The Cost and Cost-Effectiveness of Alternative Strategies to Expand Treatment to HIV-Positive South Africans: Scale Economies and Outreach Costs

Gesine Meyer-Rath, Mead Over, Daniel J. Klein, Anna Bershteyn

Abstract

The South African government is currently discussing various alternative approaches to the further expansion of antiretroviral treatment (ART) in public-sector facilities. Alternatives under consideration include the criteria under which a patient would be eligible for free care, the level of coverage with testing and care, how much of the care will be delivered in small facilities located closer to the patients, and how to assure linkage to care and subsequent adherence by ART patients.

We used the EMOD-HIV model to generate 12 epidemiological scenarios. The EMOD-HIV model is a model of HIV transmission which projects South African HIV incidence and prevalence and ARV treatment by age group for alternative combinations of treatment eligibility criteria and testing. We treat as sunk costs the projected future cost of one of these 12 scenarios, the baseline scenario characterizing South Africa's 2013 policy to treat people with CD4 counts less than 350. We compute the cost and benefits of the other 11 scenarios relative to this baseline. Starting with our own bottom-up cost analyses in South Africa, we separate outpatient cost into non-scale-dependent costs (drugs and laboratory tests) and scale-dependent cost (staff, space, equipment and overheads) and model the cost of production according to the expected future number and size of clinics. On the demand side, we include the cost of creating and sustaining the projected incremental demand for testing and treatment.

Previous research with EMOD-HIV has shown that more vigorous recruitment of patients with CD4 counts less than 350 appears to be an advantageous policy over a five-year horizon. Over 20 years, however, the model assumption that a person on treatment is 92 percent less infectious improves the cost-effectiveness of higher eligibility thresholds over more vigorous recruitment at the lower threshold of 350, averting HIV infections for between $1,700 and $2,800 (under our central assumptions), while more vigorous expansion under the current guidelines would cost more than $7,500 per incremental HIV infection averted.

Granular spatial models of demand and cost facilitate the optimal targeting of new facility construction and outreach services. Based on analysis of the sensitivity of the results to 1,728 alternative parameter combinations at each of four discount rates, we conclude that better knowledge of the behavioral elasticities would be valuable, reducing the uncertainty of cost estimates by a factor of 4 to 10.

JEL Codes: I15, I18

Center for Global Development
www.cgdev.org

Working Paper 401
April 2015

The Cost and Cost-Effectiveness of Alternative Strategies to Expand Treatment to HIV-Positive South Africans: Scale Economies and Outreach Costs

Gesine Meyer-Rath
Health Economics and Epidemiology Research Office, University of the Witwatersrand
Center for Global Health and Development, Boston University

Mead Over
Center for Global Development

Daniel J. Klein
Institute for Disease Modeling

Anna Bershteyn
Institute for Disease Modeling

The authors are grateful for the support of the HIV Modelling Consortium and the Bill and Melinda Gates Foundation. Tim Hallett and Ellen McRobie of the HIV Modelling Consortium have supported this project with great interest, patience, and understanding. Thanks goes to Rifaiyat Mahbub for help with the sensitivity analysis, to Catherine White for access to the clinic dataset, and to Kyle Wilson for help with preparing data for the clinic size calculations. We also thank reviewers from both Boston University and the Center for Global Development whose comments have improved the paper. Preliminary results of this analysis have been presented to staff at the South African National Treasury, the National AIDS Council Costing Task Team, and the National Department of Health, and the authors are grateful for all suggestions and comments received during these occasions.

CGD is grateful for contributions from the Bill and Melinda Gates Foundation in support of this work.

Gesine Meyer-Rath, Mead Over, Daniel J. Klein, and Anna Bershteyn. 2015. "The Cost and Cost-Effectiveness of Alternative Strategies to Expand Treatment to HIV-Positive South Africans: Scale Economies and Outreach Costs." CGD Working Paper 401. Washington, DC: Center for Global Development.
http://www.cgdev.org/publication/cost-and-cost-effectiveness-alternative-strategies-expand-treatment-hiv-positive-south

Center for Global Development
2055 L Street NW
Washington, DC 20036

202.416.4000
(f) 202.416.4050

www.cgdev.org

The views expressed in CGD Working Papers are those of the authors and should not be attributed to the board of directors or funders of the Center for Global Development.

Contents

Photo: Marlise Richter

Executive summary

Background

The South African government is considering alternative policies for scaling up publicly funded antiretroviral treatment (ART) for HIV/AIDS. Policies under discussion differ from previous policy in two dimensions: eligibility criterion and recruitment strategy. A number of analyses have considered the cost, cost effectiveness and/ or cost benefit of alternative eligibility criteria and recruitment strategies for a number of low- and middle income countries, including South Africa, but the detail with which cost was treated in these studies does not mirror the detail of the epidemiological projections, with authors often assuming the cost per patient-year of treatment would remain constant over time despite increases in the size of the treatment cohort by several hundred percent. In particular, economies of scale and the cost of generating the additional demand necessary for the assumed increase in recruited patients have not yet been not taken into account.

In South Africa HIV infected persons have been eligible for publicly funded ART if their CD4 count is less than a threshold of 350 cells per microliter or they test positive for tuberculosis or are children. Recruitment has been active, but has succeeded in linking to ART fewer than 70 percent of those eligible under these guidelines [34]. The baseline scenario, which we refer to as the "current guidelines eligibility, status quo recruitment" (CG.SQ) scenario, projects to 2033 the epidemiological and cost consequences of a continuation of recent policies. We treat the projected cost through 2033 of this CG.SQ scenario as a sunk cost, to which the government is unalterably committed. We investigate the impact of factors such as scale, level, and location of care on the additional cost of expanding ART treatment beyond current guidelines following eleven possible alternative policy scenarios.

Methods

Over a twenty-year projection horizon, 2014-2033, we compared the projected epidemiological consequences and facility-level costs of eleven policy scenarios to the "current guidelines, status quo" (CG.SQ) scenario described above. The eleven scenarios are combinations of four alternative eligibility criteria and two alternative recruitment strategies. Alternative eligibility strategies and their abbreviations are: A CD4 threshold of 500 or less (abbreviated as "500"), all HIV positive people (also called "universal test and treat" or "UTT"), HIV positive individuals with seronegative partners (called "discordant couples" or "DC") and HIV positive women who are pregnant (called "pregnant women" or "PW"). For the CG, the 500 and the UTT criteria, we explored both a "status quo" and a more ambitious "uniform expansion" (UE) recruitment strategy, which assumes increased testing and immediate ART initiation amongst 80% of the (eligible) population. For the DC and PW eligibility strategies, we additionally modeled an intermediate recruitment strategy called "prioritized expansion" (PE), which covers 80% of the targeted sub-population, while the rest of the population would continue to access testing and care at the "status quo".

Epidemiological data

Projections to 2033 of the number of HIV tests, HIV infections, and patients in HIV care were obtained from the EMOD-HIV model, a stochastic microsimulation model that includes incorporates reduced transmission from those on ART. The model has been calibrated to fit multiple sources of epidemiological data on the South African HIV epidemic (including prevalence by age, gender, and year; ART initiations by gender and year; CD4 counts at ART initiation; and testing by age and gender) The model provided outputs on the number of adults and children tested, in pre-ART care, on ART, in treatment failure and lost from care by CD4 cell count stratum (defined as >500, 350-500, 200-349, 100-199, and <100 cells/microliter, or corresponding pediatric CD4% and cell count values), and the number of new HIV infections for each year between 2014 and 2033.

Cost data

Data on the cost of testing and on the average outpatient and inpatient cost for an infected individual receiving no HIV care, pre-ART care, or ART came from our own bottom-up cost analyses of HIV-related care in South Africa. We separated outpatient cost into scale-independent costs (cost of drugs and laboratory tests, for which prices are mandated centrally for the entire public sector), and scale-dependent cost (staff, space, equipment and overheads) which we varied with the expected size of each clinic. For this we calculated the distribution of patients into clinics and into urban/ rural districts based on the distribution and size of ART clinics in June 2013 and assumptions about the likely growth in the size of each facility and in the number of clinics overall. Lastly, under each scenario except the baseline scenario of the current guidelines at current levels of testing, treatment uptake and retention, we also included the cost of demand creation for testing and of enabling improved retention for every additional patient who tested and initiated ART incremental to those patients in the baseline scenario in the analysis. For testing, we added the cost of a mobilization event per tested patient. In order to enable patients to present themselves for quarterly appointments at the ART clinic, we added an outreach cost per incremental patient which, at our assumed elasticities of demand, would be sufficient to attract the number of patients to that facility that are predicted by the given epidemiologic scenario. We calibrated these elasticities so that for a modest expansion the annual per-patient outreach cost would approximately equal the typical cost of four trips to a health center at ZAR 50 (USD 5) for an urban clinic and ZAR 30 (USD 3) for a rural clinic per single round trip. Outreach cost per patient can rise to a multiple of these benchmark values in high coverage scenarios or in high coverage clinics.

We then calculated cost-effectiveness (cost per infection averted) and cost utility (cost per disability-adjusted life year (DALY) averted). For the latter, we summed the total number of life years lived by HIV+ people in any type of care or health state, weighted by health-state specific disability weights, between mid-2014 and mid-2033 under each scenario and calculated the incremental number of DALYs of each scenario over the baseline scenario.

Results

Number of patients on ART

Under the current guidelines and trends in testing, linkage to care and losses to retention (CG.SQ), 2.4 million adults and 202,067 children are estimated to be on treatment by mid-2016, and 3.4 million adults and 135,424 children by 2033. If the current guidelines were kept, but testing, linkage to care and retention were improved to 80% each (CG.UE), these numbers would increase to 3.7 million adults and 236,471 children in 2016, and 5.3 million adults and 103,789 children in 2033. Under all other uniform expansion (UE) scenarios, there are more patients expected to be on treatment by mid-2016, and less by mid-2033, than under CG.UE; there is in fact an inverse relationship between the number of patients on treatment by mid-2016 and those on treatment by mid-2033 for all scenarios. This pattern of higher enrollment in early years followed by fewer patients in later years is a consequence of the epidemiological model's assumption that people on treatment are less likely to transmit HIV infection.

Development of undiscounted HIV-related cost over time

In our analysis, despite the incorporation of the prevention benefits of treatment, none of the scenarios achieves an annual cost less than the baseline of current guidelines with status quo recruitment before 2033, The cost trajectories in most scenarios flatten towards the end of the projection period, and the cost trajectories of most of the SQ scenarios (whose incremental cost over the baseline scenario per year is very small throughout) dip below the cost of the baseline scenario (CG.SQ) in the last years of the projection period, suggesting that the annual cost of any of these scenarios would eventually be lower than in the baseline and thus would eventually produce annual cost saving. The same is true for the cost of the PW.PE scenario.

Total HIV-related cost

At the discount rate of 3% and under our other central assumptions, the total cost over 20 years for the current guidelines at status quo (CG.SQ) is expected to be close to USD 36 billion, equivalent to an annual payment of US$2.4 billion. For reference, the discounted cost for the current mid-term expenditure framework of 2014-2016 is USD 3.6 billion, or about US$1.2 billion per year. The large contribution to cost of non-ART inpatient and outpatient services under the SQ scenarios is greatly reduced under the UE scenarios, but its place is filled by the large additional expenditure for testing. The two parts of ART costs we have assumed to be unrelated to the scale of a facility's work load, the cost of ARVs and labs, rise proportionately with the number of patients on treatment, as does inpatient cost for patients on ART. The component that rises disproportionately with the number of patients is the cost of outreach. Under our central assumption set, we set the elasticities of urban and rural demand to equal respectively 0.1 and 0.5. We assume cost of outreach to be zero in the baseline CG.SQ scenario and to barely appear at all under the other SQ or the PE scenarios. This is because none of these scenarios is projected to require a large percentage increase in patients. However, all five of the UE scenarios as well as the DC.PE scenario are expected to greatly increase their recruitment of patients.

Cost effectiveness

Improving testing, linkage to care and retention under the current eligibility in the CG.UE scenario would result in a total cost over 20 years of USD 14.6 billion and in a reduction of new infections by 44% to 1.9 million. This results in a cost per infection averted of USD 7,559, the highest cost per infection averted of all scenarios, including all other uniform expansion sub-scenarios.

The most cost-effective option in terms of cost per infection averted is to expand ART eligibility to pregnant women, either while maintaining the status quo for linkage and retention (Scenario PW.SQ costs $1,692 per infection averted) or while improving their linkage and retention and maintaining the status quo for everyone else (Scenario PW.PE costs $1,979 per infection averted). On the other hand, expanding eligibility for discordant couples, either as prioritised or as uniform expansion (DC.PE or DC.SQ), costs more than US$6,600 per HIV infection averted and is the least cost-effective option of all PE or SQ scenarios.

When maintaining the current status quo for testing, linkage to care and retention, expanding eligibility to discordant couples or pregnant women would have little effect on either cost or infections averted (although for the DC.SQ scenario, the increments in both cost and the number of infections averted is too small for the cost per infection averted result to be meaningful). This is because coverage of ART is already high in pregnant women compared to the general population, and current levels of partner testing are relatively low, so that few additional people would be reached by expanded guidelines unless effort is made to also encourage testing and linkage to care. Expansion of eligibility guidelines to include persons with a CD4 cell count of <500 cells/microl or all infected individuals would significantly increase both cost and infections averted, costing about $2,700 per infection averted under the SQ strategy or $5,283 per infection averted under the UE strategy. Universal test and treat guidelines would cost about 7% less per infection averted than would a <500 strategy.

Cost utility

The results of the cost-utility analysis mirror those of the cost-effectiveness analysis. As before, the most expensive and least cost-effective option, both in terms of cost per infection averted and by cost per DALY averted, is the uniform expansion of the current guidelines- every other uniform expansion scenario, including universal test and treat, becomes less expensive than the current guidelines over time, due to cost savings associated with the reduction in HIV transmission under high levels of population ART coverage.

Sensitivity analysis

We studied the sensitivity of results by first holding constant the central assumptions and considering each of several "expansion paths" from the baseline scenario to expanded eligibility and enhanced recruitment, and then by allowing all assumptions to vary over plausible ranges.

A remarkable feature of the expansion paths for several scenarios is their kinked shape. (An "expansion path" is defined as the sequence of cost and utility combinations which start from CG.SQ and proceed step-by-step to a) supplement current eligibility by including specific population X, where X = {500, UTT, DC, PW}, but with current patterns of testing and service uptake (scenario X.SQ); b) expanding testing and service provision to provide prioritized access for population X (scenario X.PE) (where available); and c) further expand testing and service provision to provide uniform access to 80% of the population (scenario X.UE).) Under the current eligibility guidelines (CG) and those expanded to include people with CD4 under 500 (500), all HIV positives (UTT) or pregnant women (PW). Initial expansion can be a good buy, but subsequent expansion to a UE scenario is always less cost-effective. We did a total of 13,824 computations of the model, 6,912 for each of two assumptions about the patients who would require outreach cost. We always assumed that outreach costs would be paid to all patients in excess of those who would seek care under the CG.SQ scenarios. For half the scenarios, we defined this excess for the facility as a whole; for the other, we calculated the increment between the two scenarios in the number of patients in each CD4 category, and applied outreach costs to those patients in health states with positive increments only. For each of these two assumptions about which patients would require outreach expenditure, we computed 12 scenarios x 4 discount rates x 3 scale elasticities x 4 distribution patterns x 12 elasticity of demand combinations. Across these combinations, half of the cost-effectiveness estimates lie between $3,000 and $8,000 per HIV infection averted, with extreme values as low as $4 and as high as $33,000 per infection averted. Most of the variation is driven by variation in the demand side parameters.

Our sensitivity analysis also shows that the distinctive concave kink in the expansion paths is robust to all tested variations in assumptions and more pronounced at lower elasticities of demand or when the number of facilities is not expanded to accommodate new patients.

Conclusions

We combined the outputs of an epidemiological and a cost model of the HIV epidemic in South Africa to calculate the incremental cost effectiveness of a range of strategies to expand eligibility beyond current guidelines. Previous research with EMOD-HIV has shown that more vigorous recruitment of patients with CD4 counts less than 350 appears to be an advantageous policy over a five-year horizon. Over 20 years, however, the model assumption that a person on treatment is 92% less infectious improves the cost-effectiveness of higher eligibility thresholds over more vigorous recruitment at the lower threshold of 350, averting HIV infections for between $1,700 and $2,800 (under our central assumptions), while more vigorous expansion under the current guidelines would cost more than $7,500 per incremental HIV infection averted.

Using the recent ART programme in South Africa as a baseline, we model the incremental cost per infection or DALY averted of each of 11 different policy alternatives for its expansion. In terms of total cost, all scenarios that maintain current trends in testing coverage, linkage to care, and retention in care ('status quo') have a very similar cost of around $36 billion over 20 years, while all scenarios that, for any given eligibility, assume uniform expansion of testing, linkage and retention for the entire population of eligible people ('uniform expansion'), have a total cost over 20 years of $50 billion USD. Within these, expanding eligibility to discordant couples (at current testing and linkage levels) and all pregnant women (at current and improved testing and linkage levels) are the least costly options, followed by expanding eligibility to all patients with CD4 cell counts < 500 cells/microl and Universal Testing and Treatment, both at the current level of testing and linkage.

The incremental cost per infection averted is between $1,600 and $2,700 for all 'status quo' recruitment scenarios, regardless of the eligibility criteria. Among these, the prioritisation of pregnant women is the most cost-effective scenario, though it has little overall impact because HIV testing rates are already high amongst pregnant women due to high coverage of antenatal care in South Africa. All 'uniform expansion' scenarios have both greater cost and greater impact, and are more costly per infection averted. The same pattern emerges in terms of cost per DALY averted, although the difference between the scenarios are exaggerated as a result of the CD4 cell count-dependency of the disability weights for people not on ART.

Expanding eligibility from the recent guideline requiring patients to have CD4 counts less than 350 to either a threshold of 500 or to include all HIV-infected people (UTT) does not increase the present value of the stream of HIV/AIDS costs through 2033. Furthermore, the health benefits of expansion are greater under either the 500 threshold or the UTT rule, so that under both of these wider eligibility alternatives cost per infection averted and cost per DALY averted are significantly less than under current eligibility rules. The uniform expansion of the current guidelines has the highest cost per outcome of all scenarios

Budgeting for universal coverage will require detailed country-specific estimates for the elasticities of supply and demand that we model here, and the calculation of not only numbers on treatment (as we have done here) but also prevalence and treatment need with far more granularity, possibly at the district or even clinic level. Furthermore, instead of subsuming all demand enhancing interventions into a single outreach cost, a more realistic model of demand would distinguish among elasticities with respect to travel vouchers, distance from home to facility, provider attitudes as well as characteristics of the patient, such as their CD4 cell count, income and education. As knowledge accumulates about the influence of such indicators, census and survey data on the exact geospatial locations of HIV infection can be used to build granular spatial models of demand which would offer the possibility of optimising the targeting of new facility construction and outreach services. These tools would help governments plan their own policies and help them in the increasingly competitive and demanding process of preparing investment cases to compete successfully for donor support.

Introduction

The South African public-sector guidelines for adult ART are in the process of being revised. A central concern is the further expansion of ART eligibility for adults beyond the current thresholds (a CD4 cell count of <350 cells/microl or active TB disease). Among the options under debate are

1. the extension of eligibility to all adults with **CD4 cell counts <500 cells/microl**;

2. the provision of ART to the positive partner in **serodiscordant couples** regardless of CD4 cell count, together with a concomitant increase in couples testing;

3. the extension of eligibility to **all pregnant women** regardless of CD4 cell count;

4. the introduction of a **universal test and treat** policy with annual testing of the population and initiation on ART of every person identified as HIV-positive regardless of CD4 cell count.[1]

A number of analyses have considered the cost, cost effectiveness and/ or cost benefit of some or all of these options for a number of low- and middle income countries, including South Africa [4-10]. A recent paper reviewed all modeled analyses of the cost or cost-effectiveness of ART, including those assuming an impact of ART on HIV transmission, and found that the detail with which cost was treated did not mirror the detail of the epidemiological projections, with authors often using uniform per patient cost despite increases in the size of the treatment cohort by several hundred percent [11].

South Africa was also included as one of four countries in a recent coordinated analysis involving twelve different HIV transmission models to inform the 2013 WHO guidelines on earlier treatment initiation [12]. Here again, cost was differed by type of regimen and health state, but factors such as economies of scale and the cost of generating the additional demand necessary for the assumed transmission impact to take hold were not taken into account.

In this analysis we treat the cost of continued care expansion through 2033 under current eligibility guidelines as a sunk cost, to which the government is unalterably committed. Under this assumption, we investigate the impact of factors such as scale, level, and location of care on the additional cost of expanding ART treatment beyond current guidelines following several possible alternative policy scenarios. For this, we apply synthesized evidence about variability of the cost of HIV testing and ART provision at different scales and levels of care, and model the additional cost of demand generation. Furthermore, we add the cost and impact of paediatric ART provision. In order to ensure comparability, we follow the methodology of the 12-model analysis as much as possible, especially in the calculation of the number of patients in care, survival, and cost-utility [12].

Methods

We compared each of the expansion options for adult ART eligibility to the current South African baseline of eligibility at a CD4 cell count of <350 cells/ microl or WHO status 3 (including TB). In order to answer questions regarding the cost effectiveness and cost utility of each of these options, we combined an existing model of the HIV epidemic in South Africa with output from an existing cost model used by the South African government for the calculation of the budget for the public-sector ART program over the last five financial years [13,14] as well as other recently published local cost data [15-17] and with disability weights from the recently updated Global Burden of Disease study [18].

[1] These options are based on the 2013 WHO guidelines [1], the results of the HPTN052 study [2], WHO PMTCT Option B+ [3] and a scenario introduced by Granich at el 2009 [4], respectively.

While this section aims to give an overview over methods, assumptions and data sources, more details on the epidemiological model can be found in Appendix A, and more details on the methods used in the economy evaluation, including mathematical derivations for each aspect, in Appendix B.

Epidemiological data

Data on the number of HIV tests and the number of patients in HIV care was obtained from the EMOD-HIV model, a stochastic microsimulation (patient-level) model developed by the Institute of Disease Modeling in Bellevue, WA, US. The model includes an impact of treatment on HIV transmission (see Appendix A for a detailed description of the model assumptions and functionality). The model has been calibrated to fit multiple sources of epidemiological data on the South African HIV epidemic (including prevalence by age, gender, and year; ART initiations by gender and year; CD4 counts at ART initiation; and testing by age and gender) and was included in the recent 12-model comparison exercise [12]. The EMOD-HIV model was altered for this analysis to include a representation of the HIV epidemic in children, including a cohort on treatment. Additionally, a tally of HIV-positive individuals lost from care at any stage (after HIV testing, during pre-ART care or after ART initiation) was added, with any patient not having linked to further care or returned for an ART visit within the last six months being counted as lost. The model provided outputs on the number of adults and children tested, in pre-ART care, on ART, in treatment failure and lost from care by CD4 cell count stratum (defined as >500, 350-500, 200-349, 100-199, and <100 cells/microl, or corresponding paediatric CD4% and cell count values), and the number of new HIV infections for each year between 2014 and 2033.

Cost data

We combined this information with data on the cost of testing and on the average outpatient and inpatient cost for an infected individual receiving no HIV care, pre-ART care, or ART from a number of sources (see Table 1). The outpatient cost of ART provision was divided into two segments, depending on the likely variation of each cost item with the size of the outpatient clinic.

Non-scale dependent cost

The cost of drugs and diagnostics was assumed not to vary by clinic size and was based on the average cost per adult or child in 2013/14 as projected by the National ART Cost Model [13,14]. This average cost was calculated by age group (1 adult and 3 child age groups) for four different types of care: treatment initiation, first-line treatment, first-line treatment failure, and second line. We included the cost of HIV counseling and testing of the entire tested population (not just HIV-positives or people on ART) as a weighted average across two different modalities, facility-based testing and mobile testing.

Table 1: Input cost parameters from South African National ART Cost Model

Type of care	Pre-ART	First-line ART	Second-line ART	Treatment initiation	Treatment failure	Source
Annual cost of outpatient care (cost of drugs and laboratory tests only)+						
Adults (>15 yrs)	53	79	416	79	10	**National ART Cost Model average for 2013/14 [13,14]**
Children <1 yr	53	145	321	91	11	
Children 1-2 yrs	53	70	316	129	11	
Children 3-5 yrs	53	131	203	98	11	
Children 6-14 yrs	53	182	302	73	11	
CD4 cell count* [cells/microl]	**>500**	**350-500**	**201-349**	**100-200**	**<100**	
Annual cost of inpatient care (urban)+						
Adults (>15 yrs), not on ART	- 57 -		80	128	188	[15]
Adults (>15 yrs), on ART	- 59 -		139	201	513	[15]
Children <1 yr, not on ART	-- 3,578 --					[16]
Children <1 yr, on ART	-- 1,660 --					[16]
Children 1-2 yrs, not on ART	- 45 -		71	123	123	**[15], adjusted for pediatric cost per patient-day equivalent based on [17]**
Children 1-2 yrs, on ART	- 47 -		124	193	335	
Children 3-5 yrs, not on ART	- 45 -		71	123	123	
Children 3-5 yrs, on ART	- 47 -		124	193	335	
Children 6-14 yrs, not on ART	- 45 -		71	123	123	
Children 6-14 yrs, on ART	- 47 -		124	193	335	
Annual cost of inpatient care (rural)+						
Adults (>15 yrs), not on ART	- 27 -		105	160	228	[15]
Adults (>15 yrs), on ART	- 94 -		239	388	732	[15]
Children <1 yr, not on ART	-- 3,578 --					[16]
Children <1 yr, on ART	-- 1,660 --					[16]
Children 1-2 yrs, not on ART	- 21 -		93	154	149	**[15], adjusted for cost per patient-day equivalent based on [17] and urban/ rural difference in adult inpatient cost**
Children 1-2 yrs, on ART	- 74 -		212	372	478	
Children 3-5 yrs, not on ART	- 21 -		93	154	149	
Children 3-5 yrs, on ART	- 74 -		212	372	478	
Children 6-14 yrs, not on ART	- 21 -		93	154	149	
Children 6-14 yrs, on ART	- 74 -		212	372	478	
Per-test cost of HIV counseling and testing+						
per adult tested (facility-based)			11			[19]
per adult tested (mobile testing)			15			Assumption
per child tested (facility-based)			43			[19]**
per child tested (mobile testing)			49			Assumption
Cost of mobilisation event			23			[19]

* or pediatric equivalent **plus cost of PCR test +All costs expressed in 2013 US dollars.

Scale-dependent cost

Unlike most projections of AIDS treatment costs, this study models the average cost of staff, space, equipment, and overheads as varying with the number of ART patients. Based on the actual size distribution of South African ART clinics in June 2013 [20] and alternative assumptions about the South African government's choice between increasing the utilization of current facilities and adding new ones (see Table 2), we projected the future distribution of patients into clinics and into urban/ rural districts. We assumed cost to vary with clinic level (primary healthcare clinics, community clinics, district or regional hospitals, tertiary hospitals) as well as with scale. In the base case analysis, the cost adjustment factor was 1.05 for clinics, and 0.79 for hospitals, based on data from a cost analysis of ART provision in 45 facilities in Zambia [21]. Based on a multi-clinic cost analysis of ART in a number of sub-Saharan African countries, including South Africa [22], we select a value of 0.8 as the central assumption for the scale elasticity and explore values of 0.6 and 1.0 in the sensitivity analysis. More details on the methods used in the estimation of scale-dependent cost is available in the section on "Expansion scenarios" further below.

Table 2: Cohort distribution assumptions

Parameter	Value	Source
Additional percentage of patients on first-line ART moving to second line each year	0.8%	National ART Cost Model average for 2013/14 to 2016/17 [13,14]
Percentage of people accessing facility-based testing (as opposed to mobile testing)	99.6%	District Health Information System data for June 2013 [19]
Distribution of pre-ART patients into levels of care		
- primary healthcare clinics	54%	
- community day/ health care centres	25%	
- district/ regional hospitals	17%	
- provincial tertiary hospitals/ national central hospitals	4%	

The future stream of costs is projected in constant 2013 USD. All cost data collected in ZAR are converted to USD at the current exchange rate of 1 USD = 9.89 ZAR.

For cost-effectiveness and cost-benefit calculations, both costs and outcomes are subsequently discounted to the present value at 3%. We explore the impact of alternative discount rates in sensitivity analyses, including a value of 5%, the current repurchase rate of the South African Reserve Bank [23].

Cost-effectiveness analysis

Using two alternative measures of effectiveness, we estimate the cost-effectiveness of each of the 11 scenarios in comparison to the current guidelines status quo (CG.SQ). For each calculated cost-effectiveness ratio, the numerator is defined as the present value of the future stream of costs under one of the 11 alternative policy scenarios minus the present value of the future stream of costs under the baseline scenario. The denominator of one set of cost-effectiveness measures is defined as the present value of the future stream of HIV infections under the baseline scenario minus the (presumably smaller) present value of the future stream of HIV infections under one of the 11 alternative policies. The cost-effectiveness ratios from this set of calculations yield the estimated cost per HIV infection averted.

For consistency with [12], the denominator of a second set of cost-effectiveness measures is defined as the present value of the future stream of disability weighted [18] (see Table 3) life years lived by HIV+ people under one of the 11 alternative policy scenarios minus the (presumably smaller) present value of disability weighted life years lived by HIV+ people under the baseline scenario. The cost-effectiveness ratios from this set of calculations yield the estimated cost per additional healthy life year and can be interpreted as the cost per DALY averted.

In all of these cost-effectiveness calculations, the same discount rate is used to compute the present values in the numerator and the denominator.

Cost-benefit analysis

Following a recent cost-benefit analysis of the investment of the Global Fund for AIDS, Malaria and Tuberculosis [24], we calculate the cost benefit ratio of each policy option by adding the averted cost of orphanhood and the value of additional productivity generated by each of the 11 alternative scenarios in comparison to the baseline scenario of the current guidelines. The cost of orphan care was based on an update to [25] and is applied to the fraction of households estimated to live below the poverty line (upper bound) of R 577 (2009 ZAR) [26]. Productivity was valued as Gross National Income per working-age person and was weighted by CD4 cell count stratum and ART status using a summary of analyses of the impact of HIV and ART on employee productivity from [24]. The assumptions used in this analysis are summarized in Table 3.

Table 3: Outcomes assumptions

CD4 cell count* [cells/microl]	>500	350-500	201-349	100-200	<100	Source
Disability weights (adults and children)						
HIV-infected, not on ART	0.053	0.053	0.221	- 0.547 -		[18]
HIV-infected, on ART			-- 0.053 --			[18]
Productivity weights (adults of working age)						
HIV-infected, not on ART	1	1		- 0.2 -		[24]
HIV-infected, not on ART	1	1		- 0.75 -		[24]
Cost of orphan care in 2013 USD						
% of orphans needing care and support			-- 56.8% --			Upper-bound poverty line [26]
Cost per child aged						[25]
0-4 yrs			-- 304 --			
5-9 yrs			-- 486 --			
10-18			-- 1,132 --			
Value of productivity in 2013 USD						
Gross National Income (GNI) per working-age person			-- 13,618 --			2012 GNI [27] divided by total population aged 20-65 from 2011 national census [28]
% of adult population assumed to be of working age			90%			[24]

Expansion scenarios

We evaluated the total and incremental cost, cost effectiveness, cost-utility and cost-benefit of each of the four eligibility scenarios mentioned above, compared to the baseline scenario of the current South African treatment guidelines. Within each scenario, we varied assumptions about how expansion would take place, based on the analytical framework used in the 12-model analysis [12]:

- "***Status quo***" (SQ): Estimated current patterns of testing and service uptake (linkage to care and retention in care, with values based on [29]) continue into the future.

- "***Prioritised expansion***" (PE): Increased testing and immediate ART uptake amongst 80% of the members of the specific subpopulation prioritised for immediate ART, while for the general population estimated current patterns of testing and service uptake continue into the future. (This sub-scenario only applies to the scenarios with discordant couples and pregnant women.)

- "***Uniform expansion***" (UE): There are substantial increases in HIV testing and linkage such that 80% of infected persons undergo annual testing and, once tested HIV positive, semi-annual CD4 monitoring (if previously tested positive), and have the opportunity to initiate ART as soon as they are eligible.

Together with the four eligibility options mentioned in the Introduction, this gives us a total of 12 scenarios for analysis:

1. Current guidelines (CG)
    a. Status quo (CG.SQ)
    b. Uniform expansion (CG.UE)
2. Eligibility at 500 (500)
    a. Status quo (500.SQ)
    b. Uniform expansion (500.UE)
3. Universal Test and Treat (UTT)
    a. Status quo (UTT.SQ)
    b. Uniform expansion (UTT.UE)
4. Discordant couples (DC)
    a. Status quo (DC.SQ)
    b. Prioritised expansion (DC.PE)
    c. Uniform expansion (DC.UE)
5. Pregnant women (PW)
    a. Status quo (PW.SQ)
    b. Prioritised expansion (PW.PE)
    c. Uniform expansion (PW.UE)

Demand- and supply-side cost considerations

As mentioned above, previous efforts to model the cost of scaling up antiretroviral therapy in South Africa have assumed zero cost for demand generation and, with the exception of our own illustrative exercise [11], have modelled ART production as if all patients in the country were served by a single gargantuan hospital, with constant unit costs by patient type. In this paper we explore the influence on cost estimation of the relaxation of these implausible demand- and supply-side assumptions.

Modeling the cost of production

We used data on the size and distribution of ART clinics in South Africa based on the District Health Information System from June 2013 [20]. We then added functionality to the model that allows the analyst to set the maximum number of clinics that will be accredited for ART over the next 20 years, the distribution of new facilities into different clinic levels, and the shape of the clinic size distribution function. Based on an increasing body of literature [21,22,30,31], we selected a scale elasticity at 0.8 for the base case analysis.

In our previous paper [11], we introduced to the literature on the modeling of antiretroviral treatment scale-up the distinction between two types of cost functions, which we termed the "accounting identity" cost function and the "flexible" cost function. Following the economics literature, we defined the accounting identity function as one which constructs an estimate of aggregate production cost by assuming that the average cost of output is fixed, so that total cost during a period is simply the product of the number of units of output during that period and the constant average cost. In contrast, we defined a "flexible" cost function as one that allows for production managers to achieve economies in response to changes in factor prices, scale of production and other relevant variables that affect their operating environment. Instead of starting with an average cost and multiplying by the units of output, the user of a flexible cost function starts by specifying that total cost at an individual facility is related to output and other facility characteristics by a parsimonious non-linear functional relationship and then divides total cost by output in order to derive average cost. Following this second more flexible approach, average cost is not necessarily constant over time and in practice tends to decline with the volume of output, displaying the pattern known as economies of scale.[2]

As econometric studies of the cost of ART in representative and/or moderately sized samples of facilities have begun to appear in the literature, evidence has accumulated against the simplest accounting identity cost functions [21,22,30]. Estimates of economies of scale are typically statistically significant, with scale elasticities, where they are reported, ranging from 0.7 to 0.9.[3] When the elasticity of scale differs significantly from unity (ie, 1), ignoring the current and future size distribution of the ART treatment facilities can substantially bias estimates of future production costs, as we showed in our previous analysis [11].

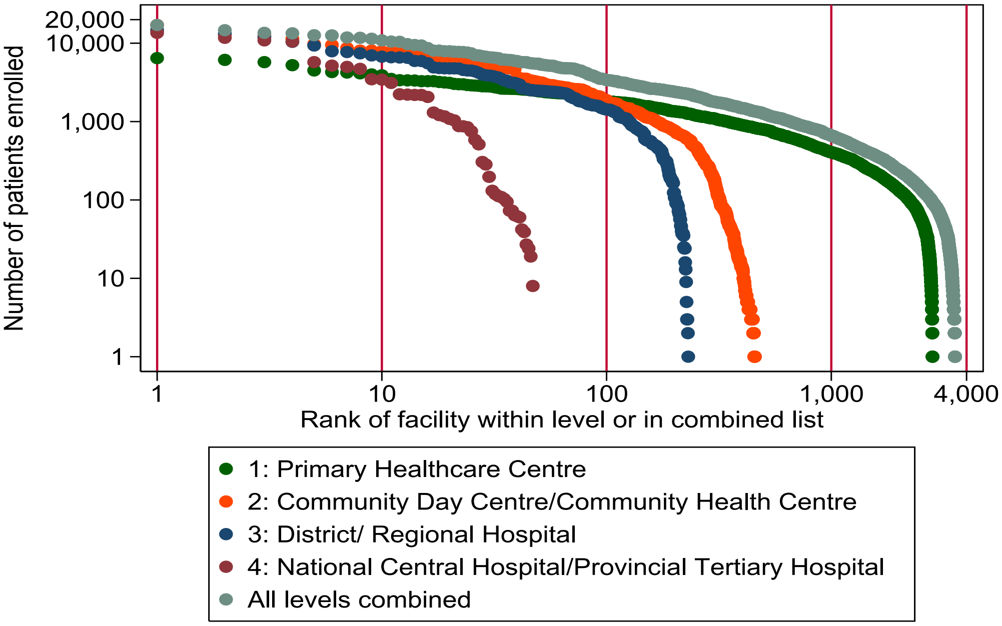
For this analysis, we benefit from a more complete census of South African ART facilities than the one used in our previous analysis [20]. Defining the scale or size of an ART treatment facility as the number of patients it treats in a given year, Figure 1 illustrates the size distribution for each of four levels of health care facility:

- Level 1: Primary healthcare centres
- Level 2: Community day centres/ Community health centres
- Level 3: District/ regional hospitals
- Level 4: National central hospitals/ provincial tertiary hospitals.

[2] Economies of scale occur when a portion of costs are fixed, and a portion variable. More generally they occur because all factors of production, not just fixed factors, are used more efficiently when they can be spread over more units of output. A flexible cost function also characterises other aspects of the production technology which are obscured in an accounting identity, such as the degree to which various inputs to the production process can be substituted for one another and whether technological change primarily benefits labor or capital. In this paper, we ignore these other aspects of a cost function.

[3] An elasticity is the percentage impact on a dependent variable of a 1% increase in an independent variable.

Figure 1. Empirical size distributions of 3,558 South African public facilities that delivered antiretroviral therapy in 2012/2013. Each point represents a facility. Axes are scaled to the logarithms of the variables. Source: Authors' construction.

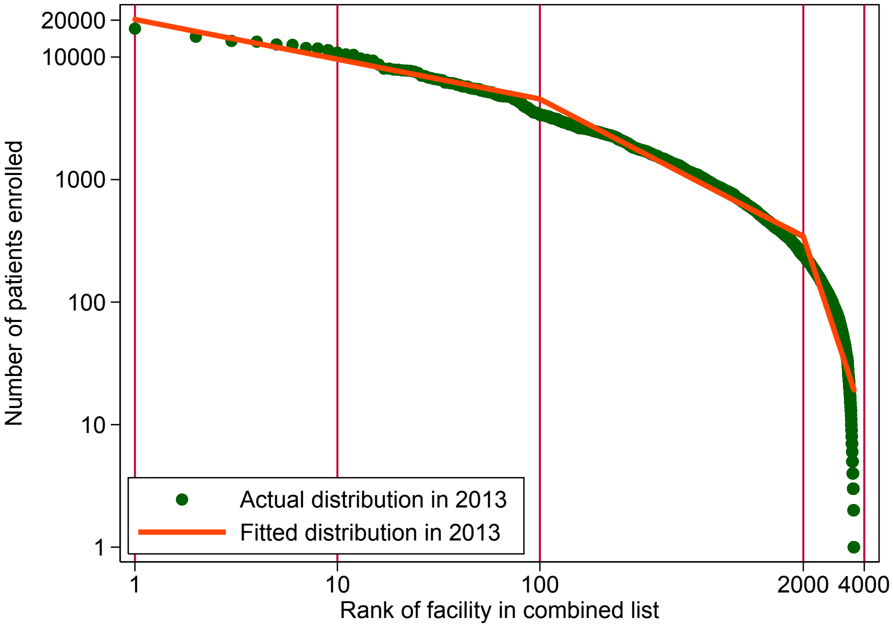


The fifth distribution, to the northeast of the other four, is of the entire combined list of all 3,558 facilities delivering ART.

Like the earlier incomplete size-distribution that we studied [11], the combined distribution in Figure 1 has an approximately constant slope in the logged variables for the first 2,000 facilities and then drops precipitously towards a facility size of 1. Figure 2 displays the piecewise linear spline fitted to the combined distribution in Figure 1, with knots at 100 and 2,000 facilities. The slopes of the two segments are -.32 for the first 100 facilities and -.85 for facilities ranked between 100 and 2000, somewhat flatter than the slopes of approximately -1 that are commonly observed for the rank-size distribution of cities and towns, suggesting that the size distribution of ART treatment facilities in South Africa has not yet matured to follow more closely that of the population itself.

If all South African facilities treated the same number of patients, these size distributions would be flat and economies of scale would not matter for projecting treatment cost. In our clinic data, facilities differed in size by more than four orders of magnitude, from those that served a single patient to the busiest facility in the sample which served 17,081. If each facility enjoys a scale elasticity substantially less than unity, the impact on total cost of adding patients to the small facilities will be larger than adding the same number of patients to large facilities. To see how much difference this can make, we explore the sensitivity of cost

Figure 2. A piecewise-linear spline with two knots explains 94% of the variation in the size-distribution of ART treatment facilities in South Africa in 2013. Source: Authors' construction.



projections to variations in how the additional patients are distributed by the current size of the facility and by the elasticity of scale that characterizes the entire system.[4]

The most ambitious ART expansion modeled in this paper (UTT.UE) expands the total patient enrollment from the approximately 2.4 million patients currently enrolled to as many as 5 million patients on treatment by the year 2033. We consider two policy dimensions of the distribution of patients across facilities. First, we model the possible expansion of the number of ART facilities, from the current 3,558 up to a maximum of 7500. Secondly, we distribute the new patients across these facilities according to three alternative patterns. We call the three distribution patterns: (1) *average*, (2) *quadratic* and (3) *proportional*. To illustrate the difference, suppose that the number of facilities is held constant at 3,558 and the number of patients is to be increased from 2.4 million to 5 million. This would be an average increase of 815 patients per facility. Under the *average* distribution pattern, the system would add 815 patients to each facility, a very small additional load for the largest facility but a massive increase for the facilities currently serving only one patient.

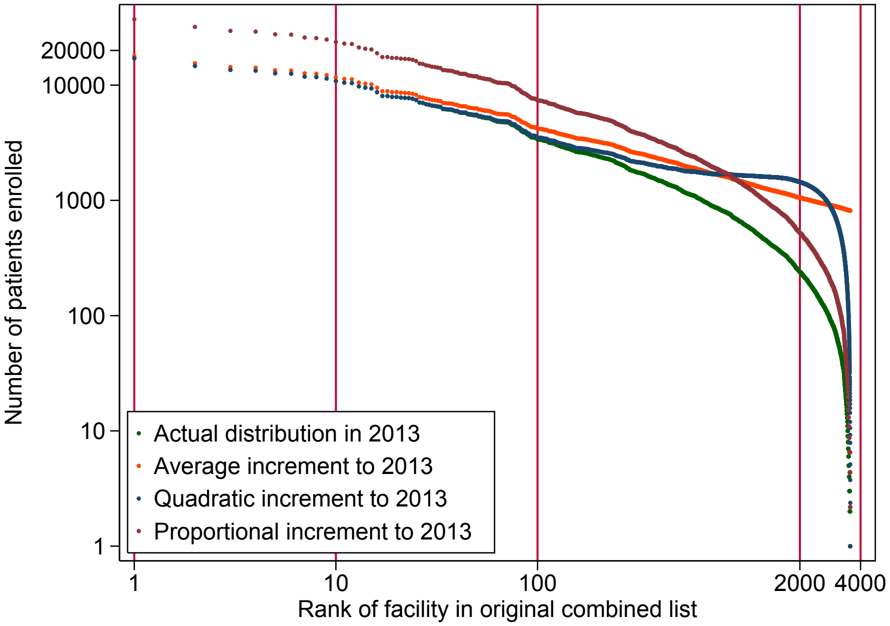
[4] In reality, the scale elasticity might vary by level of the facility, by whether it is urban or rural or by other facility characteristics. Given data on total annual ART treatment costs, number of patient-years of ART delivered and a range of other facility characteristics, in all the level 4 and a sample of 30 of each of the other levels of care, it would be straightforward to estimate a level-specific scale elasticity. Lacking such information, we assume the elasticity is constant across levels.

Since (5 - 2.4)/2.4 = 1.08, the increase to 5 million would be a 108% increase in the national number of patients. Our *proportional* distribution pattern would increase the patient load of each facility by 108%. This would mean that the largest facility, which has 17,081 patients in 2013, would expand by 18,447 patients to be serving 35,528 patients, while the facility treating only 1 patient in 2013 would be treating only 2 patients when the country reaches maximum scale-up.

Our *quadratic* distribution pattern is intermediate between the arithmetic and proportional patterns. For the quadratic pattern, we impose the constraint that neither the largest nor the smallest facility absorbs more patients, with the incremental 2.4 million patients being distributed across the middle of the range of facility ranks according to a quadratic relationship.

Figure 3 displays the impact that each of these three distribution patterns would have on the size-rank distribution of facilities, when we hold the number of facilities constant at 3,558 and expand the number of patients to 5 million.

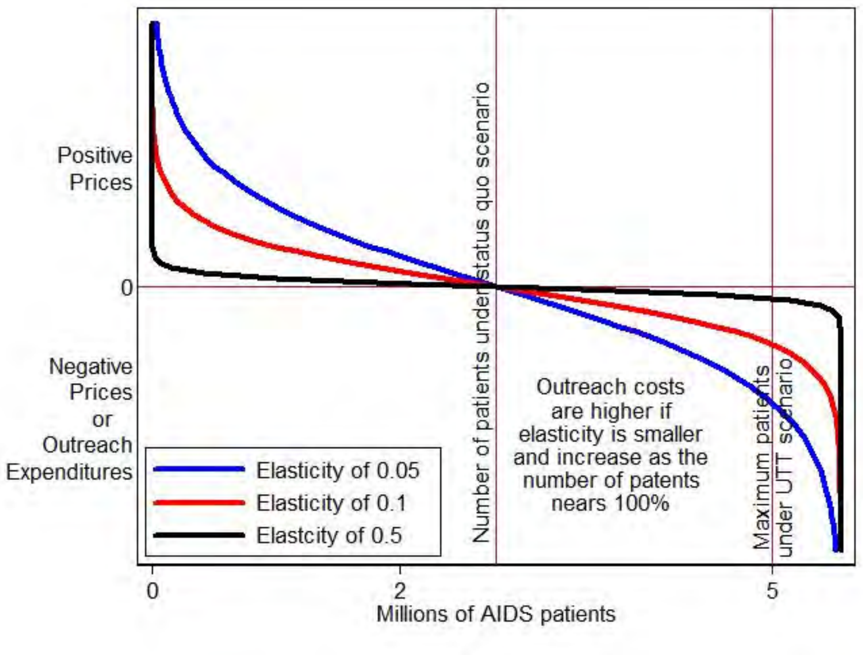
Figure 3. Three patterns for adding 2.9 million patients to the existing distribution or patients over the 3,558 facilities delivering ART in 2013.



Modeling the cost of demand generation

Experience has shown that increasing the number of patients who effectively adhere to ART requires more than simply constructing facilities and making drugs available. A growing number of studies are analyzing the leakage in the ART care cascade with the aim of designing treatment programs which minimize the loss of patients at each stage in the treatment process. While South Africa has achieved a remarkable roll-out of ART, with the number of patients in the public sector rising from none in the beginning of 2004 to more than 2.4 million in 2013, these gains have been achieved mainly among the sickest patients, for whom ART offers almost immediate health benefits. In an analysis of treatment outcomes over seven years on treatment in a large South African public-sector clinic, the median CD4 cell count at initiation rose from 82 cells/microl in 2004/05 to just 114 cells/microl in 2009/10, while loss to follow-up ranged between 19.1% and 28.6% of each annual cohort [33].

Figure 4. Demand for antiretroviral therapy services. Other things equal, the demand for antiretroviral services is greater when prices are lower or, in the absence of prices, when outreach expenditures are greater. To achieve an ambitious target of enrolled and adherent patients may require large outreach expenditures especially if the elasticity is small. Source: Authors' construction.



By making every effort to serve the patients who most needed treatment, South Africa has managed so far without the need for an extensive outreach program to support patient adherence. We assume this will continue to be the case in the baseline scenario (Current guidelines, status quo). However, to the degree that an expansion scenario requires individual facilities to recruit patients with higher CD4 counts or who are more recalcitrant for other reasons, we assume that the government will have to finance outreach activities to recruit and then retain these additional patients. To model the anticipated cost of these outreach programmes, we characterize the individual facilities as supplying a service to patients who manifest a demand for care. Like all demand relationships, the demand for ART can be expected to be small at a higher price and to increase as the price decreases. The upper half of Figure 4, above the horizontal axis, illustrates this "iron law of economics" with a demand curve plotted on axes representing the price per patient on the vertical axis and the number of patients on the horizontal axis. By continuing this demand curve below the horizontal axis, we extend this concept of the demand curve to negative prices, which we assume are spent in the form of a facility's outreach costs, and might include travel vouchers or home visits or food supplements. We assume the outreach costs increase as the number of enrolled patients increases, becoming infinitely large as enrollment approaches 100% of those eligible.

Since our model projects the growth of individual facilities, we model the outreach expenditure requirement for each facility as analogous to the Figure 4 depiction of demand in the entire country. We assume that the negligible level of outreach currently funded by the government in public facilities would, if continued, enable every facility's enrollment to expand according to the baseline scenario. Define that status quo enrollment level in facility k, year t, as $\tilde{n}_{kt}$, which is given by one of the three previously discussed scenarios for the extension of the number of clinics. Just as the outreach expenditures required to obtain 100% coverage for the entire country grow without limit, we assume the same is true at each facility.

To implement these assumptions, we include the cost of demand creation for both testing and improved retention for every patient in addition to those in the CG.SQ scenario. For testing, we added the cost of a mobilisation event per tested patient, based on an ingredients cost analysis used in the costing of South African Provincial Strategic Implementation Plans [19]. For retention, in order to enable patients to present themselves for quarterly appointments at the ART clinic, we added an outreach cost per incremental patient which, at our assumed elasticities of demand, would be sufficient to attract the number of patients to that facility that are predicted by the given epidemiologic scenario. We calibrate these elasticities so that for a modest expansion the annual per-patient outreach cost would approximately equal the typical cost of four trips to a health center at ZAR 50 (USD 5) for an urban clinic and ZAR 30 (USD 3) for a rural clinic per single round trip[5]. Outreach cost per patient can rise to a multiple of these benchmark values in high coverage scenarios or in high coverage clinics. See Appendix B for further details on our models of supply and demand.

[5] In an analysis of the transport cost of patients accessing ART in an urban and a rural clinic in South Africa, these amounts would have covered the transport cost of 100% of patients in the rural clinic, and of 90% of patients in the urban clinic [32].

Results

Number of patients on ART

Under the current guidelines and trends in testing, linkage to care and losses to retention (CG.SQ), 2.4 million adults and 202,067 children are estimated to be on treatment by mid-2016, and 3.4 million adults and 135,424 children by 2033 (see Table 4). If the current guidelines were kept, but testing, linkage to care and retention were improved to 80% each (CG.UE), these numbers would increase to 3.7 million adults and 236,471 children in 2014, and 5.3 million adults and 103,789 children in 2033.

Under all other uniform expansion (UE) scenarios, there are more patients expected to be on treatment by mid-2016, and less by mid-2033, than under CG.UE; there is in fact an inverse relationship between the number of patients on treatment by mid-2016 and those on treatment by mid-2033 for all scenarios. This result illustrates the assumed prevention benefits of ART, as higher eligibility in early years reduces HIV incidence and thus leads to fewer eligible patients in later years

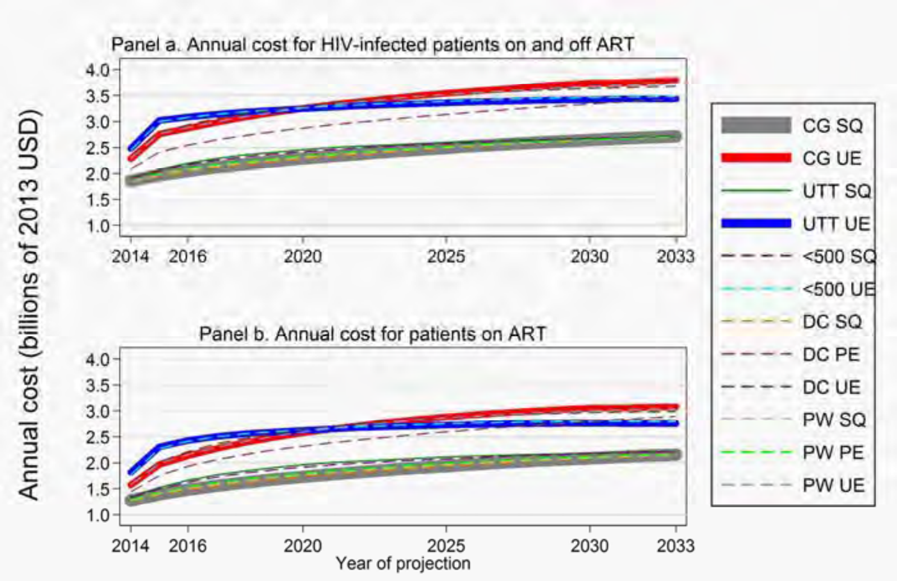
Table 4: Number of patients on ART by scenario

Scenario	Adults on ART by mid-2016	Children on ART by mid-2016	Adults on ART by mid-2033	Children on ART by mid-2033
Current guidelines (CG)				
Status quo (SQ)	2,401,552	202,067	3,402,879	135,424
Uniform expansion (UE)	3,724,236	236,471	5,266,077	103,789
Universal test and treat (UTT)				
Status quo (SQ)	2,770,414	201,347	3,613,720	99,467
Uniform expansion (UE)	4,431,985	226,155	4,947,584	52,651
Eligibility <500 (500)				
Status quo (SQ)	2,697,173	201,570	3,591,006	107,837
Uniform expansion (UE)	4,375,441	228,299	4,998,933	57,649
Discordant couples (DC)				
Status quo (SQ)	2,386,998	204,281	3,391,194	136,638
Prioritised expansion (PE)	3,309,333	199,201	4,879,992	81,912
Uniform expansion (UE)	3,868,680	234,240	5,137,060	82,751
Pregnant women (PW)				
Status quo (SQ)	2,441,042	201,883	3,413,530	125,253
Prioritised expansion (PE)	2,508,171	199,617	3,466,497	109,445
Uniform expansion (UE)	3,789,507	229,648	5,213,924	77,725

Trend of undiscounted HIV-related cost over time

Unlike in the first analysis of universal test and treat scenarios[4], none of the scenarios in our analysis is cost-saving over 20 years, though most of the status quo (SQ) sub-scenarios might be cost-saving over longer time periods. This can be seen from Figure 5 (Panel a. and b.), which tracks the undiscounted cost of HIV-related care for people on or off ART (Panel a.) and on ART only (Panel b.), by year: The cost curves in most scenarios flatten towards the end of the projection period (both panels), and the cost curves of all HIV-related care of most of the SQ scenarios (whose incremental cost over the baseline scenario per year is very small throughout) cross the curve of the baseline scenario (CG.SQ) in the last years of the projection period (Panel a.), suggesting cost savings in the annual cost from the year in which the lines cross onwards. The same is true for the cost of the PW.PE scenario.

Figure 5: Total undiscounted cost per year for all HIV-related care (Panel a.) and for patients on ART only (Panel b.)



Note: Supply side: Sigma = 0.8, Kmax = 6000, Scale-up pattern = Quadratic; Demand elasticities: Urban = 0.1, Rural = 0.5
Eligibility options: CG = Current guidelines, UTT = Universal test and treat, <500 = Initiate ART at CD4 < 500,
DC = Discordant couples, PW = Pregnant women
Expansion strategies: SQ = Status quo, PE = Prioritized expansion, UE = Uniform expansion

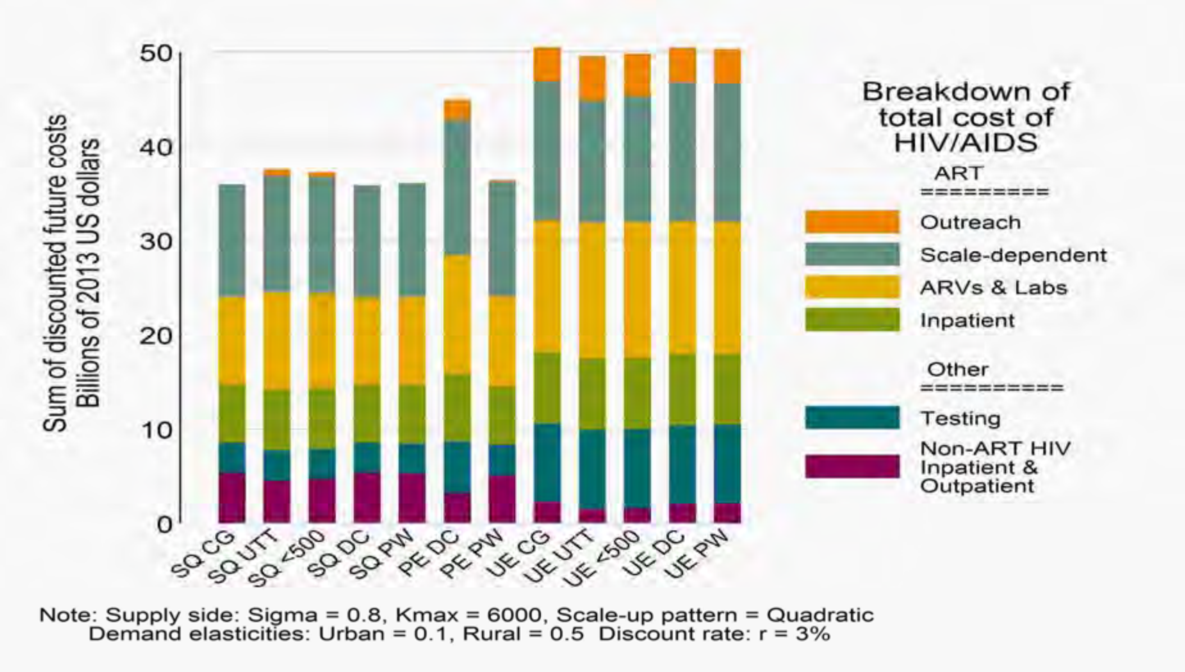
For planning and budgeting purposes, Appendix C gives the undiscounted annual cost for each of the years 2014-2033 as well as the current mid-term expenditure framework (MTEF) of 2014-2016 for each scenario.

Total HIV-related cost

At the discount rate of 3%, the total cost over 20 years for the current guidelines at status quo (CG.SQ) is expected to be close to USD 36 billion (Table 5 and Figure 6), equivalent to an annual payment of US$2.4 billion. For reference, the discounted cost for the current mid-term expenditure framework of 2014-2016 is USD 3.6 billion, or about US$1.2 billion per year. Figure 6 displays the breakdown of total cost under the central assumptions used in this paper. The scenarios are ordered by recruitment sub-scenario (from SQ to PE to UE) and then by eligibility scenario in order to highlight that, regardless of eligibility criterion, the status quo scenarios are less costly and the uniform expansion scenarios most costly.

Figure 6 displays other patterns worth noting. The large contribution of non-ART inpatient and outpatient services under the SQ scenarios is greatly reduced under the UE scenarios, but its place is filled by the large additional expenditure for testing. The two parts of ART costs we have assumed to be unrelated to the scale of a facility's work load, the cost of ARVs and labs, rise with the number of patients on treatment, as does inpatient cost for patients on ART. The component that rises the most is the cost of outreach. Under our central assumption set, we set the elasticities of urban and rural demand to equal respectively 0.1 and 0.5. We assume cost of outreach to be zero in the baseline CG.SQ scenario and to barely appear at all under the other SQ or the PE scenarios. This is because none of these scenarios is projected to require a large percentage increase in patients. However, all five of the UE scenarios as well as the DC.PE scenario must greatly increase patient recruitment, and therefore incur demand creation costs determined by the assumed testing costs and demand elasticities.

Figure 6. Total HIV-related health care costs and its components in South Africa 2014-33 (discount rate 3%)



Note: Supply side: Sigma = 0.8, Kmax = 6000, Scale-up pattern = Quadratic
Demand elasticities: Urban = 0.1, Rural = 0.5 Discount rate: r = 3%

Cost effectiveness

As shown in Table 5, improving testing, linkage to care and retention under the current eligibility in the CG.UE scenario would increase total cost over 20 years by USD 14.6 billion to a total of USD 50.5 billion and would reduce new infections by close to 50% to 1.9

million. This results in a cost per infection averted of USD 7,559, the highest cost per infection averted of all scenarios, including all other uniform expansion sub-scenarios.

The most cost-effective option in terms of cost per infection averted is to expand ART eligibility to pregnant women, either while maintaining the status quo for linkage and retention (PW.SQ) or while improving their linkage and retention and maintaining the status quo for everyone else (PW.PE). On the other hand, expanding eligibility for discordant couples, either as prioritized or as uniform expansion (DC.PE or DC.SQ), is the least cost-effective option of all PE or SQ scenarios.

When maintaining the current status quo for testing, linkage to care and retention, expanding eligibility to discordant couples or pregnant women would have little effect on either cost or infections averted. This is because coverage of ART is already high in pregnant women compared to the general population, and current levels of partner testing are relatively low, so that few additional people would be reached by expanding guidelines to these groups unless effort is made to also encourage testing and linkage to care. Expansion of eligibility guidelines to include persons with a CD4 cell count of <500 cells/microl or all infected individuals (UTT) would significantly increase both cost and infections averted. Of these two strategies, universal test and treat would be associated with about 5% lower cost per infection averted and both are about 30% more cost-effective than uniformly expanding under current eligibility guidelines.

Table 5. Cost, infections averted and cost effectiveness by scenario

Scenario	Total cost 2014-2033 [billion 2013 USD]	Total new infections 2014-2033 [millions]	Incremental cost [billion 2013 USD]	Infections averted [millions]	Incremental cost per HIV infection averted [2013 USD]
Current Guidelines (CG):					
Status quo (SQ)	36.0	4.3	comparator	comparator	comparator
Uniform expansion (UE)	50.5	2.3	14.6	1.9	7,559
Universal test and treat (UTT):					
Status quo (SQ)	37.6	3.7	1.6	0.6	2,671
Uniform expansion (UE)	49.6	1.7	13.6	2.6	5,283
Eligibility < 500 (500):					
Status quo (SQ)	37.3	3.8	1.3	0.5	2,832
Uniform expansion (UE)	49.8	1.8	13.9	2.5	5,544
Discordant couples (DC):					
Status quo (SQ)	35.9	4.3	-0.1	0.0	See note*
Prioritised expansion (PE)	44.9	2.9	9.0	1.3	6,671
Uniform expansion (UE)	50.5	2.1	14.5	2.2	6,723
Pregnant women (PW):					
Status quo (SQ)	36.1	4.2	0.2	0.1	1,692
Prioritised expansion (PE)	36.4	4.0	0.4	0.2	1,979
Uniform expansion (UE)	50.3	2.2	14.4	2.0	7,009

Note: This table uses the central assumptions. Supply side: Sigma = 0.8, Kmax = 6000, Scale-up pattern = Quadratic, Demand elasticities: Urban = 0.1, Rural = 0.5. Discount rate: 3%.

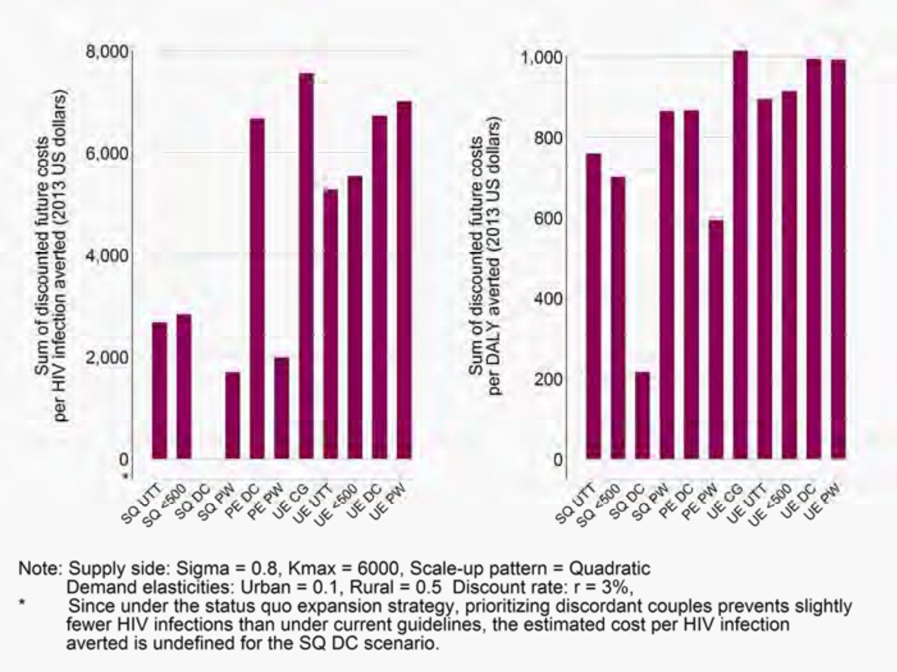
*Incremental costs and effectiveness are computed in comparison to the CG.SQ scenario. The effectiveness of the DC.SQ scenario is too close to those of the comparator CG.SQ scenario for the cost-effectiveness result to be meaningful.

The cost-effectiveness ranking of the policy options is similar when effectiveness is measured by DALYs saved instead of by HIV infection averted (Table 6 and Figure 7). Since averting an HIV infection averts several years of disability-adjusted life, it is not surprising that the cost per averted HIV infection is higher than the cost per DALY for the same intervention. The cost per averted DALY can be compared to the value of the typical person's healthy year, while the cost per averted HIV infection can be compared to the value of the person's life expectancy.[6]

[6] Because the projection horizon in this exercise is limited to 20 years, there is insufficient time to observe all the DALY benefits of averting an HIV infection. For any given expansion policy and cost assumptions, a longer planning horizon should reveal a larger ratio of DALYs averted to infections averted.

As before, the most expensive and least cost-effective option is the uniform expansion of the current guidelines, which we estimate to cost $1,015 per DALY saved. Among the five uniform expansion scenarios, the cost-effectiveness advantage of the most ambitious of these, the less than 500 and the UTT, is less pronounced by the DALY metric than by the infection averted metric. Both are about 10% more cost-effective than the uniform expansion under current eligibility guidelines.

Figure 7. Cost per HIV infection and DALY averted for 11 scenarios compared to the current guidelines status quo scenario (discount rate 3%).[7]



Note: Supply side: Sigma = 0.8, Kmax = 6000, Scale-up pattern = Quadratic
Demand elasticities: Urban = 0.1, Rural = 0.5 Discount rate: r = 3%,
* Since under the status quo expansion strategy, prioritizing discordant couples prevents slightly fewer HIV infections than under current guidelines, the estimated cost per HIV infection averted is undefined for the SQ DC scenario.

[7] The negative value associated with the discordant couple (DC) scenarios in Figure 7 occur because the effectiveness of this scenario is less than the effectiveness of the other scenarios and the costs are larger. The implausible figure of -$62,859. occurs because the number of infections averted is very slightly less than in the baseline scenario, so the denominator of the cost effectiveness ratio is a negative number close to zero.

Table 6: Disability-adjusted life years (DALYs) averted and cost utility by scenario (discount rate = 3%)

Scenario	Total cost 2014-2033 [billion 2013 USD]	Total DALYs 2014-2033 [millions]	Incremental cost [billion 2013 USD]	DALYs averted [millions]	Incremental cost per DALY averted [2013 USD]
Current guidelines (CG)					
Status quo (SQ)	36.0	592.1	comparator	comparator	comparator
Uniform expansion (UE)	50.5	606.4	14.6	14.4	1,015
Universal test and treat (UTT)					
Status quo (SQ)	37.6	594.2	1.6	2.1	759
Uniform expansion (UE)	49.6	607.3	13.6	15.2	895
Eligibility < 500 (500)					
Status quo (SQ)	37.3	594.0	1.3	1.9	701
Uniform expansion (UE)	49.8	607.2	13.9	15.1	914
Discordant couples (DC)					
Status quo (SQ)	35.9	591.6	-0.1	-0.4	216
Prioritised expansion (PE)	44.9	602.4	9.0	10.4	866
Uniform expansion (UE)	50.5	606.6	14.5	14.6	995
Pregnant women (PW)					
Status quo (SQ)	36.1	592.3	0.2	0.2	865
Prioritised expansion (PE)	36.4	592.8	0.4	0.8	593
Uniform expansion (UE)	50.3	606.5	14.4	14.5	992

Note: Supply side: Sigma = 0.8, Kmax = 6000, Scale-up pattern = Quadratic, Demand elasticities: Urban = 0.1, Rural = 0.5

Figure 8 Panels a. to d. give a graphical representation of these results for central values of the behavioral supply- and demand-side parameters. As has become the convention for displaying incremental cost-effectiveness ratios (ICERs), these charts display the 20-year cost of each scenario relative to the baseline scenario, status quo scenario on the vertical axis and the 20-year utility relative to same counterfactual on the horizontal axis. Thus scenarios represented by points with high costs and low utility will be closer to the vertical axis, while

those with low costs relative to their utility will be close to the horizontal axis. The distance of a point from the origin represents the magnitude of the costs and utility in comparison with the counterfactual in which the country remains on its present course.

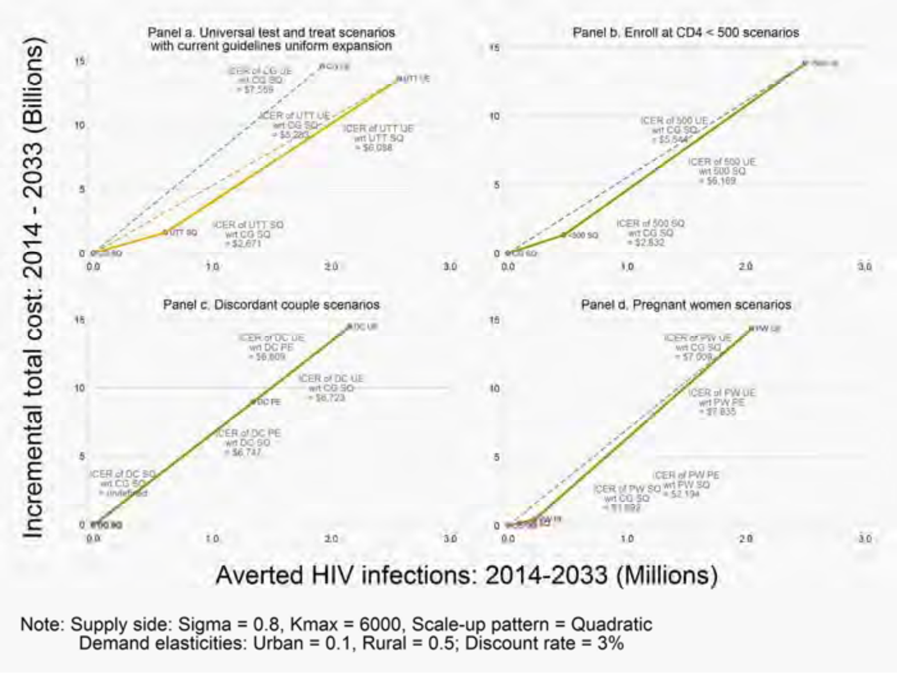
Panel a. of Figure 8 displays the cost-effectiveness of three scenarios relative to CG.SQ scenario. The uniform expansion of the current guidelines (eligibility threshold at a CD4 cell count of 350) will have the greatest cost but produce only a modest reduction in DALYs, so that its cost effectiveness, represented by the steep slope of the ray connecting the point CG.UE to the origin, is USD 7,559 per HIV infection averted over the 20 years. On cost-effectiveness grounds, this scenario is the least attractive option we analyse. It is dominated by other scenarios not only under this set of parameters, but under all the parameter combinations we have explored. In contrast, Panel a. also displays two universal test and treat (UTT) policies, both of which are more cost-effective than the CG.UE option. In comparison to scenario CG.SQ, applying the status quo recruitment policies with the UTT eligibility criteria will purchase 600,000 averted HIV infections at a cost of only USD 2,671 per averted infection. Energetically extending the same UTT eligibility criterion with the uniform expansion of service and recruitment efforts prevents additional infections at a cost of USD 6,088 per infection averted, which is less than half as cost-effective as moving to UTT.SQ, but still more cost-effective than CG.UE.

Panel b. of Figure 8 shows the results for expanding the eligibility criterion to a CD4 count less than 500 cells/microl, an intermediate policy option between the current guidelines and the universal test and treat options depicted in Panel a. While the 500 CD4 count threshold option performs very similarly to the UTT option, note that it is estimated to be a slightly more expensive way for the government to "buy" these averted infections. This is because it achieves fewer prevention benefits in the outer years than does the UTT option, though the difference as simulated in our model is small.

Panels c of Figure 8 shows that prioritized expansion to discordant couples averts 1.3 million HIV additional infections at an additional cost of USD 9 billion, for a cost-effectiveness of $6,747 per averted infection. For this discordant couple policy option, the more vigorous UE recruitment strategy is almost equally cost-effective, costing $6,809 per additional infection averted. Panel d shows that the cost-effectiveness of prioritized recruitment of pregnant women, at less than $2,200 per HIV infection averted, compares favorably with that of universal test and treat under the status quo recruitment strategy ($2,671 in panel a). However the PW PE scenario buys only a third as many additional averted HIV infections as the CG SQ strategy (200,000 as opposed to 600,000) and further effort on the pregnant women policy to uniform expansion is unattractive since it's cost per HIV infection averted is as high as that of uniform expansion under the current guidelines.

A remarkable feature of the cost-effectiveness calculations displayed in Figure 8 is the kinked shape in Panels a, b and d. This shape means that initial expansion can be a good buy, but subsequent expansion to a UE scenario, while achieving important health benefits, is less cost-effective. This pattern of rising marginal costs as an infectious disease control program nears its goal of disease elimination is familiar from the smallpox and polio eradication campaigns.

Figure 8. Incremental cost effectiveness of expansion strategies for adult ART, 2014-2033



Note: Supply side: Sigma = 0.8, Kmax = 6000, Scale-up pattern = Quadratic
Demand elasticities: Urban = 0.1, Rural = 0.5; Discount rate = 3%

Cost benefit

The cost-benefit ratio, calculated as (incremental cost - cost of orphanhood averted) / value of gained productivity, is given in Table 7. A value between 0 and 1 means that the value gained by the ART program is higher than the net cost of implementing the program. In all scenarios, there are fewer orphans compared to the current guidelines scenario; in all but the DC.SQ and the PW.SQ scenarios, productivity is gained, especially under the uniform expansion scenarios. As a result, the cost-benefit ratio is highly favourable for all scenarios, with values between 0.08 (DC.PE) and 0.32 (UTT.SQ).

Table 7. Cost of orphanhood, value of productivity, and cost benefit by scenario (discount rate = 3%)

Scenario	Total cost 2014-2033 [billion 2013 USD]	Orphanhood cost 2014-2033 [billion 2013 USD]	Value of productivity [billion 2013 USD]	Incremental cost [billion 2013 USD]	Orphanhood cost averted [billion 2013 USD]	Productivity gained [billion 2013 USD]	Incremental cost-benefit ratio
Current guidelines (CG)							
Status quo (SQ)	36.0	0.328	836	comparator	comparator	comparator	comparator
Uniform expansion (UE)	50.5	0.280	979	14.5	0.049	143.0	0.10
Universal test and treat (UTT)							
Status quo (SQ)	37.6	0.317	841	1.6	0.012	5.2	0.32
Uniform expansion (UE)	49.6	0.268	926	13.6	0.061	89.9	0.15
Eligibility < 500 (500)							
Status quo (SQ)	37.3	0.319	846	1.3	0.010	10.8	0.13
Uniform expansion (UE)	49.8	0.270	933	13.8	0.059	97.7	0.14
Discordant couples (DC)							
Status quo (SQ)	35.9	0.327	832	-0.1	0.002	-3.6	0.03
Prioritised expansion (PE)	44.9	0.283	944	8.9	0.046	108.1	0.08
Uniform expansion (UE)	50.5	0.275	960	14.5	0.054	124.7	0.12
Pregnant women (PW)							
Status quo (SQ)	36.1	0.325	835	0.1	0.004	-0.6	-0.10
Prioritised expansion (PE)	36.4	0.305	840	0.4	0.024	4.0	0.09
Uniform expansion (UE)	50.3	0.272	973	14.3	0.056	136.9	0.10

Sensitivity analysis

We did a total of 13,824 computations of the model, 6,912 for each of two assumptions about the patients who would require outreach cost. We always assumed that outreach costs would be paid to all patients in excess of those who would seek care each future year under the CG.SQ scenarios. However, when the eligibility criterion is expanded from a CD4 count of 350 to one of 500 or to all HIV positives, a given facility will eventually have fewer patients starting at low CD4 counts because it has more starting earlier, at higher CD4 counts. Thus the net increase in patients at that facility will be smaller than the increase in patients at higher starting CD4 counts. A simple assumption is that patients who would have eventually started treatment without outreach costs at a CD4 count of 350, will also be willing to start earlier without outreach costs. In this case only patients who would never have come under the current guidelines will require outreach expenditure. Alternatively outreach expenditure may be required to induce a patient with a high CD4 count to start treatment earlier even if that same patient would have started later without that inducement. In this second case, outreach costs must be paid to increase the number of patients in any CD4 category. To capture these two possibilities, for half the scenarios, we defined patients requiring outreach expenditure to be those in excess of the CG SQ scenario for the facility as a whole; for the other half, we calculated the increment between the two scenarios in the number of patients in each CD4 category, and applied outreach costs to those patients in health states with positive increments only.

For each of these two assumptions about which patients would require outreach expenditure, we computed 12 scenarios x 4 discount rates x 3 scale elasticities x 4 patient distribution patterns x 12 elasticity of demand combinations. Choosing one scenario and discount rate combination, the cost effective UTT.SQ scenario evaluated at a discount rate of 3%, the 144 supply and demand side parameter combinations are displayed in Table 8. Note that the cost-effectiveness varies from USD 1,471 to USD 10,434 per HIV infection averted across these combinations, with most of the variation due to the demand side parameters.

Table 8. Cost of the universal test and treat, with status quo expansion, scenario per incremental HIV infection averted relative to the current guidelines, status quo scenario (discount rate = 3%)

Elasticities of demand		**Build-out scenario and elasticity of scale**											
		Proportional to 4,500			Quadratic to 4,500			Quadratic to 6,000			Quadratic to 7,500		
Rural	Urban	0.6	0.8	1	0.6	0.8	1	0.6	0.8	1	0.6	0.8	1
0.05													
	0.05	10,050	10,230	10,434	7,840	7,911	7,905	6,282	6,252	6,146	5,459	5,342	5,160
	0.1	7,685	7,865	8,069	6,870	6,941	6,934	5,513	5,482	5,376	4,792	4,675	4,493
0.1													
	0.05	7,899	8,079	8,283	5,602	5,674	5,667	4,720	4,690	4,584	4,284	4,167	3,985
	0.1	5,534	5,714	5,918	4,632	4,703	4,696	3,951	3,920	3,814	3,617	3,500	3,318
	0.5	3,643	3,823	4,027	3,856	3,927	3,920	3,335	3,304	3,198	3,083	2,966	2,784
0.5													
	0.05	6,178	6,358	6,562	3,812	3,883	3,877	3,471	3,440	3,334	3,344	3,227	3,045
	0.1	3,814	3,994	4,198	2,842	2,913	2,906	2,701	2,671*	2,564	2,677	2,560	2,378
	0.5	1,922	2,102	2,306	2,065	2,137	2,130	2,085	2,055	1,949	2,143	2,026	1,844
1													
	0.05	5,963	6,143	6,347	3,588	3,660	3,653	3,314	3,284	3,178	3,227	3,109	2,927
	0.1	3,599	3,779	3,983	2,618	2,689	2,682	2,545	2,514	2,408	2,560	2,442	2,260
	0.5	1,707	1,887	2,091	1,842	1,913	1,906	1,929	1,899	1,792	2,026	1,908	1,727
	1	1,471	1,651	1,855	1,745	1,816	1,809	1,852	1,822	1,715	1,959	1,842	1,660

Note: Discount rate: 3%; Outreach costs paid to additional patients in each CD4 category.

*The bolded value of $2,671 in the seventh row and eighth column of this table corresponds to our central assumptions of demand and supply elasticities and assumed build out policy and appears in Table 5 and in Panel a of Figure 8.

Our sensitivity analysis can also be used to explore whether the kinked shape of the expansion path is due to the assumptions we have incorporated into our model of the supply and demand-side determinants of scale-up cost. Figures 9 and 10 show the sensitivity of the shape of the incremental expansion path for the UTT strategy to variations around our central assumptions of the supply and the demand behavioral elasticities respectively. These figures show that the distinctive concave kink in these expansion paths is robust to these variations.

Figure 9 shows that the proportional build-out policy of squeezing new patients into the largest facilities doubles the cost per HIV infection averted compared to any of the quadratic expansion policies, with the least expensive being those which expand to the largest number of facilities. This is because the additional outreach cost at the large facilities outweighs the savings from scale economies.

Figure 9. Sensitivity of the cost effectiveness of the UTT scenario to alternative supply-side assumptions

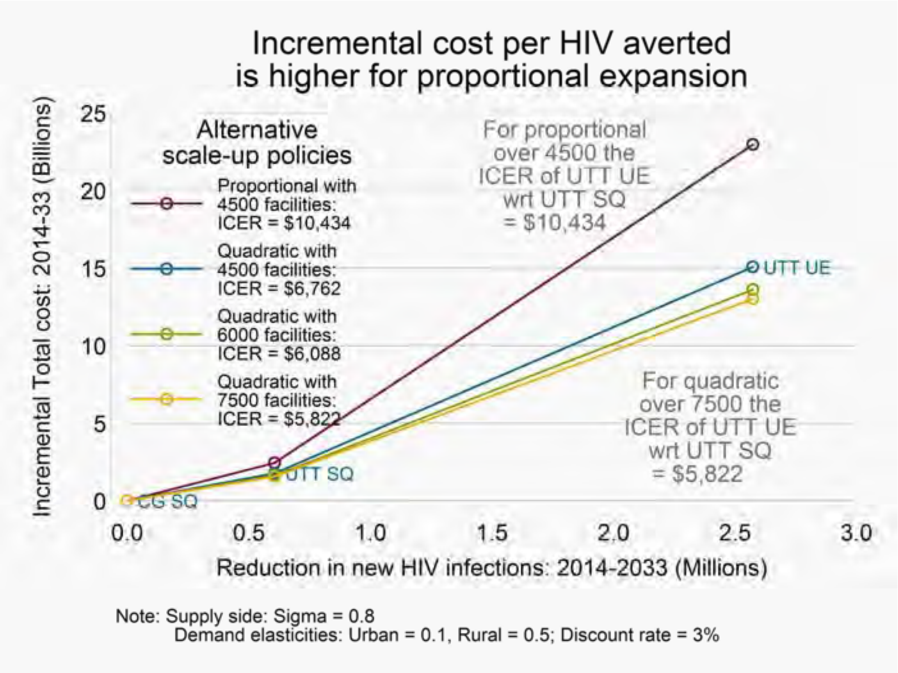
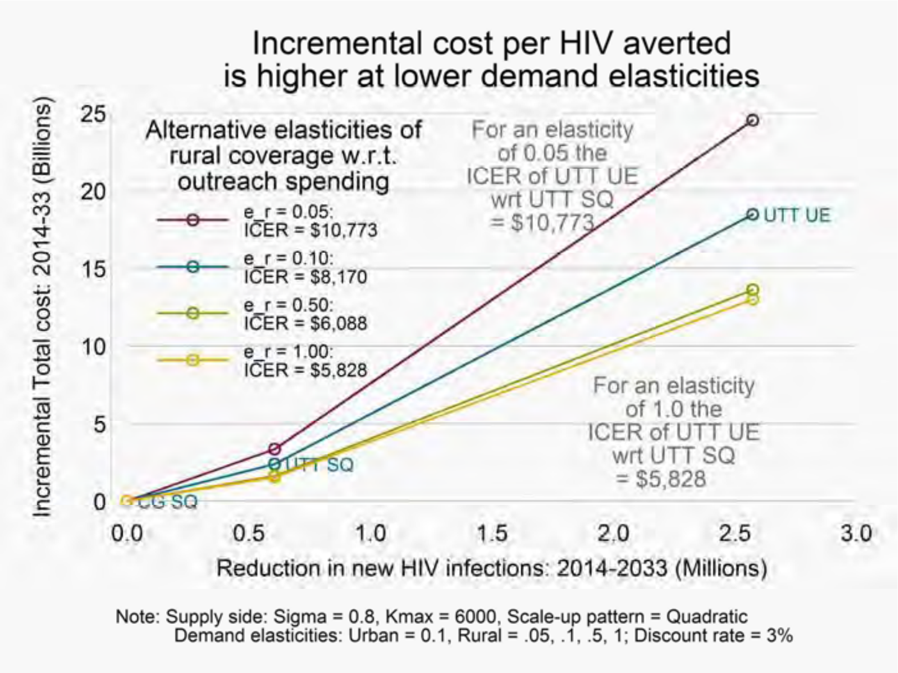


Figure 10 shows that varying the elasticity of demand for the rural population from 1 down to .05 can also double the cost per HIV infection averted and accentuate the kink in the expansion curve. The determinants of the demand for antiretroviral therapy for patients with high CD4 counts are not yet well known. If patients are relatively unresponsive to demand generation expenditures as captured here by small demand elasticities, the cost of achieving sufficiently high treatment coverage among recently infected individuals to generate HIV prevention benefits will equal exceed the highest of our estimates.

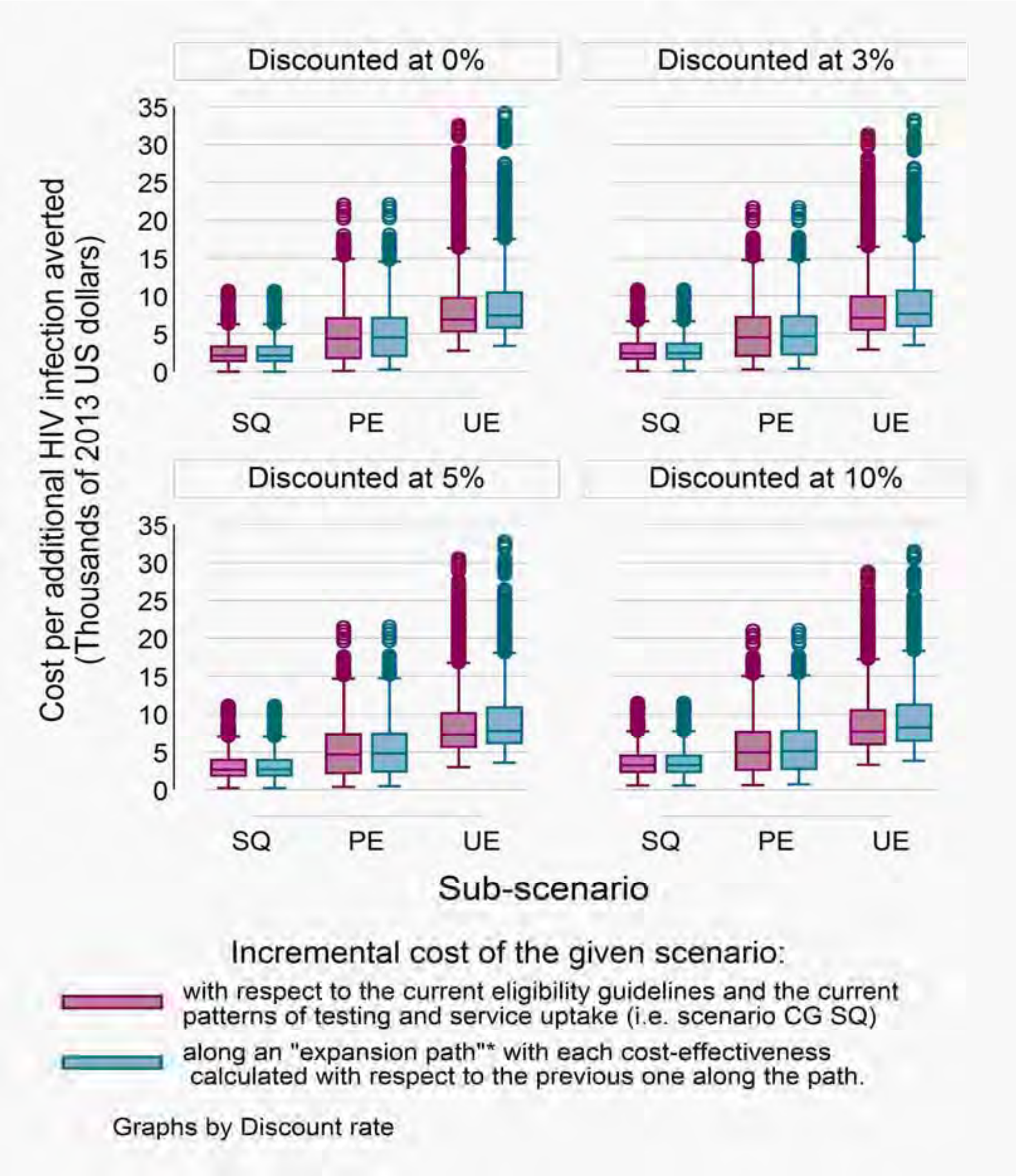
Figure 10. Sensitivity of the cost-effectiveness of the UTT scenario to alternative demand-side assumptions



Note: Supply side: Sigma = 0.8, Kmax = 6000, Scale-up pattern = Quadratic
Demand elasticities: Urban = 0.1, Rural = .05, .1, .5, 1; Discount rate = 3%

Finally Figure 11 extends the sensitivity analysis to 1,728 parameter combinations for each of the four discount rates and shows that, while the cost per DALY averted ranges widely within this parameter space, from as low as a few dollars to as high as USD10,000, the distinctive shape remains. Rather than demonstrating the benefits of increasing treatment coverage beyond the tipping point, our costing of these scenarios suggests that their cost effectiveness will degrade with scale-up, rather than improving as one might have expected. We return to this point in the Conclusions section.

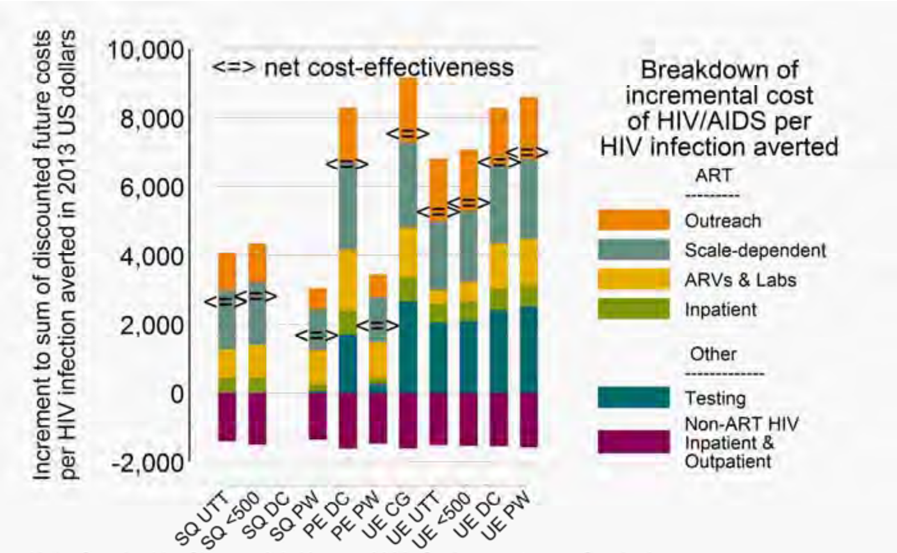
Figure 11. Results of sensitivity analysis involving 1,728 parameter combinations for each of the four discount rates.



*An "expansion path" is defined as the sequence of cost and utility combinations which start from CG.SQ and proceed step-by-step to a) supplement current eligibility by adding specific population X, but with current patterns of testing and service uptake (scenario X.SQ); b) expanding testing and service provision to provide prioritized access for population X (scenario X.PE) (where available); and c) further expand testing and service provision to provide uniform access to 80% of the population (scenario X.UE). The cost per HIV infection averted is undefined for the DC SQ scenario.

Some additional insight into the sources of variation in cost-effectiveness in this model can be gleaned from a chart showing the contribution of each of the ART components to the cost-effectiveness analysis. The height of each segment of each bar in Figure 12 is computed by dividing the cost associated with that component from the corresponding Figure 6 by the number of DALYs averted. All of the resulting ratios are positive except that for DC.SQ.

Figure 12. Net cost per HIV infection averted and its components for 11 scenarios (discount rate 3%)



Note: Supply side: Sigma = 0.8, Kmax = 6000, Scale-up pattern = Quadratic
Demand elasticities: Urban = 0.1, Rural = 0.5 Discount rate: r = 3%
The net cost-effectiveness of a scenario is the sum of the positive costs per HIV infection averted less the savings from the avoided non-ART care
For scenario SQ DC, the cost per HIV infection averted is undefined.

Comparison with previous analyses of ART expansion

As mentioned in the Introduction, this analysis is an extension of and addition to the previously mentioned 12-model comparison project [12] that was instrumental in advising the World Health Organization during the guideline revision process for the 2013 ART guidelines. The major change we introduced in this analysis was the treatment of part of the cost of treatment provision as a function of scale, and modelling the cost of increasing patient demand for ART to the required levels for 80% testing uptake, linkage to care, and retention in care. There are a number of other differences between the two analyses with regards to cost and outcomes.

First, all unit costs used in the 12-model analysis were the result of the synthesis of a number of separate cost estimates from the literature, using specialized software, of which the average cost from the National ART Cost Model (NACM) used in this analysis was only one input. Since most prices relevant to ART provision in South Africa, especially those of ARV drugs, have decreased dramatically since 2010 [13, 14] and the inputs for the evidence synthesis included a number of older estimates, the input costs used in the 12-model analysis were higher throughout. Furthermore, the distribution of patients into first- and second-line drug regimens was treated as a constant in the 12-model analysis, whereas it is based on the health-state transition matrix contained in the NACM in our analysis. Second, the 12-model analysis adds the cost of programme management (at 50% of non-ARV cost) and supply-chain management (at 20% of ARV drug costs), with the values of these mark-ups based on expert opinion. In our analysis, the cost of supply chain management is included in the ARV drug costs, and management cost, which have never been quantified for the South African ART programme, are excluded. Third, we add the cost of mobilization for every HIV-positive person being tested, while the 12-model analysis only allowed additional outreach costs for testing individuals from specific sub-populations (female sex workers, men who have sex with men, and intravenous drug users). Fourth, we allow the cost of ART provision at the level of the clinic to decrease with the scale of each clinic. Fifth, we add the cost of a transport voucher per incremental patient on ART over the current guidelines (status quo) scenario, as a proxy for increasing demand to the higher coverage levels assumed in all other scenarios.

Eaton et al [12] concluded that, "[i]n South Africa, the cost per DALY averted of extending eligibility for antiretroviral therapy to adult patients with CD4 counts of 500 cells per μL or less ranged from $237 to $1691 per DALY averted compared with 2010 guidelines." Our estimate is that this policy would cost $914 per DALY averted under our central assumptions, but could range from a low of $658 to a high of $3,706 per DALY averted depending on the elasticities of demand and supply and the national build-out policy. The 12-model analysis' findings that all expansion strategies are cost-effective for South Africa as measured against international thresholds still holds in our analysis for our central assumptions, and the ranking of expansion options by their incremental cost effectiveness is similar between the two analyses.

Conclusions

We combined the outputs of an epidemiological and a cost model of the HIV epidemic in South Africa to calculate the incremental cost effectiveness of a range of eligibility expansion strategies over current guidelines. Using the existing ART programme in South Africa as a baseline, we model the incremental cost per infection or DALY averted of each of 11 different policy alternatives for its expansion. In terms of total cost, all scenarios that maintain current trends in testing coverage, linkage to care, and retention in care ('status quo') have a very similar cost of around 36 billion USD over 20 years, while all scenarios that, for any given eligibility, assume uniform expansion of testing, linkage and retention for the entire population of eligible people ('uniform expansion'), have a total cost over 20 years of 50 billion USD. Within these, expanding eligibility to discordant couples (at current testing and linkage levels) and all pregnant women (at current and improved testing and linkage levels) are the least costly options, followed by expanding eligibility to all patients with CD4 cell counts < 500 cells/microl and Universal Testing and Treatment, both at the current level of testing and linkage.

The incremental cost per infection averted is comparable between all 'status quo' scenarios, with the prioritisation of pregnant women being the most cost-effective scenario, though it has little overall impact because HIV testing rates are already high amongst pregnant women due to high coverage of antenatal care in South Africa. All 'uniform expansion' scenarios have both greater cost and greater impact, and are more costly per infection averted. The same pattern emerges in terms of cost per DALY averted, although the differences across the scenarios are exaggerated as a result of the CD4 cell count-dependency of the disability weights for people not on ART.

Under both cost-effectiveness metrics however, cost per infection averted and cost per DALY averted, the uniform expansion of the current guidelines has the highest cost per outcome of all scenarios; if the political decision were to expand services to *all* eligible patients, defining eligibility as 'all patients with CD4 cell counts < 500 cells/microl' or simply as 'all HIV-positive patients' saves money over 20 years on simply expanding coverage to all patients eligible under the current eligibility threshold of 350 cells/ microl.

Because the scenarios are graduated in the number of people they cover, we can analyse the incremental cost-effectiveness in terms of cost per DALY averted along a hypothetical expansion path that would result from adopting a modest version of an eligibility strategy and then expanding recruitment efforts from status quo, through prioritized expansion to uniform expansion. This exercise reveals that the incremental cost effectiveness is quite favorable for small expansions, averting an HIV infection at less than $3,000 and a DALY at less than USD 800 under our central assumptions. However, in most scenarios and robust to a wide number of alternative assumptions, the incremental cost-effectiveness ratio of scaling all the way to the uniform expansion scenario is from two to five times more expensive per infection or DALY averted. This finding leads us to wonder how the ICER would behave if we had assumed even higher levels of recruitment and utilization, i.e., 95% instead of 80%. Would the difficulties of attracting these last patients prove prohibitively costly driving up the ICER? Or would an even more aggressive scale-up, although incurring higher costs per patient reached, also achieve even greater prevention benefits and thus yield more optimistic cost-effectiveness estimates?

Ultimately there is no substitute for empirical evidence production and outreach cost. How many unidentified cases of HIV never come to the center at all or come when it's too late? And how much will it cost to attract these missing patients to seek care and persist in adhering? On the production side, countries should routinely report not only the number of patients on treatment, but their distribution across facilities, so projections can incorporate economies of scale and spatial distribution. Universal coverage will also require detailed country-specific estimates for the elasticities of demand that we model here.

Furthermore, instead of subsuming all demand enhancing interventions into a single outreach cost, a more useful model of demand would distinguish among elasticities with respect to travel vouchers, distance from home to facility, provider attitudes as well as characteristics of the patient, such as their CD4 cell count, income and education.

Incorporating such a fully specified demand model into a AIDS treatment projection model has the advantage of enabling the analysis of the cost-effectiveness of multiple policy instruments. As knowledge accumulates about the influence of such policies, census and survey data on the exact geospatial locations of HIV infection can be used to build granular spatial models of demand which would enable policymakers to optimize the placement of new facility construction and outreach services. These tools would help governments plan their own policies and help them in the increasingly competitive and demanding process of preparing investment cases to compete successfully for donor support.

References for main paper

1. World Health Organization: Consolidated Guidelines on the Use of Antiretroviral Drugs for Treating and Preventing HIV Infection. Recommendations for a Public Health Approach. Geneva, June 2013
2. Cohen M, Chen YQ, McCauley M, Gamble T, Hosseinipour MC, et al. Prevention of HIV-1 infection with early antiretroviral therapy. N Engl J Med 365:493-505 (2011)
3. Rapid advice: use of antiretroviral drugs for treating pregnant women and preventing HIV infection in infants. Geneva, November 2009.
4. Granich RM, Gilks CF, Dye C, De Cock KM, Williams BG: Universal voluntary HIV testing with immediate antiretroviral therapy as a strategy for elimination of HIV transmission: a mathematical model. Lancet 373: 48–57 (2009)
5. Hontelez JAC, de Vlas SJ, Tanser F, Bakker R, Bärnighausen T, et al. The impact of the new WHO antiretroviral treatment guidelines on HIV epidemic dynamics and cost in South Africa. PLoS ONE 6: e21919. doi:10.1371/journal.pone.0021919 (2011)
6. Bärnighausen T, Bloom DE, Humair S: Economics of antiretroviral treatment vs. circumcision for HIV prevention. PNAS 109(52):21271-6 (2012)
7. Wagner B, Blower S: Costs of eliminating HIV in South Africa have been underestimated. Lancet 376: 953 (2010)
8. Wagner BG, Blower S: Universal Access to HIV Treatment versus Universal 'Test and Treat': Transmission, Drug Resistance & Treatment Costs. PLoS ONE 7(9): e41212. doi:10.1371/journal.pone.0041212 (2012)
9. Dodd PJ, Garnett GP, Hallett TB: Examining the promise of HIV elimination by 'test and treat' in hyperendemic settings. AIDS 13;24(5):729-35 (2010)
10. Hallett TB, Baeten JM, Heffron R, Barnabas R, de Bruyn G, et al. Optimal Uses of Antiretrovirals for Prevention in HIV-1 Serodiscordant Heterosexual Couples in South Africa: A Modelling Study. PLoS Med 8(11): e1001123. doi:10.1371/journal.pmed.1001123 (2011)
11. Meyer-Rath G, Over M: Modelling the Cost of Antiretroviral Treatment: State of the Art and Future Directions. PLoS Med 9(7):e1001247. doi: 10.1371/journal.pmed.1001247 (2012)
12. Eaton JW, Menzies NA, Stover J, Cambiano V, Chindelevitch L, et al: Health benefits, costs, and cost-effectiveness of earlier eligibility for adult antiretroviral therapy and expanded treatment coverage: a combined analysis of 12 mathematical models. Lancet Global Health, http://dx.doi.org/10.1016/S2214-109X(13)70172-4 (2013)
13. Meyer-Rath G, Pillay Y, Blecher M, Brennan A, Long L, et al: Total cost and potential cost savings of the national antiretroviral treatment (ART) programme in South Africa 2010 to 2017. Abstract no. WEAE0201, XVIII International AIDS Conference 2010
14. Meyer-Rath G, Pillay Y, Blecher M, Brennan A, Long L, et al: The impact of a new reference price list mechanism for drugs on the total cost of the national antiretroviral treatment programme in South Africa 2011 to 2017. Abstract no. 621, South African AIDS Conference 2011
15. Meyer-Rath G, Brennan A, Fox MP, Modisenyane T, Tshabangu N, et al: Rates and cost of hospitalisation before and after initiation of antiretroviral therapy in the urban and rural public sector of South Africa. JAIDS 62(3):322-328 (2013)
16. Meyer-Rath G, Violari A, Cotton M, Ndibongo B, Brennan A, et al: The cost of early vs. deferred paediatric antiretroviral treatment in South Africa – A comparative economic analysis of the first year of the CHER trial. Abstract no. THLBB103 (oral presentation, late breaker), XVIII International AIDS Conference 2010
17. Meyer-Rath G, Brennan A, Long L, Ndibongo B, Technau K, et al: Cost and outcomes of paediatric antiretroviral treatment in South Africa. AIDS 27(2):243-250 (2012)

18. Salomon JA, Vos T, Hogan DR, Gagnon M, Naghavi M 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 380:2129-42 (2012)
19. Cohen S: Unit cost notes for costing Provincial Strategic Implementation Plans for HIV AIDS, STIs and TB. Strategic Development Consultants, June 2013.
20. District Health Information Software, South Africa. Software from the Health Information Systems Programme (HSIP). June 2013.
21. Marseille E, Giganti MJ, Mwango A, Chisembele-Taylor A, Mulenga L, et al. Taking ART to Scale: Determinants of the Cost and Cost-Effectiveness of Antiretroviral Therapy in 45 Clinical Sites in Zambia. PLoS ONE 7(12): e51993. doi:10.1371/journal.pone.0051993 (2012)
22. Over M, Schneider MT, Velayudhan T: Explaining the variation in on-site AIDS treatment costs: the MATCH study of 161 facilities from five countries. Lancet 381,S106 (2013)
23. South African Reserve Bank: Repurchase rate. Available at http://www.resbank.co.za/Research/Rates/Pages/Repo%20Rate.aspx (accessed 9 September 2013)
24. Resch S, Korenromp E, Stover J, Blakley M, Krubiner C, et al Economic Returns to Investment in AIDS Treatment in Low and Middle Income Countries. PLoS ONE 6(10): e25310 (2011)
25. Stover J, Bollinger L, Walker N, Monasch R.: Resource needs to support orphans and vulnerable children in sub-Saharan Africa. Health Policy Plan 22(1):21-7 (2007).
26. Profile of South Africa. Application of the poverty lines on the Living Conditions Survey 2008/2009. Pretoria, 2009
27. World Bank: Gross National Income, Atlas method (current USUSD). Available at http://data.worldbank.org/indicator/NY.GNP.ATLS.CD?display=default (last accessed on 23 Dec 2013)
28. Statistics South Africa: Census 2011: Methodology and highlights of key results. Report No, 03-01-42. Pretoria, 2012
29. Rosen S, Fox MP: Retention in HIV Care between Testing and Treatment in Sub-Saharan Africa: A Systematic Review. PLoS Med 8(7): e1001056. doi:10.1371/journal.pmed.1001056 (2011)
30. Menzies NA, Berruti AA, Blandford JM The Determinants of HIV Treatment Costs in Resource Limited Settings. PLoS ONE 7(11): e48726. doi:10.1371/journal.pone.0048726 (2012)
31. Over M, The effect of scale on cost projections for a primary health care programme in a developing country. Soc Sci Med 22(3):351-60 (1986)
32. Meyer-Rath G, Kumaranayake L, Variava E, Venter F: Transport costs of patients accessing antiretroviral treatment (ART) in Gauteng and the North-West province. South African AIDS Conference 2007
33. Fox MP, Shearer K, Maskew M, Macleod W, Majuba P, Macphail P, Sanne I: Treatment outcomes after 7 years of public-sector HIV treatment. AIDS 26:1823–1828 (2012)
34. Naik R et al. Factors associated with client linkage to care following home-based HIV counseling and testing: a prospective cohort study in rural South Africa. 7th IAS Conference on HIV Pathogenesis, Treatment and Prevention, Kuala Lumpur, abstract MOAD0101, July 2013. View the abstract on the IAS conference website.

Appendix A: Model Description EMOD-HIV *v0.8*

All epidemiological simulations were performed using EMOD-HIV *v0.8*, a population model of HIV transmission calibrated to the national-level epidemic in South Africa(1). It builds upon EMOD-HIV *v0.7*, an individual-based stochastic simulation of sexual and vertical HIV transmission in a generalized epidemic setting. We have previously published the epidemiological, behavioral, and transmission parameters of EMOD-HIV *v0.7*(2).

The model simulates transmission between individuals who are paired in partnerships. Relationships are dynamically formed and dissolved at age- and gender-dependent rates governed by a pair formation algorithm (PFA), which has been described in mathematical detail(3). Briefly, the PFA establishes age- dependent rates of relationship formation necessary to reproduce the age patterns of relationships reported in a longitudinal household survey conducted in KwaZulu-Natal, South Africa (4). The PFA then fixes these rates in order to allow the age pattern of partnerships to vary with future demographic changes.

Transmission within partnerships is simulated at the level of individual coital contacts. The model accounts for "coital dilution," i.e., less frequent contacts when either partner is engaged in additional concurrent partnerships. The transmission rate per coital act depends on disease stage, condom usage, STI status, circumcision status, and other factors described previously(2).

EMOD-HIV *v0.8* includes the time-dependent scale-up of HIV treatment by simulating the "cascade of care" (Figure S1) in which testing rates, linkage rates, and eligibility criteria were allowed to vary over time according to the scale-up of testing and treatment in South Africa. Testing and linkage rates were also allowed to vary by gender to fit the number on treatment by gender and year(5).

HIV testing in EMOD-HIV can occur as a result of voluntary testing by sexually active adolescents and adults, antenatal testing at 14 weeks gestation, infant HIV RNA testing, couples testing in which seropositive individuals recruit regular partners to test at a follow-up visit, or symptomatic testing at CD4 counts below 200 cells/μL.

Rates of voluntary testing and counseling vary by calendar year and by age, with some individuals beginning regular testing shortly after sexual debut, and others beginning later in life. The rates were set to match the self-reported proportion of individuals ever tested and tested in the last year according to national-level survey data(6–8), and adjusted to match rates of ART initiation(5) and CD4 counts at ART enrollment(9, 10). For males, the probability of beginning regular testing at debut is 0.5% during the early epidemic, and grows to 25% in the present year and onward. The most rapid growth in at-debut testing occurs in 2000. The annual rate of beginning regular testing after debut also grows over time, from zero in 1998 to 4% in 2003 and 27% in 2009 and onward. For females, all these rates are increased by 30%. Unlike other forms of testing, voluntary testing can only produce one positive test result. It is assumed that individuals who test positive and subsequently fail to link or drop out of pre-ART or ART care will not resume voluntary testing after loss to follow-up. However, these individuals can re-test and link to care via antenatal testing, couples testing, or symptomatic testing.

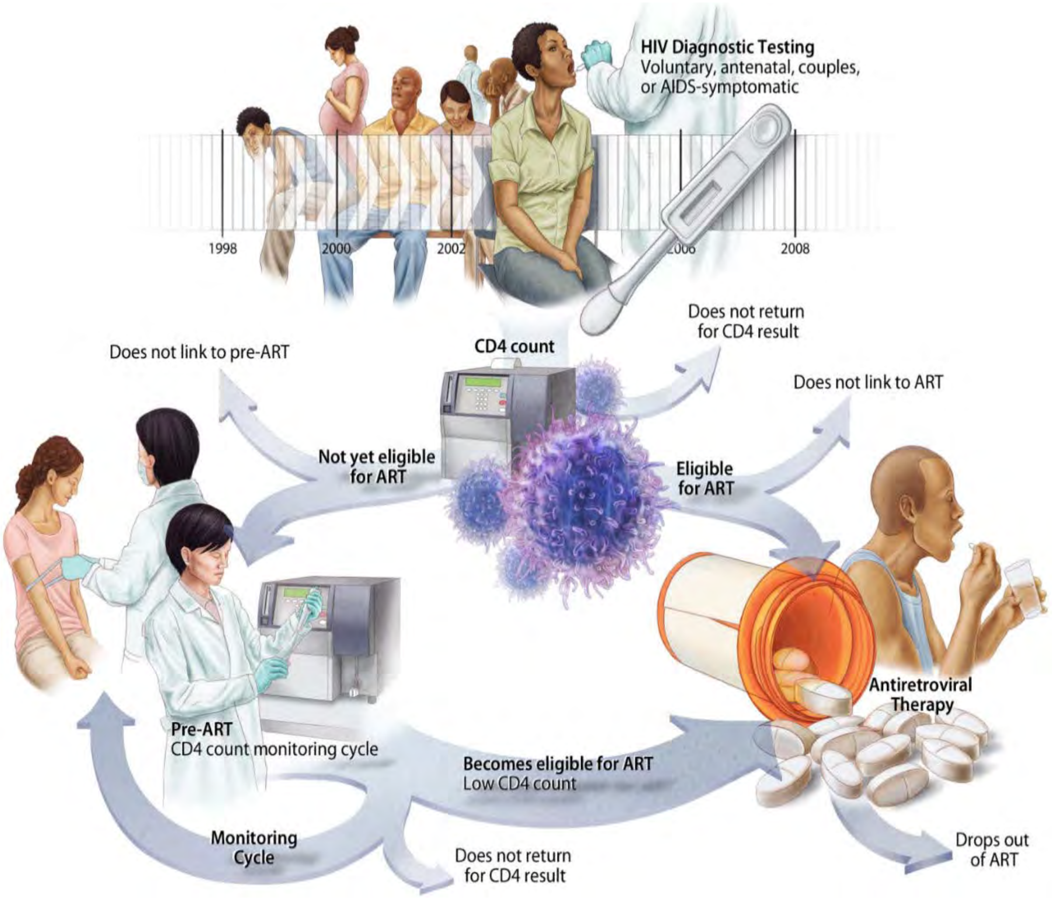


Figure S1. Illustration of the "cascade of care" modeled in EMOD-HIV *v0.8*. Individuals enter the cascade by receiving a diagnostic test motivated by voluntary, antenatal, couples, infant, or symptomatic testing. Individuals must not receive false negative test (98%) and must return for a CD4 count result (59% without an expanded access intervention) in order to be evaluated for treatment eligibility. Ineligible individuals may link to pre-ART care and return for semiannual CD4 monitoring, while eligible individuals may link to ART, with probabilities that change according to gender and over the course of ART scale-up. Eligibility guidelines also change in accordance to 2004, 2010, and 2011 changes in treatment guidelines in South Africa, and may change in 2014 to simulate a guideline change intervention.

Antenatal testing was assumed to occur at 14 weeks gestation and to vary by calendar year. The rates were estimated by multiplying the time-variable coverage of antenatal services by the time-variable rate of HIV testing and counseling in antenatal clinics(11–13). This ANC testing probability increased from 0% in 2000 to 7% in 2001, 58% in 2002, and 85% in 2006, with linear interpolation between these time points.

Couples testing was assumed to occur after an individual tests, receives a CD4 count, is deemed not yet eligible for ART, and successfully links to pre-ART care. Pre-ART monitoring visits are assumed to occur every six months. If the individual has an active sex partner at the time of the pre-ART visit, the probability of bringing the partner for HIV testing is assumed to be 10%. If the individual has multiple partners, the longest-standing partner is brought for testing. The partner receives HIV testing and counseling at this time, and thus may enter the treatment cascade via the couples testing modality. Because of low rates of linkage to pre-ART care and the low baseline rate of partner recruitment, partner testing is not a significant source of ART initiations at baseline.

Finally, individuals receive an HIV test when symptomatic, which is assumed to occur at a CD4 count below 200 cells/μL, chosen randomly between 200 and 100 cells/μL for some individuals and between 100 and 0 cells/μL for other individuals. The proportion who test due to AIDS-related symptoms above versus below 100 cells/μL was adjusted to match the number of individuals on treatment over time(5) and CD4 counts at ART initiation(9, 10). The CD4 count at symptomatic presentation is assumed to be between 200 and 100 cells/μL for 40% of males and 50% of females, and below 100 cells/μL for 60% of males and 50% of females.

Only 80% of the population could access all four modes of testing (voluntary, antenatal, couples, and symptomatic). This was the same 80% of the population that was deemed to be accessible by improvements in testing and linkage rates in the expanded access scenarios. The remaining 20% received only antenatal and AIDS-symptomatic testing, and were not affected by health care improvements in the expanded access scenarios.

Two percent of individuals receive a false negative diagnostic test(14, 15) and therefore do not link to care, but could potentially re-test at a future time. After receiving a positive HIV test, a proportion of individuals return for a CD4 result and determination of ART eligibility. The probability of doing so is 59% at baseline(16), increasing to 100% in 2014 in scenarios with expanded access to care. Changes in guidelines were assumed to have no effect on the proportion retained in this stage of the cascade, even in scenarios where eligibility was independent of CD4 count.

EMOD-HIV *v0.8* accounts for South African national guideline changes in 2004, 2010, and 2011, as well as a possible guideline change in 2014 depending on the scenario. Eligible individuals may link to ART; ineligible individuals may link to pre-ART. The probabilities of linking to pre-ART and ART were adjusted along with other parameters to match the total number enrolled in ART over time(5) and CD4 counts at ART initiation(9, 10).

The probability of returning for each subsequent 6-monthly pre-ART monitoring visit was assumed to be the same as the pre-ART linkage probability. Retention time on ART is exponentially distributed with a mean of 10 years, with a dropout probability of 9.5% in the first year.

The model was calibrated to HIV prevalence by gender, age, and year; proportion ever tested and recently tested by gender, age, year; number on ART by gender, age, and year; CD4 count at ART enrollment by gender, and total population by year. The sources of data used to fit the model are listed in Table S1.

Table S1. Outputs and sources of data used in calibration

Output	**Data**	**Source**
CD4 at ART initiation	By gender and category [<100, 100-200, 200-350] for 2009-2010	Cornell et al., *AIDS* 24(14): 2010 Cornell et al., *PLoS Med* 9(9): 2012.
Number on ART	By gender and age group 15-49 and 50+ for 2004-2012	UNAIDS Spectrum
Population	Estimate of 15-49 population for 1981-2012	ASSA 2008
Prevalence by year, gender, and age	By gender and age groups 15-49 and 50+ for years 1994-2012	UNAIDS Spectrum
Cross-sectional prevalence by gender and 5-year age group	Prevalence by age group and gender in years 2002, 2005, and 2008	Human Sciences Research Council Surveys of South Africa: 2002, 2005, and 2008
Long-term testing	Ever tested by gender in 2002, 2005, and 2008	Human Sciences Research Council Surveys of South Africa: 2002, 2005, and 2008
Recent testing	Tested past year by gender in 2006 and 2009	2009 National HIV Communication Survey of South Africa
Cross-sectional testing	Ever tested by age and gender in 2009	2009 National HIV Communication Survey of South Africa

For a given model run, the likelihoods that the model output could be consistent with each source of data (Table S1) were multiplied across the data sources to give a single score for the model run's fit to the epidemic. These fitness scores were then averaged across fifty stochastic run of the model using the same parameter settings. Because of the large number of data points included in the multiplication, typical values for a model that fit well by visual inspection were 10^{130} and higher.

To compare CD4 counts at ART enrollment to available data, we accumulated CD4 counts in the model over all ART initiations falling within the study interval of CD4 count data(9, 10). For each reported CD4 count category, we calculated the model's proportion of individuals initiating ART within this category. The likelihood score was evaluated as the value of the model's outcome in a Gaussian probability distribution characterised by the mean and standard deviation of the study data.

For the number receiving ART by year and gender, we similarly calculated the probability of the model value in a Gaussian probability distribution, but this time we used the logit-transform of both data and model values to ensure non-negative values of the distribution. Number on ART was assumed to have 0.5% error in logit space, except prior to 2002, when a 10-fold increase to 5% variance to was assumed. This is because our data source did not account for the possibility of small numbers of individuals receiving ART through the private sector. Testing rates by year and by age were similarly evaluated as a Gaussian distribution of logit-transformed data.

We calibrated prevalence by gender and year based on UNAIDS estimates, and prevalence by gender and age for three years based on household survey data. Because this provided redundant total prevalence information, survey data was used to calibrate the normalized age distribution of prevalence, while UNAIDS estimates were used for the total prevalence across all age groups.

Calibration was performed with iterative rounds of Incremental Mixture Importance Sampling, a Bayesian technique that returns samples from the posterior distribution given a likelihood function and a prior distribution for the relevant model parameters. After calibration of the pre-intervention baseline, we used the parameter configuration yielding the maximum *a posteriori* probability to simulate the scenarios.

References for Appendix A

1. Klein DJ, Bershteyn A, Eckhoff PA (Forthcoming) Dropout and Re-Enrollment: Implications for Epidemiological Projections of Treatment Programmes. *AIDS*.
2. Bershteyn A, Klein DJ, Wenger E, Eckhoff PA (2012) Description of the EMOD-HIV Model v0.7. *arXiv:12063720*. Available at: http://arxiv.org/abs/1206.3720.
3. Klein DJ (2012) in *2012 51st Annual Conference on Decision and Control (CDC)*, pp 1041–1046.
4. Ott MQ, Bärnighausen T, Tanser F, Lurie MN, Newell M-L (2011) Age-gaps in sexual partnerships: seeing beyond "sugar daddies." *AIDS* 25:861–863.
5. Adam MA, Johnson LF (2009) Estimation of adult antiretroviral treatment coverage in South Africa. *SAMJ South Afr Med J* 99:661–667.
6. Shisana O, Simbayi L, et al (2002) *Nelson Mandela/HSRC study of HIV/AIDS: South African national HIV prevalence, behavioural risks and mass media: household survey 2002* (HSRC Press)
7. Shisana O, Simbayi L, Council SAMR (2008) *South African National HIV Prevalence, HIV Incidence, Behaviour and Communication Survey, 2005* (HSRC Press).
8. Shisana O, Council HSR, Evaluation C for AD, Research and, Council SAMR, Africa) NI for CD (South (2010) *South African National HIV Prevalence, Incidence, Behaviour and Communication Survey, 2008: A Turning Tide Among Teenagers?* (HSRC Press).
9. Cornell M et al. (2010) Temporal changes in programme outcomes among adult patients initiating antiretroviral therapy across South Africa, 2002-2007. *AIDS Lond Engl* 24:2263–2270.
10. Cornell M et al. (2012) Gender Differences in Survival among Adult Patients Starting Antiretroviral Therapy in South Africa: A Multicentre Cohort Study. *PLoS Med* 9:e1001304.
11. April MD et al. (2009) HIV testing rates and outcomes in a South African community, 2001-2006: implications for expanded screening policies. *J Acquir Immune Defic Syndr 1999* 51:310–316.

12. Kranzer K et al. (2011) High Prevalence of Self-Reported Undiagnosed HIV despite High Coverage of HIV Testing: A Cross-Sectional Population Based Sero-Survey in South Africa. *PLoS ONE* 6:e25244.
13. Venkatesh KK et al. (2011) Who gets tested for HIV in a South African urban township? Implications for test and treat and gender-based prevention interventions. *J Acquir Immune Defic Syndr 1999* 56:151–165.
14. Pavie J et al. (2010) Sensitivity of Five Rapid HIV Tests on Oral Fluid or Finger-Stick Whole Blood: A Real-Time Comparison in a Healthcare Setting. *PLoS ONE* 5:e11581.
15. Bassett I et al. (2011) Screening for acute HIV infection in South Africa: finding acute and chronic disease. *HIV Med* 12:46–53.
16. Rosen S, Fox MP (2011) Retention in HIV Care between Testing and Treatment in Sub-Saharan Africa: A Systematic Review. *PLoS Med* 8:e1001056.

Appendix B: Mathematical details of the derivation of the cost and demand functions used in estimating the cost, cost effectiveness and cost benefit of expansion options for adult ART in South Africa

In this paper, we model the number of new cases of HIV and disability-adjusted life years (DALYs) averted and the cost to the South African government of scaling up the public production of antiretroviral treatment services over the period 2014 to 2033 according to a number of alternative scenarios. The scenarios differ, first, by the treatment guidelines they follow, and second by the economic assumptions regarding two distinct categories of cost: production cost and outreach cost. In this appendix we give details of the calculation of both of these cost categories, as well as the calculations used in computing DALYs averted, incremental cost effectiveness, the cost of orphanhood avoided and the value of productivity regained, and, finally, incremental cost benefit.

Modeling the cost of ART service delivery

As explained in the text, we distinguish six components of total cost. We conservatively assume that only one of these six components, that composed of facility maintenance, personnel and other overhead expenditures, is scale dependent, i.e. that its average is sensitive to the number of patient-years of ART a facility produces. We assume that the unit costs of the remaining components, ARVs, laboratory supplies, HIV testing, and inpatient care, are all insensitive to the number of patient-years of ART produced at each outpatient facility[8]. In the baseline scenario, the scale dependent component is less than a third of total cost.

Modeling economies of scale

We model the scale-dependent cost of producing ART services using a version of the simplified cost function presented in supplementary materials S2 of [11].

$$tc_{kt} = A\, n_{kt}^{\sigma} \qquad (1)$$

where the total scale-dependent cost at facility *k* in year t is a function of a constant *A*[9], the number of patient-years of treatment the facility delivers in that year, n_{kt}, and the assumed elasticity of this cost component with respect to scale, *σ*. (We abstract from other likely determinants of scale-dependent cost at a specific facility such as local prices of maintenance supplies and patient characteristics other than their health states). Values of *σ* less than one

[8] Although the cost of inpatient care is also subject to economies of scale, this is beyond the scope of our paper, as it would require information on the scale of each inpatient facility in South Africa, especially since HIV-positive patients make up only a subset of all patients assessing inpatient care.

[9] While the parameter *A* is often treated as a constant and referred to as the "unit cost", in general it varies with the mix of patient health states at facility k in year t. A health state is defined by type of care (no ART, treatment initiation, first line, first-line treatment failure, second line), CD4 cell count stratum (for inpatient cost only) and, for children, additionally by age group.

characterise a cost structure which generates economies of scale. Dividing equation (1) by the number of patients served during the year in that facility, *nkt*, yields the expression for the unit or average cost per patient:

$$uc_{kt} = atc_{kt} = {}^{A\, n_{kt}^{\sigma}}\!/_{n_{kt}} = A\, n_{kt}^{\sigma-1} \qquad (2)$$

As described in the text, the model allows the growth in clinic size to follow one of three different expansion paths, one informed by the arithmetic mean number of patients per facility which gets added to each clinic, one allocating the same proportional growth to all clinics, and lastly a pattern following a quadratic equation. The quadratic distribution pattern is intermediate between the arithmetic and proportional patterns. For this pattern, we impose the constraint that neither the largest nor the smallest facility absorbs more patients, with the incremental 2.9 million patients being distributed across the middle of the range of facility ranks according to the quadratic relationship in equation (3):

$$\Delta n_{kt} = n_{kt} - n_{k0} = a_t + b_t\, k + c_t\, k^2\ \forall\ k = 1, \dots, K_t \qquad (3)$$

where n_{k0} is the number of patients served by facility *k* in the baseline year, n_{kt} is the number of patients to be served in a future year, Δn_{kt} is the increase in the number of patients served between the two years and *k* is an index of the size-rank of the facility ordered from the largest for $k = 1$ to the smallest for $k = K_t$. The time subscripts on K_t and on the parameters capture the possibility that the number of facilities might increase over time to accommodate the growing number of patients and, in particular, to facilitate access for patients living in rural areas. To find the values of parameters a_t, b_t and c_t in year *t*, we solve the following system of three equations for those parameters:

$$\begin{gathered} \Delta N_t = \textstyle\sum_{k=1}^{K_t} \Delta n_{kt} = \sum_{k=1}^{K_t} [a_t + b_t\, k + c_t\, k^2] \\ a_t + b_t\, 1 + c_t\, 1 = 0 \\ a_t + b_t\, K_t + c_t\, K_t^2 = 0 \end{gathered} \qquad (4)$$

The first of these equations requires that ΔN_t, the total number of patients added to all facilities in the country by year *t*, be the sum of the patients added to each facility by that year. The second and third equations impose the constraints that zero patients be added to the largest and smallest facility in that year. Solving the three equations for the three unknown parameters yields:

$$\begin{aligned} a_t &= \frac{6\, \Delta N_t}{(1 - K_t)(K_t - 2)} < 0 \\ b_t &= \frac{6\, \Delta N_t\, (K_t + 1)}{K_t(K_t - 1)(K_t - 2)} > 0 \\ c_t &= \frac{6\, \Delta N_t}{K_t(1 - K_t)(K_t - 2)} < 0 \end{aligned} \qquad (5)$$

Since ΔN_t is provided by EMOD-HIV, by modeling an expansion path for the total number of facilities, K_t, we can find the values of parameters a_t, b_t and c_t in any year *t*, and then use equation (6), derived from equation (3), to distribute the patients across the K_t facilities in that year. The comparable equations for the *average* and the *proportional* scale-up programmes are equations (7) and (8).

$$n_{kt} = n_{k0} + a_t + b_t\, k + c_t\, k^2 \;\forall\, k = 1, \dots, K_t \qquad (6)$$

$$n_{kt} = n_{k0} + {}^{\Delta N_t}\!/_{K_t} \;\forall\, k = 1, \dots, K_t \qquad (7)$$

$$n_{kt} = n_{k0} \times {}^{N_t}\!/_{N_0} \;\forall\, k = 1, \dots, K_t \qquad (8)$$

Modeling the cost of outreach

For a given facility at a given level of care and in a given year, we set the upper bound of enrollment as equal to the maximum enrollment achieved in that facility under any scenario, n_k^{max}, inflated by the assumed maximum coverage attained at that maximum enrollment, ω. Thus, in a given year at a given facility, the degree of enrollment scale up can be defined by the ratio of actual enrollment to this theoretical maximum, ${}^{n_{kt}}\!/_{n_k^{max}}$. Define this proportion, *q*, for both the status quo scenario and the actual scenario as follows:

$$\begin{aligned} \tilde{q}_{kt} &= {}^{\tilde{n}_{kt}}\Big/_{\frac{n_k^{max}}{\omega}} \\ q_{kt} &= {}^{n_{kt}}\Big/_{\frac{n_k^{max}}{\omega}} \end{aligned} \qquad (9)$$

where $0 < \tilde{q}_{kt} < q_{kt} < 1$ and $0 < \omega < 1$.[10] With these definitions, we can write the demand function for ART services

$$q_{kt} = \frac{1}{1-e^{-d_t+e\,(-v_{kt})}}$$

(10)

where d_t and *e* are positive parameters and $v_{kt} > 0$ is the per patient outreach expense for each patient more than would be recruited in the status quo scenario. For example, the outreach expense could be a travel voucher, or could consist of the costs of making home visits to the patient. Equation (10) can also be written:

$$ln \frac{q_{kt}}{(1- q_{kt})} = d_t + e\, v_{kt}\ .$$

(11)

[10] [1] uses a similar logistic specification of demand.

In order to calibrate each facility's demand curve to predict the status quo demand at a zero price, we define the intercept d_t in terms of the status quo ratio of patients to the maximum number:

$$d_t = \ ln\frac{\tilde{q}_{kt}}{(1- \tilde{q}_{kt})} \, .$$

(12)

If we were predicting the utilisation level for a given level of outreach expenditure, we would use equation (11) directly. However, since the number of patients is generated by the epidemiologic models (and our selected distribution algorithm), we instead solve equation (11) to obtain the inverse demand function which provides the outreach cost per patient required to attract and retain at the specified adherence any given number of patients.

$$v_{kt} = \frac{\left[ln\frac{q_{kt}}{(1- q_{kt})} - ln\frac{\tilde{q}_{kt}}{(1- \tilde{q}_{kt})}\right]}{e}$$

(13)

The positive parameter *e* can be related to the conventional idea of an elasticity of demand. A larger value of *e*, like a larger elasticity of demand, endows this inverse demand function with greater sensitivity to the voucher or outreach expenditure, so that a given level of enrollment and adherence can be obtained at less cost per patient. Conversely a smaller value of *e* corresponds to a low elasticity and implies that a given level of patient utilization requires a larger outreach cost. With reference to Figure 4, a more elastic demand curve would be flatter than the one shown and a less elastic one would be steeper.

In the modeled scenarios, the default value of the "saturation" parameter, ω, varies between 0.7 for the primary health care facilities, most of which are in relatively dispersed in rural areas, to 0.9 for the level national and provincial reference hospitals which are located in urban areas and assumed to have saturated their local markets. In our baseline runs we set the demand elasticity parameter, *e*, equal to 1.0 for urban facilities and 0.5 for rural facilities. These values imply that, other things equal, a larger percentage increase in the travel voucher will be required to entice new patients in the more sparsely populated rural area than in the densely populated urban area.

Taken together the model of service delivery cost which incorporates economies of scale and the model of outreach cost incorporating various response elasticities allow the overall model to characterise the human behavioral responses, on the supply side, of facility and programme managers economizing scarce health care inputs and, on the demand side, of patients balancing their desire for good health against their other individual and social needs and desires. Accurate models of the future cost and success of treatment scale-up depend, *inter alia*, on reliable empirical estimates of the key parameters in these models. On the supply side, parameters in need of estimation include the elasticity of each cost component to scale and the dependency of these cost components on other features of the economic and institutional environment, such as the local prices of inputs and the structure of personnel rewards and sanctions. On the demand side, these needed parameters include the elasticity of enrollment and adherence with respect to outreach expenses, as modeled here.

Calculating incremental cost per infection or DALY averted

In this analysis, the cost-effectiveness of scenario *x* over a comparator scenario *b* (e.g. G0.SQ) is defined as the incremental cost of scenario *x* per incremental infection or per DALY averted.

1. The number of HIV infections for each scenario is generated by EMOD-HIV; the incremental number of infections is simply the difference between the number of infections in scenario *x* over scenario *b*. The incremental number of DALYs averted is calculated as the sum of all life years lived by each HIV-positive individual in a particular health state across the 20 years of projection multiplied by the utility weight (set equal to 1 - the disability weight) for this health state, according to equation (14):

$$IncDALYs_{xb} = \sum_{t\in T}\sum_{i\in I} n_{itx}\,(1-d_i)\,(1+r)^{-(t)} - \sum_{t\in T}\sum_{i\in I} n_{itb}(1-d_i)\,(1+r)^{-(t)} \qquad (14)$$

where $IncDALYs_{\text{xb}}$ is the incremental cost of scenario *x* relative to scenario *b*, *T* the total number of years *t* in the model, *I* the total number of health states *i* in the model, n_{itx} the number of HIV-infected people in health state *i* and year *t* under scenario *x*, d_i the disability weight for health state *i*, and *r* the discount rate (3% at baseline, varied to 0%, 5% and 10% in sensitivity analysis). The second term in equation (14) computes the number of DALYs for scenario *b,* and the incremental number of DALYs averted is the difference between the two terms.

2. The numerator for the incremental cost-effectiveness ratios, incremental cost, is defined by equation (2):

$$IC(hiv)_{xb} = \sum_{t\in T}\sum_{i\in I}\sum_{k\in K_t} n_{iktx}uc_{iktx}\,(1+r)^{-(t)} - \sum_{t\in T}\sum_{i\in I}\sum_{k\in K_t} n_{iktb}uc_{iktb}\,(1+r)^{-(t)} \qquad (15)$$

where $IC(hiv)_{\text{xb}}$ is the incremental cost of HIV-related care in scenario *x* relative to scenario *b*, K_t the total number of facilities delivering ART in year t, n_{iktx} the number of HIV-infected people in health state *i*, facility *k*, and year *t* under scenario *x*, and uc_{iktx} the per-person year cost of HIV-related care for this health state in this facility and year. Again, the second term in equation (15) computes the same cost for scenario *b,* and the incremental cost is the difference between the two terms. As discussed in the text and in our previous publication [11], the per-person year cost for a given year and health state and scenario, uc_{iktx}, is not a fixed constant, but rather a function of the year *t* supply- and demand-side characteristics of the *k'th* facility in which it is produced.

Calculating the cost of orphanhood and the value of productivity

In order to quantify the benefits of ART provision under any scenario of eligibility and coverage, we calculated the number of maternal orphans under each scenario and allocated the cost of orphan care to each of them, then calculated the increment for each scenario over baseline scenario *b* using equation (16):

$$IC(orphanhood)_{xb} = \sum_{t\in T}\sum_{a\in A} n_{atx}\, p(ovc)\, uc_a\, (1+r)^{-(t)} - \sum_{t\in T}\sum_{a\in A} n_{atb}\, p(ovc)\, uc_a\, (1+r)^{-(t)} \qquad (16)$$

where *IC(orphanhood)*$_{xb}$ is the incremental cost of HIV-related care in scenario *x* relative to scenario *b*, *A* the total number of age groups *a*, n_{atx} the number of maternal orphans in age group *a* and year *t* under scenario *x*, $p(ovc)$ is the percentage of orphans needing care and support (modelled on the percentage of the population living below the poverty line (upper bound) [21]) and uc_a the per-person year cost of orphan care for this age group. As above, the second term in equation (16) computes the same cost for scenario *b,* and the incremental cost is the difference between the two terms.

Treatment benefits in terms of regained productivity were calculated based on the number of adults of working age multiplied by the Gross National Income per working-age person and health-state specific productivity weights:

$$IP_{xb} = \sum_{t\in T}\sum_{i\in I} n_{itx}\, p(wa)_x\, pw_i\, GNI_{wa}\, (1+r)^{-(t)} - \sum_{t\in T}\sum_{i\in I} n_{itx}\, p(wa)_b\, pw_i\, GNI_{wa}\, (1+r)^{-(t)} \qquad (17)$$

where *IP*$_{xb}$ is the incremental productivity in scenario *x* relative to scenario *b*, n_{itx} the number of adults with HIV in health state *i* and year *t* under scenario *x*, $p(wa)_x$ is the percentage of adults being of working age, pw_i the productivity weight for health state *i*, and GNI_{wa} the Gross National Income per working-age adult. As above, the second term in equation (17) computes the same cost for scenario *b,* and the incremental cost is the difference between the two terms.

Calculating the incremental cost benefit of ART

Based on equations (15), (16) and (17), we calculate the incremental cost benefit of ART of each scenario over the baseline scenario (ICB_{xb}) as the incremental HIV-related cost minus the incremental cost of orphanhood averted and divide this by the incremental value of gained productivity IP_{xb}, following equation (18):

$$ICB_{xb} = (IC(hiv)_{xb} - IC(orphanhood)_{xb})/IP_{xb} \qquad (18)$$

References for Appendix B

1. Over M, Revenga A, Masaki E, Peerapatanapokin W, Gold J, Tangcharoensathien V, Thanprasertsuk S. The economics of effective AIDS treatment in Thailand. AIDS 21 Suppl 4:S105-16 (2007)

Appendix C: Annual undiscounted cost by year and for the 2014-2016 mid-term expenditure framework for all scenarios [2013 USD]

Undiscounted annual cost by 30 June of year	***2014***	***2015***	***2016***	***2014-2016***	***2017***	***2018***	***2019***	***2020***	***2021***	***2022***	***2023***	***2024***	***2025***	***2026***	***2027***	***2028***	***2029***	***2030***	***2031***	***2032***	***2033***
CG.SQ																					
Total cost (all patients)	1,856	1,963	2,045	**5,865**	2,117	2,181	2,235	2,286	2,330	2,374	2,417	2,455	2,490	2,528	2,559	2,591	2,618	2,645	2,671	2,694	2,713
Total cost (patients on ART)	1,272	1,387	1,477	**4,136**	1,557	1,622	1,678	1,729	1,774	1,819	1,862	1,898	1,932	1,970	2,001	2,033	2,060	2,086	2,113	2,136	2,156
Total cost (patients off ART)	395	384	373	**1,153**	363	359	355	354	351	349	347	347	346	345	343	341	340	339	336	335	332
Total cost (testing)	189	192	195	**576**	197	200	202	203	205	206	208	210	211	213	215	216	218	219	222	223	225
CG.UE																					
Total cost (all patients)	2,283	2,759	2,870	**7,912**	2,984	3,086	3,171	3,252	3,325	3,386	3,443	3,497	3,544	3,590	3,628	3,665	3,700	3,734	3,741	3,763	3,780
Total cost (patients on ART)	1,570	1,966	2,134	**5,671**	2,274	2,390	2,488	2,576	2,654	2,718	2,777	2,831	2,877	2,922	2,958	2,991	3,024	3,052	3,055	3,068	3,080
Total cost (patients off ART)	338	261	221	**821**	196	181	165	151	140	130	123	115	107	101	96	90	85	82	78	76	73
Total cost (testing)	374	531	515	**1,420**	513	516	518	525	531	538	544	551	559	567	574	583	591	601	608	618	628
UTT.SQ																					
Total cost (all patients)	1,912	2,079	2,198	**6,189**	2,290	2,356	2,408	2,457	2,496	2,522	2,544	2,572	2,594	2,614	2,629	2,645	2,658	2,672	2,686	2,696	2,703
Total cost (patients on ART)	1,339	1,523	1,661	**4,523**	1,767	1,844	1,905	1,958	2,001	2,032	2,059	2,084	2,109	2,128	2,144	2,159	2,171	2,185	2,199	2,207	2,215
Total cost (patients off ART)	384	364	342	**1,090**	326	314	302	297	290	285	278	278	274	273	270	269	267	265	262	262	258

Total cost (testing)	189	192	195	**575**	197	198	200	202	204	206	207	209	211	213	215	217	219	222	225	227	230
UTT.UE																					
Total cost (all patients)	2,479	3,023	3,087	**8,589**	3,141	3,183	3,217	3,247	3,277	3,302	3,322	3,342	3,361	3,379	3,391	3,404	3,414	3,422	3,429	3,433	3,438
Total cost (patients on ART)	1,812	2,308	2,431	**6,551**	2,505	2,557	2,593	2,624	2,652	2,676	2,695	2,712	2,727	2,739	2,747	2,754	2,759	2,760	2,760	2,756	2,754
Total cost (patients off ART)	292	184	141	**618**	122	108	101	93	86	79	73	69	63	60	56	53	50	47	44	42	40
Total cost (testing)	375	531	515	**1,420**	514	518	523	531	539	547	554	562	570	579	587	596	606	616	625	635	644
500.SQ																					
Total cost (all patients)	1,900	2,052	2,169	**6,121**	2,256	2,319	2,374	2,422	2,459	2,492	2,521	2,548	2,573	2,596	2,622	2,640	2,659	2,676	2,689	2,701	2,720
Total cost (patients on ART)	1,323	1,494	1,625	**4,443**	1,725	1,802	1,864	1,914	1,957	1,992	2,023	2,050	2,075	2,098	2,122	2,142	2,159	2,174	2,189	2,201	2,217
Total cost (patients off ART)	387	366	349	**1,102**	334	319	309	305	298	294	291	289	287	286	285	282	282	280	277	275	275
Total cost (testing)	189	192	195	**575**	197	199	200	202	204	206	208	209	211	213	215	217	218	223	223	226	228
500.UE																					
Total cost (all patients)	2,441	2,999	3,080	**8,520**	3,143	3,189	3,227	3,263	3,294	3,323	3,345	3,370	3,390	3,410	3,426	3,441	3,453	3,463	3,471	3,477	3,484
Total cost (patients on ART)	1,764	2,271	2,416	**6,450**	2,502	2,558	2,600	2,636	2,667	2,694	2,715	2,736	2,754	2,769	2,780	2,790	2,795	2,799	2,800	2,798	2,798
Total cost (patients off ART)	303	197	149	**648**	128	114	105	97	90	84	78	73	68	64	60	56	53	50	48	45	43
Total cost (testing)	375	531	515	**1,421**	513	517	522	530	538	545	552	561	569	577	586	596	604	614	623	633	644
DC.SQ																					
Total cost (all patients)	1,848	1,956	2,038	**5,842**	2,108	2,169	2,226	2,278	2,324	2,371	2,413	2,452	2,487	2,520	2,554	2,588	2,614	2,641	2,667	2,690	2,708

Total cost (patients on ART)	1,265	1,383	1,472	**4,120**	1,547	1,612	1,670	1,721	1,769	1,815	1,857	1,895	1,932	1,964	1,997	2,030	2,058	2,083	2,110	2,132	2,150
Total cost (patients off ART)	394	380	371	**1,145**	363	357	354	352	349	347	347	346	343	342	341	341	337	337	334	334	332
Total cost (testing)	189	192	195	**577**	198	200	202	204	206	209	209	211	213	214	216	218	218	221	223	224	226
DC.PE																					
Total cost (all patients)	2,083	2,414	2,544	**7,040**	2,644	2,731	2,806	2,873	2,937	2,987	3,040	3,091	3,137	3,182	3,226	3,263	3,302	3,340	3,372	3,406	3,435
Total cost (patients on ART)	1,435	1,754	1,931	**5,119**	2,057	2,161	2,246	2,318	2,385	2,441	2,495	2,547	2,595	2,641	2,685	2,722	2,760	2,795	2,827	2,860	2,888
Total cost (patients off ART)	371	318	277	**966**	251	234	221	209	204	194	188	182	175	172	166	161	157	152	148	143	138
Total cost (testing)	277	342	336	**955**	336	337	340	345	348	352	357	361	366	369	375	380	386	392	397	403	409
DC.UE																					
Total cost (all patients)	2,286	2,808	2,935	**8,030**	3,057	3,151	3,225	3,294	3,352	3,401	3,444	3,482	3,517	3,545	3,572	3,597	3,620	3,640	3,654	3,669	3,678
Total cost (patients on ART)	1,575	2,019	2,212	**5,807**	2,363	2,475	2,561	2,635	2,697	2,747	2,792	2,828	2,861	2,889	2,912	2,933	2,949	2,962	2,973	2,981	2,983
Total cost (patients off ART)	336	255	206	**798**	178	158	142	131	120	113	104	97	91	85	82	76	74	70	67	63	61
Total cost (testing)	375	534	517	**1,426**	515	518	521	528	534	541	548	556	565	570	579	588	597	608	615	624	634
PW.SQ																					
Total cost (all patients)	2,441	2,999	3,080	**8,520**	3,143	3,189	3,227	3,263	3,294	3,323	3,345	3,370	3,390	3,410	3,426	3,441	3,453	3,463	3,471	3,477	3,484
Total cost (patients on ART)	1,764	2,271	2,416	**6,450**	2,502	2,558	2,600	2,636	2,667	2,694	2,715	2,736	2,754	2,769	2,780	2,790	2,795	2,799	2,800	2,798	2,798
Total cost (patients off ART)	303	197	149	**648**	128	114	105	97	90	84	78	73	68	64	60	56	53	50	48	45	43

Total cost (testing)	375	531	515	**1,421**	513	517	522	530	538	545	552	561	569	577	586	596	604	614	623	633	644
PW.PE																					
Total cost (all patients)	1,848	1,956	2,038	**5,842**	2,108	2,169	2,226	2,278	2,324	2,371	2,413	2,452	2,487	2,520	2,554	2,588	2,614	2,641	2,667	2,690	2,708
Total cost (patients on ART)	1,265	1,383	1,472	**4,120**	1,547	1,612	1,670	1,721	1,769	1,815	1,857	1,895	1,932	1,964	1,997	2,030	2,058	2,083	2,110	2,132	2,150
Total cost (patients off ART)	394	380	371	**1,145**	363	357	354	352	349	347	347	346	343	342	341	341	337	337	334	334	332
Total cost (testing)	189	192	195	**577**	198	200	202	204	206	209	209	211	213	214	216	218	218	221	223	224	226
PW.UE																					
Total cost (all patients)	2,083	2,414	2,544	**7,040**	2,644	2,731	2,806	2,873	2,937	2,987	3,040	3,091	3,137	3,182	3,226	3,263	3,302	3,340	3,372	3,406	3,435
Total cost (patients on ART)	1,435	1,754	1,931	**5,119**	2,057	2,161	2,246	2,318	2,385	2,441	2,495	2,547	2,595	2,641	2,685	2,722	2,760	2,795	2,827	2,860	2,888
Total cost (patients off ART)	371	318	277	**966**	251	234	221	209	204	194	188	182	175	172	166	161	157	152	148	143	138
Total cost (testing)	277	342	336	**955**	336	337	340	345	348	352	357	361	366	369	375	380	386	392	397	403	409