**The assessment of experimental artificial intelligence (AI) algorithms for the diagnosis of skin tumours against human performance**

**Protocol Number:** X19-0066 & 2019/ETH00468

**Principal Investigator:** Prof. Scott W. Menzies

Sydney Melanoma Diagnostic Centre

Signature:  Date:24 April 2020

**Protocol Authors:** Scott W. Menzies

**Protocol Version 4**

**Protocol Date: 24 April 2020**

**Ethics Statement:**

The study will be conducted in accordance with the *National Statement on Ethical Conduct in Human Research* (2007), the *CPMP/ICH Note for Guidance on Good Clinical Practice* and consistent with the principles that have their origin in the Declaration of Helsinki. Compliance with these standards provides assurance that the rights, safety and well-being of trial participants are respected.

Contents

[Summary 3](#_Toc361670588)

[1. BACKGROUND AND INTRODUCTION 4](#_Toc361670589)

[1.1. DISEASE/PROPOSED INTERVENTION BACKGROUND 4](#_Toc361670590)

[1.2. RATIONALE FOR PERFORMING THE STUDY 4](#_Toc361670591)

[2. HYPOTHESIS 5](#_Toc361670592)

[3. STUDY OBJECTIVES 5](#_Toc361670593)

[3.1. PRIMARY OBJECTIVES 5](#_Toc361670594)

[3.2. SECONDARY OBJECTIVES 5](#_Toc361670595)

[4. STUDY DESIGN 5](#_Toc361670596)

[4.1. DESIGN 5](#_Toc361670597)

[4.2. EXPECTED PARTICIPANT NUMBERS 5](#_Toc361670598)

[4.3. DURATION OF THE STUDY 5](#_Toc361670599)

[4.4. ENDPOINTS 6](#_Toc361670600)

[4.5. CENTRES 6](#_Toc361670601)

[5. STUDY PARTICIPANTS 6](#_Toc361670602)

[5.1. INCLUSION CRITERIA 6](#_Toc361670603)

[5.2. EXCLUSION CRITERIA 6](#_Toc361670604)

[6. STUDY PROCEDURES 7](#_Toc361670605)

[6.1. STUDY FLOW CHART 7](#_Toc361670606)

[6.2. INVESTIGATION PLAN 8](#_Toc361670607)

[6.3. STUDY PROCEDURE RISKS 10](#_Toc361670608)

[6.4. PARTICIPANT RECRUITMENT AND SCREENING 10](#_Toc361670609)

[6.5. PARTICIPANT ENROLMENT 10](#_Toc361670610)

[6.6. INFORMATION AND CONSENT 10](#_Toc361670611)

[6.7. END OF STUDY TREATMENT/WITHDRAWAL PROCEDURE 10](#_Toc361670613)

[6.8. PATIENT WITHDRAWAL 11](#_Toc361670614)

[7. OUTCOMES 11](#_Toc361670615)

[7.1. DEFINITION OF OUTCOMES 11](#_Toc361670616)

[8. STATISTICAL CONSIDERATIONS 11](#_Toc361670617)

[8.1. SAMPLE SIZE OR POWER CALCULATION 11](#_Toc361670618)

[8.2. PROVIDE A DETAILED ANALYSIS PLAN 11](#_Toc361670619)

[9. DATA COLLECTION 12](#_Toc361670620)

[9.1. PARTICIPANT REGISTRATION 12](#_Toc361670621)

[10. QUALITY CONTROL AND ASSURANCE 12](#_Toc361670625)

[10.1. PROTOCOL AMENDMENTS **12.**](#_Toc361670628)

[11. ETHICS 12](#_Toc361670629)

[11.1. INVESTIGATOR AUTHORISATION PROCEDURE 12](#_Toc361670630)

[11.2. PATIENT PROTECTION 12](#_Toc361670631)

[12. SAFETY 13](#_Toc361670632)

[12.1. ADVERSE EVENT REPORTING 13](#_Toc361670633)

[12.2. SERIOUS ADVERSE EVENT REPORTING 13](#_Toc361670634)

[14. CONFIDENTIALITY AND STORAGE AND ARCHIVING OF STUDY 13](#_Toc361670638)

[15. TRIAL SPONSORSHIP AND FINANCING 13](#_Toc361670639)

[16. INDEMNITY 13](#_Toc361670640)

[17. REFERENCES 14](#_Toc361670641)

## Summary

**The assessment of experimental artificial intelligence (AI) algorithms for the diagnosis of skin tumours against human performance**

Protocol Version 4

OBJECTIVES

Primary objective:

To compare the diagnosis and management decision of the winning AI ISIC 2018 Challenge diagnostic algorithm/s (and others as available) compared with health professionals in the routine clinical field.

Secondary objectives:

1. To compare the AI vs Human performance in the clinic versus a previous experimental online reader study.

2. To allow the collection of a true prevalent skin lesion set with ground truth that can be used as a data set for future studies assessing AI packages.

Study design: Diagnostic test study comparing AI vs Humans in a clinical setting.

Planned sample size: N=5-7 patients (pilot whole body) N=60 (pivotal whole body examination) N=151 (individual lesion diagnosis)

Selection criteria: Patients with baseline total body photographs taken within 1-4 years of examination. Clinicians with a range of experience in the dermoscopy diagnosis of pigmented lesions.

Study procedure: Clinical and AI examination of individual lesions compared with ground truth diagnosis.

Sample size calculations and statistical power:

*Pilot*: The sample size justification is based on pragmatic considerations of feasibility of recruitment and available resources. A sample of 5-7 patients undergoing whole body examination (> 300 lesions) and n=10 with an individual diagnosis protocol is deemed sufficient to inform the study feasibility, and to provide information about the usefulness of AI in clinical setting.

*Whole body examination (management)*: A total sample size of 3900 lesions (n=60 subjects with 65 lesions per subject collected from two sites) achieves 81% power at a 5% significance level using a two-sided equivalence test of correlated proportions when the standard proportion (proportion of a positive management decision) is 0.01, with the maximum allowable difference between these proportions that still results in equivalence (symmetrical equivalence limit) being 0.005 (0.5%). *Diagnosis*: A total sample size of 151 excised lesions (collected from two sites) achieves 80% power at a 5% significance level using a one-sided equivalence (non-inferiority) test of correlated proportions when the standard proportion is 0.730 (ie. estimating that 22/30 lesions for Human and AI have the same diagnosis), the maximum allowable difference between these proportions that still results in non-inferiority (the range of equivalence) is 0.10 (10%).

Duration of study: 5 years

# 1. BACKGROUND AND INTRODUCTION

### 1.1. DISEASE/PROPOSED INTERVENTION BACKGROUND

Early diagnosis of primary melanoma remains the critical step in reducing mortality from the disease. [[1]](#endnote-1) Furthermore, an improvement in diagnostic accuracy significantly reduces costs to the healthcare system by reducing both unnecessary biopsies (currently, for every melanoma diagnosed there are 3 to 40 melanocytic lesions excise[[2]](#endnote-2)) and costs associated with managing melanoma (advanced melanoma costs $155,109 per patient per annum compared to $1,681 for early thin melanoma).[[3]](#endnote-3)

Artificial intelligence (AI) for early detection of melanoma has the potential to be transformational. The recent advancements in the field of machine learning, especially the introduction of convolutional neural networks (CNNs), has boosted the interest in this area of research.[[4]](#endnote-4) Codella et al.[[5]](#endnote-5) used ensembles of multiple algorithms to demonstrate melanoma recognition accuracies above expert dermatologists. Later on, Esteva et al.[[6]](#endnote-6) and Han et al.[[7]](#endnote-7) fine-tuned CNNs with large datasets of clinical images and demonstrated dermatologist-level accuracy for more general skin disease classification and Haenssle et al.[[8]](#endnote-8) reported expert-level accuracy for dermoscopic images (from a hand held microscope routinely used to diagnose skin tumours in the clinic) of melanocytic lesions.

A collection of dermoscopic images of the most clinically relevant types of pigmented lesions was created into a publicly available training set of 10,015 images for machine learning (HAM10000).[[9]](#endnote-9) Following this, an international competition, the **ISIC 2018 Challenge**, provided this training set as well as a test set of 1,511 dermoscopic images to the participants of the challenge with the aim to attract the best machine learning labs worldwide and to obtain reliable estimates of the accuracy of state-of-the-art machine learning algorithms.[[10]](#endnote-10) Following this competition, an open web-based reader study under the umbrella of the International Dermoscopy Society (IDS) invited their members to compare their diagnostic accuracy with that of machine learning algorithms. In this way a comparison of the most advanced machine learning algorithms with the most experienced human experts using publicly available data occurred. The average machine learning algorithm had significantly more correct diagnoses than the average human reader and the 3 best machine learning algorithms more correct diagnoses than human experts with more than 10 years domain-specific experience.[[11]](#endnote-11)

However, overfitting in machine learning algorithms remains a major impediment to the success of all AI instruments.[[12]](#endnote-12) Deep learning algorithms for image recognition require a large amount of well annotated data that is close to the real-world problems faced in the clinical setting ie. encompassing the entire spectrum of melanoma and its mimickers. When this is not achieved, overfitting of algorithms leads to failure when tested in the real world.

### 1.2. RATIONALE FOR PERFORMING THE STUDY

We wish to test various experimental AI diagnostic algorithms against human performance in a real-world setting. This will set a proof of principle that AI, when compared with human performance in an experimental setting, is predictive of what will happen in the real world.

# 2. NULL HYPOTHESIS

1. AI diagnostic packages are equivalent or non-inferior to clinician performance in a real-world setting for the diagnosis of skin tumours (see detail 4.4.2).

# 3. STUDY OBJECTIVES

### 3.1. PRIMARY OBJECTIVES

1. To compare the diagnosis and management decision of the winning AI ISIC 2018 Challenge diagnostic algorithm/s (and others as available) compared with health professionals in the routine clinical field.

### 3.2. SECONDARY OBJECTIVES

1. To compare the AI vs Human performance in the clinic versus previous experimental online reader study.

2. To allow the collection of a true prevalent skin lesion set with ground truth that can be used as a data set for future studies assessing AI algorithms.

# 4. STUDY DESIGN

### 4.1. DESIGN

Diagnostic test study comparing AI vs Humans.

### 4.2. EXPECTED PARTICIPANT NUMBERS

Pilot study: 5-7 individuals (whole body examination) and 10 individual diagnosis (biopsied lesions)

Pivotal study: 60 individuals (whole body examination) 151 individual lesion diagnosis (biopsied lesions)

### 4.3. DURATION OF THE STUDY

Duration: June 2019 – June 2024.

Participant recruitment: June 2019 – June 2024

### 4.4.1 ENDPOINTS

PRIMARY ENDPOINTS

Paired comparison of the correct management and diagnosis decision of AI vs Human.

SECONDARY ENDPOINTS

Paired comparison of a correct balanced diagnosis decision of AI vs Human.

4.4.2 HYPOTHESES

1. There is no difference (ie. there is equivalence) between the tested specialists and AI for the management of pigmented skin lesions.

2. There is no difference (ie. there is equivalence) between the tested novice clinicians and AI for the management of pigmented skin lesions.

3. There is no inferiority between the tested specialists and AI for the diagnosis of pigmented skin lesions.

4. There is no inferiority between the tested novice clinicians and AI for the diagnosis of pigmented skin lesions.

### 4.5. CENTRES

Pilot study: This will occur at the Sydney Melanoma Diagnostic Centre (SMDC) only.

Pivotal study: This will occur at SMDC (n=30 whole body, n=76 individual biopsied lesions) under Prof. Scott Menzies, A Prof. Pascale Guitera and Dr. Serigne Lo, and the Dept. Dermatology, Medical University of Vienna (n=30 whole body, n=75 individual biopsied lesions) under Prof. Harald Kittler and Dr. Philipp Tschandl.

AI algorithm development for the above studies: MetaOptima Technology Inc., Vancouver, Canada.

Data will be shared with all of the above institutions (SMDC, Dept. Dermatology, Medical University of Vienna and MetaOptima Technology Inc.).

All procedures and participants will be identical in all sites.

A Prof. Guitera is a paid consultant for MetaOptima Technology Inc. For this reason she will not be involved with the statistical analysis or writing of the first draft of the manuscript. Her conflict of interest will be formally declared on any related publication.

# 5. STUDY PARTICIPANTS

### 5.1. INCLUSION CRITERIA

A. Whole body examination

1. Patients with baseline total body photographs (TBP) taken within 1-4 years of examination

2. Gender: Male or Female

3. Age range: 18-99yrs

4. Modified Fitzpatrick I-III Skin Type[[13]](#footnote-1)

5. Willingness and ability to provide informed consent and to participate and comply with the study requirements.

B. Individual lesion examination

1. Patients undergoing an excision/biopsy of a pigmented skin lesion.

2. Gender: Male or Female

3. Age range: 18-99yrs

4. Modified Fitzpatrick I-III Skin Type

5. Willingness and ability to provide informed consent and to participate and comply with the study requirements.

### 5.2. EXCLUSION CRITERIA

1. Not meeting all inclusion criteria for either group A or B.

# 6. STUDY PROCEDURES

### 6.1.0 STUDY FLOW CHART

PATIENT RECRUITMENT & CONSENT adults with available baseline TBP taken 1- 4 years previously

Whole Body Examination

TBP performed & avatar labelled with dermoscopy images of >4mm\*\*\* lesions (ex exclusion sites)

GROUND TRUTH

of all avatar labelled lesions

**\*** Changing clinical dermatofibroma will be biopsied because of differential diagnosis of desmoplastic melanoma

AI EXAMINATION

all lesions > 4mm\*\*\* (ex. exclusion)

Diagnostic category

CLINICIAN EXAMINATION

all ground-truth **biopsied** lesions

AI EXAMINATION

all ground-truth **biopsied** lesions

Dismiss (incl. LT monitor)

Biopsy

Monitor 3 month

CLINICIAN EXAMINATION

all lesions > 4mm\*\*\* (ex exclusions)

Benign

Disagree

Agree

Diagnostic category

Change

Unchanged (3 mo) Insignificant (LT)

Unknown

Dermoscopy readers of seb K, hemangioma\*

Monitor melanocytic 3 month or long-term\*\*

Excise/biopsy suspicious

Changed on TBP

Benign

Unchanged on TBP

Diagnostic category

**\*\*** Long-term monitor is defined as ≥ 6 months

\*\*\* And >3mm if chosen for monitoring or biopsy by ground-truth assessment.

6.1.1 SIMPLIFIED FLOW CHART

Diagnostic category of biopsied lesions

Monitor 3 month

Biopsy

Dismiss (incl. LT monitor)

AI EXAMINATION

of all lesions > 4mm\*\*\*

CLINICIAN EXAMINATION

of all lesions > 4mm\*\*\*

GROUND TRUTH assessment

of all imaged lesions

AI image acquisition of >4mm\*\*\* lesions (ex exclusion sites)

PATIENT RECRUITMENT & CONSENT adults with available baseline TBP taken 1- 4 years previously

PATIENT RECRUITMENT & CONSENT adults undergoing excision/biopsy of pigmented lesions

AI EXAMINATION

CLINICIAN EXAMINATION

Diagnostic category

6.2. INVESTIGATION PLAN

**6.2.1 Methodology (Whole Body Examination)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Interventions | Enrolment Visit | Visit 1a | Visit 2 | Visit 3 month | Final Study Visit |
| Participant Consent | ✓ |  |  |  |  |
| Inclusion / Exclusion criteria | ✓ |  |  |  |  |
| Ground truth assessment |  | ✓ |  | ✓ b | ✓ b |
| AI assessment |  | ✓ |  |  |  |
| Clinicians assessment |  |  | ✓ |  |  |

a: Visit 1 can also occur at enrolment

b: Follow-up visits only occur depending upon ground-truth assessment (see study flow chart). Furthermore, ground truth may change following observations of the same patient over a longer time period according to standard practice.

**Assessible lesions:** All discrete skin lesions with a > 4 mm longest diameter *except* for clinically non-suspicious amelanotic actinic keratoses and multiple lesions consistent with an inflammatory skin eruption or ephelides. In addition, all discrete lesions > 3 mm that are chosen for excision or monitoring by ground-truth assessment. Lesions are also excluded according to the ISIC challenge data set:-- Special sites: Mucosal, nails, umbilicus, lesions with previous treatment or biopsy, insufficient image quality: bubbles, no polarization & no immersion fluid, underexposure, overexposure, lesion incompletely imaged, lesion non-centred (ie: imaging not conformant to published ISDIS imaging standards). For a comparison of the ISIC 2018 AI algorithms only pigmented lesions (defined as having any of red-blue, brown/tan, black, blue, grey colour under dermoscopy) will be chosen since this was the criteria for that challenge.

**Ground truth assessment**: The patients will be compared with baseline total body photographs (TBP) by an independent expert observer (the treating specialist). Unchanged lesions will be labelled as benign. Changed lesions will be managed either by excision (clinically suspicious), 3-month or long-term digital dermoscopy monitoring (mildly atypical to banal melanocytic lesions), or observe (seborrheic keratoses, haemangioma) according to standard clinical practice.[[14]](#endnote-13) In the observation category for *changed lesions* a high resolution dermoscopy image will be taken for an independent expert reader assessment. The protocol to determine ground truth (benign, diagnostic category) is as according to the study flow chart. Final Ground Truth is determined by a descending hierarchy of

* Histopathology (ie. biopsy changed lesion). Each histological case (electronic images of representative sections with the clinical history as found in the report) will also be reviewed by a dedicated pathologist. Sydney cases will be reviewed by Vienna and visa versa. If disagreement occurs a third pathologist will be recruited. The majority decision will be recorded as the gold standard. Melanocytic lesions will be reported using MPATH-Dx system[[15]](#endnote-14).
* Unchanged lesions on TBP = benign
* Changed melanocytic lesions undergo 3-month SDDI (excise if change)
* Subsequent unchanged 3-month digital dermoscopy OR insignificant changing long-term monitored lesions = benign[[16]](#footnote-2)
* In vivo confocal microscopy (if available).
* 2 of 2 independent readers viewing the dermoscopy images of *changed lesions* that are clinically benign (eg. Seborrheic keratoses, hemangiomas) = benign

**Diagnostic category:** For those lesions assessible for a diagnostic category (see flow chart) they will be categorized as with the 2018 ISIC challenge (1-7) as follows:

1. Melanoma

2. Melanocytic nevus

3. Pigmented Basal cell carcinoma

4. Pigmented Actinic keratosis / Bowen’s disease (intraepithelial carcinoma)

5. Benign (pigmented) keratotic lesion

6. Benign Vascular lesion

7. Dermatofibroma

8. Other (for pink lesions) … with a written response of the diagnosis

**AI assessment:** Images will be taken of all recruitable lesions using a dermoscopic attachment (MoleScope™ II, MetaOptima Technology Inc., Vancouver Canada) to a mobile phone (iPhone 8plus) and embedded in a dermatology software (DermEngine™, MetaOptima). With this dermoscopic attachment polarized images that are clear are adequate. However, if not, an antiseptic gel liquid is applied at the lesion/attachment interface. Winning AI algorithms from the ISIC 2018 Challenge will be run on each *pigmented* lesion with the resultant diagnostic category and management output (biopsy, monitor or dismiss) recorded. As well, AI assessment may be run on pink (amelanotic) lesions, although this will be a subset analysis as the ISIC 2018 Challenge was restricted to pigmented lesions.

**Clinician assessment:** Clinicians with a variety of clinical expertise (dermatology residents/registrars, specialists) will be recruited with their level of expertise (years using dermoscopy, clinician classification) recorded as occurred with the ISIC reader study. One novice clinician (dermatology resident/1st yr registrar) and one expert clinician (specialist in pigmented lesion clinics) will examine the patient.. The number of recruited on-site clinicians will vary over time according to their availability. However, those chosen to examine an individual patient will be based on availability at the time of patient recruitment. The demographics of each clinician (including experience) will be collected and reported. Clinicians will examine the whole skin of the patient according to their routine practice, but *without* taking a history or questioning the patient, and their management output for each lesion (biopsy, monitor or dismiss) recorded. After this, a subset of lesions that during ground truth assessment allowed precise diagnostic category assessment (those chosen for excision/biopsy), would be identified for the clinician to record their diagnostic category (see study flow chart).

The diagnosis and management of patients when assessing ground truth is as according to standard practice12 and independent of the AI and comparative clinician assessment. As well, the AI and comparative clinician assessment are independent of each other, independent of the ground truth *and do not influence patient management*. However, it is not routine to have either the AI measurements or multiple clinicians assessing the patient.

**Pilot study**

A pilot of n=5-7 whole body examination patients and n=10 with an individual diagnosis protocol at SMDC will be performed to assess the logistics of the study protocol. Furthermore, results of the pilot can be used to alter the AI algorithms as follows. The study is powered to achieve an equivalence of the AI algorithm and human performance using the management outputs (biopsy/monitor or dismiss). Since the ISIC challenge algorithms did not have these outputs (they had diagnostic categories) it is appropriate to allow the thresholds for the diagnostic categories that are used to create the management outputs to be altered by the pilot data. ONLY for the ISIC challenge tested algorithm will these new management output thresholds be allowed to change. The algorithm defining the diagnostic category outputs must remain as previously published. However, if any other novel algorithms are developed, the pilot data can be used to alter these prior to the pivotal study. No data from the pilot study will be incorporated into the analysis of the pivotal study.

The following algorithms from MetaOptima will be used in the pilot:

1. 7-class classification

\* ISIC winning model (see background)

\* An ensemble of Taxnet models (standard backbone network with an RNN)

\* An ensemble of Lesion Localization with Bias Unlearning models (LLBU)

2. Management decision

\* An ensemble of Taxnet models (standard backbone network with an RNN) to get ben/mal predictions and map to Dismiss | Monitor | Biopsy classes

\* An ensemble of Lesion Localization with Bias Unlearning models (LLBU) to get ben/mal predictions and map to Dismiss | Monitor | Biopsy classes

All models will be run online overnight at various intervals.

**Pivotal study**

Power calculations were adjusted following data from the whole body examination pilot (see 8.1). The pivotal study will consist of n=60 patients (n=30 SMDC, n=30 Dept. Dermatology, Medical University of Vienna). The study protocol is anticipated to be identical to the pilot unless unforeseen events occur in the latter (in which case an altered protocol will be submitted prior to commencement). The AI algorithms may change from the pilot study (see above Pilot study).

**6.2.2 Methodology (Individual Lesion Examination)**

Patients undergoing routine excision or biopsy of eligible pigmented skin lesions will undergo clinical and AI examination of those lesions only and the diagnosis (category 1-7) recorded. Management will not be recorded as leakage in decision making from the specialist decision to biopsy is an uncontrollable confounder.

### 6.3. STUDY PROCEDURE RISKS

Since patient diagnosis and management is performed according to standard practice, risks of misdiagnosis and any treatment risks are also consistent with those seen in standard patient management. The AI assessment and comparative clinician examination is non-invasive (image based) and has no associated risk since the results have no management consequence to the patient.

### 6.4. PARTICIPANT RECRUITMENT AND SCREENING

Patients: These will be identified from existing clinics (Sydney Melanoma Diagnostic Centre, Royal Prince Alfred Hospital and Dept. Dermatology, Medical University of Vienna) and according to inclusion criteria.

Clinicians: Specialists (within pigmented lesion clinics) and Dermatology residents/registrars will be recruited from Royal Prince Alfred Hospital and the Dept. Dermatology, Medical University of Vienna.

### 6.5. PARTICIPANT ENROLMENT

Potential participants (patients and clinicians) will be enrolled into the study after the informed consent process has been completed and the participant has been assessed to meet all the inclusion criteria.

### 6.6. INFORMATION AND CONSENT

An information sheet describing the study will be supplied to the patients and clinicians prior to their formal written consent.

### 6.7. END OF STUDY TREATMENT/WITHDRAWAL PROCEDURE

The final patient visit will occur after the ground truth of all assessible lesions are obtained (according to standard practice).

### 6.9. PATIENT WITHDRAWAL

Patients can withdraw at any stage of the process.

# 7. OUTCOMES

### 7.1. DEFINITION OF OUTCOMES

1***.* Correct Management Decision (MD):** Two definitions will be analysed.

 **I.** Biopsy = Correct for ground truth malignant lesions (melanoma, pigmented BCC, pigmented AK/IEC, other pigmented malignant)

Monitor (3 months) or Dismiss (includes long-term monitoring) = correct MD for ground truth benign lesions (benign as per study flow chart, or naevus, benign keratosis, benign vascular, dermatofibroma, other pigmented benign)

**II**. Biopsy or Monitor (3 months) = Correct MD for ground truth malignant lesions (melanoma, pigmented BCC, pigmented AK/IEC, other pigmented malignant)

Dismiss (includes long-term monitoring) = Correct MD for ground truth benign lesions (benign as per study flow chart, or naevus, benign keratosis, haemangioma, dermatofibroma, other pigmented benign)

2. **Correct Diagnostic Category**: Categorical classification of 7 diagnoses (see methods). Category 8 (other pink lesions) will be excluded from primary analysis but results reported in a subset analysis.

3. **Balanced multiclass accuracy**: This will be according to the ISIC 2018 challenge analysis <https://challenge2018.isic-archive.com/task3/> where the balanced diagnostic accuracy is defined as the average sensitivities obtained for each diagnostic category separately.

# 8. STATISTICAL CONSIDERATIONS

### 8.1. SAMPLE SIZE OR POWER CALCULATION FOR THE PIVOTAL STUDY

This first in human pivotal study is powered for the primary aim of comparing management and diagnosis of the clinician vs AI, according to published methods of equivalence and non-inferiority. [[17]](#endnote-15) [[18]](#endnote-16)

**1) Summary Statements****for comparing each test (AI and clinicians management) to Ground Truth – Equivalence test**

A sample size of 3900 lesions (n=60 subjects with 65 lesions per subject) achieves 81% power at a 5% significance level using a two-sided equivalence test of correlated proportions when the standard proportion (proportion of a positive management decision) is 0.01, with the maximum allowable difference between these proportions that still results in equivalence (symmetrical equivalence limit) being 0.005 (0.5%).

**2) Summary Statements****for comparing the two test clinicians vs. AI management– Non-inferiority test**

A sample of 3900 has more than 85% power to demonstrate the non-inferiority of AI algorithm for the diagnosis of pigmented skin lesions compared with the clinicians using a *one-sided* equivalence test of correlated proportions, an estimated standard proportion of 0.01 and a non in-inferiority margin of 0.005.

3**) Summary Statements for comparing the two test clinicians vs AI diagnosis to ground truth**

We have taken data from the reader study (Tschandl et al. Lancet Oncol) where there was an average of 7 lesions for every 30 that the best AI outperformed the experts (for diagnostic category 1-7). From this assumption, a sample size of 151 excised lesions achieves 80% power at a 5% significance level using a one-sided equivalence (non-inferiority) test of correlated proportions when the standard proportion is 0.730 (ie. estimating that 22/30 lesions for Human and AI have the same diagnosis), the maximum allowable difference between these proportions that still results in non-inferiority (the range of equivalence) is 0.10 (10%) ie. if the human diagnosis is less than 10% worse than the AI we classify the humans non-inferior to AI.

### For the analysis we are planning to run Generalized Linear Mixed Models with two random effects: one random effect on the participants’ level (to account for variability within participant) and another random effect at assessment level (physicians including AI algorithm).

### 8.2. PROVIDE A DETAILED ANALYSIS PLAN

A pilot study of n=5-7 patients (whole body examination) and 10 individual diagnosis cases (all biopsied) will allow the protocol to be tested in addition to a preliminary analysis of results. The pivotal study (n=60 patients whole body examination and n=151 excised/biopsied lesions) is powered to achieve an equivalence of the AI algorithm and human performance using the management outputs (biopsy/monitor or dismiss) and non-inferiority of the AI compared with the clinician for both management and diagnosis (see above). Since the ISIC challenge algorithms did not have these management outputs (they had diagnostic categories) it is appropriate to allow the thresholds for the diagnostic categories that are used to create the management outputs to be altered by the pilot data. ONLY for the ISIC challenge tested algorithm will these new thresholds be allowed to change. The algorithm defining the diagnostic category outputs must remain as previously published. However, if any other novel algorithms are developed, the pilot data can be used to alter these prior to the pivotal study.

**9. DATA COLLECTION**

### 9.1. PARTICIPANT REGISTRATION, FORMS AND PROCEDURE FOR COLLECTING DATA

All data will be recorded on the purposeful built dermatology software DermEngine by an onsite coordinating project officer.

Patients, individual lesions, clinicians and the AI packages will be assigned a unique ID code. Hence input images of individual lesions will be in the form of: PatientID\_lesion number and outputs linked to PatientID\_lesion number\_ClinicianID and PatientID\_lesion number\_AIpackageID

### 9.2. SCHEDULE FOR COMPLETION

The pilot should be completed by the end of 2019. The pivotal trial is anticipated to be finishing recruitment by the end of 2020. Ongoing assessment of algorithms could continue to 2024.

# 10. QUALITY CONTROL AND ASSURANCE

### 10.1. PROTOCOL AMENDMENTS

All protocol amendments will be submitted prior to commencement.

# 11. ETHICS

### 11.1. INVESTIGATOR AUTHORISATION PROCEDURE

Ethics approval and approved versions of the participant information and consent form will be required prior to commencement.

### 11.2. PATIENT PROTECTION

We will ensure that the study is completed in accordance with the guidelines set out in the [*National Statement on Ethical Conduct in Human Research*](http://www.nhmrc.gov.au/guidelines/publications/e72)(2007) (the *National Statement*) and the [*CPMP/ICH Note for Guidance on Good Clinical Practice*](http://www.tga.gov.au/industry/clinical-trials-note-ich13595.htm)and any other relevant legislation/guidelines.

# 12. SAFETY

### 12.1. ADVERSE EVENT REPORTING

For non-serious events these will be reported in the Annual Report.

 12.2. SERIOUS ADVERSE EVENT REPORTING

All serious adverse events will be reported immediately to the HREC. The reports will be followed by a detailed written report. Follow-up reports will identify the participant/s by unique code assigned to participants (rather than by name).

# 13. CONFIDENTIALITY AND STORAGE AND ARCHIVING OF STUDY

Research data and records will be stored electronically with secured access by the research staff. Signed consent forms and all identifiable data will be stored securely at each clinic site and separately from non-identifiable data. Trial data on the established dermatology record keeping DermEngine platform will be stored according to the standard procedure for this cloud-based platform (Cloud provider in Australia is Amazon AWS). DermEngine is compliant with the Royal Australian College of General Practitioners computer and information security standards (CISS), Australian Privacy Act, and the clinical images and mobile devices in compliance with the Australian Medical Association and the Medical Indemnity Industry Association of Australia.

Archiving of study data (excluding the input images of PatientID\_lesion number) will be for 7 years. The aforementioned deidentified input images linked to ground truth will be stored indefinitely for potential use in future retrospective AI and clinical studies (see secondary objectives and patient consent).

# 15. TRIAL SPONSORSHIP AND FINANCING

A trial project officer, and the data recording platform (MoleScope, DermEngine) will be financed by MetaOptima Technology Inc. (Vancouver, Canada).

# 16. INDEMNITY

Since patients will be managed according to standard clinic practice (ground truth assessment by attending specialist) and the interventions of the comparative clinician and AI diagnosis are non-invasive and do not affect patient management, there are no indemnity issues that do not pertain to standard clinical practice.

**17. REFERENCES**

1. Jeffrey E. Gershenwald, Richard A. Scolyer, et al. [Melanoma Staging: Evidence-Based Changes in the American Joint Committee on Cancer Eighth Edition Cancer Staging Manual](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5978683/) CA Cancer J Clin. 2017 Nov; 67(6): 472–492.  [↑](#endnote-ref-1)
2. Esdaile B, Mahmud I, Palmer A, Bowling J. Diagnosing melanoma: how do we assess how good we are? Clinical and Experimental Dermatology. 2014;39(2):129-34. [↑](#endnote-ref-2)
3. Elliott TM, Whiteman DC, Olsen CM, Gordon LG. Estimated Healthcare Costs of Melanoma in Australia Over 3 Years Post-Diagnosis. Applied Health Economics and Health Policy. 2017;15(6):805-16. [↑](#endnote-ref-3)
4. [Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proc IEEE*. 1998;86(11):2278-2324.](http://paperpile.com/b/svZpKq/7BeT) [↑](#endnote-ref-4)
5. [Codella N, Nguyen Q-B, Pankanti S, et al. Deep Learning Ensembles for Melanoma Recognition in Dermoscopy Images. *IBM J Res Dev*. 2017;61(4).](http://paperpile.com/b/svZpKq/G0Qr) <http://dx.doi.org/.> [Accessed May 28, 2017.](http://paperpile.com/b/svZpKq/G0Qr) [↑](#endnote-ref-5)
6. [Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. 2017;542(7639):115-118.](http://paperpile.com/b/svZpKq/imW5) [↑](#endnote-ref-6)
7. [Han SS, Kim MS, Lim W, Park GH, Park I, Chang SE. Classification of the Clinical Images for Benign and Malignant Cutaneous Tumors Using a Deep Learning Algorithm. *J Invest Dermatol*. 2018;138(7):1529-1538.](http://paperpile.com/b/svZpKq/0cHG) [↑](#endnote-ref-7)
8. [Haenssle HA, Fink C, Schneiderbauer R, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann Oncol*. 2018;29(8):1836-1842.](http://paperpile.com/b/svZpKq/9EAh) [↑](#endnote-ref-8)
9. [Tschandl P, Rosendahl C, Kittler H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci Data*. 2018;5:180161.](http://paperpile.com/b/svZpKq/4Ix2) [↑](#endnote-ref-9)
10. [International Skin Imaging Collaboration. ISIC 2018 Challenge: Skin Lesion Analysis Towards Melanoma Detection. ISIC 2018: Skin Lesion Analysis Towards Melanoma Detection.](http://paperpile.com/b/svZpKq/hM6W) <https://challenge2018.isic-archive.com/.> [Accessed November 11, 2018.](http://paperpile.com/b/svZpKq/hM6W) [↑](#endnote-ref-10)
11. Tschandl P et al. An open, international diagnostic study comparing the accuracy of humans and machines for skin lesion classification. Lancet Oncol. In press [↑](#endnote-ref-11)
12. Yamashita R, Nishio M, Do RKG, Togashi K. [Convolutional neural networks: an overview and application in radiology.](https://www.ncbi.nlm.nih.gov/pubmed/29934920) Insights Imaging. 2018 Aug;9(4):611-629. [↑](#endnote-ref-12)
13. As found in the ISIC challenge [↑](#footnote-ref-1)
14. Cancer Council Australia Wiki Guidelines on *Diagnostic Aids for Melanoma* https://wiki.cancer.org.au/australia/Guidelines:Melanoma/Diagnostic\_aids\_for\_melanoma [↑](#endnote-ref-13)
15. Carney P et al. Achieving consensus for the histopathologic diagnosis of melanocytic lesions: use of the modified Delphi method. J Cutan Pathol. 2016;10:830-7. [↑](#endnote-ref-14)
16. Short-term monitoring has been used by FDA as a gold standard in melanoma diagnostic devices. [↑](#footnote-ref-2)
17. Liu J et al. Tests for equivalence or non-inferiority for paired binary data. Stat Med. 2002; 21:231-245 [↑](#endnote-ref-15)
18. Liu K et al. Testing non-inferiority (and equivalence) between two diagnostic procedures in paired-sample ordinal data. [Stat Med.](https://www.ncbi.nlm.nih.gov/pubmed/14755388) 2004;23:545-59. [↑](#endnote-ref-16)