The Costs and Benefits of Alternative Approaches to Monitoring Patients on Antiretroviral Therapy: Modelling and Economic Analysis

A HIV Modelling Consortium report for the World Health Organization HIV Treatment Guidelines Development Committee

Paul Revill (University of York), Timothy Hallett (Imperial College London) & Andrew Phillips (University College London) for the HIV Modelling Consortium in collaboration with the working group on modelling of ART monitoring strategies in Sub-Saharan Africa

May 2015

Executive Summary

This study uses modeling and cost-effectiveness analysis to compare the costs and health outcomes associated with alternative patient monitoring strategies in light of the most recent data on clinical effectiveness, behavioural responses to monitoring strategies and costs. Results were reviewed by a panel of independent modellers and economists and refined through consultation with clinicians, diagnostics experts and programme managers from selected African countries. The consensus conclusions of the modelling group are as follows:

1a. If used to enable differentiated HIV care (whereby the frequency of clinic visits for patients stable on ART are reduced, with resulting cost savings), viral load monitoring is expected to be cost-effective even in the most resource-constrained settings.

1b. The costs of viral load testing and the savings in non-ART programme costs (as a result of differentiated care) are uncertain but crucial in determining whether routine viral load monitoring is cost-effective

2. Viral load monitoring using dried blood spots (DBS) is good enough and is expected to be a cost-effective way to deliver viral load monitoring (if used to enable differentiated HIV care).

3. In most cases it is expected to be cost-effective to monitor patients with viral load every 12 months. If differentiated care can be implemented with less frequent monitoring (e.g. every 24 months), and this does not result in adverse health outcomes, this would be cost-effective. Only in highly resourced healthcare systems would more frequent monitoring (e.g. every 6 months) be expected to be cost-effective.

4. There is little evidence of overall benefits associated with moving from a cut-off to define treatment failure of viral load counts > 1000 copies/ml towards either a lower or a higher cut-off.

5. When viral load monitoring is not considered feasible or is not cost-effective (e.g. if sufficient cost-savings in patient monitoring are not achievable), clinical monitoring with a confirmatory CD4 test appears a cost-effective interim approach. When current CD4 machines reach the end of their effective lifetimes, released resources can be used to transition to routine viral load monitoring to enable differentiated care.

6. If the choice remains between CD4 monitoring alternatives (before viral load testing is available but if CD4 measurement infrastructure is in place) use of a simplified switching criteria based only on the current CD4 count value is likely to lead to similar health gains and much lower costs than use of the existing WHO switch criteria.

Introduction

The 2013 World Health Organization (WHO) Consolidated Guidelines for HIV Treatment made the strong recommendation that all patients receiving antiretroviral therapy (ART) be routinely monitored using viral load monitoring (VLM) to determine the success of ART in suppressing HIV virus and to guide switching to second line drug regimens.[1] However, with only a few exceptions, the scale-up of VLM has been slow in the majority of low- and middle-income countries, particularly those in sub-Saharan Africa, under the greatest burdens from HIV.[2] Moreover, countries have adopted very different approaches to how to implement VLM. Important decisions relate particularly to the frequency of testing (the Guidelines recommend testing at 6 months following initiation and then annually, but some countries test less frequently e.g. every 2 years[3]); whether testing is done using plasma or dried blood spots (DBS); the cut-off to define treatment failure (usually defined as two consecutive VL counts >1000 copies/ml within 3 months, although lower or higher cut-offs have been considered); and the speed of scale-up in the context of other healthcare needs (e.g. should countries be investing in VLM if large coverage gaps for ART remain).

In 2014, the WHO released a Supplement to the 2013 Consolidated Guidelines with further recommendations on the provision of monitoring for patients on ART.[4] These included the withdrawal of CD4 testing for patients on ART when VLM is routinely available. They also highlighted that the way VLM is delivered depends importantly on programme context; that the choices of platform(s), decentralization of laboratory equipment, the sample type (plasma or DBS) and transport requirements, and even the degree of viral load scale-up itself all depends on resource availability (transport, laboratory and human; as well as budgetary) and the geographical make-up of ART delivery. In some cases, a 'hybrid' approach may be necessary whereby VLM is delivered differently in different parts of a country.

One of the major challenges in scaling up VLM has been its cost. Previous cost-effectiveness analyses have indicated routine VLM is unlikely to be cost-effective.[5-7] For instance, in a study to inform the 2013 Guidelines it was recommended that VLM is scaled-up only once close-to-full coverage of ART has been achieved.[7] However, recently the costs of VL testing and 2nd line ARV regimens have fallen substantially. Also, the possible role of VLM within overall ART programmes in supporting adherence[8] and enabling differentiated care for patients virologically suppressed and stable on ART[9] is now better understood. Such new insights warrant reinvestigation of the cost-effectiveness of monitoring approaches and how patient monitoring can be tailored based upon the features of ART programmes.

Background: Issues of Context

This study aims to inform investment decisions in patient monitoring alternatives in countries in a variety of different contexts and in ART programmes at different stages in the scale-up of VLM. Modelling and cost-effectiveness studies are inevitably imperfect abstractions of reality which, whilst providing valuable evidence, must be interpreted in respect of the assumptions made, the interpretations that are made of the available data and realities of differing programme contexts. Therefore it is essential that these factors be not only clearly stated but also that they have been subject to discussions with relevant stakeholders to ensure they are acceptable.

To this end, a workshop was held in Harare, Zimbabwe, in March 2015 where the modelling results were discussed with 42 delegates from a range of institutions (Ministries of Health/national laboratories; academic; international policy organizations and non-governmental organizations (NGOs)), including 19 directly involved in delivery of ART and patient monitoring in 5 countries (Zimbabwe, Malawi, Uganda, Kenya and South Africa). This workshop allowed for feedback on the modelling and cost-effectiveness work undertaken and for further discussion of issues faced by programmes in implementing patient monitoring. A report of the workshop proceedings ('Implementation issues for monitoring people on ART in low-income settings in sub-Saharan Africa') can be found in the Supplementary Appendix. The work presented subsequently reflects these discussions.

Methods and Approach

HIV and healthcare programmes in all countries face the common challenge of how to allocate limited available resources to generate greatest health gains for their populations. However, not all interventions that offer health benefits to patients can feasibly be funded and delivered. When determining how patients receiving antiretroviral therapy (ART) are monitored, programme managers and policy-makers need to ask whether the health outcomes associated with allocating resources to different monitoring alternatives (i.e. to clinical, CD4 and VL monitoring; and the different ways each can be delivered) are likely to exceed the health outcomes that would be generated if the alternatives are not implemented and instead resources are used to continue delivering other, or introduce new, HIV or healthcare interventions.

This study uses incremental cost-effectiveness analysis to assess whether the health benefits associated with a range of patient monitoring alternatives are large enough, given their costs, such that they can be deemed 'value for money'. Standard approaches to determining cost-effectiveness are used[10]. These are explained fully in Appendix 1.

The alternative patient monitoring strategies that are evaluated are shown in Table 1. Importantly, health benefits and costs associated with the alternatives need to be modelled over the long-term with downstream consequences (e.g. associated switching to 2nd line regimens and onward HIV transmission) associated with the policy also taken into account. The analyses are based upon modelling a 20-year time horizon with future costs and health outcomes discounted to present value using a discount rate of 3.5%.

A central assumption made in these analyses is that the choice of monitoring alternative affects the frequency with which patients are required to attend clinics. In many ART programmes in sub-Saharan African countries patients currently attend clinics every 2-3 months. This is much more frequent than in high income country settings. One possible reason is the lack of effective ART monitoring which would give assurance to healthcare providers that patients are sufficiently adherent to treatment and that significant drug resistance has not emerged. In this study it is assumed that only patients with a suppressed viral load (i.e. as indicated by a VL test result below the threshold used for monitoring in the last 12 months) are eligible for 'differentiated care' in which the frequency of clinic visits is reduced (e.g. to 6 monthly as compared to 2-3 monthly). Differentiated care is also expected to lead to lower non-ART programme costs (see below).

Monitoring Strategy	**What the monitoring strategy entails (for people on first line ART)**	**Failure criteria**
No monitoring	3 monthly visits to check on symptoms.	There is no possibility of patients switching to 2nd line ART
Clinical monitoring	3 monthly visits to check on symptoms.	WHO 4 condition diagnosed or 2 WHO 3 conditions diagnosed in 1 year.
Clinical monitoring, VL confirmation	3 monthly to check on symptoms. Measure viral load if WHO 4 condition diagnosed or 2 WHO 3 conditions diagnosed in 1 year.	VL > 1000 cps/mL
Clinical monitoring, CD4 count confirmation	3 monthly to check on symptoms. Measure CD4 count if WHO 4 condition	CD4 count <250 /mm^3.

	diagnosed or 2 WHO 3 conditions diagnosed in 1 year.	
CD4 count monitoring (WHO)	6 monthly CD4 count.	CD4 count < pre-ART baseline or CD4 count < 100 /mm^3
CD4 count monitoring (<200)	12 monthly CD4 count.	CD4 count < 200 after > 3 years on ART. CD4 < 100 /mm^3 after 1 > year on ART.
VL monitoring using DBS to enable differentiated care	VL measure at 6m, 12m and annually thereafter. If VL > 1000 then give adherence intervention and re-measure VL 3 months later.	VL >1000 cps/mL in confirmatory VL measure

Table 1. Evaluated monitoring strategies

Modelling the Costs and Outcomes of Alternative Monitoring Policies

The HIV Modelling Consortium involves a wide range of mathematical modelling groups and a number of them have contributed to this study. However, only one model was used to generate the results presented here - the *HIV Synthesis model* . This model is an individual-based stochastic simulation model that captures the benefits of patient monitoring to the population (i.e. through reduced HIV transmission) as well as the patient.[11, 12] It has generated results for a number of published studies on patient monitoring in resource limited settings.[7, 13, 14] Adherence, resistance to specific drugs and transmission of HIV and drug resistance; and the health consequences of these, are all explicitly modelled.

The costs of alternative strategies are estimated by multiplying the healthcare resource use expected to be incurred under each strategy (i.e. diagnostics tests delivered, number of clinic visits, use of 1st and 2nd line ART, additional healthcare use associated with WHO stage 3 and 4 events) with associated unit costs/prices. The unit costs/prices used in this study, as well as assumptions related to the performance of alternative monitoring tests and other programmatic factors with important effects on costs and expected health outcomes, are contained in Table 2.

Unit costs/prices - the assumed context is a sub-Saharan African country such as Zimbabwe	
Cost of VL DBS test ('fully loaded'*)	$22
Cost of C4 test ('fully loaded'*)	$10 [15]
Clinic visit costs without differentiated care	$80 per patient year on ART[16, 17]
Reduction in clinic visit costs with differentiated care (if VL<1000 in past year) (e.g. based on halving frequency of clinic visits from 2-3 monthly to 6 monthly)	$40 per patient year on ART
2nd line ART cost (atazanavir/r), inc. 20% supply chain cost	$288 per patient year on ART[18, 19]
Key programmatic and test performance assumptions	
Delay in ability to act on test result – VL DBS (due to time taken for blood to be sent to central location, analysed, the result be returned, and the patient attend the clinic).	3 months
Delay in ability to act on test result – clinical and CD4 (where results are available 'on the day to providers)	0 months
Probability the measure was done and result returned to the healthcare provider (VL and CD4 testing only)	0.85**
Probability of switch to second line per 3 months once failure definition met – VL DBS	0.5
Improvement in adherence if VL measured and >1000	28% chance of permanent increase. 42% chance of 6 month temporary increase.[8]

*Fully-loaded costs include all costs expected to be incurred in the delivery of the alternatives (e.g. machines, personnel, infrastructure, transport, facility management etc. in addition to test samples/reagents).

**If the measure was not done/the result returned to the healthcare provider it is assumed 25% of the cost is incurred.

Table 2. Key unit cost/prices and important programmatic and test performance assumptions

The health outcomes associated with the alternative monitoring strategies are summarised using the measure of disability-adjusted life years (DALYs)-averted - a composite measure capturing the extent to which alternatives reduce both

morbidity and premature mortality.[20] In these analyses, a year of life with asymptomatic HIV (i.e. in full health) is weighted as 0; whereas a year of life lost to premature death is weighted as 1. Disability weights are applied to time spent with the following HIV-related conditions: HIV drug toxicity (0.05), WHO stage 3 events (0.22), tuberculosis (0.40), WHO stage 4 events (0.54)[21]

Thresholds for Cost-Effectiveness

Costs and health outcomes are compared to determine the 'cost per DALY averted' (known as the incremental cost-effectiveness ration – ICER) associated with different monitoring alternatives. Determining cost-effectiveness requires comparing ICERs to a cost-effectiveness threshold (CET). The CET for a country represents the opportunity costs of resources required to fund the intervention, in terms of the health gains those resources could generate if used for alternative purposes in the healthcare system. As such, the threshold is not usually readily apparent but $500 is likely to be at the upper end of what is reasonable for the most resource-constrained countries in sub-Saharan Africa (e.g. Zimbabwe).[22] Where a country has a greater level of available resources, with close to full coverage of those in need of ART, a slightly higher CET may be appropriate (e.g. $700). In low and middle-income countries with even higher levels of resource availability and where healthcare needs of the population are widely met yet higher CETs may be appropriate (e.g. >$1000).

Key Findings

Findings are separated in what follows into (1) those results and accompanying conclusions which follow directly from the modelling/cost-effectiveness analyses undertaken; and (2) important observations and interpretations on the modelling results which also affect how they may be used for policy.

1. Modelling Conclusions

Conclusion 1(a): If used to enable differentiated HIV care viral load monitoring is expected to be cost-effective even in the most resource constrained settings.

If viral load monitoring enables differentiated care such that for those patients with suppressed viral load (i.e. below the threshold used for monitoring) the frequency of clinic visits can be reduced leading to lower non-ART costs (n.b. in the base case analysis it is assumed non-ARV programme costs fall from $80 to $40 per patient year on ART with suppressed viral load) then viral load monitoring using dried blood spots (DBS) is associated with an ICER of $326 per DALY-averted compared with no monitoring. If the cost-effectiveness threshold in a healthcare system is this amount ($326) or greater – which we expect to be likely in most ART programmes –

implementation of viral load monitoring is expected to lead to population health gains and hence be cost-effective compared to all other forms of monitoring. Hence, VLM in the context of differentiated care is expected to be a cost-effective approach to patient monitoring. However, we note that whilst a move to VLM is expected to be cost-effective, it may put additional costs on many ART programs that are currently implementing monitoring predominantly using clinical criteria.

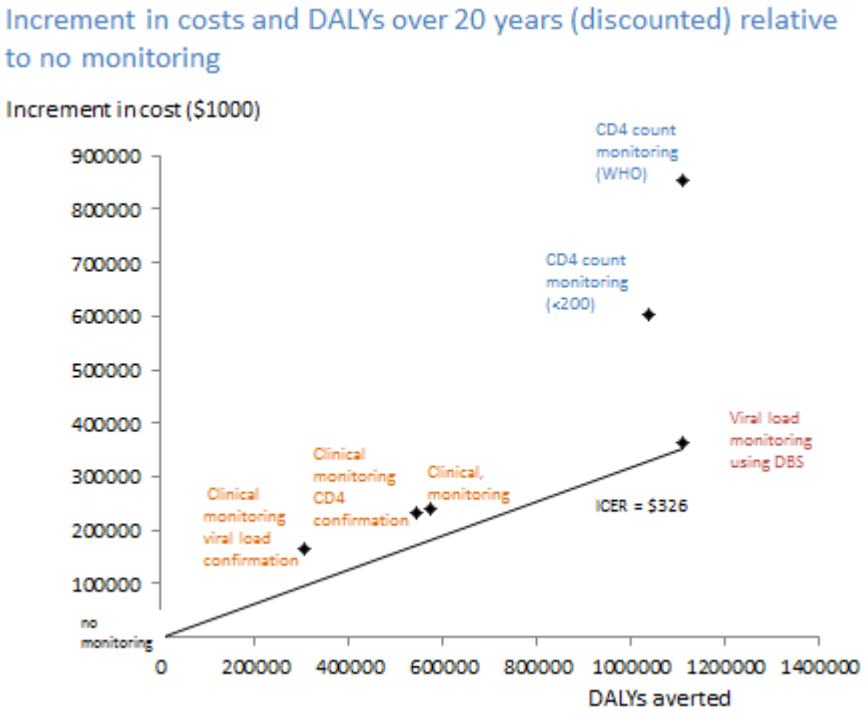

Figure 1. Cost-effectiveness of patient monitoring approaches

Conclusion 1(b): The costs of viral load testing and the savings in non-ART programme costs (as a result of differentiated care) are uncertain but crucial in determining whether routine viral load monitoring is cost-effective

The 'fully loaded' costs (including personnel, transport, infrastructure, machines; as well as samples/reagents) of viral load testing and especially the savings in non-ART programme costs enabled through viral load monitoring are uncertain and likely to differ both across and within HIV treatment programmes. To remain cost-effective compared to clinical monitoring in low resource settings the addition of viral load test costs need to be (almost entirely) offset by reduced costs of clinic visits over the year (see Figure 2). In order for VLM to be cost-effective at low viral load test costs (e.g. $12 per test; 'fully loaded'), reductions in the yearly costs of monitoring the patient must be reduced modestly ($10 per patient-year); but at higher costs (e.g. $28), the yearly-costs of monitoring patients must be reduced substantially ($40).

The extent of population health benefits generated depends upon providing viral load monitoring at test costs below and non-ART cost offsets exceeding these threshold combinations (i.e. so the ICER for VLM with differentiated care falls below the cost-effectiveness threshold). These combinations are therefore minimum requirements for cost-effectiveness.

Figure 2. Maximum viral load test cost-minimum non ART cost offset combinations required for viral load monitoring (with differentiated care) to remain cost-effective (green squares) in a healthcare setting with cost effectiveness threshold $500.

Conclusion 2: Viral load monitoring using dried blood spots (DBS) is good enough and is expected to be a cost-effective way to deliver viral load monitoring.
The use of a whole blood sample (i.e. DBS) instead of a plasma sample is not predicted to result in a marked difference in DALYs-averted, despite the greater variability in the results of DBS compared to plasma around the 1000 cps/mL threshold. The 3 month delay between taking a sample using DBS and that result being available to the healthcare provider, which is assumed in the model, has some impact on health outcomes but this is minor. Use of DBS is therefore expected to be a cost-effective approach to roll-out viral load monitoring, if used to effect the reductions in non-ART programme costs discussed above.

Conclusion 3: In most cases it is expected to be cost-effective to monitor patients with viral load every 12 months. If differentiated care can be implemented with less frequent monitoring (e.g. every 24 months), and this does not result in adverse health outcomes, this would be cost-effective. Only in highly resourced healthcare

systems would more frequent monitoring (e.g. every 6 months) be expected to be cost-effective.

If differentiated care can be implemented using viral load monitoring less frequently than every 12 months (e.g. every 24 months), and this does not lead to adverse health consequences, less frequent monitoring would be expected to be cost-effective. However, the downside health risks of differentiated care with infrequent viral load monitoring are not well understood. Further evidence on whether this approach is feasible, and the health consequences of its implementation, is required.

More frequent viral load monitoring (e.g. every 6 months) offers health gains but at much greater cost. Based upon current costs, it should only be considered an appropriate strategy in highly resourced healthcare systems (i.e. with cost-effectiveness thresholds >$1,234).

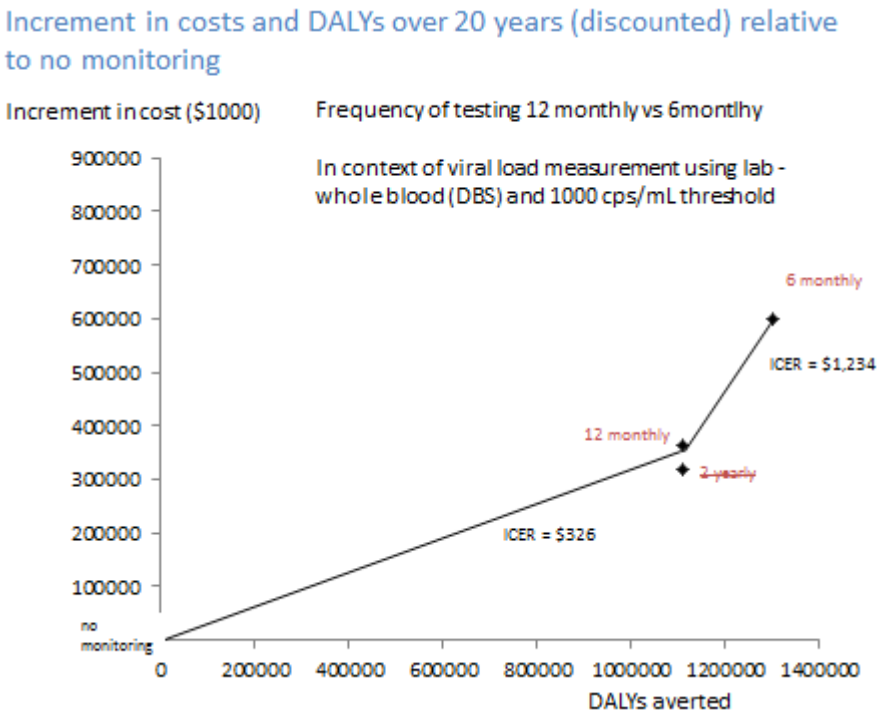

Figure 3: Cost effectiveness plane showing alternative frequencies of viral load testing.

Conclusion 4: There is little evidence of overall benefits associated with moving from a cut-off to define treatment failure of viral load counts > 1000 copies/ml towards either a lower or a higher cut-off.

In most cases it is expected to be cost-effective to monitor patients using a viral load cut-off to define failure of >1000 copies/ml. Based upon monitoring using plasma, a cut-off of 5000 copies/ml delivers health gains at a similar ICER as a 1000 cut-off (ICER=$311 compared to no monitoring/switching). A cut-off of 200 copies/ml (which would necessitate the use of plasma rather than DBS) leads to some health

gains but at disproportionately higher costs compared to a 1000 copies /ml cut-off. Assuming an approach to obtaining plasma-based testing were feasible, the use of the 200 cut-off may be an appropriate strategy in higher resourced settings with successful ART programmes (i.e. cost-effectiveness thresholds above $687), that can provide close to full coverage of 1st and 2nd line ART, but elsewhere it would not be cost-effective to do so.

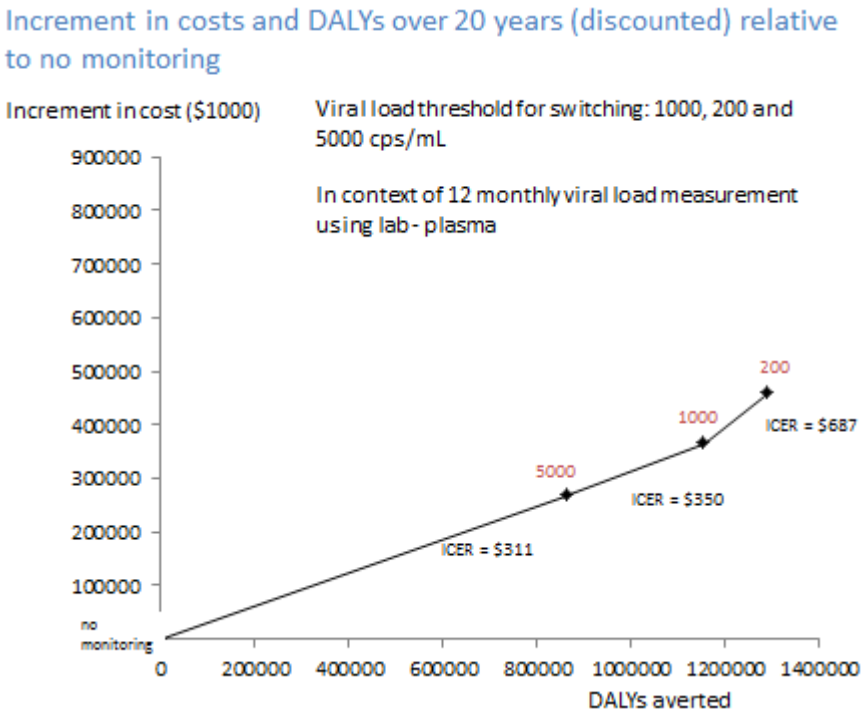

Figure 4: Cost effectiveness plane showing alternative viral load thresholds for switching.

Conclusion 5: When viral load monitoring is not considered feasible or is not cost-effective (e.g. if sufficient cost-savings in non-ART programme costs are not achievable), clinical monitoring with a confirmatory CD4 test appears a cost-effective interim approach. When current CD4 machines reach the end of their effective lifetimes, released resources can be used to transition to routine viral load monitoring to enable differentiated care.

When viral load monitoring is not considered a feasible alternative, or if the costs of viral load tests are too high and/or the reductions in clinic visit costs too low; then clinical monitoring with CD4 confirmation appears a cost-effective approach depending upon the resource environment (i.e. at cost-effectiveness thresholds above $414). It can be considered a reasonable interim approach.

With the possible exception of situations in which a comprehensive CD4 monitoring infrastructure is in place (so the marginal costs of testing are low) routine CD4

monitoring does not appear a cost-effective policy. This provides support for a policy of stopping new investment in routine CD4 monitoring capacity for people on ART when CD4 machines reach the end of their effective lifetimes.

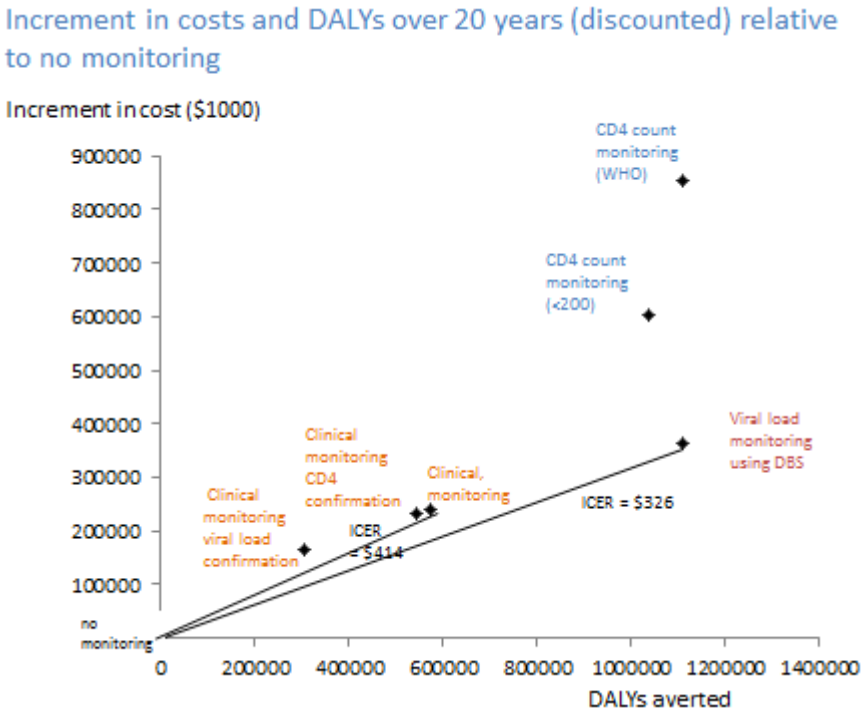

Figure 5: Cost-effectiveness plane showing the costs and effects of routine CD4 monitoring approaches compared to clinical and viral load monitoring alternatives

Conclusion 6: If the choice remains between CD4 monitoring alternatives (before viral load testing is available but CD4 measurement infrastructure is in place) use of a simplified switching criteria based only on the current CD4 count value is likely to lead to similar health gains and much lower costs than use of the existing WHO switch criteria.

In some settings the relevant policy choice remains between alternative CD4 monitoring alternatives (e.g. where CD4 infrastructure is in place and machines are available) until viral load testing is available. Switching based upon an alternative CD4 criteria of (i) first year on ART – no switching (ii) second and third years on ART - switch if CD4 < 100 (iii) after 3 years on ART switch if CD4 < 200; is much lower cost and delivers similar health gains than use of the current WHO switch criteria (so is cost-effective when only CD4 monitoring alternatives are considered). This supports the changing of the WHO CD4 switch criteria to a simpler clinical decision rule, which depends only on the current CD4 value.

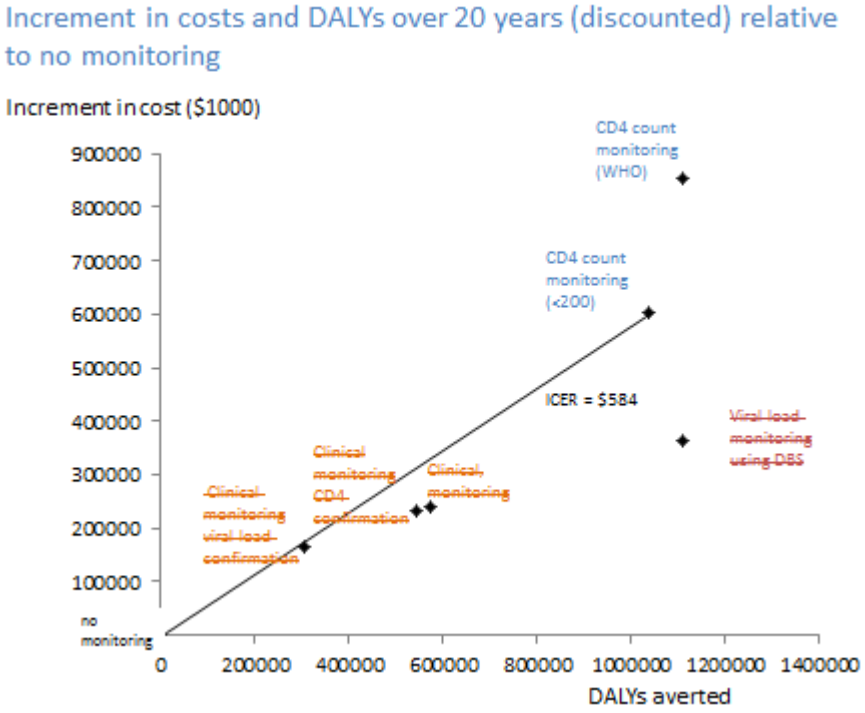

Figure 9: Cost-effectiveness plane when only CD4 monitoring alternatives are considered.

2. Issues of Context: Observations and Interpretations on the Modelling Results

The following table lists additional observations that modellers, economists and participants at the Harare workshop have highlighted as important considerations that follow from the modelling findings.

1.	It is recommended that the scale-up of viral load monitoring should be phased on the basis of further evidence of cost-savings realized through differentiated care. - Since the costs of viral load testing and the extent to which savings in non-ART costs can be realised are generally not well known further evidence on feasibility and the implications of reducing clinic visits would be valuable and is needed. Implementing routine VLM without accompanying cost savings risks imposing health losses on the population since resources could be better spent elsewhere. The costs of VLM and the attainable cost-reductions from reduced clinic visits are likely to differ by context and so evidence is likely to be required in the

	different ART programmes in which this strategy is implemented.
2.	The benefits of reduced clinic visits will depend upon how released resources will be used to generate health through other means. The analyses implicitly assume that cost reductions realized from reduced clinic visits generate health as reflected by the cost-effectiveness threshold (i.e. $500 reduced clinic visit costs avert 1 DALY when the CET of $500). The value of real resources (e.g. clinic and staff time) released may be greater or less than is reflected in long-run average costs and will depend upon how those budgetary/real resource savings are then used to improve patient health. This is likely to be context specific for which further evidence is also required.
3.	In practice, the cost of VL measurement will vary by location; and achieving the necessary cost threshold may imply different strategies. For instance: - Central, high density of PLHIV areas: centralised laboratory testing using DBS and a threshold for indication of 1000 is likely to be cost-effective for achieving access to VLM (although in some cases use of POC or plasma may provide additional advantages and be cost-effective). This will depend on country geography and other factors. - Remote, low density of PLHIV areas: reliable centralised lab based on DBS may not be possible at low cost (although should be possible in places where the DBS network for early infant diagnosis (EID) currently exists). POC VL devices at health centres (i.e. level II centres) may in the future become more widely available. Where viral load monitoring is not currently feasible in particular settings, Conclusions 5 and 6 above are likely to apply. - What factors most strongly determine these different situations would ideally be informed by costing, operational studies, and consideration of costs falling on patients under different situations. - A hybrid approach could present a potential issue of lack of equity (i.e. related to equality of access).
4.	From a clinical perspective, information on both VL and CD4 would be ideal; but the large cost of providing both makes it necessary in most settings to rely mainly on one diagnostic.
5.	The emphasis here is on DBS VL monitoring because this is the option that is feasible now and is expected to be the cheapest option and it delivers information of sufficient utility in patient

	monitoring. Confirmation with plasma VL is not necessary. But quality-control of DBS technology is certainly required.
6.	Cost and performance, and feasible level of application of POC, remain to be seen. Countries should evaluate their utility at those clinics where significant numbers of patients are on first line, where second line is available and where centralised laboratory DBS methods cannot reliably be delivered within the target cost envelope.
7.	Training of practitioners is key to the success of any monitoring strategy. The benefit of any monitoring strategy is entirely dependent upon it being executed correctly and experience indicates that a lack of training of practitioners can lead to a sub-optimal implementation of any monitoring policy. - There should be training to promote action to be taken when there is the indication to switch. Second-line would ideally be available at all clinics to maximize patient health outcomes.
8.	Data on patient preferences, clinician preferences, costs and competing demands on clinic staff time would further support the appropriate application of these recommendations to specific country program contexts.

Table 3: Key conclusions and contextual considerations highlighted as the Harare workshop on implementation issues related to monitoring patients on ART.

Conclusion

Routine 12-monthly viral load monitoring of patients on ART is expected to be cost-effective if used as a means to reduce the frequency and costs of clinic visits (i.e. to enable differentiated care). However, the costs at which viral load monitoring can be delivered and the cost reductions associated with differentiated care remain uncertain and are likely to be context specific. It is recommended that viral load monitoring is implemented in ART programmes in a phased manner on the basis of further supporting evidence on these test costs and clinic visit cost reductions.

References

1. WHO, *Consolidated guidelines on the use of antiretroviral drugs for treating and preventing HIV infection: recommendations for a public health approach*. 2013.
2. Markby, J., *Current Situation: Viral Load Testing*. 2015: Personal communication and presentation at the Implementation Issues for Monitoring People on ART in Low-Income Settings in Sub-Saharan Africa workshop, 11th March 2015, Harare, Zimbabwe.
3. Rutstein, S.E., et al., *Dried blood spots for viral load monitoring in Malawi: feasible and effective.* PLoS One, 2015. **10**(4): p. e0124748.
4. World Health Organization, *March 2014 supplement to the 2013 consolidated guidelines on the use of antiretroviral therapy drugs for treating and preventing HIV infection*. 2014.
5. Braithwaite, R.S., et al., *Alternative antiretroviral monitoring strategies for HIV-infected patients in east Africa: opportunities to save more lives?* J Int AIDS Soc, 2011. **14**: p. 38.
6. Walensky, R.P., et al., *Cost-effectiveness of laboratory monitoring in sub-Saharan Africa: a review of the current literature.* Clin Infect Dis, 2010. **51**(1): p. 85-92.
7. Keebler, D., Revill, P., Braithwaite, S.,, *Cost-effectiveness of different strategies to monitor adults on antiretroviral treatment: a combined analysis of three mathematical models.* Lancet Global Health, 2014. **2**(1).
8. Bonner, K., et al., *Viral load monitoring as a tool to reinforce adherence: a systematic review.* J Acquir Immune Defic Syndr, 2013. **64**(1): p. 74-8.
9. Duncombe, C., Rosenblum, S., Hellmann, N., Holmes, C., Wilkinson, L., Biot, M., Bygrave, H., Hoos, D., Garnett, G., *Reframing HIV care: putting people at the centre of antiretroviral delivery.* Tropical Medicine and International Health, 2015. **20**: p. 430-447.
10. Drummond, M., et al., *Methods for the Economic Evaluation of Health Care Programmes. 3rd Edn*. 2005: Oxford Medical Publications.
11. Phillips, A.N., et al., *Effect on transmission of HIV-1 resistance of timing of implementation of viral load monitoring to determine switches from first to second-line antiretroviral regimens in resource-limited settings.* AIDS, 2011. **25**(6): p. 843-50.
12. Cambiano, V., et al., *Transmission of drug resistant HIV and its potential impact on mortality and treatment outcomes in resource-limited settings.* J Infect Dis, 2013. **207 Suppl 2**: p. S57-62.
13. Phillips, A.N., et al., *Outcomes from monitoring of patients on antiretroviral therapy in resource-limited settings with viral load, CD4 cell count, or clinical observation alone: a computer simulation model.* Lancet, 2008. **371**(9622): p. 1443-51.
14. Phillips, A., et al., *Cost-effectiveness of HIV drug resistance testing to inform switching to second line antiretroviral therapy in low income settings.* PLoS One, 2014. **9**(10): p. e109148.
15. Hyle, E.P., et al., *The clinical and economic impact of point-of-care CD4 testing in mozambique and other resource-limited settings: a cost-effectiveness analysis.* PLoS Med, 2014. **11**(9): p. e1001725.
16. Tagar, E., et al., *Multi-country analysis of treatment costs for HIV/AIDS (MATCH): facility-level ART unit cost analysis in Ethiopia, Malawi, Rwanda, South Africa and Zambia.* PLoS One, 2014. **9**(11): p. e108304.

17. Siapka, M., et al., *Is there scope for cost savings and efficiency gains in HIV services? A systematic review of the evidence from low- and middle-income countries.* Bull World Health Organ, 2014. **92**(7): p. 499-511AD.
18. Bonner, K., *Personal communication with K. Bonner, Medicins sans Frontieres*. 2015.
19. Medicins San Frontiers, *Untangling the Web of antiretroviral price reductions: 17th edition - July 2014*. 2014.
20. Gold, M.R., D. Stevenson, and D.G. Fryback, *HALYS and QALYS and DALYS, Oh My: similarities and differences in summary measures of population Health.* Annu Rev Public Health, 2002. **23**: p. 115-34.
21. Salomon, J.A., et al., *Common values in assessing health outcomes from disease and injury: disability weights measurement study for the Global Burden of Disease Study 2010.* Lancet, 2012. **380**(9859): p. 2129-43.
22. Woods, B., Revill, P., Sculpher, M., Claxton, K., *Country-level cost-effectiveness thresholds: initial estimates and the need for further research*, in *CHE Research Paper 109*. 2015, Centre for Health Economics, University of York.
23. Culyer, A.J., *Hic sunt dracones: the future of health technology assessment--one economist's perspective.* Med Decis Making, 2012. **32**(1): p. E25-32.

Appendix 1:
Methods to Determine Cost-Effectiveness

To determine cost-effectiveness, the monitoring alternatives are first ranked on the basis of their effectiveness (i.e. from those leading to fewest DALYs incurred in the population to the highest DALYs incurred). Any strategies that are less effective and more costly than one (or a linear combination) of other alternatives are then not considered when estimating incremental cost-effectiveness. All remaining strategies are compared on the basis of the cost-per-DALY-averted (also known as the incremental cost-effectiveness ratio – ICER) compared to the next less effective, less costly alternative. This can be compared to the costs-per-DALY-averted (ICERs) associated with other claims on healthcare resources to determine which of the alternative monitoring policies represent value-for-money (this requires comparison to a cost-effectiveness threshold (CET); which represents the ICER(s) of such forgone interventions).

$$ICER = \frac{Incremental\ cost}{Number\ of\ DALYs\ averted} \leq CET$$

Policymakers should choose the monitoring policy that is expected to lead to greatest health gains (i.e. is most effective) as long as the ICER is less than ICERs associated with other HIV and healthcare interventions (i.e. other than monitoring alternatives) that can no longer be delivered as a result of resources being committed to the monitoring approach (the value of these foregone alternatives is represented by the CET). In this way, the monitoring policy would be expected to, not just improve the health of patients in receipt of monitoring, but also maximise health across the whole population; and can justifiably be deemed "cost-effective".

The set of strategies which could feasibly be cost-effective depending on the choice of CET can be shown on the cost-effectiveness plane. In the figures, those strategies which are cost-effective at different thresholds are shown linked by a solid line. This set of strategies is referred to as the cost-effectiveness frontier. As we move along the line past each strategy, the ICER rises and this would only be cost-effective at a higher CET than the previous strategy (as the incremental cost per DALY averted is higher than that associated with the previous strategy compared to its less effective, less costly comparator).

Further Considerations for Priority Setting

It should be noted that these analyses take the central policy challenge to be the allocation of a given healthcare budget to generate health gains in the population (i.e. a health system perspective is adopted with health as the only outcome). The analyses are intended to be inputs to deliberative processes of policy formulation and clearly other outcomes are also likely to be of concern.[23] For instance, equity (e.g. distributional fairness) and the extent to which different alternatives lead to costs (both direct and indirect, in the form of lost productivity) falling on patients and their caregivers are other criteria. It should be recognized, however, that any departure from healthcare policies expected to maximize population health by definition would be associated with some expected health losses.