Persliden J. and Sandborg M., "Use of the concept of energy imparted in diagnostic radiology," *Appl. Radiat. Isot.* **50**(1), 39-62 (1999).
115. Tapiovaara M., Lakkisto M. and Servomaa A., *PCXMC: A PC-based Monte Carlo program for calculating patient doses in medical X-ray examinations*, Report STUK-A139, Finnish Centre for Radiation and Nuclear Safety, Helsinki, Finland (1997).
116. Nagel H.D. and Stamm G., CT Expo Bucholdz, Germany, Science & Technology for Radiology, updated 2018-12-04. Available from: www.sascrad.com. (2018).
117. Virtual Phantoms Inc., VirtualDose, Albany, NY, 6.11.2019, Available from: http://www.virtualphantoms.com/.
118. Månsson L.G., *Evaluation of radiographic procedures - investigations related to chest imaging*, Gothenburg University, Gothenburg (1994).
119. Brenner D. and Huda W., "Effective dose: a useful concept in diagnostic radiology?," *Radiat. Prot. Dosimetry*, **128**(4), 503-508 (2008).
120. Börjesson S., Håkansson M., Båth M., Kheddache S., Svensson S. and Tingberg A. et al., "A software tool for increased efficiency in observer performance studies in radiology," *Radiat. Prot. Dosimetry* **114**(1-3), 45-52 (2005).
121. Håkansson M., Svensson S., Zachrisson S., Svalkvist A., Båth M. and Månsson L.G., "ViewDEX: An efficient and easy-to-use software for observer performance studies," *Radiat. Prot. Dosimetry* **139**(1-3), 42-51 (2010).
122. Svalkvist A., Svensson S., Håkansson M., Båth M. and Månsson L.G., "ViewDEX: A status report," *Radiat. Prot. Dosimetry* **169**(1-4), 38-45 (2016).
123. McCullagh P., "Regression-models for ordinal data," *J. R. Stat. Soc. B* **42**(2), 109-142 (1980).
124. Bender R. and Grouven U., "Ordinal logistic regression in medical research," *J. R. Coll. Physicians Lond.* **31**(5), 546-551 (1997).

125. Scott S.C., Goldberg M.S. and Mayo N.E., “Statistical assessment of ordinal outcomes in comparative studies,” *J. Clin. Epidemiol.* **50**(1), 45-55 (1997).