Development and external validation of a logistic regression derived algorithm to estimate a 12-month open defecation free slippage risk

Warren Mukelabai Simangolwa[1*]

[1] Independent Health Economist
Lusaka, Zambia
Email: mukewarren@gmail.com
[*]Corresponding author

Abstract: *Appropriate open defecation free (ODF) sustainability interventions are key to further mobilise communities to consume sanitation and hygiene products and services that enhance household's quality of life and embed household behavioural change for heathier communities. This study aims to develop a logistic regression derived risk algorithm to estimate a 12-month ODF slippage risk and externally validate the model in an independent data set. ODF slippage occurs when one or more toilet adequacy parameters are no longer present for one or more toilets in a community. Data in the Zambia district health information software for water sanitation and hygiene management information system for Chungu and Chabula chiefdoms was used for the study. The data was retrieved from the date of chief Chungu and Chabula chiefdoms' attainment of ODF status in October 2016 for 12 months until September 2017 for the development and validation data sets respectively. Data was assumed to be missing completely at random and the complete case analysis approach was used. The events per variables were satisfactory for both the development and validation data sets. Multivariable regression with a backwards selection procedure was used to decide candidate predictor variables with* $p < 0.05$ *meriting inclusion. To correct for optimism, the study compared amount of heuristic shrinkage by comparing the model's apparent C-statistic to the C- statistic computed by nonparametric bootstrap resampling. In the resulting model, an increase in the covariates 'months after ODF attainment', 'village population' and 'latrine built after CLTS', were all associated with a higher probability of ODF slippage. Conversely, an increase in the covariate 'presence of a handwashing station with soap', was associated with reduced probability of ODF slippage. The predictive performance of the model was improved by the heuristic shrinkage factor of 0.988. The external validation confirmed good prediction performance with an area under the receiver operating characteristic curve of 0.85 and no significant lack of fit (Hosmer-Lemeshow test:* $p = 0.246$*). The results must be interpreted with caution in regions where the ODF definitions, culture and other factors are different from those asserted in the study.*

Key words: Chiefdom, CLTS, DHIS2, Prognostic model, ODF slippage risk, ODF sustainability,

BACKGROUND

Achieving and sustaining open defecation free (ODF) status has increasingly become a shared goal for communities, interventionists, implementing organisations and governments. Poor access to sanitation and hygiene negatively impacts progress on agreed international targets on health, poverty and human dignity (Roche, et al., 2017; Hutton & Chase, 2016). In 2015, only 39% of the global population used safely managed sanitation (two of every five people lived in rural areas) and 892 million people worldwide still practice open defecation (World Health Organization , 2017). The diseases associated with poor sanitation account for about 10% of the global burden of disease (Prüss-Üstün, et al., 2008; McGinnis, et al., 2017). These include diarrhoeal diseases, acute respiratory infections, undernutrition and other tropical diseases such as helminth and schistosomiasis infection (Van Minh & Hung, 2011; Araujo Navas, et al., 2016). Diarrhoea alone accounted for 19% of the deaths in children under the age of five in Sub Saharan Africa (Mara, et al., 2010). 88% of these cases are attributable to unsafe water, inadequate sanitation, and poor hygiene (Roche, et al., 2017). Lack of access to safe sanitation costed the global economy US$222.9 billion in 2015, with associated costs linked to mortality, productivity and healthcare (Lixil, WaterAid Japan & Oxford Economics, 2016). The global economic return on sanitation spending is US$ 5.5 per US dollar invested (Hutton, 2012).

Appropriate ODF sustainability interventions are key to further mobilise communities to consume sanitation and hygiene products and services that enhance a household's overall quality of life and embed household behavioural change for healthier communities. ODF slippage occurs when one or more toilet adequacy parameters are no longer present for one or more toilets in a community (Galan, et al., 2013; Njuguna & Muruka, 2017). It is increasingly recognised that maintaining ODF status is challenging due to uncertainties in social cohesion, implementing agencies and government priorities, sustainability of toilet and handwashing technologies, sanitation financing, governance, monitoring and sanitation markets (Bongartz, 2016; Odagiri, et al., 2017). Faced with the challenge of ODF slippage, there is limited scientific evidence to guide a systematic approach on mitigating ODF slippage risk factors and covariates. ODF slippage risk prediction can be useful in guiding interventionists and communities in slippage prediction and subsequently cost-effective and timely mitigation. This is especially important as each unique combination of predicators in a prediction model can be used to estimate the probability that accommodates the stratification of risks for villages (Hendriksen, et al., 2013). A risk score is a standardised metric for the likelihood that a variable of interest will experience an outcome (Royston, et al., 2009). ODF slippage risk prediction modelling is essential to identify and provide appropriate interventions to ODF communities at high risk of slippage.

This paper aims to develop a simple systematic tool to identify villages at high risk of ODF slippage. It develops and externally validates a prognostic model to estimate 12-month slippage risk for a chiefdom. Equipped with this information, decision-makers can more wisely prioritise and allocate scarce human, financial, logistical and other associated resources for ODF sustainability interventions.

Materials and methods

Source of Data

In Zambia, for a chiefdom to be declared ODF, it must be verified and certified that all the villages have households with an adequate toilet (Zimba, et al., 2016). The four parameters governing adequacy are (1) smooth cleanable floor, (2) toilet orifice lid (or fly trap for VIP's), (3) super structure providing privacy and (4) a hand washing station with soap. Retrospective longitudinal cohort data (Riley, et al., 2013) was extracted from the district health information software for water sanitation and hygiene management information system (DHIS2 WASH MIS) registry data for both the development and validation data sets.

DHIS2 is a free and open-source framework for management of aggregated health information (Manoj, 2013). The DHIS2 is used for collection, validation, analysis, and presentation of aggregate statistical data, tailored to integrated health management activities (Asangansi, 2012). In Zambia, the DHIS2 serves as the national database platform for the Ministry of Local Government and Housing (MLGH) (Biemba, et al., 2017). The DHIS2 WASH MIS is a sanitation mobile surveillance real time monitoring tool that is used for sanitation and hygiene monitoring in Zambia (Markle, et al., 2017).

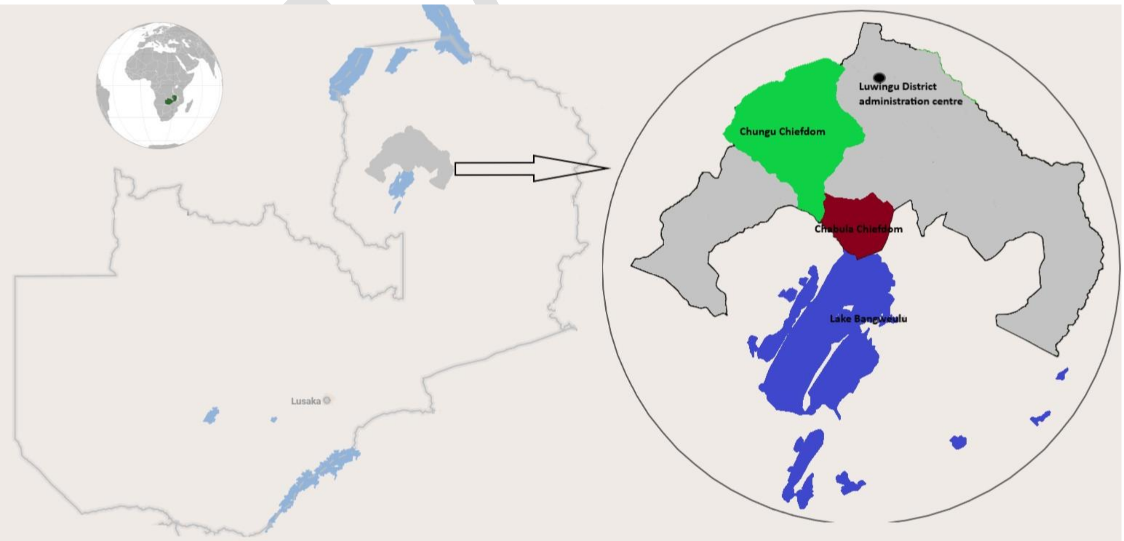
Data in the DHIS2 WASH MIS is entered from a java supported phone by community volunteers tasked to provide monthly data monitoring support to an average of 10 villages. The community volunteer amalgamates data collected from each village's sanitation action group (SAG). The village SAG is responsible for collecting paper-based household level data on parameters related to toilets. They use the water sanitation and hygiene, sanitation action group (WASH SAG) data collection form. Each community champion visits their assigned villages of supervision, during the period from first to the tenth of each month. They are expected to collect and submit aggregated data from each village's WASH SAG data collection form for the previous month by the tenth of each month. Before the community champion aggregates data and submits through their mobile phone, they randomly pick three households on the WASH SAG form for on spot verification. A technocrat, a ward government line agencies extension officer, conducts a meeting with a champion to assess the quality of the data before submission. Variables from WASH SAG data collection form are: village name, total number of households in a village, village population, and number of toilets triggered before and after CLTS, number of toilets with a lid, handwashing stations, and super structure providing privacy, smooth cleanable floor and total village population.

Interventions and village selection

Data in the DHIS2 WASH MIS for Chief Chungu and Chief Chabula were considered for development and validation of the study. Chungu and Chabula chiefdoms are two of the five chiefdoms in Luwingu district of the Northern Province of Zambia for the Bisa speaking people. Sanitation and hygiene interventions were introduced in the two chiefdoms in August 2014 through the local district council with support from its sanitation and hygiene partners. The Community-Led Total Sanitation (CLTS) approach was used as a sanitation demand creation intervention in both Chungu and Chabula chiefdoms. CLTS is designed to mobilise individuals in action to eliminate open defecation as a whole community (Harter, et al., 2018). It is a participatory approach in which facilitators visit villages and trigger awareness of sanitation and subsequently perform follow-up visits to villages to generate a community-wide effort to become ODF (Crocker, et al., 2017). To build capacity of CLTS facilitators to implement district-wide activities in Luwingu district, CLTS district and sub-district level cadres were trained and periodically followed upon in the CLTS methodology through the District Water Sanitation and Hygiene Education Committee (DWASHE) with the assistance of the CLTS national coaches. The DWASHE is a multi-stakeholder representative body for all government line agencies, donors, local nongovernmental organisations (NGOs), civil society organisations and the private sector involved in the governance and implementation of sanitation interventions at district level (Lungu & Harvey, 2009; Kanyamuna, 2010). To steer quality and effectiveness in implementation targets, strategies, standards, norms and approaches were aligned through the development and use of the Luwingu district total sanitation plan 2014-2017 through a multi-stakeholder process. All trained cadres at district and sub-district level were trained in the DHIS2 WASH MIS and provided with java supported phones, WASH SAG data forms and bicycles to enhance effectiveness in reporting. The baseline data in 2014 revealed that 38% of the population in Luwingu district had access to sanitation whereas 0% had a handwashing station with soap (SNV, 2018).

Chief Chungu and Chabula's chiefdoms were purposely sampled on merit of their ODF attainment on 20 October 2016 (Mutyoka & Makombo, 2016; Kachemba, 2016). Chungu chiefdom is led by her royal highness Chieftainess Chungu and is located 58 kilometres from the Luwingu district administration. The Chiefdom has a population of 29,840 people across 220 villages; 34 villages in Kampemba, 42 villages in Kafinsa, 38 villages in Ilambo, 52 villages in Mufili and 54 villages in Mulalashi wards (Central Stattistics Office , 2010). Villages are organised, small local communities of people in daily face-to-face interaction (Drennan & Peterson, 2006). The economy for Chungu chiefdom is agricultural with maize as the major cash crop. Chabula chiefdom is led by chief Chabula and is located 85 kilometres from the Luwingu district administration. The chiefdom has a population of 14,112 people across 117 villages; 79 villages in bwalinde and 33 villages in Ibale wards (Central Stattistics Office , 2010). Chief Chabula chiefdom's economy is a mix of agriculture and fish farming. The chiefdom touches Lake Bangweulu at Nsombo, a principal town at the northern part of the lake, which supports fish farming (Wikipedia contributors, 2018). Figure 2 shows the location of Chief Chungu and Chief Chabula chiefdoms.

Figure 1 Location of Chungu and Chabula Chiefdom's in Zambia



Village population data collected for the development model includes 67 villages of 220 villages assigned to ilambo, and Mufili wards in Chungu chiefdom. In the validation data of chief Chabula chiefdom, village population data for 56 villages of the 117 villages for Bwalinde and Ibale wards were used.

To be included for both the development and validation data sets, only the villages that had data imputed for at least 3 months of the 12 months (25%) were considered. There are no agreed estimates in literature on the amount of allowable missing data. A diverse view is held by researchers on an appropriate cut-off point for missing data but there have been suggestions as high as 20% (Chao-Ying Joanne, et al., 2006). Table 1 shows the data captured for the study.

Table 1 Village selection for the development and validation models

Chiefdom	Ward	Total number of villages	Villages with data	Villages without data	villages with >2 months data
Chungu	Kampemba	34	0	34	0
	Kanfisa	42	7	35	0
	Ilambo	38	37	1 (duplicate)	37
	Mufili	52	30	22 (3 duplicate)	30
	Mulalashi	54	15	39	0
Total		220	90	131	67
Chief Chabula	Bwalinde	79	67	12	23
	Ibale	33	33	0	33
Total		117	87	30	56

Outcome
In both the development and validation models, the outcome measured was ODF slippage. The variable ODF slippage in each village were derived by subtracting the number of monthly village level adequate toilets from the monthly number of households in the village and any none zero result was accounted for as slippage.

Predictors
The development model and validation data sets were cleaned to ensure that the number of households in a village were always greater or equal to the number of toilets that were adequate. Four cases with this inconsistency were retrieved and evaluated. The study investigated historical cases to ensure the data cleaning minimised the introduction of any new data. Champion and SAG meeting, latrine in use, latrine built after CLTS, latrine lids and latrine with smooth & cleanable floor data was extracted for each village. Other predicators include latrine privacy, number of latrines with handwashing with soap stations, number of households and total village population.

The Champion and SAG meeting reports on the monthly interaction the community volunteer has with SAG. The covariate ‘latrine in use’ describes the toilet infrastructure available in the village that are used as toilets whilst the variable ‘latrine built after CLTS’ assesses the number of latrines built after the CLTS intervention in each village. To ascertain vector transmission of faecal matter, the proxy of a latrine lid is used. In the case of a VIP, this is qualified by the provision of a functioning fly trap. The covariate, ‘latrine with smooth cleanable floor’ is aimed at ascertaining technologies to enhance use and maintenance of toilet facilities whilst ‘latrine privacy’ assesses acceptable latrine infrastructure (i.e. latrine walls, door etc.) that ensures dignity for users. Behavioural change is quantified by use of the proxy measure, ‘availability of a handwashing station with soap within 10 meters from a toilet’.

Sample Size
The study did not calculate a formal sample size but instead used all available data in the development and validation models from the DHIS2 WASH MIS data to maximise the power and generalisability of the results. There are no generally accepted approaches to estimate sample size in the derivation and validation of studies for risk prediction models. Some have suggested having at least 10 events per candidate variable (EPV) for the derivation of a model (Pavlou, et al., 2015). The EPV for both the development and validation models for this study were satisfactory (Austin, et al., 2017; Vittinghoff & McCulloch, 2007) and therefore expected to provide estimates that are robust.

Missing data

Challenges in internet, phones losing the DHIS2 WASH MIS application and failure to access villages during the rainy season were some of the unsolicited limitations to data capture and subsequent reporting. Consequently, 12% of the data was not captured for random villages in random months and is missing in the development model and 53% in the validation model. A village can be missing at one follow-up time and then measured again at one of the next, resulting in non-monotone missing data patterns. In this case, the probability of missing values was not related to the value of the observed responses and thus the data was assumed to be missing completely at random (MCAR) for both the development and validation models (Kang, 2013). In the development model however, 3% of the data was missing at random (MAR. In this case, the probability of data missing was dependent on the set of observed responses. It is established that an analysis restricted to study participants with complete data can be both biased and inefficient (Spratt, et al., 2010) and this is especially a threat to prognostic studies (Vergouw, et al., 2012). But however, it has been shown that multiple imputation methods have offered no statistical advantage over complete case analysis in some assessed scenarios (Mukaka, et al., 2016). The study used complete case analysis for data evaluation. The reduced data set for MCAR was shown to still produce valid estimates for statistical modelling as it represents a data set randomly drawn sub sample of the original data (Bennett, 2009).

Statistical Analysis
Multivariable regression with a backwards selection procedure was used to decide which of the candidate predictor variables should be included in the final prediction model, with $p < 0.05$ taken conservatively to warrant inclusion. In backward selection all the selected covariates are firstly entered at the same time into the model. Subsequently the variables with the highest p-values are removed based on of the Wald test (which allows you to calculate the significance level of a predictor). This step is repeated until there are no variables left with a p-value greater than 0.05 (Steyerberg, et al., 2001). After the backward stepwise regression was performed, the variables month, latrine built after CLTS, latrine with handwashing and latrine privacy remained in the development model. The variables, champion and SAG meeting, latrine in use, latrine lids and latrine with smooth & cleanable floor were excluded. However, the variable village population was added back in the model due to a very close p-value to $p < 0.05$ of 0.059. This inclusion is supported by literature (Hendriksen, et al., 2013).

Prognostic models in general appear to perform better in datasets used to develop the model than in new datasets. In this case, the regression coefficients and the measures of performance for the model are optimistic (Steyerberg, et al., 2004). To correct for the optimism, the study's development model parameter estimates were penalised by a shrinkage factor. To estimate the shrinkage in the model, the study used the heuristic shrinkage estimator of Van Howelingen and le Cessie (Van Houwelingen, 2001). In the heuristic shrinkage technique, the regression coefficients are multiplied by the heuristic shrinkage factor and the intercept is re-estimated (Pajouheshnia, et al., 2016) . The study's internal validation associated the amount of heuristic shrinkage by comparing the model's apparent C-statistic to the C-statistic computed by nonparametric bootstrap resampling. The bootstrap-resampled C-statistic corrects for regression towards the mean and overfitting (Steyerberg, et al., 2001). The internal bootstrap validation C-statistic for the study used the Harrell's nine steps on 200 bootstrap resamples of 688 cases each.

To record overall predictive performance in addition to discrimination, calibration was used. Discrimination is the ability of the risk score to differentiate between villages which do and do not experience an adverse event during the study period (Steyerberg, et al., 2010). This measure is quantified by calculating the area under the receiver operating characteristic (ROC) curve. A value of 0.5 represents chance, $0.7\leq$ ROC<0.8 represents acceptable discrimination, $0.8\leq$ ROC<0.9 represents excellent discrimination and 1 represents perfect discrimination (Hosmer & Lemeshow, 2000). Calibration is the ability to accurately assign the correct event probability at all levels of predicted risk (Crowson, et al., 2013). It measures the agreement between the observed and predicted risks. It is computed as the difference between the mean observed risk and the mean predicted risk. To measure calibration, the study used the Hosmer-Lemeshow goodness-of-fit test. In the Hosmer-Lemeshow goodness-of-fit test, model outputs are sorted into equally-sized groups, where the probabilities and true states in these groups are then checked by a χ^2 goodness-of-fit test (Dreiseitl & Osl, 2012). In each group, the expected number of events for ODF slippage in each group of villages was calculated as the sum of the predicted probabilities for the villages in that group. Whereas the observed number of events for ODF slippage was calculated as the sum of the number of events observed in that group (Crowson, et al., 2013). The chi-square statistic was assessed using k-2 degrees of freedom in the development model.

The study used the December 2016 revised version of Stata/IC 14.2 for windows to all its statistical analysis (StataCorp, 2015). The study followed the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement in its reporting (Collins, et al., 2015).

RESULTS AND DISCUSION

Study baseline data

A total of 803 cases were available for the development model and 676 cases for the validation data set. However, villages with some months that did not have any data in the DHIS2 WASH MIS amounting to 97 cases, were removed. Table 2 provides a summary of key study characteristics. The final development model had five covariates with an EPV=14 and the validation model had an EPV>13.

Table 2 Development and validation models key study characteristics

Characteristic	Development model: Chungu chiefdom (n =688)	Validation model: Chabula Chiefdom (n = 313)
Data collection period	October 2016 to September 2017	October 2016 to September 2017
Study design	Retrospective prognostic study	Retrospective prognostic study

Setting	Chief Chungu ODF chiefdom predominately agricultural economy	Chief Chabula ODF Chiefdom predominantly fish farming economy
Inclusion criteria	All villages reported in the DHIS2 WASH MIS which had two or more data points in the systems	All villages reported in the DHIS2 which had two or more data in the systems
Outcome	ODF or ODF slippage	ODF or ODF slippage
Included villages	67	56
ODF Status	688	313
ODF	618 (89.8%)	246 (78.59%)
ODF slippage	70 (10.2%)	67 (21.41%)

Internal validation
The ODF slippage risk model when internally validated had a high discrimination (apparent C-statistic=0.8114, 95% CI 0.75049 to 0.87240). A Hosmer-Lemeshow of 0.2377 confirmed no significant difference between observed and predicted ODF slippage. The mean for the observed risk seems to be accurately estimated by the mean for the predicted risk grouped by tenths of predicted risk.

Bootstrap validation
The difference in the true probabilities from the models prediction was at 95.9% and the percentage of fit due to noise was 4.12%. This overfitting was estimated by the heuristic shrinkage estimator. After bootstrapping with 200 resamples, the optimism-corrected C-statistic was 0.802. The predicted equation for the ODF slippage model was penalised to account for 1.2% overfitting using the Harrell method (Harrell , et al., 1996) with a heuristic shrinkage factor of 0.988.

External validation
The model from the development cohort was penalized for overfitting and applied to an external validation set. The discrimination of the model in the new data set was high (C-statistic=0.844, 95% CI 0.788 to 0.899) and the model was well calibrated with a Hosmer-Lemeshow of p=0.246 confirming no significant difference between observed and predicted ODF slippage.

The odds ratio of the final model in the prediction equation (equation 1) indicates that for each one-month increase when other covariates are held constant, the odds of ODF slippage increases by 27.7%, whereas a one person increase in the population for a community with all covariates held constant increases the risk of ODF slippage by 1%. Furthermore, for each latrine that is constructed as a result of a CLTS intervention, the odds of ODF slippage increase by 11.8% when all covariates are held constant. For each handwashing station with soap for a household, the odds of a community experiencing ODF slippage reduce by 20.2% whereas the odds increase by a factor of 4.7 for each increase in a toilet facility that has privacy, all things equal.

$$P = \frac{1}{1+e^{-x}} \quad (1)$$

where;

$x = -4.459\ (0.245t + 0.01n + 0.112\ell - 1.599h + 1.540v)$
P = ODF slippage (Adequate latrines)
t = time (months)
n = village population (people)
ℓ = number of latrines built after CLTS (latrines)
h = latrines with handwashing with soap facility (handwashing facility)
v = latrine privacy (latrine wall and door or suitable acceptable substitutes)

Examples 1 to 5 provide illustrations on the practical interpretation of these results.

Example 1: Month after ODF attainment

A village with a zero month of ODF status and a village population of 105 people (17 households) with 17 latrines built after CLTS, 17 households with handwashing with soap facilities and 17 households with latrines providing privacy, when presented with a risk of ODF slippage, would have the risk of slippage at 7%. Maintaining all prognostic factors constant, an increase in the number of months after ODF attainment to 6 months, increases the risk of ODF slippage to 26%. Whilst an increase to 12 months, the risk of ODF slippage increases to 60%.

Example 2: Village population

A village with 5 months ODF status having a village population of 60 people (10 households) with 10 latrines built after CLTS, 10 households with handwashing with soap facilities and 10 households with latrines providing privacy, when presented with a risk of ODF slippage, would have the risk of slippage at 11%. Maintaining all prognostic factors constant, an increase in village population to 120 (20 households), increases the risk of ODF slippage to 18%. Whilst a threefold increase to 180, will increase the ODF slippage to 28%.

Example 3: Latrines built after a CLTS intervention

A village with 5 months ODF status having a village population of 105 people (17 households) with 0 latrines built after CLTS, 17 households with handwashing with soap facilities and 17 households with latrines providing privacy, when presented with a risk of ODF slippage, would have the risk of slippage at 4%. Maintaining all prognostic factors constant, an increase in latrines built after a CLTS intervention by half (8 of 17 households), increases the risk of ODF slippage to 9%. When all the 17 households in the village have all their latrines built after a CLTS intervention, the risk of ODF slippage increases to 21%.

Example 4: Handwashing with soap facilities

A village with 5 months ODF status having a village population of 105 people (17 households) with 17 latrines built after CLTS, and none of the households with handwashing with soap facilities, whilst all the 17 households are with latrines providing privacy, when presented with a risk of ODF slippage, would have the risk of slippage at 100%. Maintaining all prognostic factors constant, an increase the households with latrines having handwashing with soap facilities by half (8 of 17 households), still maintains the risk of ODF slippage at 100%. When all the 17 households in the village have all their latrines with a handwashing with soap facility, the risk of ODF slippage reduces to 21%.

Example 5: Latrines providing privacy

A village with 5 months ODF status having a village population of 105 people (17 households) with 17 latrines built after CLTS, and 17 households with handwashing with soap facilities, whilst all none of the households are with latrines providing privacy, when presented with a risk of ODF slippage, would have the risk of slippage at several multiples of a 100%. Maintaining all prognostic factors constant, an increase the households with latrines providing privacy (8 of 17 households), still maintains the risk of ODF at several multiples of a 100%. When all the 17 households in the village have all their latrines providing privacy, the risk of ODF slippage reduces to 21%.

These results are aligned to literature on the influence of time after ODF attainment, population growth, CLTS interventions and toilet quality on ODF sustainability. ODF country sustainability evaluations have estimated ODF slippage rate to be 10% per year with a five years slippage return of up to 50% open defecation (Thomas, 2016; Tyndale-Biscoe, et al., 2013). Post ODF local and external community technical support has been identified as a significant factor in ODF sustainability (Tyndale-Biscoe, et al., 2013). These ODF sustainability enforcements are collaborative interactions of streamlined government line agencies, local natural leaders and chiefs (Balfour & Singh, 2015). However, there is also evidence suggesting that the ODF status of a community can still be sustained despite limited post ODF follow-up enforcement (Thomas, 2016).

This study also reinforces the findings from a study in Ghana and Ethiopia on sustainability of CLTS outcomes. The study asserts that CLTS is not an appropriate intervention in cases where the baseline toilet coverage is low and local toilet technologies are poor (Crocker, et al., 2017). Poor quality latrines can cause households to revert back to OD (Tyndale-Biscoe, et al., 2013; Mosler, et al., 2018). The CLTS methodology encourages communities to construct low cost, simple toilets leveraging on locally available materials (Mosler, et al., 2018). The implementation of CLTS are premised on the assumption that households will upgrade the initial simple low-cost toilet hardware to increasingly higher standards toilets (Khale & Ashok , 2008). Undeveloped sanitation supply chains and poor sanitation markets coupled by unstable soil conditions however, contributes to ODF slippage (Munkhondia, et al., 2018; Garn, et al., 2017). Poor sanitation markets further exacerbate ODF slippage on their influence to handwashing products and soap availability. A correlation in lack of handwashing with soap and ODF slippage was established at 8% (Shivanarain & Nancy, 2015). A cross sectional study to ascertain the association of ODF slippage and the strength of social norms in Indonesia for 587 households after a two-year ODF period estimated the slippage rate at 14.5% (Odagiri, et al., 2017).

Strengths

This study's strength is an exhaustive use of the easily accessible DHIS2 WASH MIS repository data. Furthermore, the external validation chiefdom's nomadic fish farming and peri-urban social economic setup in selected parts of communities, were distinct characteristics to that of the development model. The study further aligned its statistical plan and results to the TRIPOD statement to ensure for quality and standards. To ensure for transparency, the study used two vigorous and robust measures to correct the overfitting in the model; the heuristic shrinkage estimator and the Harrell method for bootstrapping. The final model was thus subject to correction for optimism.

Limitations

Bias of misrepresentation of data by the community cadres cannot be completely ruled out. Whilst the adverse events satisfied recommendations in literature of EPV>10, adverse events were small in the validation cohort than what is being advocated for in recent literature of at least a 100 adverse events (Collins, et al., 2015). Furthermore, recent literature has advocated for larger EPV values of between 20 and 50 (Austin & Steyerberg, 2014). Complete case analysis was undertaken in the presence of 3% MAR data. The study had a strict adherence to statistical rigor in the selection of study covariates. The relationship between covariates was not explored in this study.

The results can be considered when applying future interventions and the prioritisation aspects of service provision. However, cultural, geography, socio-economic and other factors may have a particular impact of particular predicators and should be considered in all applications of these results. Caution should be exercised when interpreting the results in contexts where the ODF definitions are deferent from those asserted by the paper.

Conclusion and implications

The study has developed and externally validated a novel population risk prediction algorithm that can predict a 12-month ODF slippage risk for communities on multiple risk factors available monthly through the DHIS2 WASH MIS. This prognostic tool represents a novel and yet simple approach to assessing the risk of ODF slippage that can be used to inform prioritization of interventions by the sanitation action groups at village level, the community champions and government extension officers, district officers, provisional officer, the national level and general implementing organisations. Future research should focus on using prospective data to develop and externally validate the ODF slippage prognostic tool in a larger EPV sample (EPV>20). Furthermore, a controlled qualitative study should be conducted to ascertain factors to explain the negative influence on ODF slippage of the increase in months after ODF, village population, quality of toilet infrastructure after CLTS and toilet privacy technologies.

Other Information

Supplementary Information

The web calculator for the risk algorithm model is accessible through the following link.

Funding

No funding was obtained to undertake the study.

Ethical considerations

Ethical approval was not necessary

References

Araujo Navas, A. L., Hamm, N. A., Soares Magalhães, R. J. & Stein, A., 2016. Mapping Soil Transmitted Helminths and Schistosomiasis under Uncertainty: A Systematic Review and Critical Appraisal of Evidence. *PLoS neglected tropical diseases,* 10(2), p. e0005208. doi:10.1371/journal.pntd.0005208.

Asangansi, I., 2012. Understanding HMIS Implementation in a Developing Country Ministry of Health Context - an Institutional Logics Perspective. *Online Journal of Public Health Informatics,* 19 December , 4(3), p. ojphi.v4i3.4302. http://doi.org/10.5210/ojphi.v4i3.4302.

Austin, P. C., Allignol, A. & Fine, J. P., 2017. The number of primary events per variable affects estimation of the subdistribution hazard competing risks model. *Journal of Clinical Epidemiology,* Volume 83, pp. 75-84. https://doi.org/10.1016/j.jclinepi.2016.11.017.

Austin, P. C. & Steyerberg, E. W., 2014. Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. *Statistical methods in medical research vol. 26,2 (2014): ,* 26(2), pp. 796-808. doi:10.1177/0962280214558972.

Balfour, N. & Singh, S., 2015. *Sustainability of ODF Practices in Kenya.* [Online]
Available at: https://www.unicef.org/esaro/UNICEF-FN-ODF-Sustainability.pdf
[Accessed 7 March 2018].

Bennett, D. A., 2009. How can I deal with missing data in my study?. *Australian and New Zealand Journal of Publich Health,* 25(5), pp. 464-469. https://doi.org/10.1111/j.1467-842X.2001.tb00294.x.

Biemba, G. et al., 2017. A Mobile-Based Community Health Management Information System for Community Health Workers and Their Supervisors in 2 Districts of Zambia. *Global Health: Science and Practice,* 5(3), pp. 486–494. http://doi.org/10.9745/GHSP-D-16-00275.

Bodner, T. E., 2008. What Improves with Increased Missing Data Imputations?. *A Multidisciplinary Journal,* 15(4), pp. 651-675. https://doi.org/10.1080/10705510802339072.

Bongartz, N. V. a. P., 2016. Going beyond open defecation free. In: N. V. a. J. F. (. P. Bongartz, ed. *Sustainable Sanitation for All: Experiences, Challenges, and Innovations.* Rugby: Practical Action, pp. 1-3. https://doi.org/10.3362/9781780449272.001.

Central Stattistics Office , 2010. *2010 census of population and housing.* [Online]
Available at:
https://www.zamstats.gov.zm/phocadownload/2010_Census/2010%20Census%20of%20Population%20National%20Analytical%20Report.pdf
[Accessed 19 December 2017].

Chao-Ying Joanne, P., Harwell, M., Liou, S.-M. & Ehman, L. H., 2006. Advances in missing data methods and implications for educational research. In: I. S. S. (Ed.), ed. *Real data analysis.* Greenwich: Information Age, p. 3178.
https://pdfs.semanticscholar.org/c557/71f7145072c84a21eefd6f0a04e65f61cd08.pdf.

Collins, G. S., Ogundimu,, E. O. & Altman , D. G., 2016. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. *Statistics in Medicine,* 35(2), p. 214–226. doi: 10.1002/sim.6787.

Collins, G. S., Ogundimu, E. O. & Altman, D. G., 2015. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. *Statistics in medicine,* 35(2), pp. 214-226. doi:10.1002/sim.6787.

Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G., 2015. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC medicine,* 13(1), pp. doi:10.1186/s12916-014-0241-z.

Crocker, J., Saywell, D. & Bartr, J., 2017. Sustainability of community-led total sanitation outcomes: Evidence from Ethiopia and Ghana. *International Journal of Hygiene and Environmental Health,* 220(3), p. 551–557. doi: 10.1016/j.ijheh.2017.02.011.

Crocker, J. et al., 2017. The true costs of participatory sanitation: Evidence from community-led total sanitation studies in Ghana and Ethiopia. *The Science of the Total Environment,* 601(602), pp. 1075-1083. http://doi.org/10.1016/j.scitotenv.2017.05.279.

Crowson, C. S., Atkinson, E. J. & Therneau, T. M., 2013. Assessing Calibration of Prognostic Risk Scores. *Statistical methods in medical research,* 24(4), pp. 1692-706. doi: 10.1177/0962280213497434.

Crowson, C. S., Atkinson, E. J. & Therneau, T. M., 2013. Assessing Calibration of Prognostic Risk Scores. *Statistical methods in medical research ,* 25(4), pp. 1692-706. doi: 10.1177/0962280213497434.

Dreiseitl, S. & Osl, M., 2012. Testing the Calibration of Classification Models from First Principles. *AMI... Annual Symposium proceedings. AMISymposium,* Volume 2012, pp. 164-9..

Drennan, R. D. & Peterson, C. E., 2006. Patterned variation in prehistoric chiefdoms. *Proceedings of the National Academy of Sciences of the United States of America,* 103(11), p. 3960–3967. http://doi.org/10.1073/pnas.0510862103 .

Galan, D. I., Kim, S.-S. & Graham, J. P., 2013. Exploring changes in open defecation prevalence in sub-Saharan Africa based on national level indices. *BMC Public Health,* 13(527), pp. doi:10.1186/1471-2458-13-527.

Garn, J. V. et al., 2017. The impact of sanitation interventions on latrine coverage and latrine use: A systematic review and meta-analysis. *International journal of hygiene and environmental health ,* 220(2 Pt B), pp. 329-340. doi:10.1016/j.ijheh.2016.10.001.

Harrell , F. E. J., Lee, K. L. & Mark, D. B., 1996. Tutorial in Biostatistics: Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine ,* 15(4), pp. 361-387. https://doi.org/10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4.

Harter, M., Mosch, S. & Mosler, H.-J., 2018. How does Community-Led Total Sanitation (CLTS) affect latrine ownership? A quantitative case study from Mozambique. *BMC Public Health,* 18(387), pp. doi:10.1186/s12889-018-5287-y.

Hendriksen, J., GJ, G., De, G. & Moons, K. G., 2013. Diagnostic and prognostic prediction models. *Journal of Thrombosis and Haemostasis,* 11(s1), p. 129–41. https://doi.org/10.1111/jth.12262.

Hosmer, D. W. & Lemeshow, S., 2000. *Applied Logistic Regression.* [Online]
Available at: https://www.researchgate.net/publication/261659875_Applied_Logistic_Regression
[Accessed 17 March 2018].

Hutton, G., 2012. Global costs and benefits of drinking-water supply and sanitation interventions to reach the MDG target and universal coverage. *Journal of Water and health,* 11(1), pp. 1-12. doi:10.2166/wh.2012.105.

Hutton, G. & Chase, C., 2016. The Knowledge Base for Achieving the Sustainable Development Goal Targets on Water Supply, Sanitation and Hygiene. *International Journal of Environmental Research and Public Health,* 13(6), p. 536. doi: 10.3390/ijerph13060536.

Jinks, R. C., 2012. *Sample Size for Multivariable Prognostic Models.* [Online]
Available at: http://discovery.ucl.ac.uk/1354112/1/PhD_master_doc%20to%20be%20submitted.pdf
[Accessed 3 March 2018].

Kachemba, N., 2016. ODF status elates Chabula chiefdom. *The Zambia Daily Mail*, 29 November, pp. http://www.daily-mail.co.zm/odf-status-elates-chabula-chiefdom/ .

Kang, H., 2013. The prevention and handling of the missing data. *Korean Journal of anesthesilogy,* May, 64(5), p. 402–406. doi: 10.4097/kjae.2013.64.5.402.

Kanyamuna, B. M., 2010. *The Impacct of Implimentating the D-WASHE Programmes in Chanyaya Community-Kafue District, Zambia: What role has National Water Policy (1994), played?.* [Online]
Available at:
http://dspace.unza.zm:8080/xmlui/bitstream/handle/123456789/1081/Kanyamuna.pdf?sequence=1&isAllowed=y
[Accessed 18 April 2018].

Khale, M. & Ashok , D., 2008. *The impact of rural sanitation on water quality and waterborne diseases.* [Online]
Available at:
http://www.communityledtotalsanitation.org/sites/communityledtotalsanitation.org/files/IHMP_Impact%20of%20Rural%20Sanitation.doc
[Accessed 17 June 2018].

Lixil, WaterAid Japan & Oxford Economics, 2016. *The true cost of poor sanitation.* [Online]
Available at: https://www.lixil.com/en/sustainability/pdf/the_true_cost_of_poor_sanitation_e.pdf
[Accessed 3 April 2018].

Lungu, C. & Harvey, P., 2009. *Multi-sectoral decentralized water and sanitation provision in Zambia: rhetoric and reality. IN: Shaw, R.J. (ed). Water, sanitation and hygiene - Sustainable development and multisectoral approaches: Proceedings of the 34th WEDC International Conference,.* [Online]
Available at: https://dspace.lboro.ac.uk/2134/28808
[Accessed 11 March 2018].

Manoj, S. M., 2013. Customising DHIS2 for Maternal and Child Health Information Management in Sri Lanka. *Sri Lanka Journal of Bio-Medical Informatics,* 3(2), pp. 47-54. doi: http://doi.org/10.4038/sljbmi.v3i2.2496.

Mara, D., Lane, J., Scott, B. & Trouba, D., 2010. Sanitation and Health. *PLoS Medicine,* 7(11), p. e1000363.doi:10.1371/journal.pmed.1000363.

Markle, L. et al., 2017. A Mobile Platform Enables Unprecedented Sanitation Uptake in Zambia. *PLoS Neglected Tropical Diseases,* 11(1), p. e0005131. http://doi.org/10.1371/journal.pntd.

McGinnis, S. M. et al., 2017. A Systematic Review: Costing and Financing of Water, Sanitation, and Hygiene (WASH) in Schools. *International Journal of Environmental Research and Public Health,* 14(4), p. 442. doi:10.3390/ijerph14040442.

Mosler, H.-J., Mosch, S. & Harter, M., 2018. Is Community-Led Total Sanitation connected to the rebuilding of latrines? Quantitative evidence from Mozambique. *PLoS One,* 13(5), p. e0197483. doi:10.1371/journal.pone.0197483.

Mukaka, M. et al., 2016. Is using multiple imputation better than complete case analysis for estimating a prevalence (risk) difference in randomized controlled trials when binary outcome observations are missing?. *Trials,* 17(341), pp. doi: 10.1186/s13063-016-1473-3.

Munkhondia, T., Simangolwa, W. M. & Maceda, A. Z., 2018. CLTS and sanitation marketing: aspects to consider for a better integrated approach. In: N. V. a. J. F. Petra Bongartz, ed. *Sustainable Sanitation for All Experiences, challenges, and innovations.* Rugby, UK: Practical Action Publishing Ltd, , pp. 100, 101. http://dx.doi.org/10.3362/9781780449272.

Mutyoka, M. & Makombo, Y., 2016. *Chabula and Chungu Chiefdom mass verification and certification report in Luwingu district.* [Online]
Available at:
http://41.77.4.165:6510/www.communityledtotalsanitation.org/sites/communityledtotalsanitation.org/files/LUWINGU%20MVC%20CC%20FINAL%20REPORT%20OCTOBER%202016%20.pdf
[Accessed 11 April 2018].

Njuguna, J. & Muruka, C., 2017. Open Defecation in Newly Created Kenyan Counties: A Situational Analysis. *Journal of health care for the poor and underserved,* 28(1), pp. 71-78. doi: 10.1353/hpu.2017.0009.

Odagiri, M. et al., 2017. Enabling Factors for Sustaining Open Defecation-Free Communities in Rural Indonesia: A Cross-Sectional Study. *International journal of environmental research and public health,* 14(1572), pp. 15-18. doi: 10.3390/ijerph14121572.

Pajouheshnia, R., Pestman, W. R., Teerenstra, S. & Groenwold, R. H., 2016. A computational approach to compare regression modelling strategies in prediction research. *BMC Medical Research Method,* 16(1), pp. 107. doi: 10.1186/s12874-016-0209-0.

Pavlou, M. et al., 2015. How to develop a more accurate risk prediction model when there are few events. *BMJ,* 351(h3868), p. doi: 10.1136/bmj.h3868.

Prüss-Üstün, A., Bos, R., Gore, F. & Bartram, J., 2008. *Safer water, better health: costs, benefits and sustainability of interventions to protect and promote health. World Health Organisation.* [Online]
Available at:
http://41.77.4.165:6510/whqlibdoc.who.int/publications/2008/9789241596435_eng.pdf
[Accessed 14 April 2018].

Riley, R. D. et al., 2013. Prognosis Research Strategy (PROGRESS) 2: Prognostic Factor Research. *PLoS Medicine,* 10(2), p. e1001380. http://doi.org/10.1371/jour.

Roche, R., Bain, R. & Cumming, O., 2017. A long way to go – Estimates of combined water, sanitation and hygiene coverage for 25 sub-Saharan African countries. *PLoS One,* 12(2), p. e0171783. doi: 10.1371/journal.pone.0171783.

Roche, R., Bain, R. & Cumming, O., 2017. A long way to go – Estimates of combined water, sanitation and hygiene coverage for 25 sub-Saharan African countries. *PLoS One,* 12(2), p. e0171783. doi:10.1371/journal.pone.0171783.

Royston, P., Moons, K. G., Altman, D. G. & Vergouwe, Y., 2009. Prognosis and prognostic research: developing a prognostic model. *BMJ,* 338(b604), p. https://doi.org/10.1136/bmj.b604.

Sharmani , B. et al., 2013. Impact of Indian Total Sanitation Campaign on Latrine Coverage and Use: A Cross-Sectional Study in Orissa Three Years following Programme Implementation. *PLoS ONE,* 8(8), p. e71438. doi:10.1371/journal.pone.0071438.

Shivanarain, S. & Nancy, B., 2015. *Sustainability of ODF Practices in Kenya.* [Online]
Available at: https://www.unicef.org/esaro/UNICEF-FN-ODF-Sustainability.pdf
[Accessed 12 March 2018].

Sinha, A. et al., 2017. Assessing patterns and determinants of latrine use in rural settings: A longitudinal study in Odisha, India. *International Journal of Hygiene and Environmental Health,* 220(5), p. 906–915. doi: 10.1016/j.ijheh.2017.05.004.

SNV, 2018. *Zambia – SSH4A Results Programme [Practice Brief].* [Online]
Available at:
http://41.77.4.164:6510/www.snv.org/public/cms/sites/default/files/explore/download/zambia-ssh4a-rp-epb-final-w.pdf
[Accessed 12 April 2018].

Spratt, M. et al., 2010. Strategies for Multiple Imputation in Longitudinal Studies. *American Journal of Epidemiology,* 172(4), p. doi: 10.1093/aje/kwq137.

StataCorp, 2015. *Stata Statistical Software.* Release 14 ed. College Station: TX: StataCorp LP.

Steyerberg, E. W., 2009. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating: By Ewout W. Steyerberg. *American Journal of Epidemiology,* 170(4), p. 528. https://doi.org/10.1093/aje/kwp129.

Steyerberg, E. W. et al., 2004. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Statistics in Medicine,* 23(16), pp. 2567-86. doi: 10.1002/sim.1844.

Steyerberg, E. W., Eijkemans, M. J., Harrell, F. E. E. J. & Habbema, D. J., 2001. Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. *Medical Decision Making,* 21(1), p. 45–56. doi: https://doi.org/10.1177/0272989X0102100106.

Steyerberg, E. W. et al., 2010. Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.),* 21(1), pp. 128-138. doi:10.1097/EDE.0b013e3181c30fb2.

Thomas, A., 2016. Strengthening post-ODF programming:reviewing lessons from sub-Saharan Africa. In: P. Bongartz, N. Vernon & J. Fox, eds. *Sustainable Sanitation for All Experiences, challenges, and innovations.* Rugby, UK: Practical Action Publishing <http://dx.doi.org/10.3362/9781780449272>, p. 84.

Tyndale-Biscoe, P., Bond , M. & Kidd, . R., 2013. *ODF Sustainability Study.* [Online]
Available at:
http://41.77.4.164:6510/www.communityledtotalsanitation.org/sites/communityledtotalsanitation.org/files/Plan_International_ODF_Sustainability_Study.pdf
[Accessed 11 March 2018].

UCLA, 2018. Multiple Imputation in Stata. *Institute for Digital Research and Education*, 13 Accessed in March, p. https://stats.idre.ucla.edu/stata/seminars/mi_in_stata_pt1_new/.

Van Houwelingen, H. C., 2001. Shrinkage and Penalized Likelihood as Methods to Improve Predictive Accuracy. *Statistica Neerlandica,* 21 December, 55(1), pp. 17-34. https://doi.org/10.1111/1467-9574.00154.

Van Minh, H. & Hung, N.-V., 2011. Economic Aspects of Sanitation in Developing Countries. *Environmental Health Insights,* Volume 5, p. 63–70. http://doi.org/10.4137/EHI.S8199.

Vergouw, D. et al., 2012. Missing Data and Imputation: A Practical Illustration in a Prognostic Study on Low Back Pain. *Journal of manupulative and therapeutics,* 35(6), p. 464–471. doi: 10.1016/j.jmpt.2012.07.002.

Virginia, R., de Albuquerque, C. & Heller, L., 2018. "The human rights to water and sanitation: challenges and implications for future priorities.". In: *In Equality in Water and Sanitation Services.* Routledge: s.n., pp. 42-61 .

Vittinghoff, E. & McCulloch, C. E., 2007. Relaxing the Rule of Ten Events per Variable in Logistic and Cox Regression. *American Journal of Epidemiology,* 165(6), p. 710–718. https://doi.org/10.1093/aje/kwk052.

WHO, 2009. *World Health Organization . Global health risks: mortality and burden of disease attributable to selected major risks.* [Online]
Available at:
https://www.who.int/healthinfo/global_burden_disease/GlobalHealthRisks_report_full.pdf
[Accessed 12 Febuary 2018].

Wikipedia contributors, 2018. *Lake Bangweulu.* [Online]
Available at: https://en.wikipedia.org/w/index.php?title=Lake_Bangweulu&oldid=838396671
[Accessed 25 May 2018].

World Health Organization , 2017. *"United Nations Children's Fund (UNICEF). Progress on Drinking Water, Sanitation and Hygiene: 2017 Update and Sdg Baselines.".* [Online]
Available at: http://41.77.4.165:6510/www.who.int/mediacentre/news/releases/2017/launch-version-report-jmp-water-sanitation-hygiene.pdf
[Accessed 12 Febuary 2018].

Zimba, R. et al., 2016. Chiengi District, Zambia Open Defecation Free After 1 Year of Community-Led Total Sanitation. *The American Journal of Tropical Medicine and Hygiene.,* 95(4), pp. 925-927. doi:10.4269/ajtmh.16-0210.