TU WIEN Informatics

Statistische Methoden zur Bewertung einer Künstliche Intelligenz (KI) Software im Diagnostischen Umfeld

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Biomedical Engineering

eingereicht von

Tek Sin Chung, BSc
Matrikelnummer 01525236

an der Fakultät für Informatik

der Technischen Universität Wien

Betreuung: Dr Renata Georgia Raidou
Mitwirkung: Univ.Prof. Dipl.-Ing. Dr.techn. Eduard Gröller

Wien, 2. Februar 2022

Tek Sin Chung	Renata Georgia Raidou

TU WIEN Informatics

Statistical Methodologies for Assessing an Artificial Intelligence (AI) Software in a Diagnostic Setting

DIPLOMA THESIS

submitted in partial fulfillment of the requirements for the degree of

Diplom-Ingenieur

in

Biomedical Engineering

by

Tek Sin Chung, BSc
Registration Number 01525236

to the Faculty of Informatics

at the TU Wien

Advisor: Dr Renata Georgia Raidou
Assistance: Univ.Prof. Dipl.-Ing. Dr.techn. Eduard Gröller

Vienna, 2nd February, 2022

Tek Sin Chung | Renata Georgia Raidou

Technische Universität Wien
A-1040 Wien ▪ Karlsplatz 13 ▪ Tel. +43-1-58801-0 ▪ www.tuwien.at

Erklärung zur Verfassung der Arbeit

Tek Sin Chung, BSc

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe, dass ich die verwendeten Quellen und Hilfsmittel vollständig angegeben habe und dass ich die Stellen der Arbeit – einschließlich Tabellen, Karten und Abbildungen –, die anderen Werken oder dem Internet im Wortlaut oder dem Sinn nach entnommen sind, auf jeden Fall unter Angabe der Quelle als Entlehnung kenntlich gemacht habe.

Wien, 2. Februar 2022

Tek Sin Chung

Danksagung

In erster Linie möchte ich mich bei meiner Betreuerin, Dr. Renata Georgia Raidou, vom Institut für Medizinische Visualisierung und Visual Analytics der TU Wien bedanken, die viel Zeit und Mühe in die Betreuung dieser Arbeit investiert hat. Ihre Anleitung und ihr Beitrag während des gesamten Projekts ermöglichte mir die Fertigstellung dieser Arbeit.

Ich möchte mich auch bei Dr. Richard Ljuhar, CEO von ImageBiopsy Lab, dafür bedanken, dass er mir die Möglichkeit gab, an einem so interessanten Projekt zu arbeiten.

Mein aufrichtiger Dank gilt insbesondere Dr. Matthew D. DiFranco, CSO von ImageBiopsy Lab, der als mein externer Betreuer fungierte. Vielen Dank, dass Sie mir Ihr Fachwissen während dieses Projekts zur Verfügung gestellt und mich während meiner gesamten Arbeit unterstützt haben.

Ich möchte mich auch bei meinen Kollegen, Dr. Zsolt Bertalan und Dipl. Ing. Ulrich Mayer bedanken, die mich ermutigt haben, ein großes Projekt wie diesen anzugehen.

Schließlich möchte ich mich bei meinen Freunden und meiner Familie bedanken, die mich während dieser Arbeit unterstützt haben.

Acknowledgements

First and foremost, I would like to thank my supervisor, Dr. Renata Georgia Raidou, from the Institute of Medical Visualization and Visual Analytics at TU Vienna for putting in time and effort to supervise this thesis. Her guidance and contribution throughout this project enabled me to complete this work.

I would also like to extend my gratitude to Dr.Richard Ljuhar, CEO of ImageBiopsy Lab, in given me the chance to work on a such an interest project.

My sincerest thanks especially to Dr. Matthew D. DiFranco, CSO of ImageBiopsy Lab, for stepping up as my external supervisor. Thank for providing your expertise during the time of this project as well as providing support throughout my thesis.

I would also like to show my gratitude to my colleagues, Dr.Zsolt Bertalan and Dipl. Ing. Ulrich Mayer, both encouraging me to take the first step in tackling a big project such as this.

Finally, special thanks to my friends and family, who has been supportive of me throughout this thesis.

Kurzfassung

Die radiologische Bestimmung des Knochenalters (KA) anhand eines Röntgenbildes der linken Hand ist nach wie vor der Referenzstandard für die Beurteilung der Skelettreife im Zusammenhang mit dem Wachstum zugrunde liegenden Erkrankungen. Aufgrund der Subjektivität und des hohen Zeitaufwands der BA-Bestimmung setzen sich KI-Algorithmen immer mehr durch. Daher empfehlen wir Methoden und statistische Empfehlungen für die Bewertung der Performance eines KI-Tools vor. Unsere Strategie wurde in einer retrospektiven Studie mit dem KI-Modell PANDA überprüft, einer vollautomatischen KI-Software, die zur Schätzung des KA auf Handröntgenbildern verwendet wird.

Wir analysierten die Röntgenbilder von 342 Patienten retrospektiv. Drei zertifizierte pädiatrische Radiologen beurteilten das KA unabhängig voneinander nach der Greulich-&-Pyle-Methode (GP). PANDA wurde anschließend verwendet, um automatische Schätzungen des KA aus demselben Satz von Bildern zu erstellen. Die Ground Truth wurde auf der Grundlage des Mittelwerts der Schätzungen ermittelt. Wir bewerteten die Übereinstimmung der KI mit den Lesern anhand von Bland-Altman-Limits of Agreement (LOA), der orthogonalen linearen Regression und mit dem Konzept der Austauschbarkeit

Die Bland-Altman-Bewertung ergab eine durschnittliche Differenz zwischen den Bewertern und der KI von -0,72 mit einem 95%CI (-1,46; 0,02) Monaten, was keinen fixen Bias anzeigt. Unter Verwendung einer orthogonalen linearen Regression wurde die Steigung zwischen den Lesern und der AI-Software mit 1,02 (95%CI: 1,00, 1,03) angegeben. Es wurde keinen proportionalen Bias festgestellt. Die Quadratwurzel des absoluten Wertes des Äquivalenzindexes der KI-Software im Vergleich zu den Bewertungen durch die Radiologen wurde mit -5,8 Monaten festgestellt. Dies bedeutet, dass die KI-Software mit den Bewertungen von Fachleuten austauschbar ist.

Die vorgeschlagenen Metriken sind nicht auf die Bewertung des Knochenalters beschränkt und können auch auf andere klinische Outputs angewendet werden, sofern es sich um eine kontinuierliche Variable handelt. Wenn man eine Bias zwischen zwei Messtechniken feststellen will, sollte eine Regressionsanalyse durchgeführt werden. Wenn es darum geht, festzustellen, ob eine Methode sicher durch eine andere ersetzt werden kann, insbesondere in der klinischen Praxis, ist Bland-Altman vorzuziehen. Gibt es keinen geeigneten Referenzstandard, mit dem verglichen werden kann, kann das Konzept der Austauschbarkeit verwendet werden. Diese statistische Methode ist nicht auf einen Referenzstandard angewiesen.

Abstract

The radiological determination of bone age (BA) from a left-hand x-ray continues to be the reference standard for skeletal maturity assessment related to short or long stature, and underlying conditions. Artificial (AI) algorithms are becoming more prevalent due to the subjectivity and time-consuming nature of BA assessment. Therefore, we proposed methods and statistical recommendations in assessing standalone performance of an AI tool. Our strategy was verified in a retrospective study using the AI model, PANDA, a fully automated AI software used to estimate bone age (BA) on hand radiographs.

We analyzed radiographs of 342 patients retrospectively. Three board certified pediatric radiologists made blind reads of BA using the Greulich & Pyle (GP) method independently. The AI-software, PANDA, was subsequently used to provided automated estimations of BA from the same set of images. The ground truth was established based on the mean of the estimations. We assessed agreement of AI with readers based on comparison of Bland-Altman limits of agreement (LOA), orthogonal linear regression and interchangeability.

Bland-Altman assessment displayed a mean difference between readers and AI to be -0.72 with 95%CI (-1.46; 0.02) months displying no fixed bias. Using orthogonal linear regression, the slope between readers and AI software was reported to be 1.02 95%CI (1.00, 1.03). No proportional bias was observed. The square root of the absolute value of the equivalence index of the AI software compared to assessments made by readers was observed to be -5.8 months. This indicates that the AI software is interchangeable with expert readers.

The proposed framework is generalizable to the other applications aside from bone age. If one wants to find bias between two techniques of measurement, regression analysis should be performed. If the purpose is to see if one method may be safely replaced by another, especially in clinical practice, Bland-Altman plot is preferred. If there is no adequate reference standard to compare to, interchangeability can be used. This statistical method does not rely on a reference standard.

Contents

CHAPTER 1

Introduction

> From the beginning it seemed reasonable to suppose that bone age assessments were something a computer could do better than a human operator.
>
> *Tanner & Whitehouse*

1.1 Introduction

The goal of this chapter is to introduce the readers to the concepts provided in the thesis. This section presents the driving force as well the overall goal, methodologies, and contribution of the thesis. To supplement, we also include an overview of the structure of the thesis.

1.2 Motivation

Artificial intelligence (AI) in the clinical setting has become more prevalent than ever [HT17]. By deriving insights from vast amounts of data, AI has the potential to transform the healthcare sector. Computer-Assisted Detection (CAD) systems can solve specific radiological tasks with high efficiency. Nonetheless, AI remains an extension of human capabilities, not a replacement. A clinical decision, a diagnosis, is to this day still the responsibility of the clinical expert. Manufacturers make sure not to take away this duty by labeling their AI devices accordingly [ACR]. In certain difficult and time-consuming tasks, AI can provide a solution within a short amount of time. This might lead to a tendency of the user to overly rely on the output without confirming the results. Hence

the user will in some cases likely be unable to "catch" an error in output with the device. This poses a risk, as some therapies employed could be significantly life-altering for the patient. Therefore a high level of performance of the AI should usually be demonstrated.

According to the Food & Drug Administration (FDA), the standard paradigm suggests that an assessment of false positives (sensitivity) and false negatives (specificity) is sufficient to assure the safety and effectiveness of an AI medical device. Performance data of such cleared devices empower such assumptions [ACR].

Although these quantities apply to classification-based outputs, where the numbers of outputs are finite when dealing with outputs of continuous nature, where the number of outputs to be measured within an interval can be uncountable *e.g.*, age estimation, this is often not optimal. Attempting to group continuous variables into bins and assessing performance as if the outputs were of categorical variables will lead to an inaccurate representation of the performance results. This type of metric will introduce some arbitrariness to the analysis and hence the loss of information [Har20]. As a general rule, it is desirable to avoid techniques that introduce arbitrary assumptions, particularly in cases where alternative techniques are available to easily avoid these assumptions. Finding these methods though is a challenging task.

Thus, in this thesis we propose alternatives methodologies and a framework for statistical performance assessment of an AI in the form of agreement, bias assessment, and interchangeability and test our clinical as well as statistical assumptions on PANDA, a bone age AI model, whose clinical output is continuous, provided by the company ImageBiopsy Lab.

By providing a framework in assessing the performance of an AI bone age model we want to give researchers the ability and knowledge to demonstrate the safety and effectiveness of their device. In addition, we intend to consider a framework that is not limited to the assessment of the bone age model itself but also applies to other areas in clinical AI, whose outputs are continuous. For this theses, we can summarize the main research question as *Which statistical strategies support researchers in demonstrating safety and performance of an AI algorithm whose output is continuous?*

1.3 Contribution

Our work is motivated by the lack of clear guidance in validating artificial intelligence as medical devices. The current standard of practice as discussed in Section 1.2is inadequate to address AI software providing numerical outputs (*e.g.*, age, angles, distances). Therefore, our contribution to the state-of-the-art with this thesis is a workflow for the statistical assessment of AI solutions in bone age assessment

In this thesis, we will provide a framework for researchers with state-of-the-art statistical techniques. This selection of methods will support any researcher in the assessment of AI software independent of the type of the numerical output of interest. Special care is given

to the enrichment of the test data set and the comparison to reference standards, *i.e.* the ground truth, to avoid pitfalls when applying the proposed statistical methodologies.

1.4 Aim of Work

This work aims to define a statistical framework used by researchers on how to assess the performance of AI algorithms, whose output is a continuous variable. For this reason, we use PANDA, a bone age estimation software provided by ImageBiopsy Lab, as a reference device to test our proposed methodologies. We looked up similar procedures that have already been implemented for bone age and other approaches that might be applied interchangeably. Subsequently, we derived the following three research questions that are of interest for all medical experts when justifying good performance assessment practice.

Q1 How does the clinical aspect of the output of interest influence the distribution and granularity of the data set to be tested on?

Q2 How can the performance of the software whose output is continuous be shown and which choices of performance metrics and performance targets are available and feasible?

Q3 How can we estimate sufficient sample size and power?

To answer the questions above, we will focus on the following tasks in the remainder of this thesis.

T1 Understanding the principles of bone age assessment, its clinical output including the intervention of AI and the relevance of the clinical assumptions into the statistical considerations (**Q1**).

T2 Exploring the current methodologies of performance assessment of bone age and possible limitations (**Q2** & **Q3**).

T3 Proposal of an improved and more robust framework for performance analysis (**Q2** & **Q3**).

1.4.1 Requirements

The outcome of this thesis should not only incorporate the tasks as above but also include a set of requirements.

R1 Generalizability

While this thesis puts its primary focus on bone age, the framework proposed should not be limited to this specific output but should also apply to any clinical output of continuous nature irrespective of the use case. Therefore methods should be sought out that can be applied to a broader perspective and not limited to the performance assessment of bone age specifically. The implementation, *i.e.*, scripts and tools used should be created with this specific aspect in mind.

R2 Scalability

One of the final goals of AI in the medical setting is to find its way to commercial use in a clinical institute. Therefore the device must pass the regulatory barrier laid out by the US and Europe, known as FDA and MDR, respectively. The regulatory field enforces a harsh standard concerning adequate methodology and performance targets with hopes that AI/ML-based Software as Medical Device (SaMD) will deliver safe and effective software functionality that improves the quality of care that patients receive. The selection of methods and definition of targets to be met to keep both the personnel in the medical as well as in the regulatory field satisfied is what the thesis will aim for.

R3 Reproducibility

The underlying concept of the proposed framework must be designed in a way that allows researchers to reproduce the methods without being bound to a specific design.

1.5 Methodological Approach

Irrespective of the clinical output in question, many of the statistical considerations done are traced back to the underlying usage and clinical assumption made for the output of interest. In this case, we are looking at bone age specifically. Understanding the nature of clinical output serves as an initial basis for all further tasks and subsequently an overview of the testing data set on which is assessed on.

Therefore, the goal of T1 is to understand the underlying specification in a bone age assessment. Here we factor in considerations such as demographics (age, gender, ethnicity) and tackle the issue of generalizability of data mainly to the census population but also limitations of the AI Model PANDA. The result of this analysis yields an overall expected distribution of the testing data set and support the argumentation of the statistical method selection.

Based on previous research, we consider the Bland-Altman analysis and the concept of interchangeability for performance assessment. Orthogonal linear regression (OLR) is used to assess any potential bias. The reasoning for the proposed methodologies is because the core issue at hand is a situation of comparing two methods that assess the same output. The output of an AI is essentially a different method providing the same

assessment, *i.e.*, a new method against a reference method. It is inevitable that opinions of AI and medical professionals differs to a certain extent. The question here is what is the difference to be expected and to what extent is this difference acceptable? The solution here is to look into agreement and interchangeability, more specific differences to be expected between the two methods. By looking at the differences in outputs instead of binning results to perform a discrete analysis of sensitivity and specificity, we avoid the pitfalls of arbitrariness and loss of information. This presents a more accurate representation of the clinical performance. The related works section emphasizes emphasize the justification why these proposed methods are the most suitable for assessing the performance of this kind further and fulfill the requirement of generalizability (T2& Req. Generalizability).

Finally, based on our proposal, we prove our assumptions in form of a multicenter study. We involve a cohort of over 300 children and three expert pediatric readers, who establish the ground truth. The device in question, the bone age software, PANDA, is tested against the current reference standard which is the radiologist, the ground truth, themselves. We apply Good Clinical Practice (GCP) as expected from the medical industry, hence implementing the framework in a scalable manner (T3 & Req. Scalability).

The contribution complementing this thesis is a statistical analysis done via a python script illustrating the outputs as seen in the results section. The underlying concept of these scripts is further described in the implementation section. The proposed allows the medical researcher to perform a fast and in-depth solution for data analysis based on the presented framework in this thesis. Though for this thesis the programming language “Python” was the choice of implementation we ensure that irrespective of the language used, the underlying concept behind is presented in the manner, *i.e.*, pseudo allowing researchers to reproduce the methods proposed without any limitations (T3 & Req. Reproducibility).

1.6 Thesis Outline

We structure the complete thesis as the following: Chapter 2 provides an outline of the medical, technical, and statistical background. The medical background explains the intricacies of bone age assessment, different methodologies used in the clinical environment in more detail, treatments resulting from a bone age assessment including its limitations. While the technical background summarizes the device of interest PANDA provided by ImageBiopsy Lab and how its AI complements the workflow of the clinical setting, the statistical background creates a foundation of the underlying statistical methodologies used in the proposal for performance assessment of a bone age model. Coming up next, Chapter 3 emphasizes the recent works in performance assessment methodologies of AI bone age models, their current advances, and limitations from a statistical point of view. There we investigate why the statistical considerations made by many of the recent advances are lacking. Based on the outcome of the presented works and the issues at fault, Chapter 4 discusses possible solutions and how our approach differs in the

form of a study design. Here we start with the estimation of an adequate sample size followed by the acquisition of data including the establishment of ground truth. We then define and justify the statistical metrics and targets used to assess the performance of the clinical output. After deciding on the proposed methods, Chapter 5 outlines a detailed explanation of the implementation process that results in python scripts and tools used. Chapter 6 demonstrates the results and possible findings of the predictive power of PANDA against medical experts performed during a formal evaluation using the proposed methods. As a result, in Chapter 7 Finally, in Chapter 8, we summarize and discuss the conclusion of this thesis, validation of our proposal in the medical and regulatory field, and possible future directions this topic can improve on.

CHAPTER 2

Background

2.1 Introduction

In order to understand the reasoning that is further explored in the later chapters, a thorough understanding of the topics encapsulating bone age needs to be laid out. This consists of understanding the meaning of bone age, its relevance in the clinical setting, the current standard of practice, the demographics on where bone age assessments are performed and finally its limitations and how the intervention of AI supports the reader.

The factors mentioned above lay out the foundation for the statistical considerations and will answer why specific methods for performance assessment are more suitable than others and explain the logic behind the compilation of the testing data set.

2.2 Clinical Background

As part of Task T1 as outline in Section 1.4, one needs to understand the clinical implications of the output provided by the AI to correctly define the distribution and granularity of the data set to be tested on. At the end of this section, the proposed solution will answer the issue concerning Research Question Q1.

2.2.1 Bone Age

Bone Age is an assessment of skeletal maturity typically based on the radiographs of the left hand and wrist and is a routine procedure in pediatric radiology departments [Mar11]. Bone age assessment relies on the predictable changes of ossification centers over time. The hand, in particular, contains many of these ossification centers for which their progression over time can be tracked by radiography. Abnormal growth is determined by comparing the bone age to the chronological age of a child and can be an indication for several conditions, such as growth hormone deficiency or hypothyroidism.

The most common methods to assess bone age are the Greulich-Pyle (GP) method, typically used by US pediatric radiologists and endocrinologists [GP50], and the Tanner-Whitehouse (TW) method, more commonly used in Europe [TW83]. The GP method is atlas-based and involves a comparison of the whole hand morphology to the typical morphology of different developmental stages. The TW method is more elaborate and involves the scoring of individual bones for maturity indicators, after which a bone age can be derived from the sum of the scores. The two methods are based on two different study populations: The GP method was originally based on an American population of high socio-economical status in the 1940s, while the TW method was originally based on a Scottish study population of low socio-economical status in the 1950s. A large-scale study found that the mean difference between the two methods is 0.39 (-2.24; 2.18) years, with the (TW method assessments being on average slightly higher than the GP method [BEK+99]. Regardless, even though the (TW method is more fine-grained, both methods are in good agreement with each other and, in part due to its simplicity and being less time-consuming, a majority of radiologists adopt the GP method [CSI17, KSO+94].

The measurement of bone age according to GP is intended to be performed on the non-dominant hand. Based on the atlas of GP, the left hand was used as the preferred way of measurement. This is due to most of the population being right-handed. This results in making the left hand more suited for analysis, as the less dominant hand is less frequently used and therefore less likely to be maimed or in any way injured. The overall concern regarding the use of radiographs of left or right hands is whether the bone ages estimated in each of the two hands of an individual child are sufficiently close so that the same estimate of bone age can be derived from either of them. The GP atlas provides references to support the conclusion that discrepancies between the two sides are too insignificant to constitute a source of error in the determination of skeletal status, thus supporting the evaluation of either hand. The decision of whether to use the left or right hand for an individual patient is left up to the physician ordering the hand bone age estimation [GP50].

2.2.2 Clinical Use - Bone Age Assessment of Radiographs

The main use of Bone Age assessment in a clinical setting is to evaluate abnormalities in development by comparing the difference of skeletal bone age to the chronological age (birthdate) of a child [BSY+07]. The assessment of chronological age is a matching process, comparing the radiograph of a subject to a defined reference that involves a sample of known sex and age [BEK+99]. The process of age estimation is a measure of biological maturity that is converted to the chronological age by comparison with a reference [Fle32]. A delayed bone age (corresponding to a younger chronological age) can be an indication of several conditions such as growth hormone deficiency, hyperthyroidism, and malnutrition. Advanced bone age is associated with elevated sex steroid levels, which happens in precocious puberty or congenital adrenal hyperplasia. Several genetic overgrowth syndromes, such as Sotos syndrome, Beckwith-Wiedemann syndrome, and Marshall-Smith syndrome, are associated with significantly advanced bone age [CSI17].

2.2.3 Ethnic Variation in Skeletal Maturation

Both GP and TW methods for determination of bone age were first defined in a primarily white population, albeit in the 40s and 50s, raising questions about the applicability of these assessments to different ethnicities. Consequently, several studies have assessed bone age in populations of diverse ethnic backgrounds, by both GP and TW methods. A recent study, using data collected from the Los Angeles Children's Hospital, reports the differences between bone and chronological age stratified by ethnicity. It found statistically significant differences for the Asian and Hispanic groups, as well as for white females, but no significant differences for either the African American group or white boys. Regardless, these differences were all less than 0.3 years, which is unlikely to be clinically significant. An age-stratified analysis reveals that bone is significantly overestimated for Asian and Hispanic children, especially in girls between the ages 10-13 and boys between 11-15 years old [ZSV+09].

2.2.4 Limitations of the GP Atlas

To summarize, the GP Atlas:

- Is a set of reference images over a range of ages (31 reference images for males and 26 for females)
- Standards are derived from a study of healthy white middle-class children in the Cleveland area from 1931-to 1942
- Reference images range from 3 months to 18 years for girls and 19 years for boys

Disadvantages of the method are that it is subjective and there are long intervals between reference standards (some more than others). Also, it only includes a very specific patient group as bone age is influenced by gender, race, living environments, social resources, and nutritional status [TFRSPdlC+07]. Practitioners are aware of its limitations and drawbacks but are also aware of its widespread adoption and clinical use as a standard for bone age assessment.

2.2.5 Role of Artificial Intelligence (AI) in the Assessment of Age Estimation

Bone Age is an assessment of skeletal maturity typically based on the radiographs of the non-dominant hand and wrist and is a routine procedure in pediatric radiology departments where bone age assessments are compared to the chronological age in light of detecting any potential endocrine and/or metabolic disorders. While this process is incremental to many disease evaluations, since the introduction of the still current golden standard by Greulich & Pyle over 60 years ago, little has changed to improve this tedious process. Nonetheless, the "Radiographic Atlas of Skeletal Development of the Hand and Wrist" by William Greulich and Idell Pyle is considered the standard in determining skeletal age in children [GP50]. Other methods, due to their technical complexity in

acquisition [PBM18, DTBP+20] or interpretation [TW83], are used less often in practice [BTSK16].

Bone age assessment or determining a "skeletal age" is a sophisticated task complicated by the complexity of evaluating the wide variations in bone mineralization tempo, shape, and size encompassed in the large number of ossification centers in the hand and wrist and hence requiring one to account and weighting for these multiple factors and as such is not a simple quantitative measurement and much more complex task [SRGO06]. As such, this type of assessment is very much prone to inter-and intra-rater variability [MIH+13]. In addition, depending on the experience and training of the radiologist, the process of estimating skeletal age based using the GP atlas can be quite time-consuming and also to a certain degree subjective.

To help radiologists handle these cases, many neural networks using AI algorithms from deep learning, image classification, and object recognition can provide imaging studies with meaningful results with almost the same accuracy as highly trained pediatric radiologists Several CAD systems already exist for the determination of bone age [MIH+13, BYW+20]. Recently, a new CAD system called PANDA was launched, which is built on artificial intelligence [IB]. The deep learning algorithm was trained with x-rays and the corresponding bone age estimation according to GP.

However, AI algorithms are known to have numerous limitations where incorrectly flagged studies might even increase the workload of radiologists and pose a risk to the patient getting treated. Given the intended use of having the bone age determined automatically in the clinical setting, the heavy reliance on the outputs, the challenging tasks, and the associated risks, a very high level of performance should be demonstrated to assure the safety and effectiveness of the device.

2.2.6 Enrichment of Testing Data Set by Clinical Indications Relevant for Bone Age

Based on the assessment of the previous chapters, the relevance of the specific demographics listed below are to be considered when sampling the data:

1. Age
2. Sex
3. Ethnicity

From the clinical perspective, the performance testing data set should contain enough granularity in terms of age and sex for assessment based on the intended patient population and intended use of PANDA. While ethnicity seems to be another important factor in a bone age assessment, the GP for estimating bone age from a hand radiograph does not rely on ethnicity. Rather, patient ethnicity may be taken into consideration by

the radiologist or referring physician when interpreting the bone age for the individual patient. Disadvantages and limitations of the GP method are well known as presented in the section2.2.3. Nevertheless, the GP method is accepted by radiologists and is still considered one of the gold standards to this day [PPMDM+20]. As such, while one must ensure that multiple ethnicities are represented in the performance data set, special considerations are not required.

Summary

To sum it up, bone age is assessed differently for male (m) and female (f) in every stage of age from 3 months to 18 and 19 years, for girls and boys, respectively. As such, the testing data sampled must ensure sufficient granularity based on the indication of PANDA's patient population regarding age and sex, which are two parameters relevant for bone age assessment.

2.3 Technical Background

This section addresses the device in question that is used to validate our statistical assumption, how the software is used in the clinical setting, and what role the medical device is fulfilling.

2.3.1 PANDA - Pediatric Bone Age and Developmental Assessment

PANDA is a fully automated, radiological image processing software intended to aid medical professionals in the estimation of pediatric bone age according to the GP method on non-dominant hand radiographs of children [IB]. The usage of PANDA is limited to bone age estimation of children aged between 24 months and 192 months (girls) or 204 months (boys). Manual estimation by comparing digital radiographs with reference images in the GP atlas is tedious and suffers from a high degree of inter-rater variability. PANDA facilitates the radiological evaluation of the bone age, where physicians can predict conditions that affect the growth of children in a consistent way. PANDA provides a swift automated method to estimate bone age. The device performs an assessment based on the whole image of the single hand radiograph, which is fed into a convolutional neural network and outputs a single bone age value. As a result, all bones visible on the radiograph, including the carpal bones are taken into account by the model.

The bone age estimation is presented on a PANDA report along with the

- chronological age (CA) derived from the DICOM input image
- difference between bone age (BA) to chronological age (BA-CA)
- standard deviation (SD) from the Brush Foundation tables based on the CA
- assessed image along with overlays indicating the assessed region.

The outputs can be viewed on any legally marketed DICOM viewer workstation. PANDA operates in a Linux environment and can be deployed to be compatible with any operating system supporting the given virtualization such as the third-party software "Docker".

2.3.2 PANDA in the Clinical Workflow

Patients with a clinical suspicion of having a growth disorder, such as delayed or advanced development, are referred by a pediatrician or pediatric endocrinologist for a PA radiograph of the non-dominant hand for initial screening of hand bone age. The resulting radiograph is saved to the PACS.

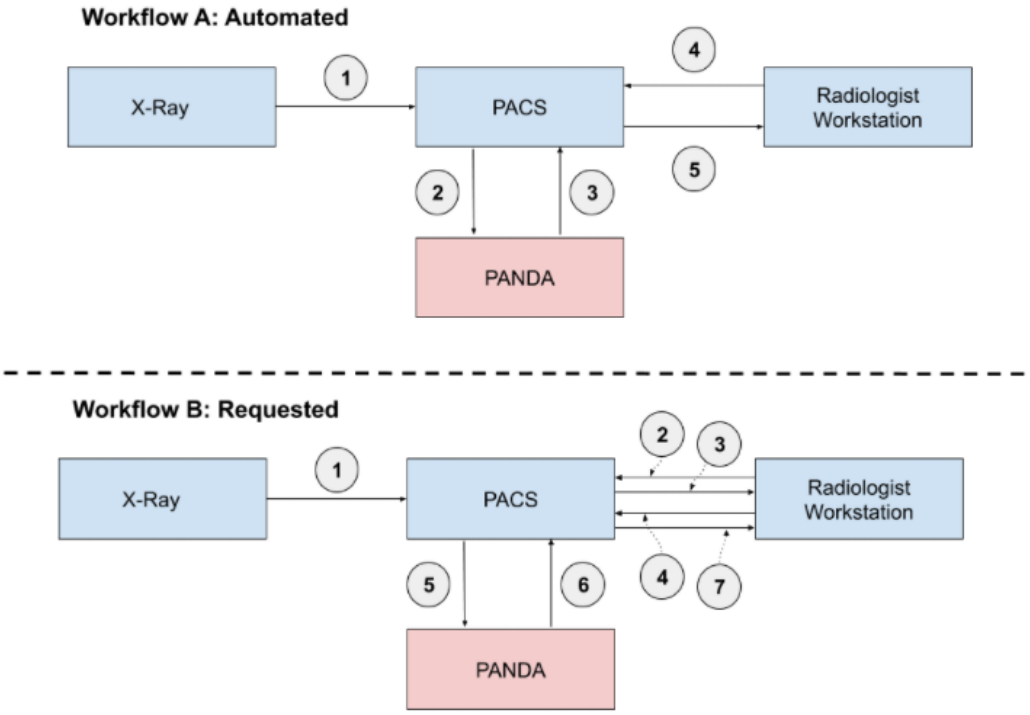


Figure 2.1: In the automated workflow A, the hand radiograph is automatically sent to PANDA, and the radiologist sees the PANDA results when they open the study for viewing. In the radiologist request workflow B, the radiologist requests a PANDA assessment while viewing the hand radiograph and receives the results in the radiology workstation alongside the original study [IB]

From this point, there are two ways for the radiologist to use PANDA:

Workflow A: Automated

1. An x-ray is sent to the PACS.

2. The PACS is pre-configured by a technician of the clinical environment to automatically direct PA hand radiographs intended for bone age assessment as DICOM files to PANDA without the radiologists in between to run the software.
3. PANDA returns its outputs as DICOM files to the PACS, and they are stored with the original study.
4. The radiologist requests the PACS to view the study on a radiology workstation.
5. The study is sent from the PACS to the radiology workstation, and the PANDA output(s) are viewable by the radiologist along with the original DICOM.

Workflow B: Radiologists request

1. An x-ray is sent to the PACS.
2. The radiologist requests the PACS to view the study on a radiology workstation.
3. The study is sent from the PACS to the radiology workstation.
4. The radiologist sends a request to the PACS to send the DICOM to PANDA.
5. The study is sent from the PACS to PANDA.
6. PANDA sends its result(s) back to the PACS.
7. The PANDA outputs are sent to the radiology workstation for viewing alongside the original DICOM.

In both cases, PANDA receives radiographs from a PACS, analyzes the images, outputs the bone age as a result, and stores the analysis results in the PACS as a DICOM file attached to the original study. Once PANDA receives the image, the analysis and the generation of reports are fully automated and entail no user interaction. The user can send images to PANDA via standardized DICOM commands or the file interface and receive reports over the same interface. In this sense, the user does not "operate" the device but simply reviews the reports presented to them and can accept or reject them within their standard reading workflow.

Summary

To summarize, PANDA automates the current manual process of the Greulich & Pyle bone age estimation method of hand radiographs. The usage of PANDA is limited to bone age estimation of children aged between 24 months and 192 months (girls) or 204 months (boys). As fixed age interval is fundamentally a limitation, the testing data set must factor in the intended use population. In addition, based on how PANDA is integrated into the clinical workflow of the radiologists, AI essentially is giving an extra opinion. To be more accurate, both, AI and the medical expert, are using different

methods to assess the same output, which in this case is bone age. Moving forwards, the question then would be a matter of comparing the two methods with each other and assessing whether the differences to be found are clinically relevant.

2.4 Statistical Background

2.4.1 Introduction

So far we have established that the role of AI in the clinical setting is automating part of the workflow within a diagnosis done by the healthcare professional. AI achieves its purpose by estimating the clinical output of interest using different techniques compared to the manual standard of practice. Strictly speaking, it is a comparison of methods. For bone age specifically, the workflow conducted by the radiologists is the established method for assessing the skeletal age of a child. However, with the latest advancements in AI, it is worthwhile to investigate whether AI can yield comparable results to a human expert. This can minimize the time and effort spent by radiologists. Therefore, it is of great importance that the AI agrees with the current standard, which has to be backed with evidence.

2.4.2 Bland-Altman Analysis

The Method Comparison Problem

In clinical practice, medical professionals often wish to assess or measure quantities such as cardiac stroke volume, blood pressure, bone age that is in many cases extremely difficult to measure directly without any adverse effects on the subject. As such an evaluation usually involves indirect methods of measurement or assessment. When new methods are proposed the evaluation is done by comparison to the established method rather than the true quantity. The established method (standard method) is also known as the "gold standard" but should not imply that the measurement was done without errors. This is because we cannot be certain that either method reflects the actual result. For bone age specifically, the radiologist is the established method for assessing the skeletal age of a child. We also know that the specialist is prone to inter-and intra-rater variability [MIH+13]. Nonetheless, today, the specialists are still considered the reference standard. In such cases, we want to see whether these methods are comparable. There will inevitably be a lack of agreement to a certain extent. What matters is the amount of disagreement,*i.e.*, how much the new method differs from the reference method. To specify, the differences generated from the two methods must be within an acceptable range from the clinical point of view. If this assumption-based clinical interpretation is full-filled, the old method can be replaced by the new method [BA99]. For example, if the differences of an automated bone age assessment via AI are within an acceptable threshold, we can rely on the assessments done by the model, as a difference smaller than the threshold would not affect clinical decision-making.

The Bland-Altman Plot

The Bland-Altman plot is a simple and very powerful method to compare two measurement techniques, where measurements/assessments of the same output on two different methods are compared. The Bland-Altman plot is a good way of showing where any disagreements occur [Gia15].

The underlying principle is based on comparing two measurements, X and Y that are assumed to be the same. We plot a graph of how the difference between two measurements (X - Y) varies with the 'true' measurement. Quite often we do not know the 'true' measurement, as it is usually unknown which of the two measurements is correct, so in the absence of a true measurement, we use the average of X and Y as the best estimate of the true measurement.

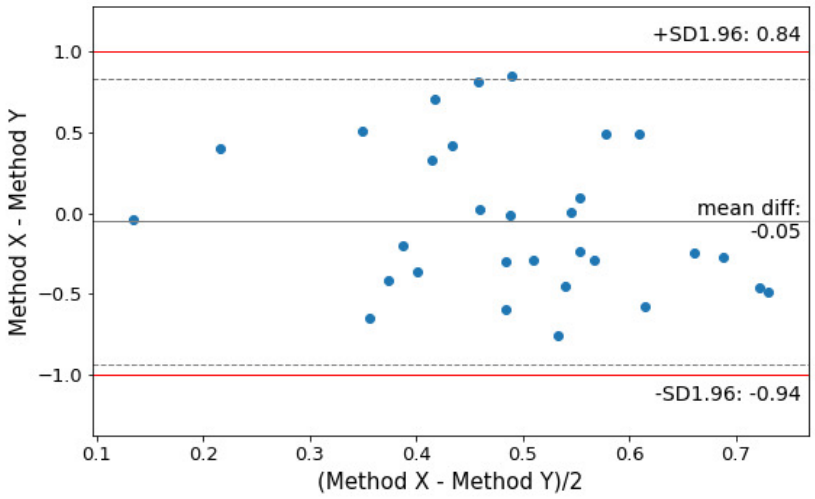


Figure 2.2: Example of the Bland-Altman-Plot computed in Python's MatplotLib - The x-axis indicates the tentative "true" measurement of the output between two measurements. The y-axis indicates the differences between the two methods. The black line displays the mean of differences (reflecting the absolute bias). The dashed line shows the Limits of Agreement. The red line the maximum acceptable limit

In this graphical method, we graph the difference (X - Y) on the vertical axis and the average (((X - Y)/2) on the horizontal axis in form of a scatter plot as seen in Figure2.2. To support interpretation, three types of additional lines are included as displayed in Figure2.2.

1. **Black Line - Mean of differences** - This indicates the expected bias when using the new method over the reference method.
2. **Black Line dashed - Limits of Agreement** - Confidence limit of agreement, shortened limits of agreement. Data points along the y-axis encapsulate 95% of the differences resulting from using the new method over the reference method.

3. **Red Line - Clinically Acceptable Threshold** - Acceptable limits defined a priory, based on clinical assumptions or standards or literature. This threshold presents the maximum difference one can accept resulting from using the new method over the other.

If the limits of agreement are within the clinically acceptable threshold, the new method is in agreement with the reference method.

2.4.3 Interchangeability

The concept of interchangeability stems from the idea of biosimilarity, a term relevant in the pharmaceutical industry, where equivalence between a test drug to a reference drug is established [OSS14]. FDA defines an interchangeable product as one that may be substituted for the reference product without the intervention of the health care provider who prescribed the product [Com]. This means drugs are considered interchangeable when the same or similar results can be produced within the same patient. To show interchangeability we compare the differences between the test and reference drugs to differences between two responses with the reference drug.

To simplify:

Excess/Reduction of Differences =

Difference (New Drug Reference Response 1 and Reference Drug Response 2) - Difference (Reference Drug Response 1 and Reference Drug Response 2)

where

Excess/Reduction of Differences < Predefined Acceptable Limit

The result is either an excess or reduction of differences, which is compared to a certain limit defined during the planning phase of a study. If the result is below the Predefined Acceptable Limit, the new drug is considered interchangeable if not superior to the reference drug.

The concept of interchangeability is currently applied by the FDA to drug testing, but can also be applied in multiple areas [Com] such as in the field of imaging. The biggest advantage of interchangeability is the independence of requiring a reference standard to compare to [OSS14]. A "true" Ground Truth will not always be available and may not be necessary to determine. While a reference standard as required by Bland-Altman is not always easy to define, the concept of interchangeability does not rely on any reference standard, but the differences created by one method over the other.

2.4.4 Sample Size and Power Study

Studying the entire population in any research is neither realistic nor viable. That is why an often large set of representative individuals is selected from the population, known

as a sample. In any clinical study, it is essential to determine the appropriate number of subjects during the planning stage of a study to provide a reliable assessment with a certain degree of statistical significance [Len01]. The importance of sample size stems from economic and ethical reasons. Undersized studies might not have the capability to produce relevant results. The lack of samples can inhibit the detection of an effect or differences between two methods when one should exist [Alt90]. Therefore, the study might result in a waste of resources when exposing subjects to harmful treatments without advancing knowledge. On the other hand, having more samples than required, an unnecessary number of subjects are potentially exposed to harmful treatments. Therefore an optimized sample size is very important.

A lot of techniques for sample size calculations are described in most conventional statistical textbooks and a growing amount of software using formulas. All these equations require one to have some idea of the results expected in a study. To simplify, the majority of formulas expect the following assumptions, resulting in four basic components:

1. **Type 1 error (alpha)** - The probability to get a false-positive result
2. **Power or Type 2 Error (beta)** - The probability to get a false-negative result
3. **The smallest effect of interest** - Difference between the studied groups one wishes to detect; mean of the sample differences.
4. **The expected variability** - Expected standard deviation of the sample differences

The estimated values are often based on either convention (related to Type 1 and Type 2 errors) or assumptions from previous literature or performed pilot studies (related Effect Size). To estimate these components, we elaborate on these four parameters and put them into perspective relevant for AI.

Principles of Hypotheses Testing

Quantitative research is driven by research questions and hypotheses, known as alternative hypotheses. Every alternative hypothesis is followed by a null hypothesis. The null hypothesis does not need to be explicitly stated because it is always the opposite of the hypothesis. The alternative hypothesis will in many cases be about some form of relationship or effect/difference between variables. The null hypothesis claims that the variables being tested are not related and show no effect/difference, that the results are the product of random chance. To demonstrate that a hypothesis is likely true researchers need to compare it to the opposite situation, *i.e.*, to reject the null hypothesis.

To put this into perspective, if a researcher asks the question "Is there a difference between the assessment of bone age between AI and the reference standard?" The alternative hypothesis would indicate: "Automated analysis with AI and the expert radiologists are in agreement." Therefore the null hypothesis would be that "Automated analysis with AI and the expert radiologists are not in agreement."

Assuming we want to back up the alternative hypothesis, we have to refute the null hypothesis. Instead of trying to prove the alternative hypothesis, the researcher must show that the null hypothesis is likely to be wrong.

Type I and Type II Errors

The null hypotheses is	True	False
Rejected	Type I error (False Positive)	Correct Decision
Not rejected	Correct Decision	Type II error (False Negative)

Table 2.1: Type 1 Type 2

The chance for an error can occur in every study. There are two major types of error – Type I and Type II errors as seen in Table2.1. Both forms of errors are concerned with the researcher's potential for making mistakes. A Type I error occurs when the researcher mistakenly rejects the null hypothesis. For example, even though treatment is not effective or shows any sort of difference, we say it is. If the null hypothesis is rejected it means that the researcher has found a relationship among variables even though there is none. So a type I error happens when there is no relationship but the researcher finds one.

A type II error is the direct opposite. It occurs when the researcher mistakenly accepts the null hypothesis. For example, a treatment is effective, but the researcher could not find the evidence. If the null hypothesis is accepted it means that the researcher has not found a relationship among variables, when one exists. So a type II error happens when there is a relationship but the researcher does not find it.

Although it is hard to predict when an error will occur, researchers can reduce the chances of making a mistake when making statistical decisions. Statistical considerations that are used to calculate the required sample size for research are linked to the chance of making an error. Researchers must evaluate the required power, estimated effect size, and acceptable significance level when establishing a sample size.

The Level of Significance α

The statistical significance of a study's findings is used to estimate how likely the results are due to chance. The α level is decided before a study and is set to a value of the maximum mistake rate that a researcher is ready to accept. To be more precise, assuming there is no difference to be detected, meaning the null hypothesis is true, how often will the researcher say otherwise, that is providing a false-positive result. Typically 0.05 is set as a very common convention, which suggests that if the null hypothesis is true, it will be rejected in only 5 out of 100 cases [CBB09, CA15].

Power 1 - β

The possibility that the researcher would correctly reject the null hypothesis when it is false, thereby avoiding a type II error, is known as power. It refers to the likelihood that your test will detect a statistically significant difference if one would exist. The lesser the probability of a type II error, the greater power research has. A type II error is more likely when power is low. Power increases as the sample size grow because more information is collected, making it simpler to reject the null hypothesis accurately. Usually, power is set at 0.8 or greater before a study begins, meaning that you should have an 80% or greater chance of finding a statistically significant difference when there is one [SH20, CBB09].

Effect Size - Minimum Difference to be Detected

The smallest clinically meaningful difference that would be valuable to identify as significant in the trial should be stated clearly by the researcher before the investigation begins [Alt90]. Larger minimum effect sizes or differences indicate the possibility that the underlying populations are separated, making it simpler to detect significance and increasing the study's power. Smaller minimum effect sizes, on the other hand, indicate that the underlying populations may overlap, and the power to detect the difference as significant will be much more difficult as type II error increases. For example, the more effective a tested treatment is, the smaller the sample size needed to detect a positive or negative effect [CSCYS21, KB10].

Variability

Sample Size is affected by the standard deviation of the population from which data is collected. A statistical test is more likely to identify a significant difference for tightly distributed data, with a small standard deviation, than for loosely distributed data, with a large standard distribution that results in greater possibilities for overlap of the distributions [KB10]. In many cases, similar to the relevant minimum difference to be detected, the population statistic may not be known before doing research. Estimates might be derived from already available data from literature or by performing a pilot study to get an estimate of such reference values.

Summary

Potential methods for assessing the performance of an AI of continuous nature include assessing the differences between the new method and the reference standard and whether these differences are acceptable. Bland-Altman analysis of agreement and the method of interchangeability is a few of the known methods. In addition to the applicable methods, the number of samples necessary to be analyzed needs to be assessed. The sample size is closely tied to statistical power and significance. A power analysis is most often used to calculate what sample size is needed. Assuming three of the four values out of the parameters — sample size, effect size (mean difference and standard deviation), significance level, or power — are known, the final parameter can be calculated. Since α is usually 0.05 and power 1-β is usually 0.8, researchers need to pay the most attention

to the effect size to calculate the needed sample size. Without an adequate number of samples, irrespective of the method used, because the test is not powered, we cannot conclude any statistical significance.

CHAPTER 3

Related Work

3.1 Introduction

After understanding the clinical aspects of bone age assessment we now explore the current practice of statistical methodologies used to assess the performance of AI models estimating bone and critically review the plausibility of these studies. This chapter presents the state-of-the-art solution and describes already presented solutions, and their shortcomings within the context of this thesis. First, initial approaches to establish a comparison between two methods will be presented and their deficiencies will be discussed. Based on the method of agreement, proposed by Bland-Altman, a more detailed description of other works will be reviewed and critically assessed. Finally, we assess the method of interchangeability based on the concept of bioequivalence and carry its methodology over to the concept of AI imaging. The exploration and analysis of publication in this field will provide a starting point from which improvements to the final framework can be made.

3.2 Comparison of Means

Initially, when we compare different methods, Method A and Method B, we derive measurements based on both methods. One method used to assess agreement was proposed by Cater by comparing means [Cat79]. Cater examined two approaches for measuring the gestational age of human babies. Gestational age was calculated from the last menstrual period but can also be derived from a score, the total maturity score (TMS), based on external physical characteristics. The author applied both methods on multiple groups of infants, stratified by the birth weight, compared the mean of each group using an unknown significance test, and concluded that the TMS is a convenient and accurate method of assessing gestational age in term babies. According to his benchmark, agreement between two methods is given when the result delivers the same

mean measurement. This method of comparison informs little about the methodologies' accuracy. Even though having similar means is a necessary condition it is insufficient to claim agreement between two methods.

To put this into perspective, we simulated two sets of measurements as seen in Table 3.1. Based on the assessment via using the mean method resulting in an average of four, one would think, Method A and Method B are equivalent. In addition, we included an analysis of significance to further emphasize the misconception. The histograms of our data as seen in Figure 3.1 is not normally distributed. Both of the variables violate the assumption of normality. This means, the conventional paired t-test is not applicable and one should use a different test to analyze this data. An alternative to the paired sample t-test is the Wilcoxon signed-rank test. The results indicating a p-value of 0.97 indicate no significant difference between both measurement methods.

The measurements displayed in the Table 3.1 demonstrate that measurement done on these paired subjects varies significantly. The histograms of both measurements (Figure 3.1 are very different which gives the sign that perhaps measuring means is not enough. Bland and Altman criticized this method as measurement errors were not considered in this assessment [MWN+21].

Measurement	Method A	Method B
1	0	3
2	8	4
3	6	4
4	0	6
5	1	3
6	7	4
7	8	3
8	1	5
9	0	5
10	6	4
11	7	3

Table 3.1: Simulated Measurements - Method A and Method B

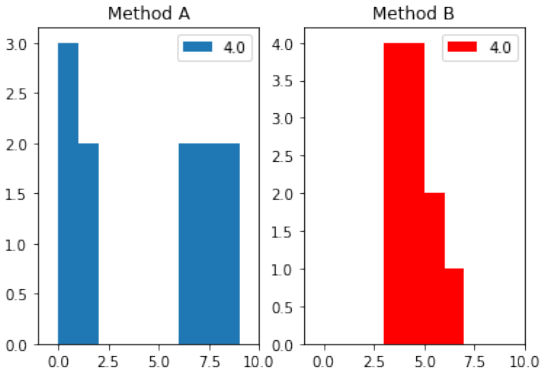


Figure 3.1: Simulated measurements done by Method A and Method B. The histogram of the measurements clearly shows that the measurements are not in agreement even though the mean is the same.

3.3 Correlation

Method comparison studies have also been analyzed using correlation coefficients. The correlation coefficient gives a value that describes any relationship between variables. Many studies have used the different techniques of the correlation coefficient to claim whether a method is in agreement or not. Keim et al. assessed two methods to measure the stroke index, a value analyzing the cardiac stroke volume [KWT+76]. The authors wanted to assess whether the stroke index obtained with dye-dilution techniques can also be obtained by impedance cardiograph. Based on their assessment with the correlation coefficient with ($R = 0.49, n = 122$, p-value less than 0.001) the authors concluded non-agreement of the two methods.

Serforntein et al. analyzed two different methods based on a scoring system to see whether the estimation of gestational age at birth agrees with each other [SJ78]. The authors claimed based on the correlation coefficient of $R = 0.85$ that these methods can be used interchangeably. One can see from the results as listed in Figure 3.2that using one method over the other would result in not only a fixed but also a proportional bias.

Other similar attempts to present evidence for method comparison were assessed by Laughlin et al. [LSF80] and Hunyor et al. [HFC78]. The authors investigated different methodologies to assess blood pressure. While Laughlin et al. presented evidence of whether clinical blood pressure monitors can be replaced with home blood pressure monitors, Hunyor et al. investigated seven different types of blood pressure (devices measuring blood pressure). Both presented evidence using the correlation coefficient. Based on the results as listed in Figure 3.3 Laughlin et al. presented low correlation, therefore, indicating that clinical blood pressure monitors cannot be replaced with home blood pressure monitors. Similar results based on correlation have been reported by Hunyor et al. testing the replaceability for seven different blood pressure monitors.

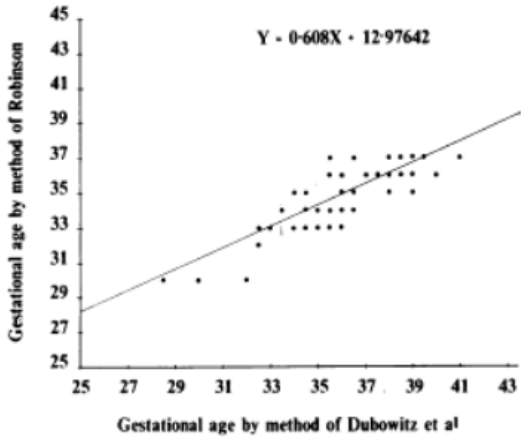


Figure 3.2: Serforntein et al. compared two methods based on a scoring system for estimating gestational age using correlation [SJ78]; The plot shows a presence of both, fixed and proportional bias. As such, a high correlation of $R = 0.85$ does not necessarily mean good agreement when comparing methods against each other.

Table 2. Correlations between home and clinic blood pressures on visit days

Visit day	*N*	Systolic	Diastolic
1	57	+0.69	+0.63
2	60	+0.83	+0.55
3	56	+0.68	+0.48
4	48	+0.66	+0.37

Figure 3.3: Laughlin et al. assessed home and clinical blood pressure monitors are replaceable and presented evidence against this by showing a low correlation between these two methods when measuring the systolic and diastolic blood pressure [LSF80].

However, the usage of the correlation coefficient is not always appropriate, as it measures the strength of a linear relationship between two methods, not the agreement between them. From the logical point of view, two approaches measuring the same thing but in a different way are expected to result in a good correlation anyway. The correlation coefficient will always be close to $r = 1$ and always be significant. One can get a good correlation even if the two measurements disagree (refer to Figure 3.2).

To put this into perspective, Figure 3.4 shows the comparison of two methods measuring systolic blood pressure. A correlation of $R = 0.94$ and $R^2 = 0.88$ would insinuate good agreement due to an excellent correlation. The plotted trend line through the data indeed shows a good linear relationship between these two measurements. However, when looking at the scatter plot, we can see that measurement 1 is almost always larger

than measurement 2, which indicates that perhaps this method of evaluation also is not sufficient to allow us to conclude whether the two measurements are in agreement. Assuming the two methods are in agreement, the points on a scatter plot of the two methods must lie close to the line of equality, not just close to the line of best fit. Hence, the correlation coefficient is not a measure of agreement; it is a measure providing information regarding the strength of a linear relationship.

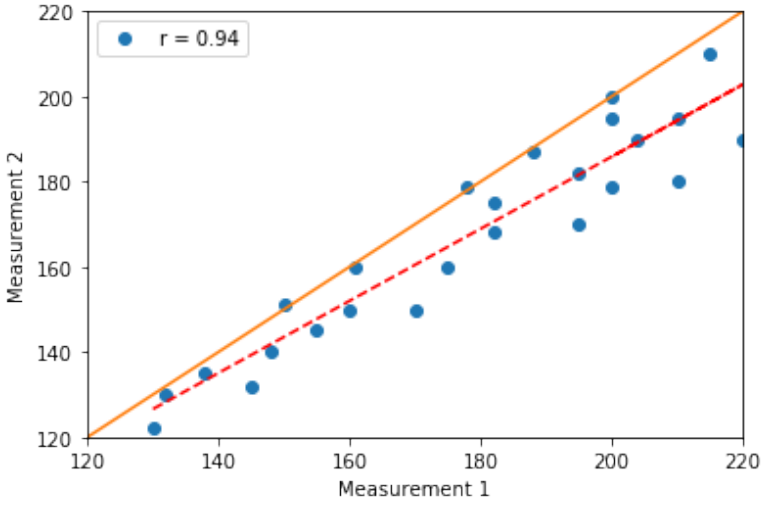


Figure 3.4: Comparison of two methods measuring systolic blood pressure; Data were taken from Bland and Altman [MWN+21] and generated using Python's Matplotlib; The red-dotted line displays the line of best fit to the data with $R = 0.94$. The yellow line indicates the line of equality with equation $y = x$.´For every point in y an equal point in x is followed as well. There is a clear difference between the line of best fit and the line of equality. Therefore, agreement cannot be claimed.

3.4 Regression

Generally, a study of linear regression is done in parallel with correlation analysis. Based on the reasoning made in the previous section, assessing points in relation to the line of identity would be more suitable than assessing agreement just based on the line of best fit. The results of one measuring method are directly connected to the results of another. As a result, it appears natural that linear regression analysis would be a valuable tool for comparing measurement methods. Regression analysis employs correlation ideas, but it goes beyond describing the degree of strength between two variables [Dog18]. Simply, the slope and intercept of the regression line can be investigated to determine if the parameters are near the line of equality, *i.e.*, a slope of one and an intercept of zero. To specify, regression coefficients of the scatter plot will provide more information about agreement because assuming measurement x and y are the same then we expect the

intercept to be zero and the slope is equal to one. This strategy, however, is not without flaws as the result from such analysis can be at times misleading.

Standard linear regression analysis known as Ordinary Least Squares (OLS) regression assumes the dependent variable is measured with error, while the independent variable is not. Depending on the choice of the dependent variable, this results in two alternative lines of best fit, which in some cases can give two substantially different approximations (refer Figure 3.5) The magnitude of the difference between the two divergent lines increases as the correlation between the two variables decreases [Hol96]. This can be avoided by employing a method that assumes errors in both variables and yields a symmetrical solution, such as Deming regression.

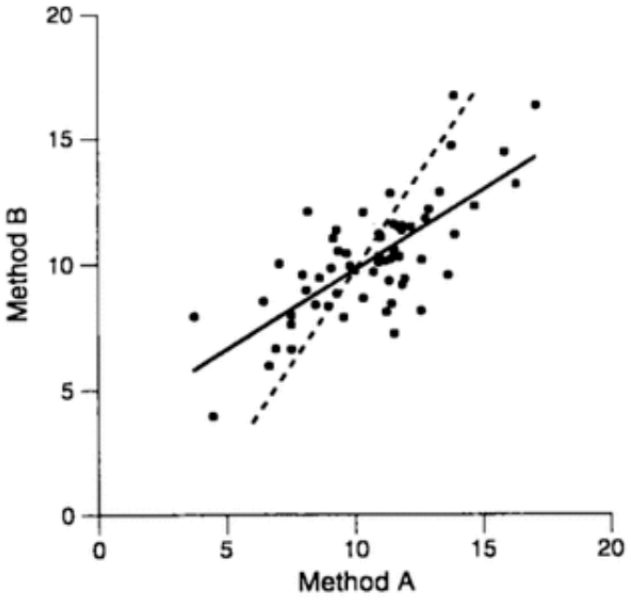


Figure 3.5: Regression lines using y as the independent variable (solid line) and x as the independent variable (dotted line) [Hol96].

In addition to the misuse of models for linear regression, difficulties also arise because regression does not yield quantified values for disagreement making parameters difficult to interpret when claiming for interchangeability. A slope of 1 and an intercept of 0 merely indicate whether a proportional or fixed bias exists. Concluding agreement based on the existence of bias is not sufficient. This does not argue completely against performing regression analysis. According to Ludbrook and Bland and Altman, if the investigator wants to calibrate one measurement against another or find bias between two techniques of measurement, regression analysis might be performed. However, if the purpose is to see if one procedure may be safely replaced by another, especially in clinical practice, the information provided in a regression analysis is insufficient [Lud09, BA86]. Agreement is not present or absent, but something that must be quantified. This is where the method

of difference, the Bland-Altman method, comes into play.

3.5 The Analysis of Differences with the Bland-Altman Approach

Assuming that the goal is to assess whether one clinical method is interchangeable with the other one should statistically study the differences of one measure over the other. Ideally, that model would yield identical findings acquired by two different methods, meaning all differences would be equal to zero. However, each measurement is accompanied by some degree of error. This is especially true, for a subjective measurement such as bone age, where even within the reference standard, *i.e.*, the radiologists show great variability among themselves [MIH+13]. This is why it is important to evaluate the magnitude of such differences. The general framework of the Bland-Altman technique has been outlined in Section 2.4.2. To summarize, the Bland-Altman Plot simply graphs every difference between two paired methods against the mean of two measurements and allows one to investigate any possible relationship between error in measurement to the true values and conclude whether the expected differences exceed any clinical relevance to determine whether the reference and the new method are interchangeable.

Although the Bland-Altman might be an intuitive method, like every other statistical method, certain boundary conditions need to be fulfilled [Dog18]. Other additional requirements might be necessary depending on the clinical use case to be addressed to avoid pitfalls when performing studies of this kind. Generally, the following things need to be addressed:

1. Assumption of normal distribution of differences
2. Adequate sample size
3. Clinical and statistical relevance

3.5.1 The Assumption of Normal Distribution

First, the requirement to fulfill the assumption of an approximation to a normal distribution is one of the fundamental challenges in the Bland-Altman analysis. The continuous variables of paired measurement themselves need not be normally distributed but their difference approximate one [Dog18]. To elaborate, the distribution of the data for bone age from the new and reference method does not necessarily need to be normally distributed but the difference between these two paired methods should be. If the requirement of a normal distribution is not met, the data can be logarithmically transformed [Gia15]. Visual inspection via a histogram or QQ-Plots but also using classical methods for testing against normally (Shapiro-Wilk, D'Agostino and or Kolmogorov-Smirnov test) can be used to verify such assumptions [GZ12a]. It is up to the researcher to decide based on the available techniques whether the distribution of the data is sufficiently normal or requires transformation.

3.5.2 Adequate Sample Size

A more critical problem inherent to every study, in general, is the lack of an adequate sample size. This problem and the impact of insufficient sample size have been addressed in Section 2.4.4. Method comparison studies need to be sufficiently sized to claim the effect detected to be statistically sound. For instance, an inappropriate number of samples might lead to a low chance of finding the actual fixed bias (indicated as the mean difference in the Bland-Altman plot) or narrower limits of agreement when comparing two methodologies, *i.e.*, false-negative results in form of misleading results and potentially claiming good performance due to the lack of samples. The author recommends including the maximum allowed difference between the two methods as a parameter in the sample size calculation when assessing for agreement [Dog18]. The Bland and Altman method provides an equation to estimate an appropriate sample size when performing method comparison studies on their website [Bla]. This equation is defined as:

$$SE = \frac{s}{\sqrt{3n}} \tag{3.1}$$

where SE is the standard error of differences between the reference and the new method, s is the standard deviation of the differences between measurements by the two methods, and n is defined as the sample size. Assuming one can estimate the accepted standard error and the expected standard deviation, the sample size can be worked out. Lu et al. criticized this approach. They argued that the equation does not account for an appropriate power and maximum allowed difference (also insinuated by Dogan et al.) between the two methods in the calculation [LZL+16]. As such, they proposed a different equation involving these two parameters in the sample size calculation. The standardized equation for estimating an appropriate sample size is shown in Equation3.2. His approach has been validated for correctness via the results from the Monte-Carlo simulation. His concept has been commercially implemented in the statistical software package tailored for use in biomedical science [Wik].

$$n = \frac{(2 + z^2_{1-\gamma/2})[tinv(1-\beta/2, n-1, t_{1-\alpha/2,n-1})]^2 S_D^2}{2(z_{1-\gamma/2} S_D - \delta)^2} \tag{3.2}$$

where $tin(\bullet)$ is denoted as the inverse of a Student's non-central t-distribution, $z_{[\bullet]}$ stands for the level of significance α, σ for the standard deviation of differences between two methods and δ being the maximum allowed difference considered acceptable.

3.6 The Current State of Research of Artificial Intelligence (AI) in Bone Age Assessment

Literature review as outlined in Section 3.6.1 investigated the current state-of-the-art in regards to performance assessment of bone age models using AI. Based on the analysis, the result shows clinical studies with multiple deficiencies, *e.g.*, lack of test samples, issues with Ground Truth, inadequate testing set. Therefore, the reported performance is questionable from both, the clinical and statistical aspects. Upon further research, the lack of studies might be due to the novelty of AI and its use case in bone age assessment. We investigated any related works in the past ten years (as of January 2022) in the scientific database "PubMed" for relevant articles using a broad search string defined as *(("bone age") AND ("artificial intelligence")) OR (("deep learning") AND ("bone age"))*. The search deemed 64 studies as potentially relevant, where the majority of the articles were published after 2017 as seen in Figure 3.6.

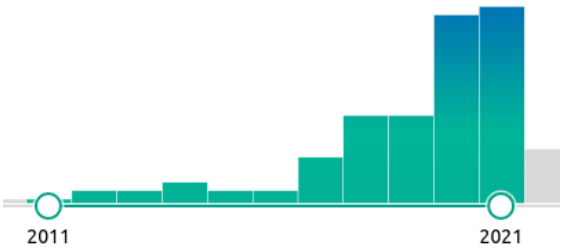


Figure 3.6: Published articles relevant for artificial intelligence in bone age assessment from 2011 to 2021. The trend shows the majority of the articles were published after 2017. This date coincides with the release of a public data set of bone age images.

This trend coincides with the release of the RSNA bone age data set, a publicly available data set as an initiative to motivate the creation of AI tools for radiology. The outcome of this challenge resulted in a study summarized by Halabi [HPKC+19]. The RSNA data set composed of 14036 clinical radiographs of the non-dominant hand was drawn from the picture archive and communication systems (PACS) of two institutions, Lucile Packard Children's Hospital at Stanford University (Palo Alto, Calif; n = 2983) and Children's Hospital Colorado (Aurora, Colo; n = 11053). These images had been interpreted by multiple pediatric radiologists, who documented skeletal age in the radiology report based on a visual comparison to the GP atlas.

Therefore, we assume due to the scarcity of available data, research in this field was very limited. After assessing for titles and abstracts only 15 articles remained relevant for this thesis. The studies excluded were articles focusing on the creation of the model itself instead of performance, reviews, and surveys discussing the potential use case of AI in medical imaging and assessing bone age on a different anatomical region. Due to similarities between the remaining study, we discuss two of these clinical more specifically. These two studies cover for the majority the current state-of-the-art in

assessing performance of AI in bone age.

3.6.1 Statistical and Clinical Relevance in Bone Age Assessment

A statistically relevant result based on any hypothesis test does not necessarily always lead to a clinically relevant result. To simplify, an outcome of a statistical analysis might yield significant results from the statistical point of view but from the clinical point of view, these results might not be significant enough to trigger a change in clinical management [Rub21]. Others might accidentally claim clinically relevant findings based on inappropriate use of the statistical tests applied to the data. In the following section, we dive further into what has been established in the field of AI, specifically bone age and further elaborate and question the clinical relevancy the authors claim to make.

Larson et al. developed and validated a bone age AI model for the RSNA Bone Age Machine Learning Challenge [LCL+18], initiated by Halabi et al. [HPKC+19]. The study was an initiative to facilitate and demonstrate the practical use case of AI in medical imaging. This team assessed the performance of the bone age model by using the mean absolute difference (MAD) as a metric for measuring accuracy and Bland and Altman's Limits of Agreement as a measure for agreement. The MAD is an intuitive measure of variability that indicates the mean of the absolute values of differences between observations of the predicted value (AI) and their true value (Reference Standard, Ground Truth). According to Larson et al. the ground truth/the reference standard was established by the mean of the reviewers' bone age estimates as an objective reference standard bone age cannot be determined. Their device was tested on a data set of 200 radiographs stratified by gender (100 males, 100 females). The distribution of the testing data set is shown in Table 3.2.

	Bone Age (years)																		
	<3	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**All**
female	4	3	3	2	6	10	4	9	10	14	12	10	2	8	3	0	0	0	100
male	3	0	1	5	4	3	4	6	5	13	18	13	11	6	7	0	1	0	100
All	7	3	4	7	10	13	8	15	15	27	30	23	13	14	10	0	1	0	200

Table 3.2: Bone age and sex stratification of the RSNA testing dataset [LCL+18]

Larson et al. report the performance of their model with a mean difference in bone age estimates of 0 years, reporting essentially no bias, and a MAD of 0.63 years, meaning on average the estimates of their model differ from the Ground Truth of the testing data set of about 6 months. These results are also accompanied with an analysis of agreement, the Bland-Altman plot, showing the maximum differences one would expect when comparing AI against the Ground Truth, in this case, the mean of the reviewers estimates, as seen in Figure 3.7. Though the limits of agreement or not displayed specifically in the plot, one can assess visually that the expected differences lie in the range of 1.5 years. Unfortunately, Larson et al. does not define the maximum allowed difference, *i.e.*, a pre-defined clinical agreement limit, where differences below this threshold are clinically negligible.

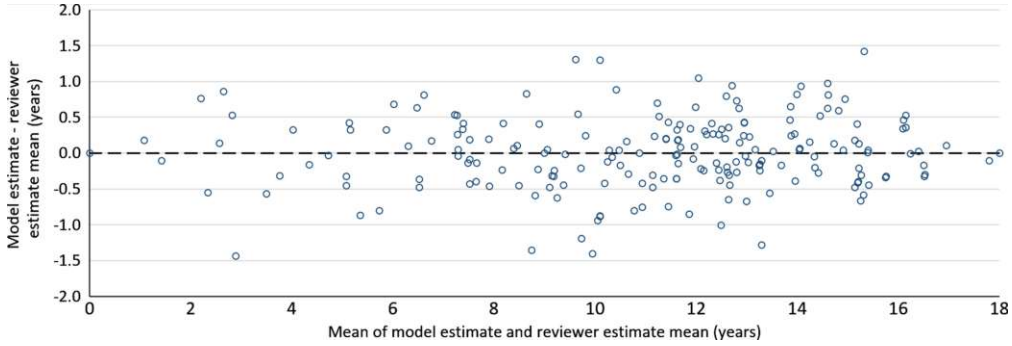


Figure 3.7: Bland-Altman Analysis of Larson et als AI model against the Ground Truth [LCL+18]. The x-axis displays the mean estimation between the model and the observers' mean estimate (Ground Truth) indicating the tentative "true" value. The y-axis shows the difference between the estimation of the AI and the Ground Truth indicating the disagreement between the two methods.

The results showed excellent performance of the AI model. Upon reviewing the statistical analysis though, certain issues might raise concerns about the clinical relevancy presented in this paper. Even though the statistical tests show promising results, it is unclear whether the results showing significant values may perhaps be a statistical fluke as the authors did not provide any sample size calculation. The assumption that a sample size of 200 images is sufficient to indicate that the performed tests are essentially not powered. In addition, no confirmation that the distribution of the differences approximates a normal distribution was provided to apply the Bland-Altman analysis. Figure 3.8 describes Larson et al.'s testing set via a histogram which approximates a normal distribution with a slight skew to the right. Based on this we assume that the differences will also approximate a normal distribution a explain the legitimacy of using the analysis for agreement.

Finally, upon further review, the distribution on the test data set in Figure 3.8 and more detailed in Table 3.2 clearly shows that the data does not contain enough granularity for the assessment of certain age groups especially for younger age groups (<3) and older age groups (>16). Suffice to say, multiple age/sex categories in the testing data set are under-represented or not represented at all. These shortfalls indicate proves that statistical evidence does not necessarily coincide with clinical relevance. Essentially Larson et al.'s statistical approach shows good performance overall but the evidence shown is skewed towards the ages with higher samples and therefore does not reflect the actual performance over all ages. Larson et al. also acknowledge in the paper that the model itself does not perform well on patients of younger ages. We assume that this is because they attempt to reflect the applied clinical practice where the underlying population distribution of children, who are inclined to have a bone age assessment, are normally distributed. The sampling strategy of stratifying only by age might not the ideal method in this case as, specifically for bone age as explained in Section 2.2, the

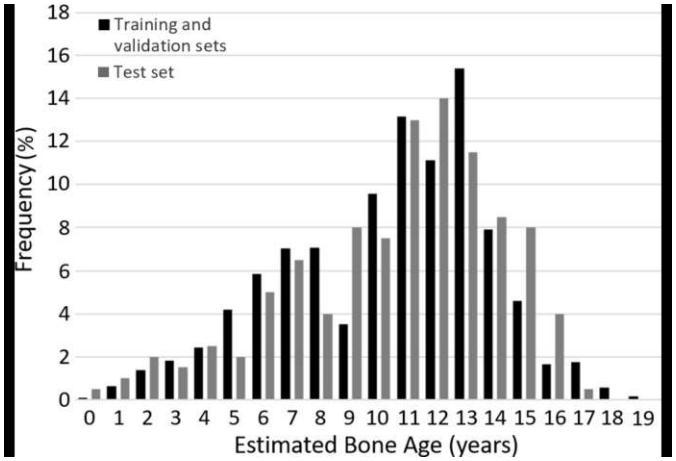


Figure 3.8: Distribution of the testing set used to assess performance. The distribution approximates a normal distribution [LCL+18].

testing data sampled must ensure sufficient granularity in age and sex.

A similar statistic is presented in a retrospective study conducted on a German cohort of 514 patients with various indications for a bone age assessment by Booz et al. [BYW+20]. The researcher tested a commercially available AI software estimating bone age to assess the performance on a standalone data set. The authors tested their data for normality and report the results, in addition to the MAD and Bland-Altmans Limits of agreement, an analysis of correlation. Though not specifically addressed with proper power analysis, a sample size of over 500 patients should be sufficiently sized for the sake of the study. Similar to the Larson study, the testing data reflect a normal distribution. As such the performance shown in this paper tends to reflect the age groups with higher representation in the data set.

Another study conducted by Kim et al. assessed the performance of their AI on a Korean cohort of 200 children [KSY+17]. They addressed the issue of testing on a normally distributed data set by performing stratified random sampling by age. The Bland-Altman analysis showed high differences between the age of $12-15$ year old children. The authors listed several limitations concerning the results of this study. They reported results based on a limited sample set indicating that no power analyses were performed. In addition, estimating the performance based on a cohort from a single clinical site raises concerns over the generalizability of the overall population in terms of ethnicity.

We understand that simulating real-world clinical practice should be the first approach, where performing random sampling would be the most dependable method. But when it comes to testing for performance it is important to find a balance between efficacy and effectiveness. This boundary is not easy to find. Larson attempted to stratify by gender, whereas Booz sampled completely randomly, both resulting in a similar distribution of the testing set. We assume because age is such a high contributing factor in the assessment

of bone age, stratifying by age should be the higher priority, *i.e.*, to put every age group as per indented patient population of the AI on the same level of scrutiny simulating not a normal distribution but instead a uniform one.

Another popular metric of differences used by many authors is the mean absolute deviation (MAD) [LCL+18, BYW+20, RLY+19, HPKC+19, PBM+20, YGX21, KKQ+20, WCG+21], as well as the root mean square error (RMSE)[TLS+18, KSY+17]. Both types of mean assess the absolute deviation of the AIfrom the Ground Truth. The difference between the two types of mean is, compared to the MAD, the MAD penalizes outliers more due to squaring the differences before estimating the absolute of means. The MAD on the other hand is more intuitive to interpret. Many authors claim good performance of their AI model based on either or both types of means, suggesting this metric to be state-of-the-art in assessing performance. Upon further investigation, all of them do not justify why these deviations can be considered acceptable. Our assumption from the clinical point of view is that deviations up to 6 months may be insignificant [SLK+20]. The GP atlas provides reference hand images in 0.5–1 year increments. For instance, the reference images for females between the age of 10 to 14 are 10, 11, 12, 13, 13.5, and 14; As such, a bone age that is half a year lower or higher than the Ground Truth may not be clinically significant. Looking at the quality of the respective study design, many of these trials raise certain deficiencies. Among others a justification of adequate sample size [LCL+18, HPKC+19], inadequate distribution of samples [TLS+18, BYW+20, WGC+20, PBM+20, YGX21, KKQ+20] or concerns regarding a proper Ground Truth the device's performance is compared to [KSY+17]. These pitfalls raise concerns regarding the reported performance the authors claim in their studies.

3.7 The Analysis of Differences Without a True Reference Standard Using the Concept of Interchangeability

As already introduced in Section 2.4.3, the idea of interchangeability proposed by Obuchowski et al. has already existed in the realms of pharmaceutical products where one assesses the switchability between test (new) and reference drug (reference standard) [OSS14]. To summarize, a drug is interchangeable when both test and reference drug produces the same results with a certain degree of acceptable error. The new drug must at least not be inferior to the reference drug. Obuchowski et al. applied this idea to diagnostic and imaging tests motivated by the fact of constant innovation in the medical imaging field and procedures where one needs to know whether a new diagnostic test can replace or perform as well as the existing test, *i.e.*, without adversely affecting the patient. The author explains, one of the main advantages of testing for interchangeability is the apparent necessity of a reference standard to compare to. This is because a valid reference standard, a true Ground Truth, will not be always available.

Obuchowski et al. applied this concept to a study in the measurement of the acetabular version for patients with femoroacetabular impingement using two modalities, CT and

MRI. The state-of-the-art for measuring this parameter is using computed tomography (CT). Patients suspecting to have this measure taken during their preoperative planning also have their magnetic resonance imaging (MRI) exam taken in many cases. Both modalities allow the measuring of the acetabular version, where the benefit of using only the MRI would reduce the amount of exposure one would get from radiation. For the MRI to replace the CT, the former modality must not be inferior to the reference modality. The CT on the other hand, even though state-of-the-art, is not the true gold standard. Fortunately, the concept of interchangeability does not rely on a reference standard. The theory assesses the differences in any possible random scenarios and investigates whether the excess of differences created by the new methodology are considered acceptable or not. The authors define the null and alternative for testing for interchangeability as follows:

$$H_0 : \gamma = E(Y_{iTjk} - Y_{iRjk'})^2 - E(Y_{iRjk} - Y_{iRjk'})^2 > \theta i \tag{3.3}$$

$$H_1 : \gamma \leq \theta i \tag{3.4}$$

where Y_{iTjk} denotes the result with the new test (T) (in this case MRI) by reader j for sample i on occasion k, and Y_{iRjk} denotes the result with the existing reference modality (in this case CT) by reader j for sample i on occasion k. The difference denoted as $\gamma$is then compared to an accepted excess of differences ϑin a random reader scenario. The random reader scenario simulates an environment where the differences between every pair of readers are considered in the measure of differences. This ensures any potential inter-and intra-rater-variability is included in the assessment as well. Assuming the excess of differences γcreated by the new modality is within the accepted difference limit ϑ, we can safely assume that the new modality is interchangeable with the reference modality. The Obuchowski study with J = 3 readers and i = 22 hip images shows that replacing MRI measurements with CT measurements would result, as the author states "*indifference in measurements of 2.0°-3.1° in excess of the differences that we would expect to see just using CT*". Therefore they concluded that MRI is interchangeable with CT in the measurement of the acetabular version for patients with femoroacetabular impingement.

This concept re-purposed by Obuchowski et al. from the pharmaceutical side to imaging and diagnostic tests might also find application in AI, where one of the core tasks in medical AI is to solve imaging problems where a true reference standard might not always be available, specifically for bone age. The method of interchangeability might be one of the metrics to assess the performance of AI.

3.8 Conclusion

To this day, many have attempted to assess agreement using multiple approaches such as mean, correlation, or regression. The Bland-Altman method has established itself in

multiple areas of medicine as an appropriate technique for comparing methods against each other and may help researchers to compare a new method against another one or a reference standard. Recently, the analysis for agreement has started to see the application in the field of AI, specifically in the assessment of bone age. While research papers attempted to evaluate the performance of such models, many lack important aspects in terms of sample size or adequate representation in the testing set based on the clinical relevancy to justify good performance of their model. These shortfalls should be addressed in our future performance testing. Finally, while a reference standard as required by Bland-Altman is not always easy to define, another theory, the concept of interchangeability, does not rely on any reference standard, but the differences created by one method over the other. The independence from requiring a reference standard makes this statistical method for assessing performance very powerful.

CHAPTER 4

Realization of Study Design

4.1 Introduction

The absence of a clear guideline in how to evaluate the performance of an AI estimating continuous outputs, in this case, bone age makes this topic very interesting for researchers. Emphasized by the lack of an adequate study design as seen in the current state-of-the-art in bone age assessment, referenced in Section 3.6.1, a framework consisting of possible methods for performance assessment would benefit the researcher. We addressed the prerequisites required by Task T1 and Task T2, as defined in Section 1.4, in Section 2.2 and 2.4, respectively. To achieve Task T3, we proposed an updated framework based on the pitfalls discussed in Chapter 3, considering the following tasks as listed below:

- Study Design
- Objectives and Hypothesis
- Statistical methods & Performance targets
- Sample Size & Power Study
- Sampling Strategy & Generalizability

Based on the assumptions above, we escalated our statistical consideration into a full-fledged clinical study to provide evidence of a working framework. Therefore, ImageBiopsy Lab provided us with their medical software, the bone age AI model, PANDA, CE-marked since 2019 [IB]. The focus of this thesis is to apply our proposed statistical techniques and modeling on PANDA to present additional evidence related to standalone performance. Outputs based on the statistical considerations are implemented via Python scripts.

4.2 Study Design

This section presents an overview of the intended workflow of the clinical study. As part of Requirement R2 outlined in Section 1.4, medical device software must present evidence of performance in form of a clinical study. This includes the selection of methods and definition of targets to be met to keep both the personnel in the medical as well as in the regulatory field satisfied. The investigations in the following sections afterward will provide evidence to back up the proposed study design.

4.2.1 Acquiring the Images: Sampling

The US clinical site has access to existing bone ages reports from multiple subsidiaries from which the standalone performance test data set of 345 images as described in Section 4.6 is drawn. Sampled bone age studies will be anonymized and stored in a clinical management system. Access is restricted to authorized users on a per-project basis.

4.2.2 Study Personnel: Acquiring the Radiologists' Readings

Three fellowship-trained pediatric radiologists will provide readings utilizing the GP method through a clinical trial management system. Each radiologist will be an American Board of Radiology-certified radiologist with sub-specialty certification/Certificates of Added Qualifications (CAQs) in pediatric radiology. Each observer will have at least 5 years post-fellowship experience in the interpretation of pediatric bone age exams. As discussed in Section 4.4.1, the mean of the three assessments will represent the ground truth to test the hypothesis of agreement between PANDA and readers. The clinical management system consisting of a viewer and custom-built annotation tool is used for facilitating readings of the images. An example of a reading scheme is provided in Figure 4.1. The radiologists will be blinded to the chronological age of the subject as well as the bone age assessments from the clinical report including the bone age readings among each other. This measure is taken to ensure that the radiologist doing the reading is assessing as unbiased as possible [BDB+01]. The annotations will be added as metadata to the stored cases.

4.2.3 Model Validation: Acquiring PANDA's Readings

Automated analysis of the radiographic images with the PANDA software was accomplished via an internal clinical pipeline by the installation of PANDA in a Docker container integrated into the clinical management system of the study site. PANDA will be executed on the standalone performance test data set of 345 images. The results will be provided back for data analysis.

4.2.4 Statistical Analysis

Statistical methods to be assessed are described in Section 4.4. The analysis will be based on Bland-Altman, regression, and the concept of interchangeability. Outlier detection

Figure 4.1: Reading of bone age Part 1 - Overview of the interface to provide to enable the bone age assessment from the observers on the clinical site. Users will render a bone age assessment by providing the radiographic age in years and months to the corresponding plate according to the GP clinical reference standards.

will be performed using the modified z-score [IH93]. This detection method is defined for a given measurement x_i as

$$z_i = \frac{x_i - \bar{x}}{1.4825 * MAD} \tag{4.1}$$

where MAD is the median absolute deviation. Measurements with a modified z-score above 3.5 or below -3.5 are considered potential outliers. These samples will be visually inspected to determine the root cause of the deviation. Based on the result of the outlier assessment, the sample may be excluded from the statistical analysis. Based on the clinical relevance of bone age as described in Section 2.2, we also assessed the performance of the AI model on different sexes.

4.3 Objectives and Hypothesis of the Clinical Investigation

Any proper and sound research study requires objectives derived as a hypothesis. A well-defined research hypothesis contributes greatly to the solution of the research problem. Therefore we lay out the objective of interest and develop the hypothesis that will serve as the baseline of the study.

4.3.1 Objective

In this multi-center, retrospective study the investigators are targeting to study the agreement of an AI model with the reference standard, which comprises of PANDA compared to expert radiologists' estimation of bone age. To this end, archived X-ray images from multiple clinical centers in the USA will be sampled and evaluated for the estimation of bone age using the PANDA software and the readings from the radiologists. The primary objective is to validate the performance of the AI model (PANDA) in making automated estimation of bone age. This is done by comparing the automated measurements of hand radiographs (output of PANDA) with the measurements made by expert radiologists. The evidence displayed must show non-inferior performance compared to the current reference standard.

4.3.2 Hypothesis

This study will assess performance via Bland-Altman and interchangeability to assess whether PANDA and expert radiologists bone age are in agreement and interchangeable:

Agreement:

1. H_0*:* Automated analysis with PANDA and the Ground Truth are not in agreement. Alternatively expressed as
   $H_0 : 95\%(\mu_{Panda} - \mu_{Rads}) \leq -\Delta \mid 95\%(\mu_{Panda} - \mu_{Rads}) \geq -\Delta$
2. H_1*:* Automated analysis with PANDA and the Ground Truth are in agreement. Alternatively expressed as
   $H_1 : -\Delta \leq 95\%(\mu_{Panda} - \mu_{Rads}) \leq +\Delta$

where μ stands for the respective assessment, PANDA and the Ground Truth, and Δ indicates the clinically maximum allowed difference. Two methods are in agreement if 95% of occurring differences (Upper and lower bound of 95% CI limits of agreement) are within Δ.

Proportional Bias:

1. H_0*:* Automated analysis with PANDA and the Ground Truth is proportionally biased. Alternatively expressed as
   $H_0 : B_1 \neq 1$
2. H_1*:* Automated analysis with PANDA and the Ground Truth is not proportionally biased. Alternatively expressed as
   $H_1 : B_1 = 1$

where B_1 stands for the slope based on orthogonal linear regression of PANDA and the Ground Truth. If the 95% CI for slope does not contain 1, there is statistically significant evidence for proportional bias between the two methods [Lud97].

Interchangeability:

1. H_0**:** Automated analysis with PANDA is inferior to the assessments of three expert radiologists. The new method is not interchangeable. Alternatively expressed as $H_0 : \gamma = E(Y_{iTjk} - Y_{iRjk'})^2 - E(Y_{iRjk} - Y_{iRjk'})^2 > 0$
2. H_1**:** Automated analysis with PANDA is interchangeable to the assessments of three expert radiologists. The new method is interchangeable. Alternatively expressed as $H_1 : \gamma = E(Y_{iTjk} - Y_{iRjk'})^2 - E(Y_{iRjk} - Y_{iRjk'})^2 \leq 0$

where γ stands for the resulting difference when interchanging one method with the other, PANDA and the expert readers. A detailed description of the hypothesis is provided in Equation 3.3. Setting $\theta \leq 0$ would indicate equivalent or superior performance when interchanging the AI with the expert observers.

4.4 Statistical Methods and Performance Targets - Ground Truth, Agreement, Regression & Interchangeability

As part of Task T1 and Task T2 defined in Section 1.4, Chapter 2 explained the basics of the current state-of-the-art statistical tools used in the assessment of bone age assessment, where Chapter 3 addressed the legitimacy and pitfalls when using the said methods for assessing the performance of the AI model. Based on the assessment in the previous chapters and as part of fulfilling the defined requirements (Requirement R1 — R3), we propose the following statistical methods and performance targets as part of Task T2. At the end of this section, the proposed solution answers the issue concerning Research Question Q2.

4.4.1 Ground Truth: Establishing a Reference Standard

For this study, the Ground Truth for the comparison of methods will be the mean of bone age assessments made by each of the three human reviewers per image. The number of participating observers are based on the clinical standard, known as Blinded Independent Central Review (BICR), deeming three readers sufficient [SRM+21] As described in Section 2.2, bone age assessment is prone to high inter-rater variability. We will take measures to minimize inter-rater variability as much as possible. Recruiting experienced and training personnel will minimize variability to a high extent. Therefore the mean should be a more accurate representation of the actual bone age assessment compared to for instance the median, as outliers are less impactful and less prone to occur.

In addition to the above measures taken, we will verify the reliability of the Ground Truth in the form of the intraclass correlation (ICC). The intraclass correlation coefficient is a widely used reliability index in analyses of interrater reliability. It is constructed by the variance components for reader (3 participating readers), case (age group), and random errors for an ANOVA model. Based on the 95% CI of the ICC estimate, values less than 0.5, between 0.5 and 0.75, between 0.75 and 0.9, and greater than 0.90 are indicative of

poor, moderate, good, and excellent reliability, respectively, based on Koo et al [KL16]. Multiple forms of the ICC exist. For this study specifically, we will assess reliability via mean-rating (k = 3), absolute-agreement, 2-way mixed-effects model justified as follows according to the definition by Koo et al.:

- **mean rating**: mean value of 3 raters as an assessment basis, the experimental design of the reliability study involves 3 raters,
- **absolute agreement**: different raters assign the same estimate to the same subject.
- **2–way–mixed–effects model**: selected raters are the only raters of interest

4.4.2 Agreement: Bland-Altman Method

The performance of PANDA on bone age assessments will be assessed for agreement and absolute bias with expert radiologists using Bland-Altman plots. 95% confidence interval (CI) of limits of agreement (LOA) between the two methods (denoted as a blue interval in Figure 4.2 is calculated based on the average reading of radiologists (= Ground Truth) vs. PANDA. The 95% CI of LOA is compared to the standard boundary (denoted as a red line in Figure 4.2, the maximum allowed difference Δ. Δ will be defined as the average LOA between any reader combination, *i.e.*, radiologist vs radiologist. Assuming we recruit experienced and trained personnel to reduce variability as much as possible, the readings of the radiologists will be considered clinically relevant. Therefore differences occurring between any single radiologist pair participating in the study would reflect the differences happening in the real world. Averaging all the differences of each reader pair, *i.e.*, their average LOA will bring one closer to the expected maximum difference Δ. Agreement and therefore good performance is shown when the upper bound of 95% CI of the upper LOA and the lower bound of 95% CI of the lower LOA are within the maximum allowed difference. In addition, the mean difference in the Bland-Altman plot indicates the presence of a fixed bias. Fixed bias is present when the intercept differs significantly from zero. The CI of the mean difference must include 0 to emphasize the lack of significant fixed bias [Lud97].

4.4.3 Proportional Bias: Regression Analysis

Section 3.4 addressed the misuse of linear regression when concluding for agreement. It also established the method as a useful indicator for assessing bias. Therefore, regression of PANDA's measurement vs. the Ground Truth will assess any age-specific proportional bias based on the slope. The standard linear regression, ordinary least squares regression, as addressed in Section 3.4 is not a suitable model due to only taking errors of one variable into account. As such, the regression model utilized will be orthogonal linear regressions, a model assuming errors in both, the independent and dependent variable [BR90]. Proportional bias is present when the slope differs significantly from unity. The CI of the slope must include 1 to emphasize the lack of significant proportional bias [Lud97]. Figure 4.3 visually presents the presence of proportional bias.

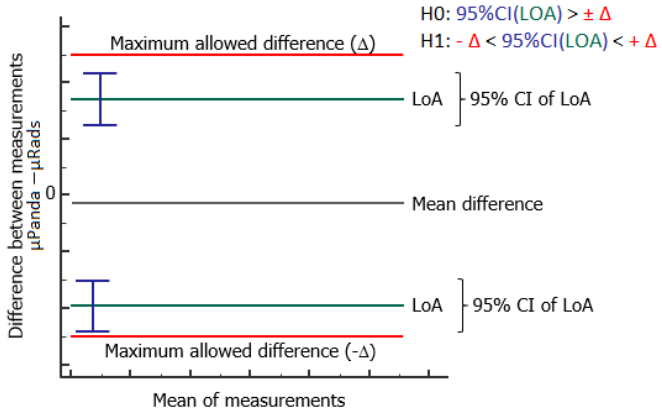


Figure 4.2: Description of reference values - measurement of agreement between PANDA and mean of the radiologists' assessment. The 95% CI of limits of agreement (LOA) of PANDA (blue) will be compared to the maximum allowed difference (red - limits of agreement among the radiologists) [Sch21].

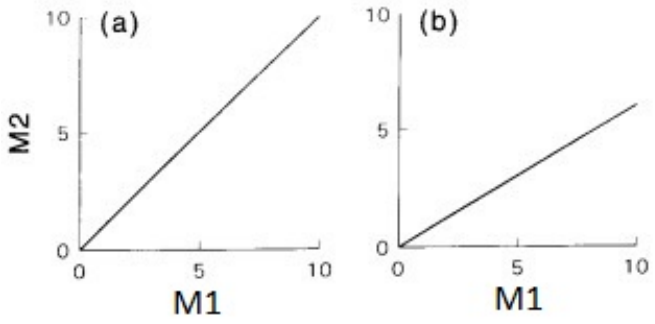


Figure 4.3: Ideally, two interchangeable methods (M1 and M2) do not present any proportional bias, as seen in (a). Proportional bias is present when the slope differs significantly from unity (b). [Lud97].

4.4.4 Interchangeability: Analysis of Expected Differences With Absence of a Reference Standard

In addition to providing evidence of the performance of PANDA via agreement and regression, an assessment of interchangeability utilizing the concept from Obuchowski et al. is conducted [OSS14]. The benefits of not relying on a reference standard have been discussed in Section 4.4.

To summarize, when tests are compared with each other, in this case, PANDA and the radiologists, one expects that both assessment methods produce the same clinical result in any given patient. To show the interchangeability of two modalities, one compares the differences in assessments from one method over the other. To specify, the

differences between PANDA and the radiologists are compared to differences between two assessments of the radiologists under different occasions in time. The assessment yields an estimated equivalence index γ with a 95% confidence interval. The equation for the test for interchangeability then goes as listed below

$$\gamma = E(Y_{iTjk} - Y_{iRjk'})^2 - E(Y_{iRjk} - Y_{iRjk'})^2 \tag{4.2}$$

where Y_{iTjk} denotes the result with the new test (T) modality, that is PANDA, by radiologist j, which provides the same result for every read, for image i on occasion k, and Y_{iRjk} denotes the result with the existing reference modality, that is the radiologists, by radiologist j for image i on occasion k.

However, in this study, the readers will only read each image once and as such provide no replicates. Therefore, we adapt the definition of γ as following:

$$\gamma = E(Y_{iT} - Y_{iRj})^2 - E(Y_{iRj} - Y_{iRj'})^2 \tag{4.3}$$

The subscripts for the replicate/occasion and the subscript for reader in device output are removed. On the right-hand side of this equation, the first mean square explains the deviation between the device output with the assessment from the radiologist; while the second mean square explains the deviation among the assessments from different radiologists. Thus, when calculating the estimation of γ, the first mean square average over all the readers and cases and the second mean square average over all the pair of readers and the cases, the result shows the excess of difference resulting from the comparison. Figure 4.4 visualizes the concept of interchangeability intuitively.

If $\gamma>0$, it means the deviation between the device output and the assessment from the truthers is larger than the deviation among the assessments from the truthers. We then need to assess whether these differences are acceptable within the clinical practice. Setting the equivalence limit to 0 or smaller provides evidence that the new AI-supported methodology is superior to the current reference standard.

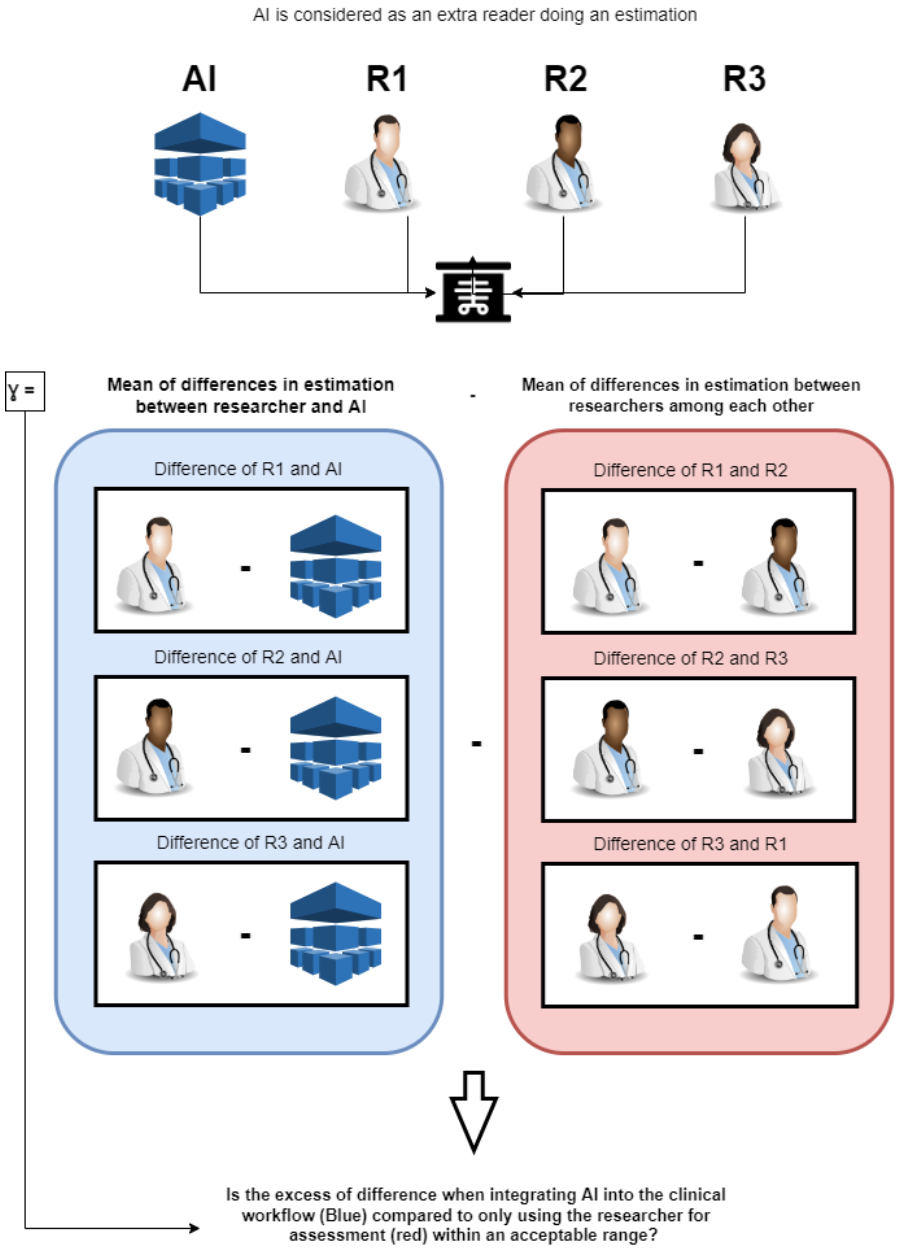


Figure 4.4: The concept of interchangeability is tailored to the use case of AI in the clinical setting. The AI is considered as an additional reader. The first mean square on the left-hand side explains the deviation between the device output with the assessment from the radiologist; while the second mean square on the right-hand side explains the deviation among the assessments from different radiologists.

4.5 Sample Size Calculation & Power Study

Section 2.4.4 addressed the importance of having a sufficient sample size and performing power studies to produce clinically relevant results from a clinical investigation as part of answering Research Question Q3 and Task 3 addressed in Section 1.4. Multiple formulae exist to assess an adequate sample size. For this thesis, we are applying the Bland-Altman method to assess agreement between two methods, *i.e.*, AI compared to the mean of the radiologists' assessment. Section 3.5.2 discussed Lu et al's approach for sample size estimation when using the Bland-Altman method. As such, to estimate the number of images for performance testing we utilize the formulae of Lu et al. as described in Equation 3.2, based on the following parameters:

1. a predetermined level of alpha α, beta β
2. the mean (μ) and standard deviation (σ) of differences between two methods
   - model & radiologist
3. clinical acceptable limits (δ) - the maximum allowed difference between
   - radiologist & radiologist (inter-reader variability)

In clinical studies, the parameters above are usually estimated from data in existing literature or are obtained from the results of a pilot study. As sample size estimation can only be approximated anyway, the estimates do not need to be exact [Bla15]. For this study, we will rely on data from existing studies.

4.5.1 Establishing the Reference Values

The study from Larson et al. as described in Section 3.6.1 presents an ideal case for estimating parameters necessary for calculating the sample size. Even though we pointed out flaws related to this study, due to the similarities between ours and the author's case in terms of study design and the AI model in question, we can consider the results of their study sufficiently close to use for sample size estimation.

Figure 4.5 summarizes the study results of the Larson study. This figure shows a comparison of their AI model against their observers and a comparison of their observers among each other. To provide reference values for the mean difference, the standard deviation of the differences, and the clinical acceptable limits, we utilize these clinical reads summarized below [LCL+18].

Level of Significance α and Power of the Study

Generally, the significance level alpha α is set to 0.05. Depending on the level of significance one wants to achieve, based on the risk of the device or drug, one can define a significance of 0.1 or 0.01. For this study, we define $\alpha = 0.05$. This means we are willing

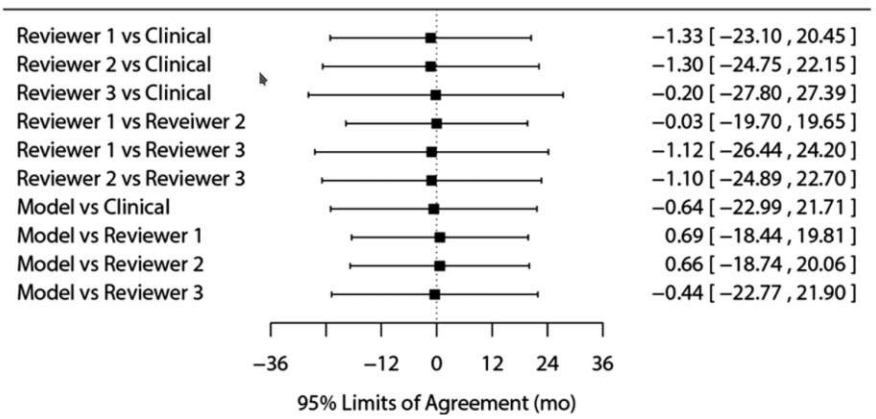


Figure 4.5: Reference values (Mean difference, Standard deviation of differences & Clinically acceptable threshold) from the Larson study were used to determine the sample size

to accept a 5% risk of concluding a significant result when there is no significant result (false positive). The power, specified as 1 - β, where β stands for the probability of type II errors, is generally set to 0.80. For this study, the power is set to 0.85. This means we are willing to accept a 15% risk of not detecting a significant result when there is a significant result (false negative).

Mean Difference & Standard Deviation of Differences

The mean difference and the standard deviation of differences between the model and the radiologist are estimated based on the **average of the clinical reads** between Reviewer 1-3 and Model as reported in Figure 4.5. We will exclude the reads of the clinical reports. The clinical reports consist of assessments from different readers and would increase the overall reader variability. In addition, the credentials of the radiologists are unknown. Both these pitfalls might affect the estimation of the parameters adversely. Therefore excluding the clinical report would lead to a closer approximation of the parameters for sample size calculation. Based on data in the literature provided in Figure 4.5 and summarized in Table 4.1, the mean difference used for the sample size estimation is the average of the mean differences between the estimates of reviewers compared to the performance of the model. μ **was found to be 0.3 months**.

	Reviewer 1	Reviewer 2	Reviewer 3	Mean of Radiologists
Model [months]	0.69	0.66	-0.44	0.30

Table 4.1: Mean difference of the three human reviewers' estimates in months, compared with that of the model. Reference values are taken from Figure 4.5 of Larson et al's paper [LCL+18]. The average of these three estimates form the basis used for the sample size estimation with $\mu = 0.3$ months.

The parameter for estimating the standard deviation of the differences can be estimated based on the mean difference and one end of the limits of agreement (LOA = mean difference ± 1.96 * standard deviation). Rearranging the standard formula we can estimate the standard deviation of differences (standard deviation = (LOA - mean difference)/1.96). For each observer (Reviewer 1-3), the standard deviation of differences is calculated as seen in Table 4.2 based on the results from the Larson study as reflected in 4.5. The average of the three standard deviations of difference σ **is expected to be 10.35 months** and serves as one of the parameters used to estimate the sample size.

	Mean Difference & Limits of Agreement [months]	Standard Deviation [months]	Mean Standard Deviation [months]
Model vs. Reviewer 1	0.69 [-18.44, 19.81]	9,76	10.35
Model vs. Reviewer 2	0.66 [-18.74, 20.06]	9.90	
Model vs. Reviewer 3	-0.44 [-22.77, 21.90]	11.40	

Table 4.2: Calculation of the mean standard deviation of the three human reviewers' estimates in months, compared with that of the model. Reference values are taken from Figure 4.5 of Larson et al's paper [LCL+18].

Maximum Allowed Difference

The standard boundary is defined as the maximum allowed difference between two modalities – new and reference modality. Typically a reference standard of acceptable differences exists to compare to. However, in some cases, this is not always available. How far apart measurements can be without leading to problems is a question of clinical judgment. Statistical methods cannot answer such a question.

To find an acceptable difference in bone age assessment we will rely on real word practice. We will utilize the intra-rater variability of the expert readers. Clinical assessments made by experienced personnel are to be considered legitimate assessments. Therefore any differences occurring between the observers should be considered acceptable, As such, we can use the limits of agreement among each observer pair. The average limits of agreement of the three reader pairs will determine the acceptable boundary.

Based on the reference values in Figure 4.5 and summarized in Table 4.3, we first assessed the average of the mean differences of each observer pair (Reviewer 1-3) resulting in an average of mean differences $\mu_{meandifference} = -0.75$ months. Applying the same principle as described for estimating the parameter of the standard deviation of differences in Section 4.5.1, we estimated the standard deviation of differences from each observer pair and calculated the mean. This results in a standard deviation of the mean inter-observer difference among the radiologists $\mu_{stddiff} = 11.69$ months. Given the definition of 95% limits of agreement (LOA = mean difference ± 1.96 * standard deviation of differences) with $\mu_{meandifference} = -0.75$ months and $\mu_{stddiff} = 11.69$ months, the upper and lower limits of agreement based on all observer pair is 22.16 and -23.66 respectively $(-0.75[-23.66, 22.16])$. The maximum difference among the radiologists, either the lower or upper limits of agreement will be used as the maximum accepted difference. We

designate the highest difference within the limits of agreement as the maximum allowable difference. Therefore **the maximum allowed boundary δ, is 23.66 months**. The results are summarized in Table 4.3

	Mean Difference & Limits of Agreement [months]	Standard Deviation [months]	Average of Mean Differences & Average of Mean Standard Deviation
Reviewer 1 vs. Reviewer 2	-0.03 [-19.70, 19.65]	10.04	[-0,75; 11.69]
Reviewer 1 vs. Reviewer 3	-1.12 [-26.44, 24.20]	12.91	
Reviewer 2 vs. Reviewer 3	-1.10 [-24.89, 22.70]	12.14	

Table 4.3: Mean limits of agreement of the three human reviewers' estimates in months. Reference values are taken from 4.5 of Larson et al's paper [LCL+18].

Based on the parameters (mean difference μ, the standard deviation of differences σ and the maximum allowed difference δ) listed above and utilizing the formulae of Lu et al. as outlined in Equation 3.2, the minimum number of images to guarantee the power of the statistical test (0.85) for standalone performance testing was calculated to be at least **333 images**.

4.6 Sampling Method

In Section 3.6.1 we discussed the pitfalls of entirely relying on random sampling. To summarize, assuming the underlying population of patients having a bone age assessment taken is normally distributed, the analysis would lack sufficient granularity for younger and older age ground. The lack of samples for younger and older age groups would make the result clinically irrelevant. At the end of this section, the proposed solution will answer the issue concerning Research Question Q1 as outlined in Section 1.4.

That is why, the standalone performance testing data set should contain enough granularity in terms of both, age and sex, for assessment based on the intended patient population of PANDA as described in Section 2.3.1. PANDA is intended for 24 months (2years) – 204 months (17 years). This results in a total of 15 age groups. It is important to note that the goal of the standalone performance testing is to show that PANDA works equally well overall ages as defined in the intended patient population. By equally distributing samples among the 15 age groups, enough granularity in the performance data can be provided over all ages. As such stratified sampling i.e. dividing members of the population into homogeneous subgroups, in this case in ages, will be conducted to achieve this objective. Based on this strategy (stratification based on ages) the data can now be randomly sampled from multiple clinical sites. This strategy allows the provision of a robust performance over the ages of the intended patient population while at the same time indirectly considering an approximate distribution of other parameters such as ethnicities and sexes.

It is important to note that following this sampling strategy will not result in any sampling bias during the sampling process. Aside from the criteria set in age no insight regarding

Age Group	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
start month	24	36	48	60	72	84	96	108	120	132	144	156	168	180	192
end month	35	47	59	71	83	95	107	119	131	143	155	167	179	191	204
n images	23	23	23	23	23	23	23	23	23	23	23	23	23	23	23

Table 4.4: Distribution of the images from the standalone performance testing dataset in months. Based on the intended age population a total of 15 age groups is defined.

other relevant parameters (refer Section 2.2 for bone age such as race, gender, condition (normal, delayed, advanced), and underlying diseases will be available when sampling the data. This means no stratified sampling based on sex will be performed, as the second major parameter for bone age. Nonetheless, the standalone performance data set should provide a balanced distribution of male and female patients overall. Based on the US census in 2018, simple random sampling should be sufficient to ensure equal distribution among sex [Bur21]. This methodology of sampling the standalone performance yields sufficient granularity as indicated in the device's patient population regarding age and sex as well as providing an adequate representation of the pediatric US population.

To conclude, Section 4.5 determined a minimum number of images 333 samples to be sufficient for the study. To provide adequate performance over all ages, we will provide equal distribution among every age group increasing the total number of images in the standalone performance testing data set to 345 images for 15 age groups as reflected in Table 4.4.

CHAPTER 5

Implementation

5.1 Introduction

This chapter presents the implementation of the proposed statistical tools described in section 4.1. The data analysis as outlined in section 4.1 utilizes python scripts. First, section 5.2 will provide the details of implementing the estimation of the adequate sample size for the Bland-Altman method based on the methodology of Lu et al [LZL+16]. This is done as part of Research Questions Q3 as required in section 1.4. Following that, we provide a step-by-step guide in regards to the implementation of the statistical methods as part of the fulfillment of Task 2 as part of section 1.4.

To facilitate reproducibility of the results as per R3 of the requirements outlined in section 1.4.1, we ran the scripts in a virtual environment with Python 3.8 with the following core libraries:

- NumPy 1.20.1
- pandas 1.2.4
- SciPy 1.6.2

In addition, providing a description of the proposed method in form of pseudo code as part of Requirement R3, emphasizes the criteria set by Requirement R1 making the methods language independent and therefore generalizable.

5.2 Estimation of Sample Size

As outlined in equation 3.2, the standard formula can be used to estimate the required amount of samples needed, provided the mean difference $\mu = 0$. In many cases $\mu > 0$.

Alternatively, Lu et al. suggested reaching the required estimate by using an iterative method based on their power formula

$$power = 1-(\beta_1+\beta_2) = 1-(prob(t_{1-\alpha/2,n-1}, n-1, \tau_1)+prob(t_{1-\alpha/2,n-1}, n-1, \tau_2)) \quad (5.1)$$

where the type II error β consists of two parts, β_1 & β_2 indicating the type II errors for both, upper limit of LOA and lower limit of LOA, respectively. Both values of error are estimated by $probt[\bullet, n-1, \tau_1]$ & $probt[\bullet, n-1, \tau_2]$, respectively. They denote the comulative distribution function of a Student's non-central t-distribution with $n-1$ degrees of freedom and the respective non-centrality parameter τ_1 & τ_2.

Both parameters, τ_1 & τ_2, are therefore estimated as follows.

$$\tau_1 = \frac{\delta - \mu - z_{1-\gamma/2}\sigma}{\sigma\sqrt{\frac{1}{n} + \frac{z^2_{1-\gamma/2}}{2(n-1)}}} \quad (5.2)$$

$$\tau_2 = \frac{\delta + \mu - z_{1-\gamma/2}\sigma}{\sigma\sqrt{\frac{1}{n} + \frac{z^2_{1-\gamma/2}}{2(n-1)}}} \quad (5.3)$$

where $z_{[\bullet]}$ stands for the level of significance α, μ and σ for the mean and standard deviation of differences between two methods, respectively, and δ being the maximum allowed difference considered acceptable.

Assuming we can pre-determined these parameters above, we can estimate using equation 5.2 and 5.3, the respective parameters τ_1 & τ_2. These parameters are then plugged inserted into equation 5.1 including the sample size n to estimate the required power.

In case the sample size is the parameter of interest, we can re-purpose the power formula in a binary search algorithm to look for desired n as outlined in Algorithm 5.1

As such, the script implemented uses the following parameters as input:

- μ expected mean of differences
- σ expected standard deviation of differences
- δ maximum allowed difference between 2 methods
- $\gamma = 0.05$ by default; confidence level of LOAs; 95% of data are inside the LOAs
- $\alpha = 0.05$ by default; confidence level of LOA Confidence Intervals. 95% LOA CI's contain the true population percentile.
- maximum sample size n_{max} before stopping the binary search

Algorithm 5.1: Binary search to estimate sample size

Input: Maximum sample size n_{max}, preferred power β
Output: Number of required samples $n_{estimate}$

```
Set LowestSampleSize ← 0;
Define HighestSampleSize ← n_max = 10000;
while LowestSampleSize ≤ HighestSampleSize do
    PrelimSampleSize ←
     Middle of LowestSampleSize &HighestSampleSize;
    Estimate Power using equation 5.1 using PrelimSampleSize as starting point
    if Power estimated via PrelimSampleSize < preferred power β then
        set LowestSampleSize = PrelimSampleSize ;
    else if Power estimated via PrelimSampleSize > preferred power β then
        set HighestSampleSize = PrelimSampleSize ;
    else
        return n_estimate = PrelimSampleSize ;
    end
end
```

5.2.1 Verification of the Implemented Method

Lu et al's sample size algorithm has been implemented into multiple commercially available statistical software. One such software, MedCalc, uses the method by Lu et al to calculate the said sample size. The company provides an example of calculated sample sizes using reference inputs on their website [Wik]. The example data used to verify the implemented search algorithm 5.1is provided in Figure 5.1. Verification yielded a positive outcome as results as seen in the table presented in Figure 5.1 reflected the values in our testing.

5.3 Implementation of the Bland-Altman method

The Bland-Altman method entails examining pairs of measurements from a group of subjects summarized into a scatter plot. The y-axis represents the difference between two measurements, while the x-axis represents the mean of the two values. The mean of differences μ and SD of differences σ are then calculated, allowing an estimation of the upper and lower limits of agreement. Finally, both, μ and σ are also included in the scatter plot as a horizontal line. The mean difference and limits of agreement are values stemming from a single sample, thus may not reflect the values of the entire population. Therefore estimation of the confidence interval of the above measures provide a more accurate representation. The calculation of the CI of the mean difference and the limits of agreement are displayed in equation 5.4 and 5.5, respectively. μ relates to the mean of differences, while S_d relates to the standard deviation of differences for n samples. $tinv(\bullet)$ stands as an indicator for the defined point of the t distribution (95%) with n - 1

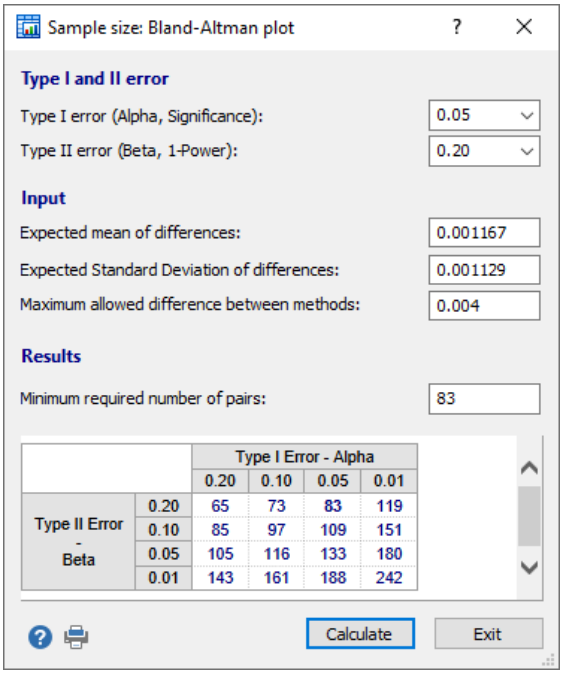


Figure 5.1: Reference values based on the MedCalc Software to verify algorithm 5.1 [Wik]

degree of freedom.

Suppose a data frame with the following structure as outlined in Table 5.1, where *BoneAgePreds* represents the prediction of the AI model and *GroundTruth* indicates the mean of three observers for each subject *patientID*. The concept for performing an analysis of agreement is presented as seen in Algorithm 5.2.

$$CI_{meandifference} = \mu \pm \frac{S_D}{\sqrt{n}} * tinv(1 - \alpha/2, n - 1) \tag{5.4}$$

$$CI_{LOA} = LOA_{upper,lower} \pm S_D * \sqrt{\frac{1}{n} + \frac{1.96^2}{2(n-1)}} * tinv(1 - \alpha/2, n - 1) \tag{5.5}$$

patientID	BoneAgePreds	GroundTruth
1	modelPrediction_pat1	gt_1
2	modelPrediction_pat2	gt_2
3	modelPrediction_pat3	gt_3
n	modelPrediction_patn	gt_n

Table 5.1: Example data frame used to present the analysis of agreement via the Bland-Altman method.

Algorithm 5.2: Implementation of the Bland-Altman method

Input: Data frame as defined in Table 5.1
Output: Bland-Altman Plot

1 Calculate the mean and standard deviation differences of the AI & the ground truth by comparing the assessment of every read against the model's prediction.
2 $\mu \leftarrow Mean(BoneAgePreds - BoneAgeReads)$;
3 $\sigma \leftarrow SD(BoneAgePreds - BoneAgeReads)$;
4 LOA $\leftarrow \mu \pm 1.96 * \sigma$;
5 Estimate the 95% CI based on equation 5.4 & 5.5
6 Define a maximum allowed difference δ.
7 Plot the data and draw the horizontal lines for 95% CI μ, upper and lower 95% LOA, and maximum allowed difference δ.

5.4 Implementation of the Orthogonal Linear Regression

Similar to the standard linear regression, ordinary least squares (OLS), Orthogonal linear regression draws a line of best fit of the data. While the OLS regression fits a line based on the assumption that only the dependent variable is subject to measurement errors, orthogonal linear regression assumes error in both, dependent and independent variables. Pythons SciPy package provides a function to estimate the slope and intercept based orthogonal linear regression [ort]. We utilize this function to implement our regression method to assess proportional bias. Suppose a data frame with the following structure as outlined in Table 5.1. A description of the values in the data frame is provided in Section 5.3. The basic concept for performing an analysis of proportional bias is presented as seen in Algorithm 5.3.

Algorithm 5.3: Implementation of the orthogonal linear regression

Input: Data frame as defined in Table 5.1
Output: Proportional Bias from Regression Plot

1 Estimate the slope and intercept using the function from SciPy based on the output of the AI and the Ground Truth
2 $slope, intercept \leftarrow OrthogonalLinearRegression(AI, GroundTruth)$;
3 Bootstrapping *slope* will yield 95% CI of *slope*
4 Draw the scatter plot and illustrate the line of best fit based on *slope* and *intercept* using the equation $y = slope * x + intercept$

5.5 Implementation of the Concept of Interchangeability

The description of the concept of interchangeability has been explained in section 2.4.3 and graphically summarized in Figure 4.4. Suppose a data frame with the following structure as outlined in Table 5.2, where *BoneAgeReads* represents the reads of each

patientID	reader	BoneAgeReads	BoneAgePreds
1	1	read1	modelPrediction_pat1
1	2	read2	modelPrediction_pat1
1	3	read3	modelPrediction_pat1
2	1	read1	modelPrediction_pat2
2	2	read2	modelPrediction_pat2
2	3	read3	modelPrediction_pat2
continue for n images			

Table 5.2: Example data frame used to present the concept of interchangeability.

individual observer and *BoneAgePreds* indicate the prediction of the AI model for each *patientID*. To note, the value *modelPrediction* is the same for every *patientID*, as the model consistently produces the same output per *patientID*.

As such, the underlying algorithm to assess the performance using the concept of interchangeability is presented in Algorithm 5.4.

Algorithm 5.4: Implementation of the concept of interchangeability

Input: Data frame as defined in Table 5.2
Output: 95% CI Equivalence Index γ

1 Calculate the mean of squared differences of the AI & observers by comparing the assessment of every read against the model's prediction.
2 Mean of squared difference AI $\leftarrow Mean(BoneAgePreds - BoneAgeReads)^2$;
3 Calculate the mean of squared differences of the observers by comparing the assessment of every reader against each other *i.e.* R1, R2, R3 $\rightarrow$ R1vR2, R1vR3, R2vR3.
4 **for** $patientID \leftarrow 1$ **to** *maximum iterations* **do**
5 | Inter-rater-variability $\leftarrow$ Get difference of every observer combination
6 **end**
7 Mean of squared difference Observers $\leftarrow Mean$(Inter-rater-Variability)^2;
8 Equivalence Index $\gamma \leftarrow$ Mean of squared difference AI - Mean of squared difference Observers
9 Bootstrapping with 10000 repeats as suggested by the authors will yield 95% CI of γ

CHAPTER 6

Results

6.1 Introduction

The chapter presents the results of the proposed validation framework highlighted in section 4.1 and therefore answers the research questions as outlined in section 1.1. To recapitulate the study design, the standalone performance data set was sampled from multiple US clinical sites affiliated with a tertiary care center, which serves large portions of various US states. Images were sampled and ground-truthed following the process as described in section 4.6 and 4.4.1, respectively. The total number of images in the standalone performance testing data set was 345 single-hand images. In this data set, there are readings of hand images by three board-certified radiologists for bone age consisting of a total of three reads per image. The mean of the three radiologists established the ground truth. These images were then analyzed by PANDA and the corresponding reports were collected and analyzed for statistical performance as proposed in section 4.4. The methods include a comprehensive statistical assessment based on Bland-Altman and regression and an assessment via the concept of interchangeability. The following results will be presented and discussed in this section.

1. Standalone performance test data set
    a) Study population
    b) Removal of outliers
2. Reliability of the Ground Truth
3. Agreement
    a) Bland-Altman plot of the standalone performance data set
    b) Analysis of agreement

4. Regression analysis

5. Calculation of the equivalence index

6.2 Study Population

Radiographs of the left hand from the institute for pediatric radiology of a tertiary care university hospital were accessible (5541 exams). The patient population of the institute is predominantly Caucasian. Indications for all children and adolescents was the assessment of bone age by the Greulich and Pyle atlas. Only one exam per patient was used in the study. Based on the technical indications of PANDA as described in 2.3.1 we excluded children younger than two years, as well as ages above 17 years. Stratified random sampling was performed to select 23 patients for each year of life, resulting in a study sample of 345 patients with a mean chronological age of 9.77 ± 5.05 ranging between 2 -– 17 years consisting of 178 males (mean chronological age, 10.1 ± 5.22) and 167 females (mean chronological age, 9.46 ± 4.86). The collation of the standalone performance testing set is presented in Figure 6.1 The underlying diseases and indications for the children presented as follows: 56.2% (n=194) Endocrine, nutritional and metabolic diseases, such as small stature, high stature or precocious puberty; 15% (n=52) hereditary related factors such as scoliosis; 11.3% (n=39) normal condition; 6.4% (n=22) mental and behavioral disorders such as developmental delay and 10% (n=38) other conditions.

6.2.1 Failure Rates

PANDA successfully produced output for 345 images resulting in a failure rate of 0/345 = 0%. Failures in this context are considered to be images where PANDA produces an error report rather than a bone age output report.

Exclusions

Out of a total of 345 images, 1 image was excluded from the test set. The reason is provided below. The subject in question is a three years old white females suffering from accelerated growth. The participant was excluded as her radiograph was considered outside the scope of the study. The radiograph consisted of an image of a right hand. Two out of three readers did not provide a bone age assessment for this image because it was outside the scope of the study.

Outliers

The outlier detection via the method of Iglewicz & Hoaglin yielded 4/344 potential outliers from the statistical point of view. Outliers were determined by calculating the robust z-score on the differences between PANDA and ground truth. A robust z-score above 3.5 or below -3.5 was treated as a potential outlier. When an outlier was deemed to indicate a problem that was outside the scope of the intended use or imaging requirements

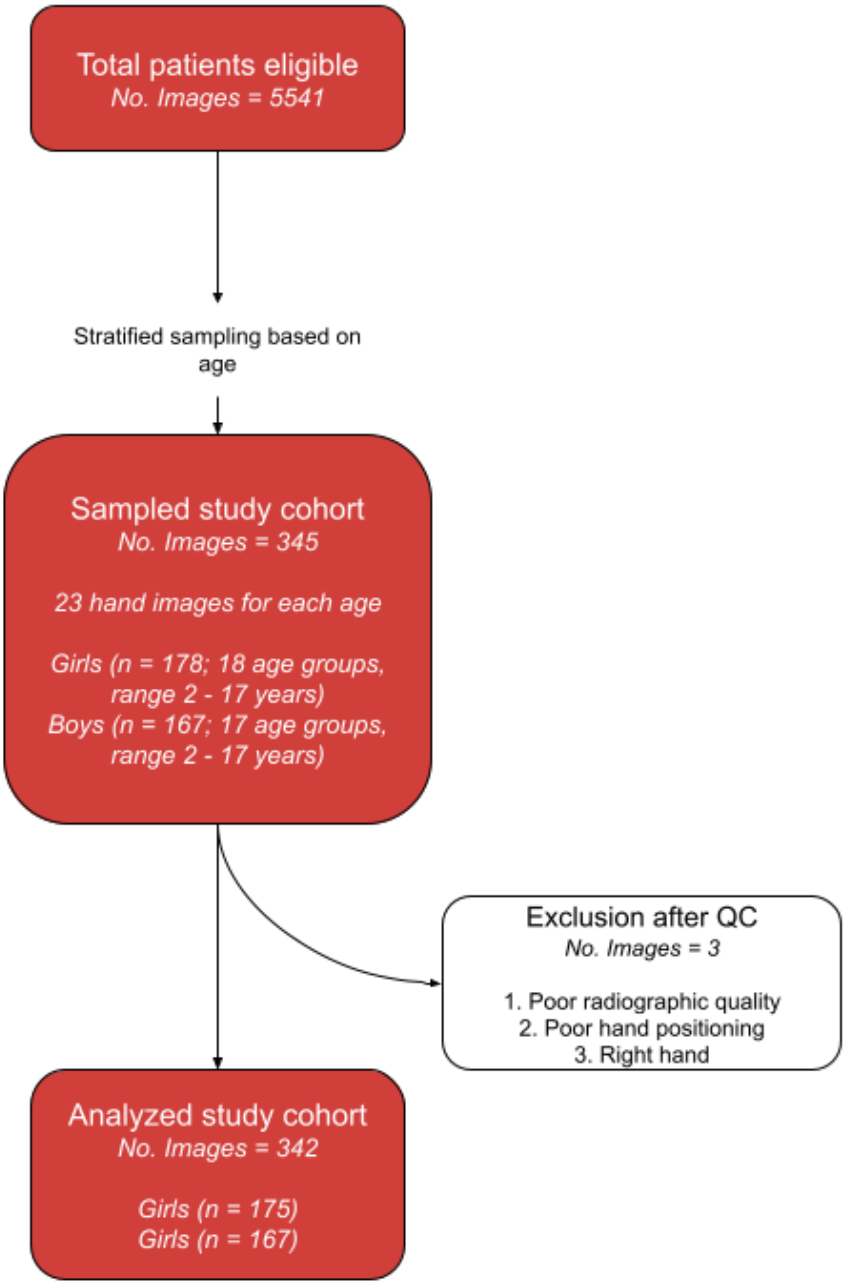


Figure 6.1: Study Sample constitution. From an eligible set of individuals from a clinical center, 345 patients were selected stratified by age between 2 – 17 years old. 23 images were allocated for each bin. After quality control, the cohort eligible for the study were 342 subjects, 175 girls and 167 boys.

then the image was removed from the data set for standalone performance testing. Upon visual inspection by the clinical investigator from the study site, 2/4 potential outlier images were confirmed to be outliers and were removed from the standalone performance test set. The participants excluded were considered outside the scope of the study. The exclusions were justified by not conforming to the image requirements set by PANDA due to poor radiographic quality or poor hand positioning. The concerned subjects were both two years old white females, one suffering from premature thelarche and the other showing signs of poor growth.

Providing unintended inputs to the medical AI software, PANDA, can result in unexpected results, which might significantly deviate from the Ground Truth. These artifacts would affect the reported statistics adversely if they were included in the data analysis. The performance reported would not reflect the true results as the outcome is skewed based on the influence of the artifacts.

The other two potential outliers were conforming to image and data requirements of PANDA following a visual inspection by the clinical investigator and hence were not excluded from the study. After controlling for outliers due to image quality such as poor radiographic quality or poor patient positioning, standalone performance testing was conducted on a set of 342 images as described in Figure 6.1. The total number of patients included in the standalone performance testing, therefore, exceeded the predefined limit of minimum sample size as described in section 4.5 (n=333).

6.3 Reliability of Ground Truth

To address the concerns of reliable ground truth, as to whether the mean of the three reads is considered reliable as ground truth, we assessed the intra-class correlation (ICC) amongst the three expert observers to determine the reliability of the expert radiologists in rating bone age on the standalone performance test set.

Post-hoc power analysis of the ICC was performed in line with techniques described in Walter et al. [WED98]. Using 0.05% significance, 85% power and 3 raters, 21 samples per case were estimated to be sufficient to observe minimum acceptable reliability of 0.85. For this study, a case was considered one age group as described in 4.4.1.

ICC estimates and their 95% confidence intervals were calculated based on a mean-rating ($k = 3$), absolute-agreement, 2-way mixed-effects model. The definition of good reliability is described in Section 4.4.1. To reiterate, based on the 95% CI of the ICC estimate, values less than 0.5, between 0.5 and 0.75, between 0.75 and 0.9, and greater than 0.90 are indicative of poor, moderate, good, and excellent reliability, respectively, based on Koo et al. [KL16]. The values for the majority of the respective age groups show an ICC of over 0.90 as presented in Table 6.1, indicating excellent reliability. The age group for the group of nine year old is slightly lower with an ICC of 0.88, well within good reliability. The lower end of the CI for all age groups shows a downwards trend in reliability starting

Age group	Age Interval	ICC
1 (n=21)	24 to 35 months	0.93 [0.86; 0.97]
2 (n=22)	36 to 47 months	0.96 [0.92; 0.98]
3 (n=23)	48 to 59 months	0.97 [0.94; 0.99]
4 (n=23)	60 to 71 months	0.90 [0.79; 0.95]
5 (n=23)	72 to 83 months	0.94 [0.89; 0.97]
6 (n=23)	84 to 95 months	0.93 [0.86; 0.97]
7 (n=23)	96 to 107 months	0.92 [0.85; 0.97]
8 (n=23)	108 to 119 months	0.93 [0.86; 0.97]
9 (n=23)	120 to 131 months	0.88 [0.77; 0.95]
10 (n=23)	132 to 143 months	0.90 [0.81; 0.96]
11 (n=23)	144 to 155 months	0.94 [0.88; 0.97]
12 (n=23)	156 to 167 months	0.97 [0.94; 0.99]
13 (n=23)	168 to 179 months	0.97 [0.94; 0.99]
14 (n=23)	180 to 191 months	0.97 [0.94; 0.99]
15 (n=23)	192 to 204 months	0.96 [0.93; 0.98]

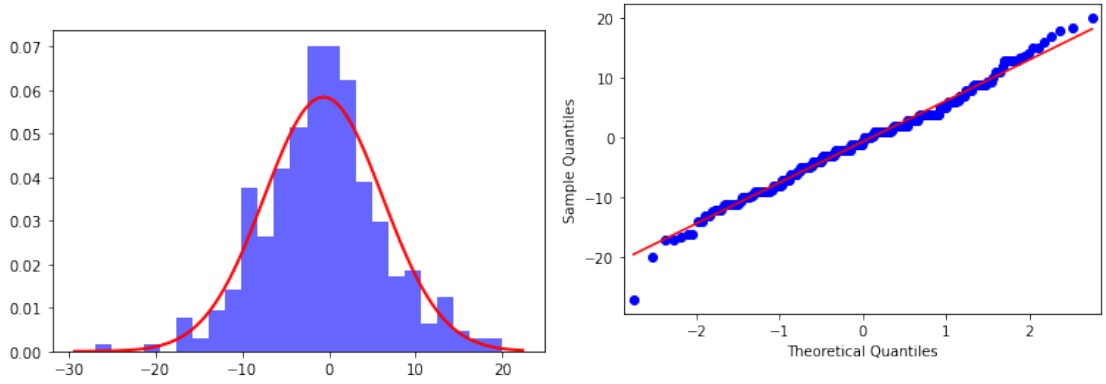
Table 6.1: Results of reliability testing for the mean of the observer assessment as the ground truth via the ICC. The model used is the mean-rating (k = 3), absolute-agreement, 2-way mixed-effects type. The ICC, supplemented by its 95% CI is estimated for each age group. The results show good reliability overall cases based on the definition of reliability as set by Koo et al. [KL16].

from five years to 12 years old. This indicates disagreements between the truthers to some degree within these age groups but negligible based on the criteria set by the ICC.

6.4 Test for Agreement: Bland-Altman's Limits of Agreement

6.4.1 Test for Normality

We outlined in section 3.5.1 that testing for agreement using the Bland-Altman method requires the data to approximate a normal distribution. Specifically, the data based on the differences between the two methods, bone age estimations from PANDA, and the ground truth should approach a normal distribution. An assessment for normality should be performed both, visually and with significance tests [GZ12b]. Therefore we first visualized the data via the histogram and the Q-Q-plot as seen in Figure 6.2 to investigate the distribution of the sample. The histogram illustrates the distribution of differences between PANDA and the Ground Truth. The curve roughly approximates a bell-like shape as seen in Figure 6.2a. The differences between PANDA and the Ground Truth are drawn in the Q-Q-plot. A perfect normal distribution follows a diagonal line on a 45-degree angle. Deviations from the diagonal line indicate deviations from normality



(a) Histogram of the difference between the bone age estimations for PANDA and the ground truth

(b) Q-Q-plot of the difference between the bone age estimations for PANDA and the ground trut h

Figure 6.2: Testing for normality by the means of visualization of the data via Histogram and Q-Q-Plot. Plots indicate potentially a Gaussian Distribution.

[MPS+19]. An approximation of this is seen in Figure 6.2b indicating potentially a Gaussian distribution. Few small deviations can be seen at the bottom left of the plot, which indicates that the data is not perfectly normally distributed. These visual checks were accompanied by statistical tests using the D'Agostino's and Shapiro-Wilk's Test for normality [GZ12b], resulting in p-values of 0.03 and 0.03, respectively. While D'Agostino suggests the data is bell-shaped and Shapiro not. The disagreement of some tests is based on the nature of significant tests being overly sensitive to large sample sizes or not sensitive enough for low sample sizes [MPS+19]. With the support of the illustrated histogram and Q-Q-plot as recommended by the authors, we treated the data as sufficiently Gaussian and proceeded with the analysis of agreement.

6.4.2 Test for Agreement

As established per Section 4.4.1 and verified in Section 6.3, the ground truth for this study is considered to be the mean of bone age assessments made by each of the three human reviewers for each image. Agreement is shown when both, the upper and lower end of the 95% CI of upper and lower limits of agreement (LOA), respectively, of PANDA against the ground truth, are within the average LOAs of the radiologists themselves.

The mean difference reflects the fixed bias indicated as the red dashed line in Figure 6.3. A fixed bias is present when 95% CI for mean difference (illustrated as the red gradient in Figure 6.3) does not include 0. The mean difference between the average of the three readers and PANDA was -0.72 months ($95\%CI : [-1.46; 0.02]$), indicating no significant fixed bias (see Table 6.2). The upper and lower LOA between PANDA and the Ground Truth is 12.98 months ($95\%CI : [11.72; 14.25]$) and -14.42 months ($95\%CI : [-15.69; -13.15]$), respectively, as indicated by the grey-dashed line with the

CI displayed as the grey gradient in Figure 6.3 and detailed in Table 6.2.

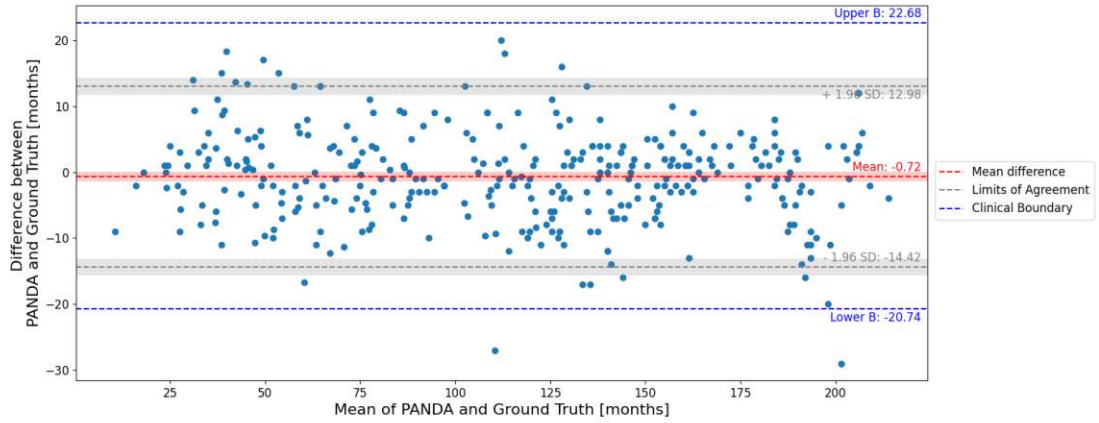


Figure 6.3: The Bland-Altman plot demonstrates agreement between PANDA and the average assessment of radiologists (Ground Truth) of the study population. The shaded red area depicts 95% CI for mean differences and accounts for the absolute bias. The shaded grey area displays 95% CI for Limits of Agreement. The dotted blue lines indicate the mean limits of agreement amongst expert radiologists and serve as acceptance thresholds for PANDA. The LOA based on PANDA and the Ground Truth did not surpass the maximum allowed difference. This indicates that PANDA agrees with the observers.

PANDA vs. Ground Truth				
Mean Difference [months]	PANDA Lower LOA [months]	PANDA Upper LOA [months]	Radiologists Mean Lower LOA [months]	Radiologists Mean Upper LOA [months]
-0.72 (-1.46; 0.02)	-14.42 (-15.69; -13.15)	12.98 (11.72; 14.25)	-20.74	22.68

Table 6.2: Results of testing for agreement between PANDA and the expert readers in table format.

The maximum allowed difference is the average of the inter-rater-variability among each reader pair as listed in Table 6.3 Given the definition of 95% limits of agreement (LOA = average difference ±1.96 * standard deviation), with an average of mean differences of -0.97 months and a mean of the standard deviation of differences of 11.07 as seen in Table 6.3, the upper and lower limits of agreement is 22.68 and -20.74, respectively (-0.72 [-20.74, 22.68]). These LOA are the average maximum difference among the radiologists and establish the maximum allowed boundary limits marked as the blue dashed lines in Figure 6.3 .

The alternative hypothesis state, two methods are in agreement if

$$H_1 : -\Delta \leq 95\%(\mu_{Panda} - \mu_{Rads}) \leq +\Delta$$

	Mean Difference [months]	Standard Deviation [months]	Average of Mean Differences & Average of Mean Standard Deviation [months]	Average LOA [months]
Reviewer 1 vs. Reviewer 2	0.44	10.45	[0.97; 11.07]	[-20.74; 22.68]
Reviewer 1 vs. Reviewer 3	1.46	12.81		
Reviewer 2 vs. Reviewer 3	1.01	9.97		

Table 6.3: Mean Difference and Standard Deviation of each observer pair in months. The mean of the observer pair LOA's establishes the maximum allowed difference.

As seen in Table 6.2, the upper and lower boundary of the 95% CI of the upper and lower LOA between PANDA and mean observers did not exceed the upper and lower LOA of the mean observers.

Alternatively, we express this as:

$H_1 : -20.74$ months≤ -15.69 months $\wedge$ 14.25months ≤ 22.68 months

Mathematically, this demonstrates that the assessment of the model agrees with the assessment of the average reads of observers.

The analysis of agreement shows that relying solely on radiologists to estimate bone age will result in differences between -20 months to 23 months of under- and overestimation, respectively. Using PANDA would reduce these differences to -16 months to 14 months of under- and overestimation, respectively. Figure 6.3 displays the Bland-Altman plot illustrating our description above. The upper and lower bound of the shaded grey area indicating the upper and lower bound of the 95% CI of LOA does not surpass the maximum allowed boundary presented as the blue-dashed line.

Subgroup Analysis - Sex

Based on the clinical relevance of bone age we also assessed the performance stratified by gender, for boys and girls, respectively. The summarized results are displayed in Table 6.4 and Figure 6.4 and 6.5 for male and female populations, respectively.

	Sample	Mean Difference [months]	Limits of Agreement [months]		Maximum Boundary [months]	
			Upper LOA	Lower LOA	Lower B.	Upper B.
Cohort for girls/boys	**Boys(n = 167)**	1.38 [0.34, 2.42]	14.73 [12.95; 16.51]	-11.97 [-13.75; -10.19]	24.80	-17.40
	Girls(n = 175)	-2.73 [-3.71, -1.75]	-15.59 [-17.26; -13.92]	10.13 [8.46; 11.80]	19.28	-22.55

Table 6.4: Performance of PANDA against ground truth for girls/boys.

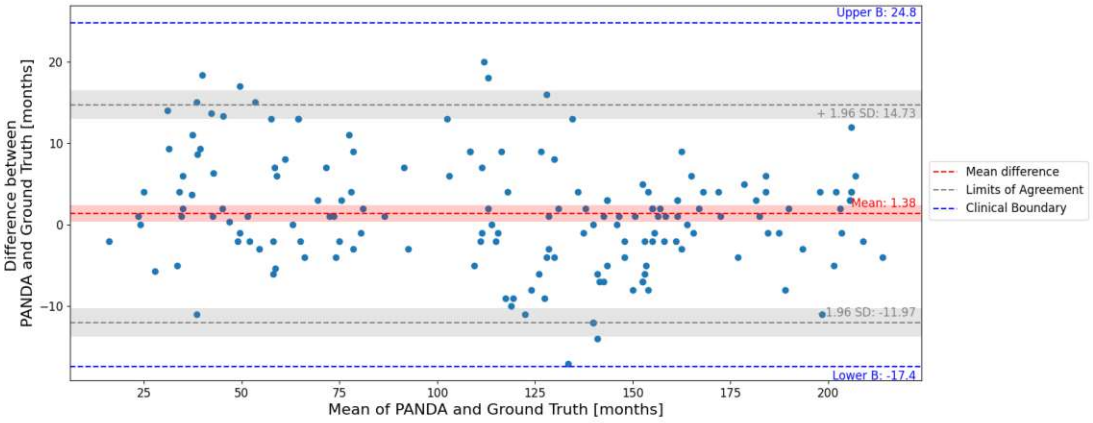


Figure 6.4: Bland-Altman plot demonstrates agreement between PANDA and the Ground Truth of the male study population. The shaded red area depicts 95% CI for mean differences and accounts for the absolute bias. The shaded grey area displays 95% CI for Limits of Agreement. The dotted blue lines indicate the mean limits of agreement amongst expert radiologists and serve as acceptance thresholds for PANDA. The LOA based on PANDA and the Ground Truth did not surpass the maximum allowed difference. This indicates that PANDA agrees with the observers.

As presented in Table 6.4 and visualized in Figure 6.4 and Figure 6.5, the upper and lower boundaries of the 95% CI of the upper and lower limits of agreement, respectively (shown as the grey gradient), of the male and female subgroup fall within the respective maximum boundaries (illustrated as the blue-dashed line) and therefore demonstrate good agreement. Mathematically, we express this for males:

$H_1 : -17.40$ months ≤ -13.75 months $\wedge 16.51$ months ≤ 24.80 months

and for female:

$H_1 : -22.55$ months ≤ -17.26 months $\wedge 11.80$ months ≤ 19.28 months

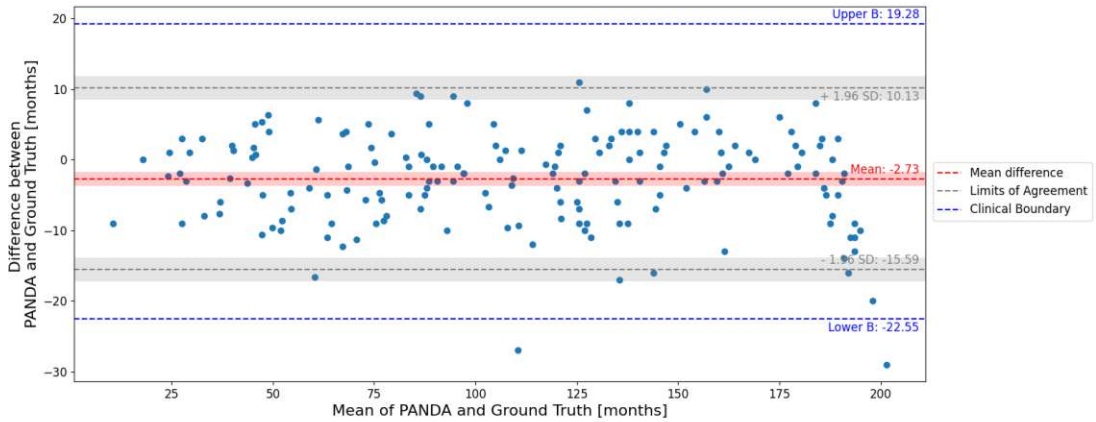


Figure 6.5: Bland-Altman plot demonstrates agreement between PANDA and the Ground Truth of the female study population. The shaded red area depicts 95% CI for mean differences and accounts for the absolute bias. The shaded grey area displays 95% CI for Limits of Agreement. The dotted blue lines indicate the mean limits of agreement amongst expert radiologists and serve as acceptance thresholds for PANDA. The LOA based on PANDA and the Ground Truth did not surpass the maximum allowed difference. This indicates that PANDA agrees with the observers.

Based on the results in Table 6.4, the 95% CI of the mean difference does not include 0, therefore indicating significant fixed bias for both sexes. The assessment shows that PANDA has a slight tendency to overestimate bone age for boys of approximately one month and underestimate bone age for girls of about three months.

6.5 Test for Presence of Proportional Bias - Slope of Regression

Proportional bias was tested using orthogonal linear regression between PANDA and the Ground Truth, and assessing whether the slope of the regression line is near 1. The slope of the regression reflects the proportional bias. A proportional bias is present when 95% CI for slope does not include one. Estimates and confidence intervals obtained for slope and intercept are presented in Table 6.5 and visualized in Figure 6.6.

The alternative hypothesis state, the two methods are not proportionally biased if

$H_1 : B_1 = 1$

The 95% CI for the slope indicated by the red line in Figure 6.6 is reported to be $B_1 = 1.02(1.00, 1.03)$ for the entire cohort. The slope includes $B_1 = 1$. Therefore no significant proportional bias is present.

PANDA vs. Ground Truth	
Slope	**Intercept**
1.02 (1.00; 1.03)	-0.72 (-1.46; 0.02)

Table 6.5: Intercept and slope of orthogonal linear regression of the study population. The 95% CI for the slope includes 1 and therefore indicates no proportional bias.

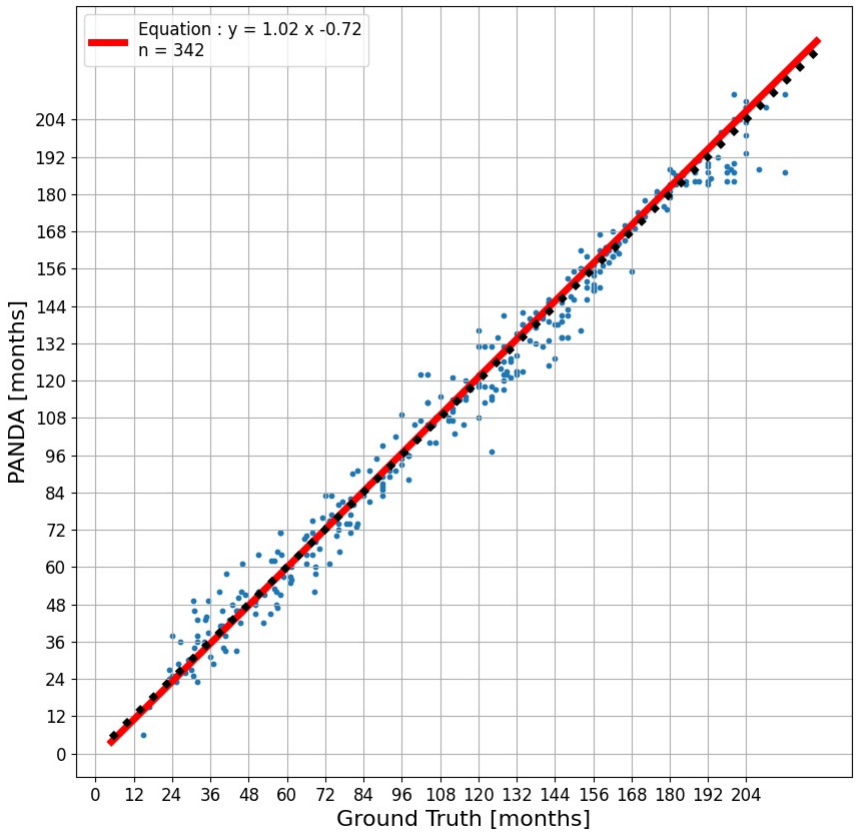


Figure 6.6: Regression analysis between the ground truth (mean assessment of radiologists) and PANDA. The dashed-black line describes the line unity with the slope $B_1 = 1$ and intercept $B_0 = 0$. The red line illustrates the line of best fit based on the orthogonal regression model. The slope indicates the presence of a proportional bias. 95% CI for slope includes 1 and demonstrates no significant proportional bias.

6.5.1 Subgroup Analysis - Sex

Based on the clinical relevance of bone age we also assessed the performance stratified by gender, for boys and girls, respectively. The summarized results are displayed in Table 6.6 and Figure 6.7 and 6.8 for male and female populations, respectively.

PANDA vs. Ground Truth		
	Slope	**Intercept**
Boys (n=167)	1.02 [1.00; 1.03]	1.38 [0.34; 2.42]
Girls (n=175)	1.01 [0.99; 1.03]	-2.73 [-3.71; -1.75]

Table 6.6: Intercept and slope of orthogonal linear regression of male and female the study population.

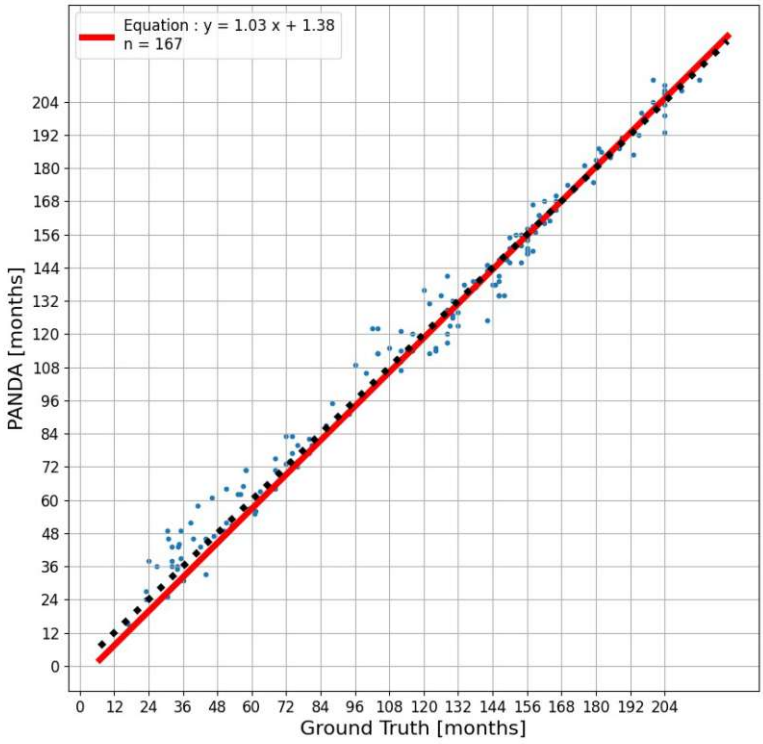


Figure 6.7: Regression analysis between the ground truth (mean assessment of radiologists) and PANDA for the male study population.

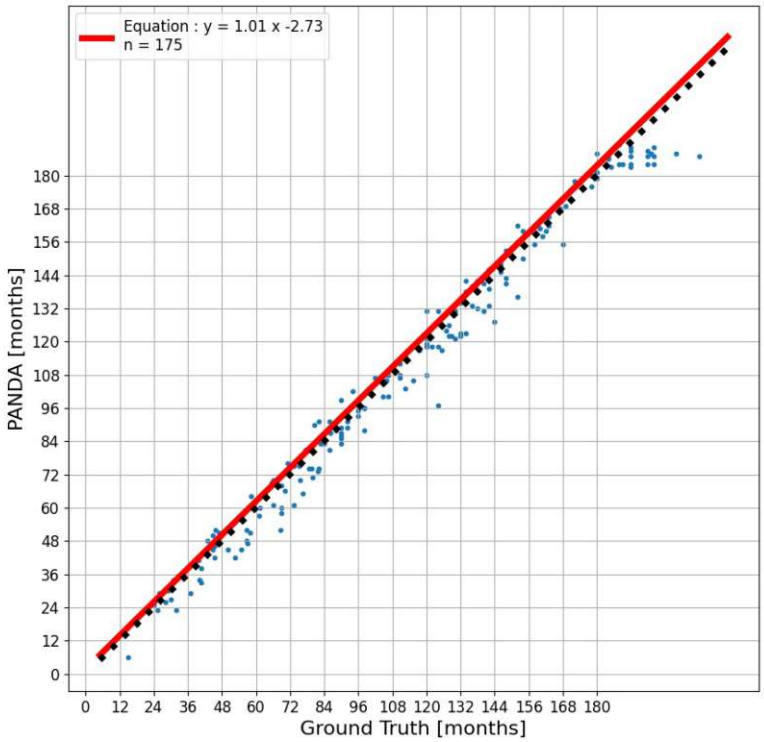


Figure 6.8: Regression analysis between the ground truth (mean assessment of radiologists) and PANDA for the female study populations.

As presented in Table 6.6 and visualized in Figure 6.7 and Figure 6.8 for the male and female cohort, respectively, the dashed-black line describes the line unity with the slope $B_1 = 1$ and intercept $B_0 = 0$. The red line illustrates the line of best fit based on the orthogonal regression model. A proportional bias is present when 95% CI for slope does not include one. The 95% CI for the slope indicated by red line in Figure 6.6 is reported to be $B_1 = 1.02(1.00, 1.03)$ and $B_1 = 1.01(0.99, 1.03)$ for the male and female cohort, respectively. The slope includes $B_1 = 1$ for both population. Therefore no significant proportional bias is present.

6.6 Test for interchangeability - The Equivalence Index

The results of the interchangeability test are presented in Table 6.7. The mean squared difference between PANDA and the assessments made by random readers was 91.2; the mean squared difference between two assessments made by random readers (Radiologist 1, 2, and 3) was 125.4. The estimated equivalence index is -34.2. To make the result interpretable, the square root of the absolute value is taken, resulting in an index of γ = -5.8, with a 95% CI of -7.1 to -4.8 (via a bootstrap percentile CI). The equivalence index is below the pre-specified acceptance criteria of $\gamma \leq 0$. The negative equivalence index indicates that switching PANDA in place of an expert reader would lead to an average reduction in differences between bone age estimates as compared to expert readers alone. Based on the results of the interchangeability test, PANDA is non-inferior to the readers and therefore is interchangeable with the readers.

PANDA vs. Random Readers (Radiologist 1, 2 & 3) - Mean squared difference [$months^2$]	**Random Readers (Radiologist 1, 2 & 3) among each other Mean squared difference** [$months^2$]	**Equivalence Index** [$months^2$]	**Signed square root of absolute value of Equivalence index** [$months$]
91.2	125.4	-34.2	-5.8 (-7.1; -4.8)

Table 6.7: Results of testing for interchangeability between PANDA and the expert readers.

CHAPTER 7

Discussion

7.1 Results and Limitations of the Standalone Performance Testing Set

In Section 2.2.6 we established that the important parameters for bone age assessment are age and sex. To answer Research Question Q1 *How does the clinical aspect of the output of interest influence the distribution and granularity of the data set to be tested?* as part of the aim of the thesis in Section 1.4, we evaluated the performance of our AI model on the data set stratified by age from multiple centers in the US. Based on our assumption of the US census as discussed in Section 4.6, simple random sampling resulted in approximately equal distribution among sex. Our sampling strategy resulted in a uniform distribution as seen in Table 6.1 except for outliers considered in section 6.2.1. This strategy allowed testing on a sample with enough granularity and adequate representation over all ages and sex as defined by the intended patient population of PANDA.

Thus, the answer to the question is, bone age specifically is assessed differently for male (m) and female (f) in every stage of age from 3 months to 18 and 19 years, for girls and boys, respectively. As such, the testing data sampled must ensure sufficient samples in age and sex, the two parameters relevant for bone age assessment. Table 4.4 presents the distribution of the data to be tested on.

We solved the answer to Research Question Q3 *How can we estimate sufficient sample size and power?* by investigating sample size estimation methods available based on our hypothesis testing for agreement. An accepted approach and implemented in multiple statistical software, e.g. MedCalc, PASS, is based on Lu et al's method. The general equation provided by Lu et al's paper to estimate the sample size (refer Equation 3.2 included a boundary condition assuming the mean difference $\mu = 0$. As a certain amount of bias should be expected from either side, the AI or the expert readers, the mentioned

formula could not be applied to our case. Therefore, as described in Section 5.2, we utilized the author's power formula as seen in Equation 5.1 and applied an iterative method to estimate the required sample size based on the desired power. This approach has also been suggested by the authors. The concept of the binary search algorithm has been outlined in Algorithm 5.1. The implemented method has been verified against the expected inputs and outputs based on the reference values from the statistical software, *MedCalc*, as seen in Figure 5.1. These values are listed on their website [Wik]. The sample size for the standalone performance data set was estimated using reference parameters based on literature from the Larson study, as presented in 4.5. Applying the parameters to the script resulted in a minimum acceptable sample size of 333 images for a power of 85%.

Thus, the answer to this question is the following: Sample size is closely tied to statistical power and significance. A power analysis is most often used to calculate what sample size is required. Assuming three of the four values out of the parameters — sample size, effect size (mean difference and standard deviation), significance level, or power — are known, the final parameter can be calculated. In clinical studies, these parameters are usually estimated from data in existing literature or are obtained from the results of a pilot study. Our study relied on data from existing studies, specifically the Larson study.

Our data set is collected to address the performance of the entire population suspected to have a bone age assessment limited to the patient population of PANDA. In some cases, one might be more interested in how the AI performs for specific age groups. Specifically for younger and older age groups where subjects experience rapid growth is in some cases interesting to assess. Our data set as listed in Table 4.4 provides only 23 images per age group. The number of images will lack the power to provide statistically sound results. This is one limitation of the data which we tested on.

7.2 Results and Limitations of the Reliability of the Ground Truth as the Mean

To ensure that our device met its intended use of providing a radiological bone age estimate with accuracy and precision that was clinically meaningful, performance must be measured against an appropriately validated ground truth. If the ground truth was to be obtained through a mean of radiological bone age estimates performed by expert radiologist readers, as performed for this study, we had to ensure that the variability of the estimates is low enough. This is done to provide confidence in the validity of the ground truth against which PANDAs' performance was measured.

To ensure inter-reader variability does not negatively impact the ground truth, we considered reader qualifications, training, and recruitment of multiple readers in the truthing process. In addition, we also provided the Intra-class Correlation (ICC) among the truthers to determine the reliability of the expert radiologists after the data has been truthed. The ICC value provided should reflect the inter-reader variability.

As mentioned in Section 4.4.1, the ICC value is constructed by the variance components for the reader, case, and random errors for an ANOVA model. We ensured that the case variability is smaller than the reader variability. This is to ensure the ICC value will be dominated by the reader variability (based on the expert radiologists) and not the case variability (bone age). If the case variability is much larger than the reader variability, the ICC value will be dominated by the case variability and cannot reflect the reader variability. Therefore we performed analysis for each age group based on the majority of the plates reflected in the GP Atlas.

The results provided indeed show good reliability using the mean of the experts reads as the ground truth. With an ICC of over 0.90, Koo et al. defined this statistic as excellent reliability.

7.3 Results and Limitations of Testing Methodology

We evaluated multiple methods for assessing the performance of AI models with outputs with continuous variables. For this thesis, we used bone age assessment as an example.

To answer to Research Question Q2 *How can the performance of the software whose output is continuous be shown and which choices of performance metrics and performance targets are available and feasible?*, we proposed and executed a study including a statistical analysis using the Bland-Altman plot, regression & the concept of interchangeability.

The Bland-Altman plot estimates presented in Section 6.3 relies on a reference standard, the ground truth, to compare to. We established and validated our assumption of the reference standard as the mean of the observers' estimate. The limits of agreement *i.e.* the extreme differences to be expected from using PANDA over the radiologist between -14.42 and 12.98 months, were below the maximum acceptable boundary we defined as the limits of agreements based on the reader inter-rater variability as seen in Table 6.2 with a difference between -20.74 and 22.68 months.

Testing the hypothesis for agreement as described in Section 4.3.2 and supported by the results in Section 6.4.2 presents

$H_1 : -20.74 \text{ months} \leq -15.69 \text{ months} \wedge 14.25 \text{months} \leq 22.68 \text{ months}$

Based on this assessment we can safely reject the null hypothesis H_0. Therefore, we can conclude that PANDA agrees with the current reference standard. We have as such provided evidence that integrating the AI into the clinical workflow, supports the radiologist in reading bone age more accurately of approximately 6 months on average.

Subgroup analysis based on sex shows that the mean differences, an indicator for fixed bias, reported in Table 6.4 between PANDA and expert readers vary from 1.38 months for boys to -2.73 months for girls. These differences could reveal tendencies of either PANDA or expert readers towards slight over-or under-estimations of GP bone age, but we do not see the differences reported here rising to a level of clinical significance. We assume the presence of a fixed bias is influenced by the precision of GP bone age estimates. The

outputs provided differs between PANDA (nearest month model) and experts (nearest plate from the atlas or between two plates). The result is a binning effect, as expert reader estimates are made in 3-month intervals.

One limitation we found during testing for agreement was our defined maximum boundaries as 22.68 and -20.74 months (upper and lower, respectively), the difference one would expect from the clinical practice. Similar results have been report by Larson et al with cases up to 24 months [LCL+18]. These differences reflect the rare (5% based on the definition of the limits of agreement) and extreme cases that one can expect when relying solely on readings of the experts. However, the proposed maximum acceptable difference among the radiologists is large enough to invalidate the clinical meaning of the reading for the younger age group. Being allowed about 1-2 years off the ground truth especially for the younger age group might be concerning. This can be explained as the threshold is based on the study population examined resulting in the establishment of the maximum allowed difference. In our analysis, the cases range from 2 years to 16 years. The boundary generalizes to the studied group and can therefore not accurately reflect the boundary tailored to the younger age group specifically. However, as seen in Figure 6.3 these cases of such differences occurring are rare. Nonetheless, this limitation for the younger age group should be further investigated. To address the matter of generalizability as required by Section 1.4.1 we proposed a method that does not rely on a ground truth to compare to *i.e.* the concept of interchangeability.

We presented in this study a method of evaluating clinical acceptance of a bone age AI algorithm using interchangeability, which incorporates the randomness of clinical reads made by different experts and simulates the impact of adding the AI algorithm into this workflow. An interchangeable reader would go unnoticed, in that inter-reader differences would not change when that reader began making bone age estimates. In this study, the introduction of the AI bone age algorithm PANDA led to a reduction in inter-reader differences of approximately 6 months on average based on the results in Table 6.7.

Testing the hypothesis for interchangeability as described in Section 4.3.2 and supported by the results in Section 6.6 presents

$H_1 : \gamma \leq -5.8$

Based on this assessment we can safely reject the null hypothesis H_0. Therefore, we can conclude that PANDA is interchangeable with the current reference standard.

Thus to summarize, if the researcher wants to calibrate one measurement against another or find bias between two techniques of measurement, regression analysis might be performed. However, if the purpose is to see if one procedure may be safely replaced by another, especially in clinical practice, the Bland-Altman plot is preferred. If there is no adequate reference standard to compare to, required for the Bland-Altman analysis, interchangeability can be used. This statistical method does not rely on a reference standard to compare to. It assesses any excess or reduction in differences from interchanging one method over the over.

7.4 Reflection on the Scientific Outcomes of the Thesis

This section reflects on how this thesis contributes to the state-of-the art, overall. We assess to what degree we have fulfilled the overall goals as set in Section 1.4 based on the evidence in our results.

As stated in Section 1.2, the current state-of-the-art does not address how the performance of AI in medical imaging whose output is continuous *e.g.*, bone age, distance, or angles on radiographs, should be assessed. Due to the recent application of AI methods in medical imaging, there is currently no reference standard or clear guidance for assessing the performance of AI-based technology in healthcare with continuous values.

Specifically, for the estimation of bone age, a difficult and time-consuming task, where AI can provide a solution within a short amount of time, not much research relating to performance assessment has been done in this field, as explained in Section 3.6. To the best of our knowledge, this is the first comprehensive methodology to fill this gap in literature.

The overall aim of the entire research project was to explore and execute suitable statistical models in a clinical study to assess the performance of an AI software used in healthcare. In Chapter 2, we determined the role of AI as supplementary, a second opinion, using different means to assess the same output, compared to the gold standard, the radiologist. Therefore, we concluded this to be a method comparison study. The Bland-Altman method has established itself in multiple areas of medicine as an appropriate technique for comparing methods against each other, and may help researchers to compare a new method against another one or a reference standard. Our research as outlined in Chapter 3 show that the analysis via Bland-Altman has recently begun to see the application in the field of AI, specifically in the assessment of bone age. We also investigated other statistics that do not rely on a reference standard, *i.e.*, the concept of interchangeability, already applied in the pharmaceutical field. In addition, we have also addressed the issue of the presence of bias and proposed the analysis using regression.

While research papers attempted to evaluate the performance of such models, many lack important aspects in terms of sample size or adequate representation in the testing set, based on the clinical relevancy, to justify good performance of their model. These shortfalls were addressed in our performance testing as outlined in Chapter 4. We created a study protocol that could deliver meaningful results considering but not limited to study population and methods, sample size, and power. Our study design consisted of partially writing scripts to analyze and visualize the data, as well as addressing any statistical considerations relating to sample size and power. The details have been presented in Chapter 5.

Our results presented in Chapter 6 support the following claim. If the researcher wants to calibrate one measurement against another or find bias between two techniques of measurement, regression analysis should be performed. If the purpose is to see if one method may be safely replaced by another, especially in clinical practice, the Bland-

Altman plot is preferred. If there is no adequate reference standard to compare to, a requirement for the Bland-Altman analysis, interchangeability can be used. This statistical method does not rely on a reference standard to compare to.

7.5 Reflection on the Tasks and Requirements of the Thesis

In Section 1.4, we defined tasks and requirements to address the answers to the questions (Q1 - Q3) interesting for researchers. To this end, we will discuss to what degree they were accomplished.

Task 1 *Understanding the principles of bone age assessment, its clinical output including the intervention of AI and the relevance of the clinical assumptions into the statistical considerations* has been addressed in Section 2.2, resulting in the solution as presented in Section 7.4 for Research Question Q1. We established in Section 2.2 that bone age is age and sex dependent. Therefore, the standalone performance testing data set should contain enough granularity in terms of both, age and sex based on the clinical implications of bone age and the intended patient population of PANDA. The underlying distribution must be uniform. By equally distributing samples, enough granularity in the performance data can be provided.

Task 2 *Exploring the current methodologies of performance assessment of bone age and possible limitations* addressed part of Research Question Q2 and Q3. In Chapter 2 and further emphasized in 3 we have established that the Bland-Altman Method for agreement, regression and the concept of interchangeably statistical methods that should be used to assess performance. These methods addresses the issue of generalizability and scalability, as required by Requirement R1 and Requirement R2. The statistical techniques mentioned above can be applied to any clinical output of continuous nature irrespective of the use case [SSA+21]. It is not bound to the assessment of bone age specifically, therefore fulfilling Requirement R1. These methods are established techniques applied by many fields of clinical research [Dog18, Com]. As such, the proposed methods accomplishes the requirement as defined in Requirement R2.

We addressed Task 3 *Proposal of an improved and more robust framework for performance analysis* by applying these methods resulting from Task 2 into a clinical study as outlined in Chapter 4. We ensured that the study was sufficiently powered as described in Section 4.5, defined a more robust sampling method as outlined in Section 4.6, and established an adequate reference standard and performance targets to compare to in Section 4.4.1 and 4.4, respectively. We also ensured the fulfillment of Requirement R3 by providing the information for the statistical methods, sample size calculation in a reproducible manner as presented in Chapter 5.

To conclude, our thesis has adequately addressed every task presented in this work while incorporating the constraints as the set by the requirements.

CHAPTER 8

Conclusion & Future Works

Already 20 years ago Tanner and Whitehouse suspected *"From the beginning, it seemed reasonable to suppose that bone age assessments were something a computer could do better than a human operator"* [Car02]. As this study demonstrates, the two great authorities in bone age assessment will be proven right in the long run. This chapter summarizes the work presented and provides a conclusion to the investigation conducted. In addition, an outlook for possible improvements and other future topics are given.

8.1 Summary

The main research question defining this thesis was *Which statistical strategies support researchers in demonstrating safety and performance of an AI algorithm whose output is continuous?* As demonstrated in this thesis, one can use Bland-Altman analysis for agreement, regression or interchangeability as defined and applied in Chapter 4 and Chapter 6, respectively, to demonstrate safety and performance of an AI algorithm whose output is continuous. When applying these statistical tests, one needs to ensure that the test is sufficiently powered in order to claim significant results (Section 4.5). Finally, clinical implications behind the output in question will determine the distribution of the sample that is tested on. We addressed this issue for bone age as presented in Section 2.2. This can differ from other clinical outputs and should be evaluated beforehand.

In this retrospective study, we assessed an automated computerized solution for bone age assessment and compared its performance against expert readers. The role of AI in the clinical setting is currently supplementary. A time-saving tool providing a different opinion compared to its current referenced standard, the human operator. Therefore, it is essential to assess whether two different approaches agree with each other and whether the differences resulting from using one method over the other will result in clinically relevant. We present in this study methods of evaluating clinical acceptance of a bone age AI algorithm using the measure of agreement based on the Bland-Altman analysis

and interchangeability. While agreement relies on a reference standard to compare to, a ground truth will not always to available or can be applied. Therefore we presented an alternative solution utilizing the concept of interchangeability. This method does not rely on a reference standard, as it incorporates the randomness of clinical reads made by different experts and investigates whether replacing the AI with an expert would result in unacceptable differences.

Based on the current state-of-the-art we proposed using the analysis of agreement via Bland-Altman and interchangeability as adequate techniques to assess performance. To assess the effectiveness of the methodologies, the proposed statistical analysis was escalated to a clinical investigation.

We evaluated both approaches on a data set sampled from a multi-center clinical site stratified by age to evaluate the performance for each age with the same weight. Three pediatric experts in reading bone provided assessments to each sample. Ground truth was established using the mean of three estimations. An important step was to provide a power analysis to the study. This is done to ensure that the results acquired from the investigation are powered and the results are significant.

For the analysis of agreement, we defined the average inter-rater variability based on the average LOA of each reader pair from expert radiologists as the maximum allowed difference considered acceptable. The results of the experiment support the conclusion for good agreement showing differences occurring using the AI are lower than the differences occurring among the expert observers. The results from assessing for interchangeability indicate that using the AI in place of an expert reader would lead to an average reduction in differences between bone age estimates as compared to expert readers alone. Both approaches, therefore, support the idea that looking at the differences and assessing whether such differences are acceptable in the clinical practice determines whether a device is in agreement or interchangeable.

The main contribution to this thesis is a workflow for the statistical assessment of AI solutions in bone age assessment. In conclusion, the results presented in this study show promising results for the proposed methodologies. Both metrics are not restricted to the assessment of bone age and can also be applied to other output of interest provided the output is a continuous variable. Whether one assesses performance using Bland-Altman's limits of agreement or interchangeability strongly depends on the clinical aspect of the output of interest. It is up to the researcher to decide which method is adequate based on the use-case of the device.

Our proposed approach is indeed generalizable to the other applications aside from bone age. We have also applied our framework to length and angle measurements of the lower extremity, a different diagnostic output, whose output is a also continuous variable [SSA+21]. Aside from assessing for agreement with the Bland-Altman-Approach, we also assessed for non-inferiority using the concept of interchangeability by quantifying the equivalence index γ.

8.2 Future Work

Finally, we present possibilities for further improving our framework including future investigations.

Diverse and Larger Data Set

From the result of the performance testing, it seems that there is no significant bias between the AI's prediction and the ground truth and also no apparent trend of difference over the magnitude. Bone age is influenced by gender, race, living environments, social resources, and nutritional status. While the performance testing data set satisfies the minimum requirements from the statistical and clinical aspect, further evaluation on a more diverse and larger data set specific for age, race, and sex might provide even more insight on whether a bias of the algorithm exists.

Clinically Acceptable Difference

Additional measures could be taken to further improve the quality of the ground truth. This inherently also affects the maximum threshold δ established based on the mean of the LOA of the observer pairs. Even though actions were taken to reduce variability as much as possible based on the observers' experience, competence, and training, differences among the observers are still very large. Aside from the inherently flawed manual process of reading bone age, it appears that the broad clinical boundaries may be still impacted by substantial intra-reader variations in bone age estimations. This could be one of the possible reasons for the large threshold. To address the concern of high intra-reader variability that may be contributing to clinically potentially questionable boundaries, we may consider using more than three readers. This measure could reduce any potential selection bias for the readers. Alternatively, the ground truth can be determined by allowing the three truthers to reach a consensus on the estimated bone age if the intra-reader variability,*i.e.*, the difference in bone age assessment among each other exceeds a certain amount.

List of Figures

List of Tables

List of Algorithms

Glossary

AI Artificial Intelligence. 1–5, 7, 10, 13, 14, 29–35, 37, 39–41, 44–46, 55, 56, 60, 71–79, 82

BA Bone Age. 11

CA Chronological Age. 11

CAD Computer Aided detection. 1, 10

DICOM Digital Imaging and Communications in Medicine. 11–13

FDA Food & Drug Administration, agency responsible for protecting and promoting public health. 2, 4, 16

GCP Good Clinical Practice. 5

GP Greulich and Pyle. 8–11, 29, 33, 38, 39, 73, 82

MAD Mean Absolute Deviation, a metric to assess accuracy, similar to the standard deviation it measures the absolute difference between two measures and averages it. This metric provides information about the absolute difference of two measures irrespective of whether the difference is positive or negative.. 30, 32, 33, 39

MAD Root Mean Square Error, a metric to assess accuracy, similar to the standard deviation it measures the square of the absolute difference between two measures and averages it. Due to the differences being squared, this metric penalizes outliers. This metric provides information about the absolute difference of two measures irrespective of whether the difference is positive or negative.. 33

MDR Medical Device Regulation. 4

OLR Orthogonal Linear Regression. 4

PA Projection of an X-ray in which beam path is from the back (posterior) to the front (anterior) to the body. 12, 13

PACS Picture archiving and communication system, technology providing storing and easy access to images and reports of multiple modalities. 12, 13

PANDA abbreviation for Pediatric Bone Age and Developmental Assessment, an AI model that estimates bone age, a clinical output considered a continuous variable. xiii, 2–6, 10–13, 37, 38, 40–44, 49, 57, 58, 60–67, 70–74, 76, 81–83, 85, 86

RSNA Radiological Society of North America, a peer-reviewed scientific journal that has been operating since 1923. 29, 30, 85

SaMD Software as Medical Device. 4

SD Standard Deviation. 11, 53, 55

TW Tanner and Whitehouse. 8, 9

Bibliography

[ACR] FDA cleared AI algorithms. https://models.acrdsi.org/. Accessed: 2021-09-02.

[Alt90] D.G. Altman. *Practical Statistics for Medical Research*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 1990.

[BA86] J. Martin Bland and Douglas G. Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, pages 307–310, 1986.

[BA99] J Martin Bland and Douglas G Altman. Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, 8(2):135–160, 1999. PMID: 10501650.

[BDB+01] Matthew Berst, Lori Dolan, Marta Bogdanowicz, Max Stevens, Shirley Chow, and Eric Brandser. Effect of knowledge of chronologic age on the variability of pediatric bone age determined using the greulich and pyle standards. *AJR. American journal of roentgenology*, 176:507–10, 03 2001.

[BEK+99] R K Bull, P D Edwards, P M Kemp, S Fry, and I A Hughes. Bone age assessment: a large scale comparison of the greulich and pyle, and tanner and whitehouse (tw2) methods. *Archives of Disease in Childhood*, 81(2):172–173, 1999.

[Bla] Martin Bland. How can i decide the sample size for a study of agreement between two methods of measurement? https://www-users.york.ac.uk/~mb55/meas/sizemeth.htm. Accessed: 2021-12-04.

[Bla15] M. Bland. *An Introduction to Medical Statistics*. Oxford medical publications. Oxford University Press, 2015.

[BR90] Paul T Boggs and Janet E Rogers. Orthogonal distance regression. *Contemporary Mathematics*, 112:183–194, 1990.

[BSY+07] Bora Büken, Alp Alper Safak, Burhan Yazıcı, Erhan Büken, and Atilla Senih Mayda. Is the assessment of bone age by the greulich–pyle method reliable at forensic age estimation for turkish children? *Forensic Science International*, 173(2):146–153, 2007.

[BTSK16] Micheál Breen, Andy Tsai, Aymeric Stamm, and Paul Kleinman. Bone age assessment practices in infants and older children among society for pediatric radiology members. *Pediatric Radiology*, 46, 08 2016.

[Bur21] US Census Bureau. Age and sex composition in the united states: 2018, Oct 2021.

[BYW+20] Christian Booz, Ibrahim Yel, Julian L. Wichmann, Sabine Boettger, Ahmed Al Kamali, Moritz H. Albrecht, Simon S. Martin, Lukas Lenga, Nicole A. Huizinga, Tommaso D'Angelo, Marco Cavallaro, Thomas J. Vogl, and Boris Bodelle. Artificial intelligence in bone age assessment: accuracy and efficiency of a novel fully automated algorithm compared to the Greulich-Pyle method. *European Radiology Experimental*, 4(1):6, January 2020.

[CA15] MO Columb and MS Atkinson. Statistical analysis: sample size and power estimations. *BJA Education*, 16(5):159–161, 08 2015.

[Car02] Helen M. L. Carty. Assessment of skeletal maturity and prediction of adult height (tw3 method).: 3rd edition. edited by j. m. tanner, m. j. r. healy, h. goldstein and n. cameron. pp 110. london, etc: W. b. saunders, 2001. isbn: 0-7020-2511-9. £69.95. *Journal of Bone and Joint Surgery-british Volume*, pages 310–311, 2002.

[Cat79] J I Cater. Confirmation of gestational age by external physical characteristics (total maturity score). 54(10):794–795, October 1979.

[CBB09] Suprakash Chaudhury, Amitav Banerjee, and J Bhawalkar. Hypothesis testing, type i and type ii errors. *Industrial psychiatry journal*, 18:127, 07 2009.

[Com] Office of the Commissioner. Biosimilar and interchangeable biologics: More treatment choices.

[CSCYS21] Ceyhan Ceran Serdar, Murat Cihan, Doğan Yücel, and Muhittin Serdar. Sample size, power and effect size revisited: simplified and practical approaches in pre-clinical, clinical and laboratory studies. *Biochemia Medica*, 31, 02 2021.

[CSI17] Ana L. Creo and W.Frederick Schnwenk II. Bone age: A handy tool for pediatric providers. *Pediatrics*, 16, 2017.

[Dog18] Nurettin Oezguer Dogan. Bland-altman analysis: A paradigm to understand correlation and agreement. *Turkish Journal of Emergency Medicine*, 18(4):139–141, 2018.

[DTBP+20] Jannick De Tobel, Jeroen Bauwens, Griet Parmentier, Ademir Franco, Nele Pauwels, Koenraad Verstraete, and Patrick Thevissen. Magnetic resonance imaging for forensic age estimation in living children and young adults: a systematic review. *Pediatric Radiology*, 50:1–18, 11 2020.

[Fle32] H. Flecker. Roentgenographic observations of the times of appearance of epiphyses and their fusion with the diaphyses. *Journal of anatomy*, 67 Pt 1:118–164.3, 1932.

[Gia15] Davide Giavarina. Understanding bland altman analysis. *Biochemia medica*, 25(2):141–151, 2015.

[GP50] William Walter Greulich and Sarah Idell Pyle. *Radiographic Atlas of Skeletal Development of the Hand and Wrist*. Stanford University Press, University of Michigan, 1950.

[GZ12a] Asghar Ghasemi and Saleh Zahediasl. Normality tests for statistical analysis: A guide for non-statisticians. *International journal of endocrinology and metabolism*, 10:486–489, 12 2012.

[GZ12b] Asghar Ghasemi and Saleh Zahediasl. Normality tests for statistical analysis: A guide for non-statisticians. *International journal of endocrinology and metabolism*, 10:486–489, 12 2012.

[Har20] Frank Harrell. Categorizing continuous variables. `https://discourse.datamethods.org/t/categorizing-continuous-variables/3402`, 2020.

[HFC78] SN Hunyor, JM Flynn, and C Cochineas. Comparison of performance of various sphygmomanometers with intra-arterial blood-pressure readings. *British medical journal*, 2(6131):159—162, July 1978.

[Hol96] Sally Hollis. Analysis of method comparison studies. *Annals of Clinical Biochemistry*, 33(1):1–4, 1996. PMID: 8929061.

[HPKC+19] Safwan S. Halabi, Luciano M. Prevedello, Jayashree Kalpathy-Cramer, Artem B. Mamonov, Alexander Bilbily, Mark Cicero, Ian Pan, Lucas Araújo Pereira, Rafael Teixeira Sousa, Nitamar Abdala, Felipe Campos Kitamura, Hans H. Thodberg, Leon Chen, George Shih, Katherine Andriole, Marc D. Kohli, Bradley J. Erickson, and Adam E. Flanders. The RSNA pediatric bone age machine learning challenge. *Radiology*, 290:2:498–503, 2019.

[HT17] Pavel Hamet and Johanne Tremblay. Artificial intelligence in medicine. *Metabolism*, 69:S36–S40, 2017. Insights Into the Future of Medicine: Technologies, Concepts, and Integration.

[IB] Ib lab panda ce.

[IH93] B. Iglewicz and D.C. Hoaglin. *How to Detect and Handle Outliers*. ASQC basic references in quality control. ASQC Quality Press, 1993.

[KB10] Prashant Kadam and Supriya Bhalerao. Sample size calculation. *International journal of Ayurveda research*, 1:55–7, 01 2010.

[KKQ+20] Sven Koitka, Moon S. Kim, Ming Qu, Asja Fischer, Christoph M. Friedrich, and Felix Nensa. Mimicking the radiologists' workflow: Estimating pediatric hand bone age with stacked deep neural networks. *Medical Image Analysis*, 64:101743, 2020.

[KL16] Terry Koo and Mae Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15, 03 2016.

[KSO+94] D G King, D M Steventon, M P O'Sullivan, A M Cook, V P L Hornsby, I G Jefferson, and P R King. Reproducibility of bone ages when performed by radiology registrars: an audit of tanner and whitehouse ii versus greulich and pyle methods. *The British Journal of Radiology*, 67(801):848–851, 1994. PMID: 7953224.

[KSY+17] Jeong Rye Kim, Woo Hyun Shim, Hee Mang Yoon, Sang Hyup Hong, Jin Seong Lee, Young Ah Cho, and Sangki Kim. Computerized bone age estimation using deep learning based program: Evaluation of the accuracy and efficiency. *American Journal of Roentgenology*, 209(6):1374–1380, 2017. PMID: 28898126.

[KWT+76] H. J. Keim, J. M. Wallace, H. Thurston, D. B. Case, J. I. Drayer, and J. H. Laragh. Impedance cardiography for determination of stroke index. *Journal of Applied Physiology*, 41(5):797–799, 1976. PMID: 791918.

[LCL+18] David B. Larson, Matthew C. Chen, Matthew P. Lungren, Safwan S. Halabi, Nicholas V. Stence, and Curtis P. Langlotz. Performance of a Deep-Learning Neural Network Model in Assessing Skeletal Maturity on Pediatric Hand Radiographs. *Radiology*, 287(1):313–322, April 2018.

[Len01] Russell V Lenth. Some practical guidelines for effective sample size determination. *The American Statistician*, 55(3):187–193, 2001.

[LSF80] Karen D. Laughlin, Donald J. Sherrard, and L Fisher. Comparison of clinic and home blood pressure levels in essential hypertension and

variables associated with clinic-home differences. *Journal of chronic diseases*, 33 4:197–206, 1980.

[Lud97] John Ludbrook. Comparing methods of measurement. *Clinical and Experimental Pharmacology and Physiology*, 24(2):198–203, 1997.

[Lud09] John Ludbrook. Confidence in altman-bland plots: A critical review of the method of differences. *Clinical and experimental pharmacology & physiology*, 37:143–9, 09 2009.

[LZL+16] MJ Lu, WH Zhong, YX Liu, HZ Miao, YC Li, and MH Ji. Sample size for assessing agreement between two methods of measurement by bland-altman method. *The International Journal of Biostatistics*, 12:2, 2016.

[Mar11] David D et al. Martin. The use of bone age in clinical practice – part 1. *Hormone research in paediatrics*, 76:1–9, 2011.

[MIH+13] Marjan Mansourvar, Maizatul Akmar Ismail, Tutut Herawan, Ram Raj, Sameem Abdul Kareem, and Fariza Nasaruddin. Automated bone age assessment: Motivation, taxonomies, and challenges. *Computational and mathematical methods in medicine*, 2013:391626, 12 2013.

[MPS+19] Prabhakar Mishra, ChandraM Pandey, Uttam Singh, Chinmoy Sahu, and Amit Keshri. Descriptive statistics and normality tests for statistical data. *Annals of Cardiac Anaesthesia*, 22:67–72, 01 2019.

[MWN+21] Mohammad Ali Mansournia, Rachel Waters, Maryam Nazemipour, Martin Bland, and Douglas G. Altman. Bland-altman methods for comparing methods of measurement and response to criticisms. *Global Epidemiology*, 3:100045, 2021.

[ort] Orthogonal distance regression (scipy.odr)¶.

[OSS14] Nancy A. Obuchowski, Naveen Subhas, and Paul Schoenhagen. Testing for interchangeability of imaging tests. *Academic Radiology*, 21:11, 2014.

[PBM18] Andrew Pennock, James Bomar, and John Manning. The creation and validation of a knee bone age atlas utilizing mri. *The Journal of bone and joint surgery. American volume*, 100:e20, 02 2018.

[PBM+20] Ian Pan, Grayson Baird, Simukayi Mutasa, Derek Merck, Carrie Ruzal-Shapiro, David Swenson, and Rama Ayyala. Rethinking greulich and pyle: A deep learning approach to pediatric bone age assessment using pediatric trauma hand radiographs. *Radiology: Artificial Intelligence*, 2:e190198, 07 2020.

[PPMDM+20] Monika Prokop-Piotrkowska, Kamila Marszałek-Dziuba, Elżbieta Moszczyńska, Mieczysław Szalecki, and Elzbieta Jurkiewicz. Traditional and new methods of bone age assessment – an overview. *Journal of Clinical Research in Pediatric Endocrinology*, 13, 10 2020.

[RLY+19] Xuhua Ren, Tingting Li, Xiujun Yang, Shuai Wang, Sahar Ahmad, Lei Xiang, Shaun Richard Stone, Lihong Li, Yiqiang Zhan, Dinggang Shen, and Qian Wang. Regression convolutional neural network for automated pediatric bone age assessment from hand radiograph. *IEEE Journal of Biomedical and Health Informatics*, 23(5):2030–2038, 2019.

[Rub21] David A. Rubin. Assessing bone age: A paradigm for the next generation of artificial intelligence in radiology. *Radiology*, 301(3):700–701, 2021. PMID: 34581631.

[Sch21] Frank Schoonjans. Sample size calculation: Bland-altman plot, Apr 2021.

[SH20] Jacob Shreffler and Martin R Huecker. Type i and type ii errors and statistical power. 2020.

[SJ78] G L Serfontein and A M Jaroszewicz. Estimation of gestational age at birth. comparison of two methods. *Archives of Disease in Childhood*, 53(6):509–511, 1978.

[SLK+20] Nan-Young Shin, Byoung-Dai Lee, Ju-Hee Kang, Hye-Rin Kim, Dong Oh, Byung Lee, Sung Kim, Mu Lee, and Min-Suk Heo. Evaluation of the clinical efficacy of a tw3-based fully automated bone age assessment system using deep neural networks. *Imaging Science in Dentistry*, 50:237, 09 2020.

[SRGO06] A. Schmeling, W. Reisinger, G. Geserick, and A. Olze. Age estimation of unaccompanied minors: Part i. general considerations. *Forensic Science International*, 159:S61–S64, 2006. International IOFOS Symposium on Forensic Odontology 2006 and 3rd International Conference on Reconstruciton of Soft Facial Parts 2006.

[SRM+21] A. M. Schmid, D. L. Raunig, C. G. Miller, R. C. Walovitch, R. W. Ford, M. O'Connor, G. Brueggenwerth, J. Breuer, L. Kuney, and R. R. Ford. Radiologists and Clinical Trials: Part 1 The Truth About Reader Disagreements. *Ther Innov Regul Sci*, 55(6):1111–1121, 11 2021.

[SSA+21] S. Simon, G. M. Schwarz, A. Aichmair, B. J. H. Frank, A. Hummer, M. D. DiFranco, M. Dominkus, and J. G. Hofstaetter. Fully automated deep learning for knee alignment assessment in lower extremity radiographs: a cross-sectional diagnostic study. *Skeletal Radiol*, Nov 2021.

[TFRSPdlC+07] JM Tristán Fernández, F Ruiz Santiago, A Pérez de la Cruz, G Lobo Tanner, MJ Aguilar Cordero, and F Collado Torreblanca. [the influence of nutrition and social environment on the bone maturation of children]. *Nutricion hospitalaria*, 22(4):417—424, 2007.

[TLS+18] Shahein Tajmir, Hyunkwang Lee, Randheer Shailam, Heather Gale, Jie Nguyen, Sjirk Westra, Ruth Lim, Sehyo Yune, Michael Gee, and Synho Do. Artificial intelligence-assisted interpretation of bone age radiographs improves accuracy and decreases variability. *Skeletal Radiology*, 48, 08 2018.

[TW83] J.M. Tanner and R.H. et al. Whitehouse. *Assessment of skeletal maturity and prediction of adult height (TW2 method)*. Academic Press Inc; Subsequent edition, 1983.

[WCG+21] Fengdan Wang, Wangjiu Cidan, Xiao Gu, Shi Chen, Wu Yin, Yongliang Liu, Lei Shi, Hui Pan, and Zhengyu Jin. Performance of an artificial intelligence system for bone age assessment in tibet. *The British Journal of Radiology*, 94(1120):20201119, 2021. PMID: 33560889.

[WED98] S. D. Walter, M. Eliasziw, and A. Donner. Sample size and optimal designs for reliability studies. *Statistics in Medicine*, 17(1):101–110, 1998.

[WGC+20] Fengdan Wang, Xiao Gu, Shi Chen, Yongliang Liu, Qingxin Shen, Hui Pan, Lei Shi, and Zhengyu Jin. Artificial intelligence system can achieve comparable results to experts for bone age assessment of chinese children with abnormal growth and development. *PeerJ*, 8, 2020.

[Wik] Medcalc. `https://en.wikipedia.org/wiki/MedCalc`. Accessed: 2021-12-04.

[YGX21] Shoujian Yu, Jianbang Ge, and Xiaolin Xia. *Bone Age Assessment Based on Deep Convolution Neural Network*, page 70–75. Association for Computing Machinery, New York, NY, USA, 2021.

[ZSV+09] Aifeng Zhang, James W. Sayre, Linda Vachon, Brent J. Liu, and H. K. Huang. Racial differences in growth patterns of children assessed on the basis of bone age. *Radiology*, 250(1):228–235, 2009. PMID: 18955510.