Feasibility and validity of use a telephone-administered Wound Healing Questionnaire for detection of surgical site infection following abdominal surgery in low and middle-income countries: an international, multicentre cohort Study Within a Trial (TALON)

Introduction

Purpose of this Statistical Analysis Plan

This document is the Statistical Analysis Plan (SAP) for the TALON study within a trial, and should be read in conjunction with the TALON protocol. This SAP details the proposed analyses and presentation of the data for the main paper(s) reporting the results for the TALON trial. The results reported in these papers will follow the strategy set out here. Subsequent analyses of a more exploratory nature will not be bound by this strategy, though they are expected to follow the broad principles laid down here. The principles are not intended to curtail exploratory analysis (e.g., to decide cut-offs for categorisation of continuous variables), nor to prohibit accepted practices (e.g., transformation of data prior to analysis), but they are intended to establish rules that will be followed, as closely as possible, when analysing and reporting data. Any deviations from this SAP will be justified in the final publication of the TALON Study. The analysis will be carried out by an appropriately qualified researcher, who will ensure integrity of the data during their data cleaning processes.

Background

The current 'gold standard' for assessment of surgical site infection (SSI) during the 30-days after surgery is in-person review according to Centre for Disease Control Criteria (1). However, in-person assessment is labour and time intensive, and requires patients to take additional time-off work and incur costs of travel. More efficient follow-up pathways are required, for example over the telephone, that are of equal quality to in-person wound assessment. The overall aim is to evaluate the feasibility and validity of telephone administration of patient-reported questionnaire for wound follow-up in low resource settings.

Study design

TALON is a multi-centre, international, non-randomised Study Within a Trial exploring the feasibility and validity of remote, digital follow-up pathways for surgical site infection assessment. TALON is embedded within three international randomised controlled trials: (1) FALCON, a pragmatic multicentre factorial randomised controlled trial testing measures to reduce surgical site infection in low- and middle-income countries; (2) Sterile Glove and Clean Instrument Change at the Time of Wound Closure to Reduce Surgical Site Infection (ChEETAh) trial and (3) Perioperative respiratory care and outcomes for patients undergoing high risk abdominal surgery (PENGUIN).

Aims

The aims of TALON are:

(1) To evaluate the diagnostic accuracy of telephone administration of the wound healing questionnaire in evaluation of abdominal surgical site infection across LMICs.

(2) To assess the feasibility of delivery of the telephone WHQ by a non-surgeon researcher within global surgery trials.

(3) To assess the feasibility of a video evaluation of the wound as a diagnostic adjunct for telephone-based wound follow-up.

Reference diagnostic test

The 'gold standard' reference diagnostic test for surgical site infection (SSI) during the 30-days after surgery for evaluation in this study is in-person review according to US Centre for Disease Control Criteria (1). The following definition will be used in the host trials to identify deep incisional or superficial incisional SSIs:

- The infection must occur within 30-days of the index operation

AND

- The infection must involve the skin, subcutaneous, muscular or fascial layers of the incision

AND

- The patient must have at least one of the following:

    o Purulent drainage from the wound

    o Organisms are detected from a wound swab

    o Wound opened spontaneously or by a clinician AND, at the surgical wound, the patient has at least one of: pain or tenderness; localised swelling; redness; heat; systemic fever (>38°C).

    o Diagnosis of SSI by a clinician or on imaging

The assessment will be made by an appropriately trained clinician (surgeon, trainee or nurse).

Index diagnostic test

The index diagnostic test under evaluation is the adapted Wound Healing Questionnaire (WHQ) for global surgery trials. The Bluebelle WHQ was developed and validated in the UK (English language) to assess post-discharge infections following abdominal surgery (HTA: 12/200/04) and is attractive for use in randomised trials. The WHQ was designed to be completed either by healthcare professional, or self-reported by patients (2), and as such has been described as a 'universal-reporter' outcome measure (UROM) (3). In a UK validation study, the WHQ demonstrated good reliability and excellent discrimination (4-6). The WHQ has now been adapted using mixed qualitative and quantitative research methods for use in global surgery trials. Items 1 to 10 are scored between 0 and 2 (Not at all, A little, A lot), and Items 11 to 19 are scored between 0 and 1 (No, Yes). These are added together create an overall score between 0 and 29. It will be administered over the telephone by a non-surgeon researcher (junior doctor or research nurse) between 27 and 30-days after surgery (i.e., *before* the reference diagnostic test) as the index diagnostic test in this study.

Data categorisation

The TALON study protocol was designed so that patients return to hospital at 30-days after surgery for the 'gold standard' reference test (in-person assessment). However, this is unlikely to always be possible due to patient (mobility, deterioration, psychological reasons) and environmental (cost, transport links, COVID-19 risk) factors. Patient data within the TALON study will therefore be categorised in two ways based upon whether or not they returned for 30-day in-person follow-up:

- *Group 1*. A TALON wound healing questionnaire has been attempted **AND** the patient returned for 30-day in-person follow-up (**paired data**)

- *Group 2*. A TALON wound healing questionnaire has been attempted but the patient **DID NOT** return for 30-day in-person follow-up (**unpaired data**)

This statistical analysis plan will be split into two subheadings under each heading, to describe the (1) **feasibility** and (2) **validity** analyses.

Patient populations

Feasibility

Data from *Group 1* and *Group 2* will be used to explore the feasibility of a telephone pathway for remote follow-up (i.e. all patients for whom a TALON wound healing questionnaire was attempted).

Validity

Only data from *Group 1* will be used in the validation analysis (i.e. data available for both the reference standard test and the index test).

Outcomes

Feasibility

- Telephone contact rate: patients successfully contacted by the telephone as of proportion of all patients where contact was attempted

- Return rate: the proportion of telephone WHQ returned, with and without in-person follow-up (retention benefit), and reasons not returned.

- Data completion rate: proportion of missing data within each form and for each questionnaire item

- Patient satisfaction: patient's self-reported satisfaction with undergoing telephone follow-up.

Validity/accuracy

- Proportion of surgical site infections that are correctly identified by the telephone WHQ

Analysis

Descriptive data (patient, disease and operation characteristics and aggregate outcome data) will be presented using simple summary statistics (i.e. numbers and percentages for binary data and means (or medians) and standard deviations (or inter-quartile ranges) for continuous normal (or non-normal) data.

Overall flow of patients through the study will be presented in *Figure 1*. This will include (as a minimum) the number of total of patients recruited into the study, the proportion categorised into Group 1 and Group 2, the proportion available for analyses of feasibility and validity, and finally the proportion of patients with missing data either from the TALON study or the host trial.

Feasibility

Baseline demographics (including patient, disease, and operation-specific factors) of all patients included in the feasibility analysis will be presented overall and by included country (*Table 1*) and patient home location (urban versus rural, *Table 2*). Details of the telephone follow-up pathway including the number of attempts for telephone contact, type of phone (including access to a video phone), language used, and use of translation will be presented overall and by country (*Table 3*) and patient home location (*Table 4*). Finally, feasibility outcomes will be presented using simple summary statistics by both country (*Table 5*) and hospital setting (urban versus rural, *Table 6*). Patients that died before the time of telephone follow-up will be described but excluded from the analysis of feasibility outcomes. Differences across countries and patient home location groups on baseline demographics, pathway features and feasibility outcomes will be tested using Student's T-test for parametric data and the Mann-Whitney U test for non-parametric data. The χ2 test will be used for categorical data.

Baseline demographics of patients that did (*Group 1*) and did not (*Group 2*) undergo in-person follow-up in addition to attempted telephone follow-up will be compared. This will allow exploration of whether certain patient groups were more challenging to reach for telephone follow-up (*Table 7*). Differences between *Group 1* and *Group 2* on baseline demographics will be tested using Student's T-test for parametric data and the Mann-Whitney U test for non-parametric data. The χ2 test will be used for categorical data.

The proportion of patients with and without in-person follow-up by month throughout the study period will be presented in *Figure 2*, both overall (*Table 8*) and by country (*Table 9*). This will be used to explore the impact of the SARS-CoV-2 pandemic on the ability of sites to perform 30-day in-person follow-up.

For patients followed-up before local emergence of SARS-CoV-2 (determined using World Health Organisation data), a multivariable binary regression model will be used to explore

factors associated with unsuccessful telephone contact (**Group 1** and **Group 2** data). Clinically plausible patient, disease, operation, and country/location specific factors will be selected *a priori* for inclusion in adjusted model. The multivariable model will be presented in **Table 9**.

Validity

Cross-tabulations of the reference test diagnosis ('no SSI' or 'SSI') against a binary outcome variable derived from the total score of the index test (created by a cut-off score; a WHQ total score of less than or equal to a specified value) will be presented using descriptive statistics (**Table 10**).

Criterion validity will be examined against the reference (face-to-face SSI assessment) to evaluate the performance WHQ in discriminating between individuals with and those without SSI. Sensitivity and 1-specificity values of the WHQ for different cut-off scores will be used to plot a receiver operating characteristic (ROC) curve. From derivation data in the UK, cut points of 6 to 8 were found to have the optimal sensitivity and specificity. We will calibrate the tool with a range of cut points from 5 to 9 in order to explore differences in the AUROC with differing WHQ calibration across included countries.

The overall ability of the WHQ to discriminate between individuals with and those without SSI will be measured by the area under the ROC curve (AUC), with uncertainty presented using 95% confidence intervals. ROC curves will be presented overall (with optimal score calibration calculated by country) (**Figure 3**) and at different cut-offs (pre-defined as between a WHQ score between 5 and 9) overall and by country (**Figure 4**).

Diagnostic test accuracy will be calculated as AUROC, Sensitivity, Specificity, Positive Predictive Accuracy, Negative Predictive Accuracy overall (again with optimal score calibration by country) and at different cut-offs (pre-defined as between a WHQ score

between 5 and 9) (**Table 11**). Uncertainty in these estimates will be presented using 95% confidence intervals.

Differences in the AUROC across key subgroups will be explored to see if the WHQ performs differently across different patient and risk groups. Exploratory subgroups include: (1) country; (2) upper-middle income versus low-middle and low-income countries, defined by the World Bank Human Development index; (3) patient home location (urban versus rural), defined by the patient in the TALON telephone questionnaire; (4) patient age (<70 versus $\geq$70 years); (5) emergency versus elective surgery. Subgroup analyses will be summarised by using the AUROC characteristic, using the overall optimally calibrated model defined in Figure 4 and displayed in **Table 12**. Differential AUROC estimates across subgroups will be tested for significance by calculating the P-value for interaction.

Missing data

The overall rate of missing item response data is anticipated to be very low**.** Missing data will be included in flowcharts and relevant summary tables, allowing denominators to remain consistent in calculations. A sensitivity analysis for the primary validation model will be performed with missing item response data imputed using Multiple Imputation by Chained Equations if the level of missingness is above 5% overall (i.e. per questionnaire) or for any individual item. Patients with a missing value for the reference standard result will be excluded from analysis.

Diagnostic test accuracy will be presented again as AUROC, Sensitivity, Specificity, Positive Predictive Value, Negative Predictive Value overall (again with optimal score calibration overall and by country) and at different cut-offs (pre-defined as between a WHQ score between 5 and 9) (**Table 12**). Uncertainty in these estimates will be presented using 95% confidence intervals.

Other exploratory analyses

Further exploratory analyses have been pre-defined for the purpose of enriching the understanding of the validation analysis. These include:

- The Rasch unidimensional measurement model examines the extent to which individual items in the WHQ capture a unidimensional wound healing construct and therefore the extent to which summation of scores from the WHQ are legitimate and will deliver reliable measurement. Fit to the Rasch model identifies anomalies in the data. Fit will be established using a variety of indicators and fit statistics e.g., targeting of the population to the WHQ, including fit of items, differential item function (DIF) (e.g. gender, language, country) and dependency on other items. Mis-fitting items not conforming to the requirements of the Rasch model will be removed. Scoring categories may need to be reviewed if disordered thresholds are identified as part of the Rasch analysis. The Rasch Unidimensional Measurement Model software (RUMM2030) will be used to analyse the PROM. An exploratory analysis will analyse the impact of using Rasch transformed score cut-offs versus simple raw score cut-offs on the AUROC receiver operating characteristic curve characteristic (7-9).

- A decision curve analysis to explore threshold values and support implementation of the telephone follow-up questionnaire into global surgery trials and/or practice.

Either one, both, or neither of these exploratory analyses may be presented in the primary study report depending the complexity of between-country and between-language patterns that emerge from the data.

Reporting

This statistical analysis will be reported with reference to SAMPL (Statistical Analyses and Methods in the Published Literature) guidelines (10), Patient-Centred Outcomes Research Institute's (PCORI) methodology standards (11), STARD guidelines for diagnostic test accuracy studies (12), and COSMIN guidelines for patient reported outcomes research (13).

Analyses will be performed using R Studio V3.1.1 (R Foundation for Statistical Computing, Vienna, Austria) and RUMM2030 (RUMM Laboratory LTD, Perth, WA, Australia).

Tables and figures

- Figure 1. Flow of participants recruited to TALON from recruitment through to feasibility (*Group 1 and Group 2*) and validity (*Group 1*) analyses

Feasibility

- Table 1. Baseline demographics by country *(Group 1 and Group 2 data)*

- Table 2. Baseline demographics by patient home location (urban versus rural) *(Group 1 and Group 2 data)*

- Table 3. Telephone follow-up pathway by country *(Group 1 and Group 2 data)*

- Table 4. Telephone follow-up pathway by patient home location (urban versus rural) *(Group 1 and Group 2 data)*

- Table 5. Feasibility outcomes by country *(Group 1 and Group 2 data)*

- Table 6. Feasibility outcomes by patient home location (urban versus rural) *(Group 1 and Group 2 data)*

- Table 7. Differences in baseline demographics between patients followed-up in person *(Group 1)* and with no in-person follow-up *(Group 2).*

- Table 8. Differences in proportions of patients with *(Group 1)* and without *(Group 2)* in-person follow-up by month throughout the study period

- Table 9. Factors associated with unsuccessful telephone contact in the multivariable model *(Group 1 and Group 2)*

- Figure 2. Proportion of all patients with telephone follow-up that also received 30-day in-person follow-up month by month over the study period (Overall, and by country) *(Group 1 and Group 2)*

Validity

- Table 10. Cross-tabulation of patients Wound Healing Questionnaire (WHQ) score and whether or not they received a diagnosis of SSI at the in-person assessment 30-days after surgery *(Group 1)*

- Table 11. Summary of diagnostic test accuracy characteristics overall (calibrated and optimised at country level) and at different pre-defined cut-offs *(Group 1).*

- Table 12. Area under the receiving operator curve characteristic for the WHQ across key subgroups *(Group 1).*

- Figure 3. Receiver operating characteristic curves for Wound Healing Questionnaire overall (calibrated and optimised at country level) and at different pre-defined cut-offs *(Group 1).*

- Figure 4. Sensitivity analysis for the primary model (ROC) using Multiple Imputation by Chained Equations to impute missing questionnaire and item data *(Group 1).*

References

1.   Horan TC, Andrus M, Dudeck MA. CDC/NHSN surveillance definition of health care-associated infection and criteria for specific types of infections in the acute care setting. Am J Infect Control. 2008;36(5):309-32.

2.   Reeves BC, Andronis L, Blazeby JM, Blencowe NS, Calvert M, Coast J, et al. A mixed-methods feasibility and external pilot study to inform a large pragmatic randomised controlled trial of the effects of surgical wound dressing strategies on surgical site infections (Bluebelle Phase B): study protocol for a randomised controlled trial. Trials. 2017;18(1):401.

3.   Macefield R, Brookes S, Blazeby J, Avery K. Development of a 'universal-reporter' outcome measure (UROM) for patient and healthcare professional completion: a mixed methods study demonstrating a novel concept for optimal questionnaire design. BMJ Open. 2019;9(8):e029741.

4.   Macefield RC, Reeves BC, Milne TK, Nicholson A, Blencowe NS, Calvert M, et al. Development of a single, practical measure of surgical site infection (SSI) for patient report or observer completion. J Infect Prev. 2017;18(4):170-9.

5.   Bluebelle-Study-Group. Validation of the Bluebelle Wound Healing Questionnaire for assessment of surgical-site infection in closed primary wounds after hospital discharge. Br J Surg. 2019;106(3):226-35.

6.   Elliott D, Bluebelle-Study-Group. Developing outcome measures assessing wound management and patient experience: a mixed methods study. BMJ Open. 2017;7(11):e016155.

7.   Andrich D. Rating scales and Rasch measurement. Expert Rev Pharmacoecon Outcomes Res. 2011;11(5):571-85.

8.   Tennant A, Conaghan PG. The Rasch measurement model in rheumatology: what is it and why use it? When should it be applied, and what should one look for in a Rasch paper? Arthritis Rheum. 2007;57(8):1358-62.

9.   Tennant A, Penta M, Tesio L, Grimby G, Thonnard JL, Slade A, et al. Assessing and adjusting for cross-cultural validity of impairment and activity limitation scales through differential item functioning within the framework of the Rasch model: the PRO-ESOR project. Med Care. 2004;42(1 Suppl):I37-48.

10.   Lang TA, Altman DG. Basic statistical reporting for articles published in biomedical journals: the "Statistical Analyses and Methods in the Published Literature" or the SAMPL Guidelines. Int J Nurs Stud. 2015;52(1):5-9.

11.   Gaglio B, Henton M, Barbeau A, Evans E, Hickam D, Newhouse R, et al. Methodological standards for qualitative and mixed methods patient centered outcomes research. Bmj. 2020;371:m4435.

12.   Cohen JF, Korevaar DA, Altman DG, Bruns DE, Gatsonis CA, Hooft L, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. BMJ Open. 2016;6(11):e012799.

13.   Mokkink LB, de Vet HCW, Prinsen CAC, Patrick DL, Alonso J, Bouter LM, et al. COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. Qual Life Res. 2018;27(5):1171-9.