The Nkateko Trial

Statistical Analysis Plan

Approved by TSC:

Signature (Chair): _____ Date: ___/___/_____

Trial Statisticians:

Eustasius Musenge, Senior Biostatistician[1]
Tobias Chirwa, Professor of Biostatistics, Head of the Division of Statistics and Epidemiology[1]
Sandra Eldridge, Professor of Biostatistics[2]

Principal Investigators

Margaret Thorogood, Professor of Epidemiology, & Honorary Professor[3],
Jane Goudge, Director of the Centre of Health Policy[1],

Co-Investigators (Alphabetical)
Eustasius Musenge, Senior Biostatistician
Melanie Betram, Health economist
Rhian Twine, Head of LINC Office, MRC/Wits Agincourt HDSS Unit
Sandra Eldridge, Professor of Biostatistics,
Stephen Tollman, Director, MRC/Wits Agincourt HDSS Unit
Tobias Chirwa, Head of the Division of Epidemiology and Biostatistics,
Xavier Gómez-Olivé, Research Manager, MRC/Wits Agincourt HDSS Unit

This document was an adaption of the Mental Health & Neuroscience Clinical Trials Unit (MH&N CTU) Analysis Strategy template (version 1.5; 13/02/2008) and was drafted by Eustasius Musenge and Tobias Chirwa

[1] School of Public Health, University of the Witwatersrand
[2] Queen Mary University of London
[3] Warwick Medical School University of Warwick, UK

Table of Contents

1 INTRODUCTION

1.1 Purpose of Statistical Analysis Plan

This document states in detail the plans for the analysis of the main outcome of the study and of the quantitative data that will be collected in the Nkateko trial. The quantitative outcome data will be reported in one or more principal papers on the results of the trial. Subsequent papers, which will be more exploratory or descriptive, will not be bound by this strategy and will be clearly identified as such.

1.2 Members of the Writing Committee

The following constitute, in alphabetical order after the lead writer, the members of the writing committee of the analysis plan:

- Eustasius Musenge
- Felix Limbani
- Jane Goudge
- Margaret Thorogood
- Sandra Eldridge
- Tobias Chirwa
- Xavier Gómez-Olivé

1.3 Summary

South Africa has a high and rising prevalence of hypertension, many people are unaware they have the condition and pharmaceutical management is often inadequate. Until recently, primary care clinics focused on management of acute conditions, but recent government initiatives are shifting the focus to management of chronic disease, including HIV and hypertension. This cluster-randomised controlled trial will test the effectiveness of a new clinic-based lay health worker to supplement government initiatives and support care of chronic disease.

1.4 Changes from planned analysis in the protocol

No changes are planned from the analysis described in the protocol.

2 STUDY OBJECTIVES AND ENDPOINTS

2.1 Study Objectives

2.1.1 Primary Objective

To compare the effectiveness of a quality improvement intervention involving use of clinic based lay health workers to ‘usual care’, in improving access to care, adherence to treatment, and management of hypertensive patients, in rural South Africa

2.2 Outcome Measures

2.2.1 Primary Outcome

A population level measure of hypertension control will be derived from cross-sectional surveys carried out before and after the intervention. This primary outcome will be the change between the two surveys in the percentage of people in the population who have elevated blood pressure that is combined with other factors resulting in a risk profile that indicates moderate or greater added risk of cardiovascular disease. More detailed definitions of outcomes are described in section 5.7.

2.2.2 Secondary Outcomes

a) Change in proportion of the population with undiagnosed hypertension (see table 10 in the appendix)
b) Change in the proportion of the population reporting they had had their blood pressure measured (see table 10 in the appendix),
c) Change in the proportion of the population reporting that they are using medication for hypertension (see table 10 in the appendix)
d) Change in the proportion of the population at different levels of blood-pressure-related cardiovascular risk by age group and sex (see tables 5, 6, 7, 8 and 9 in the appendix)
e) Change in the proportion of people in the population reporting that they have attended a clinic in the last year (see table 10 in the appendix)
f) Retention in care of people with diagnosed hypertension defined by the proportion of appointments kept during the study period (see table 11 in the appendix)

3 STUDY METHODS

3.1 Overall Study Design and Plan

We will conduct a cluster-randomised trial in tandem with a detailed realist evaluation, drawing on global experience of evaluating complex interventions[1,2]. The units of randomisation will be health facilities and their surrounding catchment populations. We propose to randomise eight clinics, four of which will receive the intervention. To achieve our primary outcome, we aim to both increase the proportion of the population under active management for their hypertension and reduce the level of blood pressure in those patients already receiving care. For this reason, the outcome of the trial will be measured at population level, and we estimate that the trial has a power of above 80% to detect an 11% to 13% reduction in the proportion of the population at moderate or greater cardiovascular risk as a result of their blood pressure and other risk factors. Realist evaluation will provide data on adaption of the intervention to the context, individual and organizational processes of change as well as contextual factors that influence outcomes[3]. We are collecting data at two cross-sectional points 24 months apart: baseline (when intervention is introduced) and at the end of intervention at each clinic to provide data on the sustainability of the intervention. The detailed study methods are described in the following sections.

3.2 Study Population

The trial will be based in the Agincourt sub-district of Mpumalanga Province, South Africa. Since 1992 the MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt) has collected population data, with vital events (pregnancy outcome, death, migration) updated yearly. The total population under surveillance is about 90,000 people (52,592 older than 18 years) who live in 15,500 households in 26 villages within a rural, former Bantustan area, with high labour migration.

Inclusion criteria for clusters (clinics): publically funded primary care clinics serving a population which includes people living in the Agincourt HDSS area.

Inclusion criteria for individuals participating in the cross-sectional surveys: Adults (male and female) aged over 18 years, and permanently resident in the Agincourt HDSS area. The sample was selected proportional to the full Agincourt population with oversampling on the older population.

Exclusion criteria for individuals participating in the cross-sectional surveys and the process evaluation interviews: women who report that they are pregnant, individuals who are unable to give informed consent or are unable to respond to the questionnaire.

3.3 Method of Randomisation

Randomisation took place in the community at a meeting of clinic staff and members of local clinic committees. Eight primary health care facilities were randomised. After showing sheets of paper with the names of the clinics to the meeting, they were put into sealed envelopes, several community members were invited to shuffle the envelopes, which were then put into a bag and seven other community members each in turn chose one envelope, and the chosen clinic was allocated to a slot in the order of intervention clinic, control clinic, intervention clinic, etc. Appendix A gives the summary notes for the randomisation meeting that was held.

3.4 Treatment masking (Blinding)

This is a trial of a health service intervention where blinding of participants was not possible. In addition, the field workers who collected the baseline and outcome measures live in the area and use the clinics, so they may be aware of which clinics had extra lay health workers. We will be recruiting and training the field team for the survey that will provide the outcome measure towards the end of the intervention. We will ensure that the field workers are told only that this is a survey on hypertension and clinic use. The hypothesis of the trial will not be disclosed. The statistician responsible for the main analysis will remain blinded to which clusters are intervention and non-intervention as far as possible.

3.5 Sample size determination

The two cross-sectional surveys, which will be used to assess the effectiveness of the intervention, will each include at least 4000 participants, giving approximately 500

people in each cluster. We adopted the use of the coefficient of variation (standard deviation of the cluster means divided by the overall mean) as used in similar study settings when we cannot get a good intra-cluster variation[4-6]. For a background prevalence of 36% (i.e. proportion of patients at moderate or greater risk of cardiovascular disease) and a coefficient of variation of 0.132 (error margin 4.5% (0.132±0.045)) based on data collected in the same site in 2010, different scenarios of proportions of moderate or greater risk patients at the end of trial in the control and intervention arms and their associated power are shown in Table 1. Based on the scenario of a 15% difference (i.e. 36% in control vs 21% in intervention), the highest power will be 97.4% and we will have power of above 80% to detect an 11% absolute reduction of people at moderate or greater cardiovascular risk. These calculations assume that the coefficient of variation will be similar in the two groups and that effects of the interventions are similar across clusters.

Table 1: Power matrix of different scenarios

	Control	**"36%" No change**
	CV	**0.132**
Intervention	"5% difference" 31%	**20.7**
	"10% difference" 26%	**68.6**
	"12% difference" 24%	**85.4**
	"15% difference" 21%	**97.4**
	Control	**"34%" 2% change**
	CV	**0.132**
Intervention	"3% difference" 31%	**10.5**
	"8% difference" 26%	**52.7**
	"10% difference" 24%	**75.4**
	"13% difference" 21%	**94.2**
	Control	**"32%" 4% change**
	CV	**0.132**
Intervention	"1% difference" 31%	**4.3**
	"6% difference" 26%	**35.2**
	"8% difference" 24%	**58.4**
	"11% difference" 21%	**87.7**

4 DATA COLLECTION

4.1 Baseline

The baseline survey will collect data by questionnaire on self-reported medical history (hypertension, diabetes, stroke, heart failure, angina, heart attack), family history of heart problems and stroke, smoking, most recent use of primary care clinics in the last 12 months, and their preferred clinic. Participants' blood pressure, total cholesterol, random blood glucose (using an Omron automatic blood pressure machine [model M6W]) and waist circumference will be measured by trained field staff. The sample for the survey will be drawn from the Agincourt HDSS census database. In addition to the information collected from the participants, the following personal and household level information will be drawn from the census database: sex, household, date of birth, marital status, education, employment status, household asset score and nationality of origin.

4.2 Follow up

The follow-up survey will use a different sample, which will be selected using the same age/sex distribution. The data collection will be the same as that listed above for the baseline survey.

4.3 Timing of Data collection

The baseline survey took place September – November 2013, following which the intervention started in February 2014 and last for 21 months. The follow-up survey took place September to November 2015.

4.4 Assigning population survey participants to clusters

The survey asks two questions, one on usual clinic and the other on last clinic used. For the main analyses, the "usual clinic" variable was be used to assign individuals in the survey to a cluster. Some individuals had a usual clinic that was not participating in the study. These individuals will be excluded from the analyses but their results will be briefly described.

4.5 Clinic link data

During the intervention period, a study data clerk asked every chronic patient arriving at both control and intervention clinics, for consent to collect data from their clinical file and to link it to their census record. Following consent, all visits for that patient were then recorded.

5 GENERAL ISSUES FOR STATISTICAL ANALYSIS

5.1 Blinding of the Statistical Analysis

Clinic identifiers were encrypted in the dataset, so that the primary analysis is carried out blinded to which group of clinics are the intervention clinics.

5.2 Population Surveys

Two population surveys were carried out (one before the intervention starts and one after it finishes) to estimate the primary outcome of the trial. In each survey, a random sample of 5000 people aged over 18 years were selected from census records, to allow for 20% attrition, so that at least 4000 individuals were included in each survey, contributing approximately 500 people in each of the 8 clusters. From previous experience in this research setting, we expect good participation (~80%). The sample was disproportionately stratified to ensure adequate representation of males and older people. This is necessary because the population pyramid is heavily weighted to younger people and there are fewer men than women amongst older adults due to labour migration.

5.2.1 Intent-to-treat and Per protocol population

We will conduct an intention-to-treat analysis with respect to clinic. Since we were not recruiting individuals and collecting data from cross-sectional surveys the concept of intention-to-treat does not apply to these individuals in the usual way. Individuals were asked to specify their usual clinic and the last clinic that they attended. For the primary analysis they will be analysed in the clinic they specify as “usual”, assuming therefore that the intention was that they received the intervention. Those individuals who specify a “usual” clinic outside the 8 intervention and control clinics will not be included in this analysis. We will undertake a sensitivity with individuals analysed in the “last” clinic that they attended – see also section 6.5.

5.3 Database

5.3.1 Data quality

Quality control process started in the field. Fieldworkers check their own forms for completeness. Thereafter, the Team Supervisors re-check the forms and return those not complete to the field workers for completion. Forms designated as complete by the Team Supervisors are then passed to a Quality Checker who further checks them for completeness and consistency. All forms identified with problems by the Quality Checker are returned to the Project Site Manager. The Project Site Manager discusses the problems with the field teams and gives the forms back to the Fieldworkers to correct. After correction, the forms are sent again to the Quality Checker. Forms that satisfactorily pass the Quality Checker’s checks are afterwards sent for data entry.

The data is being entered into a Microsoft SQL server relational database through a user interface written in Visual Basic.net. Two Data Capturers independently enter the data from each form. On entry, the data from each form undergoes validation and consistency checks that are built into the data entry software. Forms whose data fail the checks are returned to the field teams for correction. After data entry, the Quality Checker runs a program that compares the first and second entry records and corrects discrepancies between them. Corrected records and records that had no mistakes are copied into a verified entry record in the database.

Data are entered while fieldwork is still in progress so that mistakes could be rectified immediately. All the forms had a unique barcode identifier that is used to tracking the movement of the forms at each point of the quality control, data entry and data verification processes.

5.3.2 Database freeze

The data base will be frozen once all the data from both surveys have been entered and all validity checks have been completed. Date of database freeze to be confirmed

5.4 Analysis software

All the statistical analysis will be done using Stata versions 13 and 14, from within which "do" (STATA program) files for data management, preliminary analysis and main analysis will be done with comments for ease of follow-ups. Further, log files will be used to keep track of output data.

5.5 Methods for withdrawals, loss to follow-up and Missing Data

We will not have withdrawals nor loss to follow-up since this is a survey cross-sectional design. However, participants selected for the sample may not be present when we visit the household the first time, in which case we will carry out up to two more call backs. In the case where field call backs and tracing have not been successful to ensure all key data have been collected, multiple imputation methods will be used to generate outcomes. For example, blood pressure level might be missing for some cases. We will conduct a complete case analysis and also an analysis where missing blood pressure levels are imputed (imputed analysis) and compare whether there are apparent variations in the results obtained. The same applies for other key variables like self-reported hypertension.

Multiple imputation methods for cluster randomised trials with few clusters are not well developed, although this is a growing research field. A systematic review done recently showed none of the studies reviewed had a multiple imputation procedure that handled clustering and very few techniques also used these in regression models that truly adjusted for the clustering effect[7]. We intend to implement some of these procedures in our analysis as part of sensitivity analysis, but the main analysis will be based on the complete cases.

5.6 Method for handling Outliers

At the data entry stage, ranges are set for key outcome variables as collected in the questionnaires. The data entry screens and databases will be created in reference to the questionnaire so the range checks are consistent with what was expected. For example, variable to record hypertension will be coded as 1 if "yes" and 0 if "no". Any values, besides these two, will be considered out of range. For continuous variables, it is a challenge to determine outliers. For example, for blood pressure and blood glucose levels, the team will have to determine acceptable levels, and as such these cannot easily be set at data entry stage.

If a value is flagged as being out of range the questionnaire is returned to the field team. If there is no apparent error, then the out-of-range variable is considered by two senior members of the research team and a decision is made in each case on whether the

value is plausible (for example a known diabetic with an out-of-range blood glucose), If the value was felt to be implausible the variable is being entered as 'missing'.

In terms of methods in analysis for handling this, the approaches we will adopt will be to work with and without outlier variables; and also adopting multiple imputation procedures able to cater for clustering.

5.7 Derived and Computed Variables

This section is directly linked to objectives of the study (Section 2 and, specifically, section 2.2 on outcomes).Two variables we will need for many of the outcomes listed below are mean systolic hypertension and mean diastolic hypertension. Each of these will be calculated by discarding the first of the three measures taken (because the first one is usually artificially high due to the person's reaction to having their blood pressure measured), and then calculating the mean of the second and third measures. The same goes for mean pulse.

The **primary outcome** (as outlined in Table 2) requires us to calculate 'the percentage of people in the population who have elevated blood pressure that is combined with other factors resulting in a risk profile that indicates moderate or greater added risk of cardiovascular disease'. Appendix C (section 15.3) shows STATA commands for generating this primary outcome variable.

We have listed six secondary outcomes (Section 2.2.2). The first of these is a change in the **proportion of the population with undiagnosed hypertension.** We will need a new binary variable for undiagnosed hypertension. This should be calculated using a combination of mean systolic and diastolic BP plus the answers to questions 2 and 3 – (on the questionnaire). Thus, someone with undiagnosed hypertension will have answered "No" to question 2 and question 3 below, and will have EITHER a mean systolic pressure equal to or greater than 140 OR a mean diastolic pressure equal to or greater than 90 (or both). Question 2 states that "*Have you ever been told by a doctor, nurse or other health worker that you have raised blood pressure or hypertension?*" Question 3 is "*Have you received any drugs for high blood pressure prescribed by a doctor, nurse or other health worker in the last two weeks?*"

The second secondary outcome checks for **change in the proportion of the population reporting they had had their blood pressure measured**. This will be derived from Question 4 – "*Have you ever had your blood pressure measured by a doctor, nurse or other health workers?"* The third secondary objective is on **change in the proportion of the population reporting that they are using medication for hypertension.** Derived from Question 3 – "*Have you received any drugs for high blood pressure prescribed by a doctor, nurse or other health worker in the last two weeks?.* These two are direct variables from the questionnaire and do not need any further computation. For the primary analysis of this outcome, only those who name specific drugs will be categorized as being prescribed drugs for high blood pressure.

The fourth is change in the **proportion of the population at different levels of blood-pressure-related cardiovascular risk by age group and sex**. In this case we will also

need **blood-pressure-related cardiovascular risk** divided into the different levels of no added risk, low added risk, moderate added risk and high added risk (see Table 2).

The secondary objectives under section 2.2.2, e and f (restated below) do not need any additional computations, and will be derived from clinic link data. These two outcomes are:

e) Change in the proportion of people in the population recorded as having attended a clinic in the last year.
f) Retention in care of people with diagnosed hypertension defined by the proportion of appointments kept during the study period.

Note that the clinic link data will provide information on the proportion of people recorded as having attended a clinic in the last year, rather than, as originally stated, reporting they had attended.

5.8 Description of clinic link data

The clinic link data will provide information on attendance at any one of the eight trial clinics over a period of around 18 months by residents of the Agincourt research census site. Consenting individuals attending a clinic are identified in real time on the census data base (using fuzzy matching) by trained data clerks posted in all the clinics. Information about all clinic attendances of identified individuals, including diagnosis and medication prescribed, is recorded. These linked data will enable us to compute by clinic:

- The proportion of people in the population recorded as having attended each clinic in the last year
- Proportion of appointments kept during the study period by those individuals with a diagnosis of hypertension.

The clinic link data also provide the potential to describe socio-demographic differences between those who attend and those who do not attend the clinics – an investigation of association with secondary outcomes (bullet points above) using the logistic regression models, taking into account the clustering in the data.

5.9 Description of available data

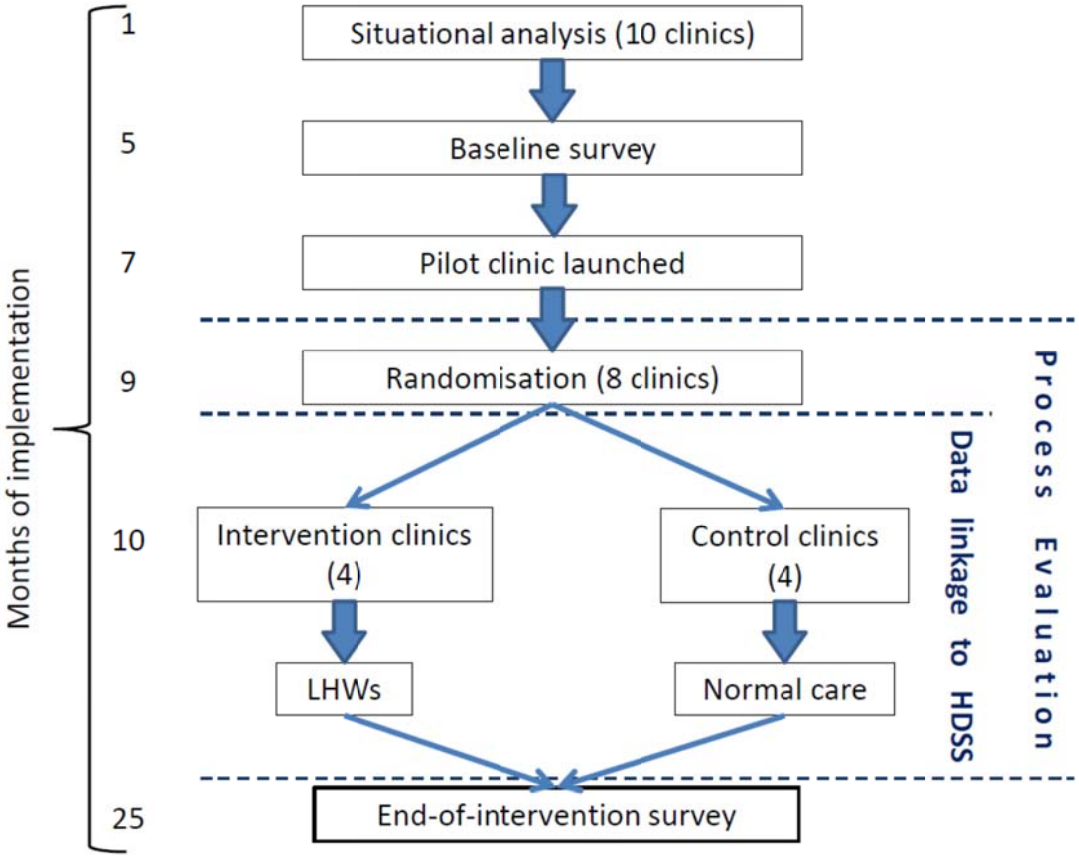
Descriptive statistics for continuous variables such as age (in years) will be done using relevant measure of central tendency (mean) and measures of spread (standard deviation) of the data accounting for clustering. For continuous variables which are normally distributed, we will report the means and standard deviations adjusted for clustering. For example, on the variables age or blood pressure levels; we can report the mean ± standard deviation when data are normal. Apart from presenting these measures, data will also be presented graphically using some of the following: histograms, box and whisker plots, lowess plots, scatter plots and kernel density plots.

Categorical variables such as gender will be described through frequency tabulations by reporting the number of observations and missing values, where necessary, for instance on gender: male (n, %), female (n, %). Bar charts and pie charts will also be used to give

visual descriptive of the categorical variables. We will estimate population prevalence by standardizing this to the Agincourt age groups distribution.

5.10 Flow Diagram for Trial

The following figure shows the schematic view of the data flow and randomised between the control and intervention cluster arms.



5.11 Representativeness of sample

For each of the two population surveys, we intended to select a random sample of 5000 people aged over 18 years from the demographic surveillance database. The baseline survey has been completed, with a sample of 4903 persons and a response of 3905 (79.6%). The samples for the two surveys will be disproportionately stratified to ensure adequate representation of males and older people. This is necessary because the population pyramid is heavily weighted to younger people and there are fewer men than women amongst older adults due to labour migration [8,9]. The samples will be a fair representation of the resident adult population in Agincourt at both the baseline and end of intervention visit. Because labour migrants are living away most of the time and are not available for interview (and only rarely use local primary care services) they are not

included in the surveys. Those individuals that are identified as labour migrants in the census database will not be selected for the surveys, but there will be some people who have become labour migrants since the last census update, and these people will inevitably be missed in the survey.

5.12 Baseline comparability of randomized groups

Once the database is complete with data from both population surveys entered, we will compare the respondents' characteristics by randomised group and by individual clinic to assess the comparability of the groups. We want to ensure that individuals in the control and intervention clusters are similar by describing their socio-demographic factors; these will be described for both the baseline and the follow-up survey. If similar, this will strengthen our findings to the extent that other factors constant, we can attribute differences after 24 months to the intervention strategies in place. We will also compare with the intervention and control group those individuals surveyed who did not attend a participating clinic – we expect these numbers to be small.

6 ANALYSIS OF PRIMARY OUTCOME

6.1 Definition of outcome measure

The primary outcome will be the change in the percentage of people in the population who have elevated blood pressure that is combined with other factors resulting in a risk profile that indicates moderate or greater added risk of cardiovascular disease. as indicated by the shaded cells in Table 2. This includes individuals with either: a systolic blood pressure of 160 and above, diastolic of 100 and above, systolic blood pressure of 140-159 plus one or more risk factors, or diastolic blood pressure 90-99 plus one or more risk factors. These data will be obtained from two population surveys (one before the start of the trial and one after the trial ends).

Table 2. Modified South African Guideline: Stratification of cardiovascular risk in patients with hypertension (defined as SBP>139 or DBP>89)

Blood pressure (mmHg)	**Presence of risk factors or other conditions**			
	No risk factors	One or two risk factors	Three or more risk factors or diabetes	Associated clinical conditions
SBP 140 -159 or DBP 90-99	Low Added Risk	Moderate Added Risk	High Added Risk	Very High Added Risk
SBP 160 -179 or DBP 100-109	Moderate Added Risk	Moderate Added Risk	High Added Risk	Very High Added Risk
SBP 180 + or DBP 110+	High Added Risk	Very High Added Risk	Very High Added Risk	Very High Added Risk

6.2 Descriptive statistics for outcome measure

An overall frequency tabulation of the primary outcome cross-classified by group will be produced, based on the study design.

6.3 Primary analysis

The analysis of the primary outcome (binary, as defined above) will be done in three steps Thus, firstly, a group level analysis of the overall effect of the intervention using the test of differences in two proportions (within each group at baseline and after follow-up and between intervention and control clusters) will be conducted adjusting for the design effect and clustering correction factors for each group. We will not use the usual Pearson's Chi-Square test which assumes independence as our data are clustered. Intra-class correlation (ICC) adjusted Pearson's Chi-square versions will be used as those used by Reed. These will involve creating and pilot testing a STATA program to perform these adjusted Chi-Square based analyses for binary outcome data[10-12].

6.4 Assumption checks

Post logistic regression derived residuals for pre and post analysis will be tested for normality using tests such as Shapiro-Francia, Kolmogorov-Smirnov test or the Skewness Kurtosis test. The F-ratio test will be used to test for equality of variances between the two groups being compared. Other model fit diagnostics will be performed to assess for outliers, leverage and influential data taking cognizance of the clustering.

6.5 Other analysis supporting the primary analysis (including sensitivity analyses)

Apart from looking at the observed effect of the intervention, a sensitivity analysis will be conducted:

1. The primary analysis on hypertensive drugs described in section 5.7 excludes from the analysis those who do not name any specific drugs we will include these in a sensitivity analysis.

2. We will undertake sensitivity with individuals analyzed in the "last" clinic that they attended.

3. We will adjust for potential confounding factors using the two stage regression model[13]. Firstly, two logistic regression models for control and intervention clinics that include covariates will be fitted separately, with individual level covariates such as gender and age. Secondly, based on residuals from the first stage we will aggregate outcomes to cluster level and cluster/clinic level factors (e.g. clinic size) will be used in regression modeling at the cluster level to test the effect of the intervention.

4. In order to compare the results from the two stage regression an alternative procedure will be used, the mixed effects model adjusting for covariates.

   Model sensitivity to the individual-level data will be assessed on different temporal covariance structures (exchangeable, auto covariance and unstructured) on the final model. Such assessment will be made possible because of the many appointments with the 21 months of follow-up.

   Thus, endeavors to ensure data are complete will be explored rigorously taking advantage of this longitudinal nature of data and, in the process a report for missing data

by variable will be compiled. This will help to classify the missing data as either missing at random (MAR) or missing not at random (MNAR). Assuming the data are MAR the proposed mixed logistic models should sufficiently handle these. If otherwise, as with MNAR we will employ data imputation procedures accordingly, taking caution not to distort the patterns observed from the data. However, as this is an area in which research is ongoing, this will be an exploratory analysis, not reported in the main trial report, and drawing on latest research. The dummy tables for these and details on how imputation on clustered data will be done, will not be provided in this analysis plan.

7 ANALYSIS OF SECONDARY OUTCOMES

7.1 Secondary analysis

Secondary outcomes will be analyzed as described above for the primary outcome. They will include: proportion of population in each cluster screened (data from the population survey), adherent to medication and retained in care (from records of clinic activity). Adherence will be defined by using the records of pharmacy refills. The secondary objectives will investigate changes in proportions independently and dependently. The independent tests for proportions (Pearson's Chi-Square test) will be used for assessing between control and intervention/clusters change at the same time point.

7.2 Assumption checks of secondary outcomes

The following procedures will be done on the final multiple variable logistic regression model:

- Checking influential observations
- Checking multi-collinearity of explanatory variables
- Checking model specifications and perform any necessary interactions
- Testing for overall model goodness of fit

7.3 Adverse Events

Any incidents or accidents which are related to the conduct of this trial including the delivery of the trial intervention, or which affect trial staff while at work will be recorded and copies of the report sent to all three members of the Management Team and the Chair of the Trial Steering Committee. The Chair of the Steering Committee, in consultation with the Management Team, will decide whether any further investigation or action is necessary. There is no external body looking at adverse events.

8 REFERENCES

1. Ssengooba F, McPake B, Palmer N. Why performance-based contracting failed in Uganda--an "open-box" evaluation of a complex health system intervention. *Social science & medicine* 2012; **75**(2): 377-83.
2. Craig P, Dieppe P, Macintyre S, et al. Developing and evaluating complex interventions: the new Medical Research Council guidance. *Bmj* 2008; **337**: a1655.
3. Pawson R, Tilley N. Realist Evaluation. London: Sage Publications Ltd; 1997.
4. Hayes R, Bennett S. Simple sample size calculation for cluster-randomized trials. *International Journal of Epidemiology* 1999; **28**(2): 319-26.
5. Pronyk PM, Hargreaves JR, Kim JC, et al. Effect of a structural intervention for the prevention of intimate-partner violence and HIV in rural South Africa: a cluster randomised trial. *Lancet* 2006; **368**(9551): 1973-83.
6. Murray DM, Varnell SP, Blitstein JL. Design and analysis of group-randomized trials: a review of recent methodological developments. *Journal Information* 2004; **94**(3).
7. Díaz-Ordaz K, Kenward MG, Cohen A, Coleman CL, Eldridge S. Are missing data adequately handled in cluster randomised trials? A systematic review and guidelines. *Clinical Trials* 2014: 1740774514537136.
8. Kahn K, Collinson MA, Gomez-Olive FX, et al. Profile: Agincourt health and socio-demographic surveillance system. *International journal of epidemiology* 2012; **41**(4): 988-1001.
9. Collinson MA, White MJ, Bocquier P, et al. Migration and the epidemiological transition: insights from the Agincourt sub-district of northeast South Africa. *Global health action* 2014; **7**: 23514.
10. Donner A, Donald A. The statistical analysis of multiple binary measurements. *Journal of clinical epidemiology* 1988; **41**(9): 899-905.
11. Donner A, Klar N. Methods for comparing event rates in intervention studies when the unit of allocation is a cluster. *American journal of epidemiology* 1994; **140**(3): 279-89; discussion 300-1.
12. Reed JF, 3rd. Adjusted chi-square statistics: application to clustered binary data in primary care. *Annals of family medicine* 2004; **2**(3): 201-3.
13. Hayes R, Moulton L. Cluster Randomised Trials. Dordrecht Netherlands: Chapman & Hall/CRC Interdisciplinary Statistics; 2009.

9 APPENDICES

9.1 Appendix A: SOP for Randomization Meeting

Objectives

- Ensure that all attendants understand the project and the next steps
- To randomly select 4 out of 8 clinics to receive the Nkateko Intervention
- To ensure that the clinic staff and the wider Agincourt population are confident that this process is truly random and is not influenced by any members of the Nkateko team
- To have a written and photographic record of the procedure for use in future papers, talks, etc

Invitees

- 1 member from each of the 8 clinic committees
- Facility manager from each clinic
- Both primary health care supervisors
- Members of the Agincourt Community Advisory Group Non-communicable disease task team
- 1 representative from each village

Prepare beforehand:

- Eight pieces of paper (suggest A5 size) each with the name of one clinic in large letters
- Eight opaque envelopes made from reasonable quality paper (so they will withstand shuffling many times) (
- Flip-chart with eight spaces for writing names, labelled sequentially as 'Clinic 1 Intervention', 'Clinic 2 Comparison', Clinic 3 Intervention', Clinic 4 Comparison' and so on.

Procedure

- Explain Nkateko
- Explain how we will randomise
- Randomisation procedure
    - Make sure everyone can see the eight clinic names and then put each one into an envelope and seal it.
    - Ask at least three (preferably more) different people in the audience to shuffle the envelopes
    - Ask seven people in turn to select an envelope and open it. Each time, as the envelope is opened, fill in the next space on the flip chart. The last space is filled by the remaining clinic
    - Make a record of the result and take a photo of the completed flip chart.
- Explain next steps

Notes from Nkateko randomisation meeting

DATE: 22 January 2014
VENUE: **MRC/Wits Agincourt Offices, Agincourt village**
PURPOSE: **Randomisation of clinics for Nkateko study**

Translator – Audrey Khosa
38 people attended

Opening prayer and welcome – Rhian Twine
The Nkateko Project

Overview – Felix Limbani

Felix gave a background to the study – using a PowerPoint presentation (available from Felix for reference. He mentioned reasons for the study and that the intervention fits into the new ICDM model. We will need to work with the clinics to make sure that the lay health workers work improves chronic care in the clinics. Our meeting today is to decide which 4 clinics will have lay health workers and which 4 will not.

Evaluation - Felix Limbani

Last year we went to find out what is happening in the clinics and we did a population based survey of blood pressure. When the LHWs are in the clinic we will evaluate their work, and then after two years we will go back to the community and do another population based survey to see if their blood pressures have changed. Felix and his team's role is to evaluate the project while Zola will work to develop the LHW intervention

Lay health worker programme – Sr Zola Myakayaka

Zola described what the lay health worker will do in the clinic. Duties might differ between clinics as clinics are different (size, number of staff, business etc). Will workshop with clinic staff to decide how the lay health workers should be working with the clinic staff. Zola's role in the programme is to develop the work of the LHW in the clinic.

The randomisation process – Mr Felix Limbani

Felix explained how we are going to randomise the clinics.

1. Reminder that this is research
2. We will look at each clinic separately so that we can tell why the programme worked well in one clinic but not in the other if this is the case. This is especially true for this project as the LHW programmes will be different in each clinic.
3. We did not want to just decide in the office which clinic to chose, so we are doing it randomly and publicly. Every clinic should have an equal chance to be selected as an intervention clinic.
4. Described difference between control and intervention clinics.
5. Showed the eight pieces of paper with the clinic names on them
6. Showed the 8 unmarked envelopes
7. Put one clinic name into its own envelope
8. Put all the envelopes into the box
9. Two or three people will come and shake the box
10. Then we will get 7 people to come up and each chose an envelope.
11. First, third, fifth and seventh clinics taken out will be intervention
12. Second, fourth, sixth and eight clinics taken out will be control

Everyone agreed that it is a fair process before we started.

1. Zola put all the names into envelopes in front of everyone
2. All envelopes were put in a box
3. A volunteer came up to shake the box
4. Audience suggested we fold the envelopes so that they moved more easily in the box
5. We opened the box and folded the envelopes smaller
6. Second person came and shook the box
7. First person chose –
8. Another person shook the box

9. Second person chose -
10. Another person shook
11. Third person chose –
12. Someone always came up to shake in between another person coming to chose and the results are

Next steps – Sr Zola Myakayaka and Felix

1. Advertise for LHWs at clinics and surrounding villages
2. Interviews of applicant by Nkateko team and clinic staff and clinic committee
3. Appointment
4. Workshops with clinics to design the programme
5. Training

Closing prayer and refreshments – Rhian Twine

9.2 Appendix B: Analysis dummy tables

Table 1: Comparing socio-demographic variables **at baseline** for controls and intervention clinics

Variable		Control baseline	Control endpoint	Intervention baseline	Intervention endpoint	Neither intervention or control Baseline	Neither intervention or control Endpoint
Sex	Male						
	Female						
Age (yrs) mean (std)							
Marital status	In a union						
	Not in a union						
Highest Education Attained	No schooling						
	Primary						
	Secondary						
	Tertiary						
Household asset score	Most poor						
	Poor						
	Least poor						
Smoking at least every day	Never						
	Previous						
	Less than one a day						
	One or more a day						

Table 2: Descriptive clinical data comparing control, intervention, baseline and end of intervention

Risk factors		Control baseline Prop (n)	Control endpoint Prop (n)	Intervention baseline Prop (n)				Intervention endpoint Prop (n)
Family history of CVD* (M<55yrs, W<65yrs)	Yes							
	No							
Men's waist circumference M>94cms,	Yes							
	No							
Women's waist circumference W>80cms	Yes							
	No							
Glucose	Below 11 and not fasting							
	Between 6-11 and fasting							
	Above 11							
Self-reported diabetes	Yes and on treatment							
	Yes but not on treatment							
	No							
Total cholesterol	Above 5.1							
	Below or equal to 5.1							
Self-reported coronary heart disease								
Self-reported heart failure								
Self-reported Stroke or TIA								
Blood pressure	SDP below 140 and DBP below 90							
	SBP 140 -159 or DBP 90-99							

Risk factors		Control baseline Prop (n)	Control endpoint Prop (n)	Intervention baseline Prop (n)				Intervention endpoint Prop (n)
	SBP 160 -179 or DBP 100-109							
	SBP 180 + or DBP 110+							

Table 3: Proportion of individuals for whom the 'usual' clinic was the same as the 'last visited' clinic.

		Control		Intervention		Clinics outside area	
		Usual and last clinic are both control	Usual clinic is control, but last clinic is intervention	Usual and last clinic are both intervention	Usual clinic is intervention but last clinic is control	Usual and last clinic are both outside area, or last clinic is control	Usual clinic is outside area, but last clinic is intervention
Total							
Gender	Female						
	Male						
Age	Under 35						
	35-less 55						
	55 and over						

Table 4: Modified South African Guideline: Stratification of cardiovascular risk in patients with hypertension (defined as SBP>139 or DBP>89)

Blood pressure (mmHg)	**Presence of risk factors or other conditions**			
No risk of hypertension	No risk factors	One or two risk factors	Three or more risk factors or diabetes	Associated clinical conditions
SBP 140 -159 or DBP 90-99	Low Added Risk	Moderate Added Risk	High Added Risk	Very High Added Risk
SBP 160 -179 or DBP 100-109	Moderate Added Risk	Moderate Added Risk	High Added Risk	Very High Added Risk
SBP 180 + or DBP 110+	High Added Risk	Very High Added Risk	Very High Added Risk	Very High Added Risk

Use Table 4 to allocate individuals to level of risk, which is then used to complete Table 3

Table 5: Proportion of people with risk factors in the control clinics at baseline

Blood pressure (mmHg)	**Presence of risk factors or other conditions**			
No risk of hypertension	No risk factors	One or two risk factors	Three or more risk factors or diabetes	Associated clinical conditions
SBP 140 -159 or DBP 90-99				
SBP 160 -179 or DBP 100-109				
SBP 180 + or DBP 110+				

Table 6: Proportion of people with risk factors in the control clinics at end of intervention

Blood pressure (mmHg)	**Presence of risk factors or other conditions**			
No risk of hypertension	No risk factors	One or two risk factors	Three or more risk factors or diabetes	Associated clinical conditions
SBP 140 -159 or DBP 90-99				
SBP 160 -179 or DBP 100-109				
SBP 180 + or DBP 110+				

Table 7: Proportion of people with risk factors in the intervention clinics at baseline

Blood pressure (mmHg)	**Presence of risk factors or other conditions**			
No risk of hypertension	No risk factors	One or two risk factors	Three or more risk factors or diabetes	Associated clinical conditions
SBP 140 -159 or DBP 90-99				
SBP 160 -179 or DBP 100-109				
SBP 180 + or DBP 110+				

Table 8: Proportion of people with risk factors in the intervention clinics at end of intervention

Blood pressure (mmHg)	**Presence of risk factors or other conditions**			
No risk of hypertension	No risk factors	One or two risk factors	Three or more risk factors or diabetes	Associated clinical conditions
SBP 140 -159 or DBP 90-99				
SBP 160 -179 or DBP 100-109				
SBP 180 + or DBP 110+				

Table 9: Change in level of cardiovascular risk of study population (primary and secondary outcomes)

Level of cardiovascular risk	Control proportion(n)			Intervention proportion(n)			Statistical Test statistic assessing difference between cluster change and p-value
	Baseline Prop (n)	End Prop (n)	Within cluster change and p-value	Baseline Prop (n)	End Prop (n)	Within cluster change and p-value	
No or low risk							
No risk							
Low risk							
Moderate or higher risk							
Moderate risk							
High risk							
Very high risk							

Repeat for men and women

Table 10: Changes in hypertension risk, blood pressure, medication and attendance within clusters

	Control proportion(n)			Intervention proportion(n)			Statistical Test statistic assessing difference between cluster change and p-value
	Baseline Prop (n)	End Prop (n)	Within cluster change and p-value	Baseline Prop (n)	End Prop (n)	Within cluster change and p-value	
People with moderate or higher risk***							
Undiagnosed hypertension							
Self-reported blood pressure measured							
Self-reported using medication for hypertension							
Self-reported attended a clinic in the last year.							

Table 11: Clinic link record of number of visits by sex, age, patient attendance and availability of medication over 4 time periods (each 5 months long)

Variable		Control clinics				Intervention clinics			
		1	2	3	4	1	2	3	4
Total number of visits									
Number of patients									
Sex	Male								
	Female								
Age	Mean age								
Patient attendance (Proportion of visits attended within 3 or less days of appointment date)	Visit with 3 days of appointment								
	Visit NOT within 3 days of appointment								
	No appointment								
Given full set of medication	Yes								
	No								

9.3 **Appendix C: STATA code for computing the primary outcome**

```
   *--------------------------------------------------------------------*
   *          PRIMARY OUTCOME CREATION AND RELATED VARIABLES            *
   *--------------------------------------------------------------------*

   *Age and sex: men aged over 55 and women aged over 65 have a risk factor
for age
   gen age= (VisitDate - DoB)/365.25
   recode age (18/29.999=1 "18-29") (30/39.999=2  "30-39") (40/49.999=3  "40-
49") (50/59.999=4  "50-59") (60/69.999=5  "60-69") (70/79.999=6  "70-79")
(80/max=7  "80+"), gen(age_grp) label(agegrp)

   gen AgeRiskGroup=(age>55 & Gender=="M" & age!=. |age>65 & Gender=="F" &
age!=.)
   tab AgeRiskGroup age_grp

   /*Smoking: A smoker will have one risk factor for smoking if he/she
   •  answers ‘yes’ to Q20
   •  answers ‘yes’ to Q21
   •  answers ‘daily’ to question 22
   */
   gen smoking=(Q20_EverSmokedCigarettes==1 & Q21_CurrentlySmoking==1 &
Q22_SmokingFreqLast30Days<=2 & Q20_EverSmokedCigarettes!=.|
Q22_SmokingFreqLast30Days<=2 & Q20_EverSmokedCigarettes!=.)

   *Dyslipidemia If cholesterol level in  Q35 is > than 5.1mmol/L the person
has risk factor for dyslipidemia
   gen dyslipedemia=(Q34_TotalCholesterol>5.1 & Q34_TotalCholesterol!=.)

   /*Family history of CVD. A participant has a risk factor for family
history if who
   •  answers ‘yes’ to Q19 and
   •  either answers  ‘yes’ to Q19b or ‘yes’ toQ19c
   (only one risk factor regardless of how many relatives)
   */
   gen family_risk=(Q19_StrokeHeartFMember==1 & Q19_StrokeHeartFMember!=. |
Q19_StrokeHeartFather==1 & Q19_StrokeHeartFMember!=.|Q19_StrokeHeartMother==1
& Q19_StrokeHeartFMember!=.)
   label var family_risk "Family history of CVD"
   /*Waist circumference
   •  A man has a risk factor for waist circumference if it is >94cms in Q36
   •  A woman has a risk factor for waist circumference if it is >80cms in
Q36
   */
   gen waist=(Q36_WaistCircumference>94 &
Gender=="M"|Q36_WaistCircumference>80 & Gender=="F")

   gen waist01=.
   replace waist01=1 if Q36_WaistCircumference<94 &
Gender=="M"|Q36_WaistCircumference<80 & Gender=="F"
   replace waist01=2 if Q36_WaistCircumference>=94 &
Q36_WaistCircumference<102 & Gender=="M"|Q36_WaistCircumference>=80 &
Q36_WaistCircumference<88 & Gender=="F"
   replace waist01=3 if Q36_WaistCircumference>=102 &
Q36_WaistCircumference<112 & Gender=="M"|Q36_WaistCircumference>=88 &
Q36_WaistCircumference<98 & Gender=="F"
   replace waist01=4 if Q36_WaistCircumference>=112 &
Q36_WaistCircumference!=. & Gender=="M"|Q36_WaistCircumference>=98 &
Q36_WaistCircumference!=. & Gender=="F"

   /*Definition of diabetes (for the purpose of this study)
```

```
   The definition of diabetes can come either from a recorded high blood
sugar or a record of being on treatment for diabetes. So someone has diabetes
if
   •  their blood sugar in  Q34(?) > 11.1 mmol/L
   OR
   •  if they have answered 'yes' to Q 4,5 and 6
   •  and are using one of the drugs listed in questions 6a,b, or c.
   The person needs to have been told they have diabetes, AND have taken
medicine for it in the last two weeks AND the medicine should be one
   of those listed in Q6a (if it is an 'other drug' then Xavi and I will need
to make a decision about it).
   */
   gen Q6a_Glibenclamide_bin=(Q6a_Glibenclamide<=2)
   gen Q6a_Gliclazide_bin=(Q6a_Gliclazide<=2)
   gen Q6a_Metformin_bin=(Q6a_Metformin<=2)

   gen Q6_drugs_taken=
Q6a_Glibenclamide_bin+Q6a_Gliclazide_bin+Q6a_Metformin_bin

   gen diabetes=(Q35_GlucoseInBlood>11.1 &
Q35_GlucoseInBlood!=.)|(Q5_RaisedBloodSugar==1 & Q6_drugs_taken>=1)

   /*To calculate mean systolic blood pressure, the first systolic reading
will be discarded and then the mean
   systolic pressure will be calculated as the mean of the second and third
readings.
   The mean diastolic blood pressure and pulse will be similarly calculated.
   Participants will be grouped according to blood pressure level:
   4:Stage two plus hypertension: systolic >180 and/or diastolic >110
   3: Stage two hypertension: systolic > 160 and <180 and/or diastolic >100
and <110
   2: Stage one hypertension: systolic > 140 and <160 and/or diastolic >90
and <99
   1: Normal blood pressure systolic < 140 and diastolic < 90
   */
   gen systolic_bp_mean= 0.5*(Q28a_Systolic2+Q29a_Systolic3)
   gen diastolic_bp_mean= 0.5*(Q28b_Diastolic2+ Q29b_Diastolic3)
   gen heart_rate_mean= 0.5*(Q28c_HeartRate2+ Q29c_HeartRate3)

   gen bplevel=.
   replace bplevel=1 if systolic_bp_mean<140 & diastolic_bp_mean<90
   replace bplevel=2 if (systolic_bp_mean>=140 &  systolic_bp_mean<160) |
(diastolic_bp_mean>=90 & diastolic_bp_mean<100)
   replace bplevel=3 if (systolic_bp_mean>=160 &  systolic_bp_mean<180) |
(diastolic_bp_mean>=100 & diastolic_bp_mean<110)
   replace bplevel=4 if (systolic_bp_mean>=180 &  systolic_bp_mean!=.) |
(diastolic_bp_mean>=110 & diastolic_bp_mean!=.)
   label define bplabel 1 "Normal blood pressure" 2 "Stage One hypertension"
3 "Stage Two hypertension" 4 "Stage Two Plus hypertension"
   label values bplevel bplabel
   * Creation of the minor cvd risk factors variable
   gen
minor_cvd_riskfactors=waist+family_risk+dyslipedemia+smoking+AgeRiskGroup
   replace minor=. if Q0a_InterviewOutcome>2
   * Creation of Stroke, Heart Failure and Heart Attack variables
   gen stroke_ever=(Q7_EverHadStroke==1 & Q7_EverHadStroke!=. |
Q8_EverReceiveStrokeRx==1 & Q8_EverReceiveStrokeRx!=.)
   gen heart_failure=(Q10_HeartFailure==1 & Q10_HeartFailure!=.|
Q11_ReceivedHeartFailureRx==1 & Q11_ReceivedHeartFailureRx!=.)
   gen attack_heart=(Q16_HeartAttack==1 & Q16_HeartAttack~=.|
Q17_ReceivedHeartAttackRx==1 & Q17_ReceivedHeartAttackRx~=.)

   * Creation of the associated clinical factors variable
   gen associated_clinical_factors= stroke_ever+heart_failure+attack_heart
```

```
   replace associated=. if Q0a_InterviewOutcome>2

   *Table Two Analysis plan codes
   gen riskfactors_category=.
   replace riskfactors=1 if minor_cvd_riskfactors==0
   replace riskfactors=2 if minor_cvd_riskfactors>=1 &
minor_cvd_riskfactors<=2
   replace riskfactors=3  if minor_cvd_riskfactors>=3 &
minor_cvd_riskfactors!=.| diabetes==1
   replace riskfactors=4  if associated_clinical_factors>=1 &
associated_clinical_factors!=.
   label define risk 1 "No risk factors" 2 "One or two risk factors" 3 "Three
or more risk factors or diabetes" 4 "Associated clinical conditions"
   label values risk risk
   tab bplevel risk,


   set type double
   gen cvdnkateko=.
   replace cvdnkateko=0 if bplevel==1 | minor_cvd_riskfactors==0 & bplevel!=.
   replace cvdnkateko=1 if bplevel==1 & minor_cvd_riskfactors==0 &
associated_clinical_factors==0
   replace cvdnkateko=2 if (bplevel==2 & minor_cvd_riskfactors>=1) &
(bplevel==2 & minor_cvd_riskfactors<=2) ///
   |(bplevel==3 & minor_cvd_riskfactors==0)| (bplevel==3 &
minor_cvd_riskfactors>=1) & (bplevel==3 & minor_cvd_riskfactors<=2)
   replace cvdnkateko=3 if (bplevel==2 & minor_cvd_riskfactors>=3 &
minor_cvd_riskfactors!=.) ///
   |(bplevel==4 & minor_cvd_riskfactors==0)| (bplevel==3 &
minor_cvd_riskfactors>=3 & minor_cvd_riskfactors!=.)
   replace cvdnkateko=4 if (bplevel>=2 & bplevel!=. &
associated_clinical_factors>=1 & associated_clinical_factors!=.) ///
   |(bplevel==4 & minor_cvd_riskfactors>=1 & minor_cvd_riskfactors<=2)|
(bplevel==4 & minor_cvd_riskfactors>=3 & minor_cvd_riskfactors!=.)

   label var cvdnkateko "Levels of CVD Risk Modified SA2015 Nkateko"
   label define cvdnkateko 0 "No risk" 1 "Low risk" 2 "Moderate risk" 3 "High
risk" 4 "Very high risk"
   label values cvdnkateko cvdnkateko
```