BMJ Global Health

Economic evaluations of health system strengthening activities in low-income and middle-income country settings: a methodological systematic review

Nathaniel Hendrix [ID] ,[1] Xiaoxiao Kwete,[1,2] Sarah Bolongaita [ID] ,[1] Itamar Megiddo [ID] ,[3] Solomon Tessema Memirie [ID] ,[4] Alemnesh H Mirkuzie,[5] Justice Nonvignon,[6] Stéphane Verguet [ID] [1]

Handling editor Lei Si

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjgh-2021-007392).

Check for updates

For numbered affiliations see end of article.

Correspondence to
Professor Stéphane Verguet;
verguet@hsph.harvard.edu

ABSTRACT

Objective Health system strengthening (HSS) activities should accompany disease-targeting interventions in low/middle-income countries (LMICs). Economic evaluations provide information on how these types of investment might best be balanced but can be challenging. We conducted a systematic review to evaluate how researchers address these economic evaluation challenges.

Methods We identified studies about economic evaluation of HSS activities in LMICs using a two-stage approach. First, we conducted a broad search to identify areas where economic evaluations of HSS activities were being conducted. Next, we selected specific interventions for more targeted literature review. We extracted study characteristics using the Consolidated Health Economic Evaluation Reporting Standards (CHEERS) checklist. Finally, we summarised authors' modelling decisions using a framework that examines how models are developed to emphasise generalisability, precision, or realism.

Findings Our searches produced 1978 studies, out of which we included 36. Most studies used data from prospective trials and calculated cost-effectiveness directly from these trial inputs, rather than using simulation methods. As a group, these studies primarily emphasised precision and realism over generalisability, meaning that their results were best suited to specific settings.

Conclusions The number of included studies was small. Our findings suggest that most economic evaluations of HSS do not leverage methods like sensitivity analyses or inputs from literature review that would produce more generalisable (but potentially less precise) results. More research into how decision-makers would use economic evaluations to define the expansion path to strengthening health systems would allow for conceptualising impactful work on the economic value of HSS.

Key questions

What is already known?
► Health system strengthening (HSS) is an important step towards increasing the capacity and quality of care in low-income and middle-income countries.
► Economic evaluation of HSS activities can inform decisions about how HSS should be balanced with investments in disease-targeting interventions but is challenging to perform.

What are the new findings?
► Our systematic review summarises and analyses the methodological choices that researchers have used to address the challenges of conducting economic evaluation on HSS activities.

What do the new findings imply?
► We show that a substantial share of economic evaluations of HSS activities do not report the use of simulation methods, uncertainty analyses or distributional analyses.
► The use of these methods could increase the reliability and generalisability of this research.

INTRODUCTION

Despite the great potential of health system strengthening (HSS) activities to substantially reduce the morbidity and mortality burden in low/middle-income countries (LMICs),[1] the examination of their value for money has received relatively little attention from researchers. Methods for the economic evaluation of disease-targeting programmes and technical interventions are quite mature, but the variety and idiosyncrasies of health systems make the assessment and economic evaluation of HSS activities more challenging.[2] Even though it is clear that HSS is critical, policymakers often lack empirical evidence to inform how they should prioritise HSS relative to disease-targeting interventions or which specific activities of HSS they should prioritise and fund.

HSS refers to investments in the infrastructure of healthcare delivery or in improving interactions between components of a health

system.[3 4] The goals of investing in HSS are generally to improve the capacity, efficiency, or quality of healthcare delivered, or to expand the range of services offered.[5] Many examples show how health workforce training, physical infrastructure, supplies and coordination between healthcare providers can improve health outcomes for individuals at all stages of life.[6] This has proven especially true when outbreaks of infectious diseases occur: in diseases such as Ebola and COVID-19, fragile health systems can exacerbate the epidemics' impacts.[7 8]

For these reasons, there has been acknowledgement of the interdependence between cross-cutting, 'horizontal' investments in HSS and disease-targeting, 'vertical' approaches to improving health services access and health outcomes in LMICs.[9] Economic evaluation of HSS activities would help to optimise the balance between these two types of investments—cross-cutting or 'horizontal' HSS investments versus standalone or 'vertical' disease programme investments—so as to both respond to urgent needs for curative and preventative interventions and to build a more robust health system. HSS plays a unique role in LMICs vis-à-vis high-income countries: while the latter can generally implement new disease-targeting interventions while assuming that the delivery mechanism is already in place, LMICs, where health systems are weaker, must often design and build these delivery mechanisms prior to investment in new interventions.[10–12]

Economic evaluation methodologies such as cost-effectiveness analysis (CEA) have primarily assessed the value of narrowly defined (eg, disease-targeting) interventions. Further methodological developments, though, could improve the usefulness of economic evaluation for HSS activities, and be used to inform decisions about how much should be invested in disease-targeting interventions versus HSS activities, for example.[13] Compared with its use in assessing disease-targeting interventions, economic evaluation of HSS can be more challenging. This is, in part, because HSS often produces a range of multifaceted impacts across various diseases and conditions, which can be difficult to measure in their entirety.[14] Less is known about the broad and overlapping health impacts of HSS, compared with the disease-specific impacts of targeted interventions. As such, measuring only those benefits that are most obvious may make HSS activities seem like less favourable investments relative to disease-specific interventions, whose impacts are easier to define; on the other hand, attempting to account for every impact of an HSS activity may render its economic evaluation impossible. It can also be difficult to determine whether conclusions about the value of a given HSS activity can be generalised to other settings, or whether they apply only to a specific context. Finally, health systems' architectures are highly context-specific, which makes the generalisability and comparability of HSS activities across settings difficult.

In this study, we systematically reviewed published economic evaluations of HSS activities to improve our understanding of the methods currently being used to overcome the challenges cited above. We also collected information on which outcomes were selected and evaluated to determine how researchers address the challenges of identifying, estimating and reporting the most important outcomes associated with HSS activities. To analyse the methodologies used in the included studies, we employed a well-established comprehensive checklist of quality for health economic evaluations.[15] We then summarised the array of published studies using an existing framework that uses the trade-offs modellers make in conducting their evaluations to determine how the findings could best be interpreted and translated into practice. Finally, we provide a few recommendations for future research on the conduct of economic evaluations of HSS activities.

METHODS
Search strategy
Because HSS is a term with broad meanings and because our interest was primarily methodological, we used a two-stage approach to the literature search. The first stage served to provide an overview of the types of HSS activities included in economic evaluations. Our goal in this first stage was not to be comprehensive and exhaustive, but rather to broadly illustrate a growing area for CEA. During this first stage, we also identified specific activities for inclusion in more targeted literature searches. From the results of the first search, we selected three illustrative activities that represented three potential types of HSS activity—that is, tool, workforce strengthening and platform development—and that seemed to be well-represented in the literature. We then conducted the second stage of literature review, where we performed a more exhaustive search for economic evaluations on the three selected potential types of HSS activity.

To conduct the first, wide-ranging stage of our literature review, we used Medical Subject Headings (MeSH) terms, which collect a range of subjects under a common term. We searched MEDLINE for articles published in 2010 or later using a selection of MeSH terms designed to identify economic evaluations of HSS activities in LMICs (see online supplemental appendix section S1 for complete search strings). The initial search strategy was executed in October 2020 and was followed by a snowball search of references in the articles selected for full-text review. From this first search, we selected electronic medical records (EMRs) to represent tool-based HSS activities, task-shifting to represent workforce activities and home-based maternal and neonatal visits to represent delivery platform development.

For the second, targeted stage of our literature review, we worked with a medical librarian to develop search strings focused on MeSH terms as well as specific words in titles and abstracts that would gather all the economic literature on these activities (see online supplemental appendix section S2 for specific strings used). We

conducted this search in November 2021, including a snowball literature search.

Inclusion and exclusion criteria

Studies were included if they reported both economic and health outcomes associated with HSS activities in one or more LMICs. We did not place any restrictions on the type of health outcome reported. However, we required that an incremental cost-effectiveness ratio (ICER) be reported in or calculable from the study, whether it was provided by the authors or not. We reviewed only publications available in English.

We followed Hauck *et al*[5] in defining HSS as any investment in 'human and capital resources' on which the delivery of specific health interventions depends. As such, we excluded any study that focused on an intervention that did not somehow expand the capabilities of some segment of the health system. While some definitions of HSS include public education and increased demand due to improved health literacy, we have excluded those interventions from the current study in favour of focusing on the physical and workforce infrastructure of healthcare delivery.[16]

One reviewer (NH) screened titles and abstracts for inclusion in the first stage of the review, while two reviewers (NH and SB) screened titles and abstracts in the second stage. Two reviewers (NH and SV) independently screened the full text of studies that had not been excluded in the prior stage. Discrepancies were resolved by discussion between the two reviewers.

Data extraction and analysis

We developed a list of data to extract from each study based on the CHEERS checklist.[15] This included information on the intervention and its comparator; setting; perspective (eg, payer, society); time horizon; model type (if any); source(s) of disability (or utility) weights, if used; source(s) of cost and effectiveness estimates; sensitivity analyses (if any); and ICER. One reviewer (NH) independently extracted data from the studies, which was assessed for completeness by a second reviewer (SV). We also used the CHEERS checklist to assess the quality of included studies. As our focus was on the methods employed in these studies, we used only the methods portion of the CHEERS checklist in both the data extraction and quality assessment portions of our study.

For the purposes of classifying types of HSS activities, we modified an existing taxonomy developed by *The Lancet Global Health* Commission on High Quality Health Systems.[1] Our modified taxonomy included the following classes: (i) governance, which deals with medical and payment policies, as well as intersectoral interventions, such as sanitation infrastructure; (ii) platforms, which reflects the physical facilities available to healthcare staff and the services they offer; (iii) workforce, which we define as investment in the human capital of health workers and (iv) tools, which includes information technology and devices that improve care delivery. Each intervention was assigned a category through discussion between the two reviewers.

We also classified the health outcomes reported into three categories: disability, mortality and other. This categorisation is similar to the outcomes used by the Global Burden of Disease study.[17] Deaths and life years lost or gained were counted as mortality-related outcomes. The constructed metrics of quality-adjusted life years (QALYs) and disability-adjusted life years (DALYs), which include both mortality and morbidity outcomes, were counted as morbidity-related outcomes for simplicity. Any outcome not classifiable into the previous two categories was placed into the 'other' category.

As an exploratory analysis of trends in cost-effectiveness between the different types of HSS activities, we reframed, when possible, selected health outcomes as estimated DALYs averted. We first assumed that QALYs gained were equivalent to DALYs averted. This is a very rough approximation that we felt could however be acceptable for our purpose here due to the similarity in their calculation methods.[18] We next approximated life-years gained to DALYs averted by assuming that no disability weight would be applied to the period of improved survival. Finally, we applied this same assumption to deaths averted by converting it to life years gained, using roughly the difference between the mean target age at which the intervention took place and the life expectancy for Japanese females (ie, the highest life expectancy in the world) as a reference age for life expectancy. Most of these studies already included discounting, and so we did not apply a separate discount rate. Because of the complexity of estimating the impact of disease cases averted, improved guideline adherence, or other similar health outcomes, we did not include these in the ICER comparisons.

Overview of cost-effectiveness literature

We concluded our analysis by summarising broad choices made by researchers performing economic evaluations on HSS activities. Economic evaluation requires choices about what to include and what not to include. These trade-offs are necessary for the creation of tractable models and should be informed by the purposes for which models are created. We summarised broad choices made by researchers performing economic evaluations and CEAs on HSS activities to gain insight into how researchers envisioned their work being used. For this portion of our analysis, we used a framework developed by Levins that seeks to classify models in terms of the trade-offs that they make.[19] This framework identifies three major goals of model development: generalisability, precision and realism. Generalisability is the ability to use a model across different settings, which includes estimating value both in larger populations in the same setting (generalisability) and across different settings (transportability); precision is how much uncertainty there is around the model's results; and realism is the degree to which model inputs correspond to the real
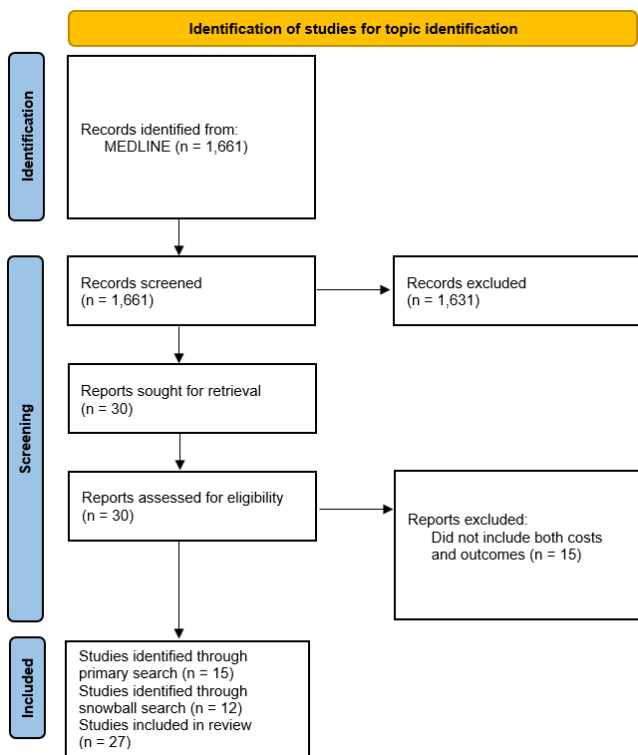
Identification of studies for topic identification

Identification

Records identified from:
MEDLINE (n = 1,661)

Screening

Records screened
(n = 1,661) → Records excluded
(n = 1,631)

Reports sought for retrieval
(n = 30)

Reports assessed for eligibility
(n = 30) → Reports excluded:
Did not include both costs
and outcomes (n = 15)

Included

Studies identified through
primary search (n = 15)
Studies identified through
snowball search (n = 12)
Studies included in review
(n = 27)

Figure 1 Preferred Reporting Items for Systematic Reviews and Meta-Analyses diagram showing the flow of the first stage of the systematic study review process.

world. Researchers are able to emphasise no more than two of these goals as they develop their models.

Identifying researchers' choices in these terms illuminates how their results can be interpreted and translated into practice.[20] For example, researchers who emphasise generalisability and precision may use equation-based models with simplifying assumptions from which broad theories can be developed.[19] Studies developed to provide generalisability and realism often make similar simplifying assumptions, but integrate more empirical data from the real world. However, their lack of precision—often in the form of wide confidence intervals—means that they are best suited to providing relative or qualitative results. Finally, researchers who choose to focus on precision and realism often produce studies that can provide reliable and accurate predictions for specific settings but lack broad generalisability. There is no standardised method of classifying studies using this framework, therefore we recorded our subjective impressions of the implicit choices made by study authors.

Patient and public involvement
Neither patients nor the public were involved with the design and conduct of this systematic review.

RESULTS
The first stage of our search identified 1661 publications (figure 1) and was conducted with the goal of identifying specific HSS activities that may have been the subject of

several economic evaluations. Following title and abstract screening, 30 studies remained. We identified an additional 12 studies through snowball search of the selected studies' references and conducted a full-text review of each. A total of 27 studies were included (see table 1 for selected characteristics of included studies and online supplemental appendix table S1 for full study characteristics). All the studies rejected during full-text review met our definition of HSS but did not report information on both costs and health benefits.

The second stage of our search focused on three specific activities: use of EMRs, home-based maternal and neonatal visits, and task-shifting. We reviewed 30 EMR-related titles and abstracts, of which three remained in the final analysis after review of complete papers. Our search resulted in 192 titles and abstracts related to home-based maternal and neonatal care, of which we included five in the final analysis. Finally, we reviewed 95 titles and abstracts on the subject of task-shifting and included eight papers in our analysis.

Because some papers identified in the first search also appeared in the second, we arrived at a total of 36 unique papers.

Summary characteristics of studies from the first stage
Studies in the first search primarily focused on 'platform' and 'workforce' interventions, with eight studies representing each category (figure 2). The cross-cutting nature of many HSS activities made it difficult to ascertain whether any particular group would benefit from some activities (eg, EMRs), but 17 of the 27 studies focused on activities that would primarily benefit maternal and child health, an important focus area in LMICs. Studies were produced at a relatively constant pace over the past 10 years, with no noticeable acceleration or deceleration in their rate of publication.

Of the 27 studies, only nine used simulation-based modelling methods such as decision trees, Markov models or agent-based models to estimate intervention cost-effectiveness. Decision trees and public health tools (such as the Lives Saved Tool[21 22]) were the most commonly used model types. A small number of studies used Markov modelling and agent-based models. Most studies, regardless of their analytic methods, used primary data to estimate intervention impact and effectiveness. Among primary data sources, cluster randomised trials (CRTs) were the most common. Studies that included secondary data primarily used literature review rather than expert opinion. Most studies clearly indicated the methods of costing used. Among costing methods, nine studies used microcosting (ie, 'bottom-up' costing) alone, making it the most common costing methodology encountered, followed by a mix of microcosting and gross costing (ie, 'top-down') methods,[23] which seven studies used.

Most studies reported disability and mortality outcomes (online supplemental appendix figure S1). Eight studies that reported disability outcomes used weights from the Global Burden of Disease study,[24] which was the most

Table 1 Selected characteristics of the included studies (n=36)

First author	Year published	Setting	Interventions	Class of HSS	Comparator	Effectiveness measure	Thematic emphases
Arora	2017	India and Bangladesh	Telephone management of pressure ulcers for people with spinal injuries	Platforms	Control arm of randomised controlled trial	Patient QALYs pre/post	Precision, realism
Barasa	2012	Kenya	Dissemination of guidelines for paediatric emergency care	Workforce	Control arm of cluster randomised trial	Average % improvement in 14 indicators between intervention and control sites; DALYs estimated by external source	Generality, realism
Broughton	2013	Niger	Quality improvement to prevent postpartum haemorrhage	Governance	Same site(s) before intervention	PPH cases, DALYs, deaths averted; compliance with birth management and newborn care standards	Precision, realism
Burn	2014	All LICs	Pulse oximeter use during surgery	Tools	Modelled usual care	DALYs averted (discounted; disability due to non-fatal hypoxic brain injury not included)	Generality, precision
Buttorff	2012	India	Task-shifting for mental healthcare (provided by lay health workers)	Workforce	Control arm of cluster randomised trial	QALYs produced; work days gained; psychiatric symptom score	Precision, realism
Chanda	2011	Zambia	Task-shifting to community health workers for treatment of uncomplicated malaria	Workforce	Unblinded trial based on patient-preferred treatment	Properly identified and treated cases	Precision, realism
Chen	2012	Nicaragua	Volunteer orthopaedic surgeon trips	Platforms	Expert evaluation of likely counterfactuals	DALYs averted	Precision, realism
Colbourn	2015	Malawi	Women's community groups; quality improvement in maternal/child health	Governance	Control arm of cluster randomised trial	Maternal, neonatal, and perinatal mortality; DALYs averted (attributable to reduced mortality only)	Precision, realism
Datiko	2010	Ethiopia	Task-shifting to health extension workers for treatment of tuberculosis	Workforce	Control arm of cluster randomised trial	Fully treated cases	Precision, realism
Do Prado	2011	Brazil	Task-shifting to home-based guardians for directly observed treatment of tuberculosis	Workforce	Unblinded trial based on patient-preferred treatment	Fully treated cases	Precision, realism
Driessen	2012	Malawi	Implementation of electronic medical records	Tools	Modelled usual care	Length of hospital stays, time transcribing medical records, laboratory use	Precision, realism
Gertler	2014	Argentina	Pay for performance incentives for use and quality of maternal and child health services	Governance	Control arms in natural experiment	Ten tracers indicating quality of maternal and child health services; DALYs for neonatal death and low birth weight only	Precision, realism

Continued

Table 1 Continued

First author	Year published	Setting	Interventions	Class of HSS	Comparator	Effectiveness measure	Thematic emphases
Hayford	2014	Bangladesh	Extended clinic hours, vaccinator training, active surveillance and community participation in vaccination	Workforce	Same site(s) before intervention	Fully immunised children	Precision, realism
Howlader	2011	Bangladesh	Strengthening maternal care in community health centres	Platforms	Modelled usual care	Maternal mortality	Generality, precision
Jansen	2014	India	Facilitating improved follow-up for cleft palate surgeries	Platforms	Same site(s) before intervention	Number of follow-up appointments kept	Precision, realism
Jo	2019	Bangladesh	Enhanced home-based maternal and neonatal care	Platforms	Control arm of cluster randomised trial	Neonatal deaths averted; DALYs averted	Precision, realism
LeFevre	2013	Bangladesh	Improved training and mobilisation in neonatal community care with or without the addition of home-based care	Platforms	Control arm of cluster randomised trial	Neonatal deaths averted; DALYs averted	Generality, realism
Lewycka	2013	Malawi	Training of women's groups and peer counsellors for maternal and infant health	Workforce	Control arm of cluster randomised trial	Years of life lost	Precision, realism
Luangasanatip	2018	Thailand	Hand-washing promotion among ICU healthcare workers	Workforce	Modelled usual care	QALYs gained; deaths averted	Generality, precision
Manasyan	2010	Zambia	Training midwives in essential newborn care	Workforce	Same site(s) before intervention	Lives saved; DALYs averted	Precision, realism
Mathewos	2017	Ethiopia	Home-based maternal and neonatal care	Platforms	Control arm of cluster randomised trial	Neonatal deaths averted; DALYs averted	Precision, realism
Nandi	2015	India	Scale-up of home-based neonatal care	Platforms	Modelled usual care	Incidence of severe neonatal morbidity; neonatal deaths averted	Generality, realism
Patouillard	2011	Ghana	Task-shifting to community health workers for intermittent preventative treatment for malaria	Workforce	Control arm of cluster randomised trial	Fully treated children	Precision, realism
Pitt	2016	Ghana	Home-based neonatal care	Platforms	Control arm of cluster randomised trial	Neonatal mortality, (discounted 3% per year) life-years saved	Precision, realism
Prinja	2018	India	Mobile electronic medical record programme with reminders for guidelines based care	Tools	Control arm of cluster randomised trial	Infant deaths averted; episodes of illness averted; DALYs averted	Precision, realism
Puett	2012	Bangladesh	Task-shifting to community health workers for treatment of severe acute malnutrition	Workforce	Control arm of cluster randomised trial	Deaths averted; DALYs averted	Generality, realism
Sabin	2012	Zambia	Training traditional birth attendants in neonatal emergencies	Workforce	Control arm of cluster randomised trial	Deaths averted; DALYs averted	Precision, realism

Continued

Table 1 Continued

First author	Year published	Setting	Interventions	Class of HSS	Comparator	Effectiveness measure	Thematic emphases
Seaman	2020	GBD-region specific averages of all LMICs within region	Hep B vaccination with controlled temperature chain protocol and compact prefilled auto-disable syringes	Tools	Modelled usual care	DALYs averted	Generality, precision
Shepard	2020	Zimbabwe	Results-based financing for maternal and child care	Governance	Control arm of cluster randomised trial	QALYs based on lives saved, according to LiST tool	Generality, realism
Somigliana	2011	Uganda	Obstetrics-focused ambulance service	Tools	Expert evaluation of likely counterfactuals	Life years saved	Precision, realism
Stenberg	2019	67 LMICs	Strengthening primary care and supporting, cross-sectoral interventions as part of universal health coverage	Governance	Modelled usual care	Deaths averted; gains in life expectancy at birth	Generality, precision
Tripathy	2010	India	Women's community groups teaching new mothers about health of self and infant	Platforms	Control arm of cluster randomised trial	Lives saved; life-years saved	Precision, realism
Vossius	2014	Tanzania	Training in neonatal resuscitation	Workforce	Same site(s) before intervention	Lives saved; DALYs averted	Precision, realism
Xue	2012	China	Implementation of electronic medical records	Tools	Same site(s) before intervention	Length of hospital stay; infection rate; mortality	Precision, realism
Yan	2014	China	Task-shifting to community organisations for HIV detection and care	Workforce	Same site(s) before intervention	Cases of HIV detected	Precision, realism
Zeng	2018	Zambia	Results-based financing for maternal and child care	Governance	Control arm of cluster randomised trial	Lives saved; QALYs based on lives saved	Precision, realism

DALYs, disability-adjusted life years; ICU, intensive care unit; LICs, low-income countries; LMICs, low-income and middle-income countries; QALYs, quality-adjusted life years.
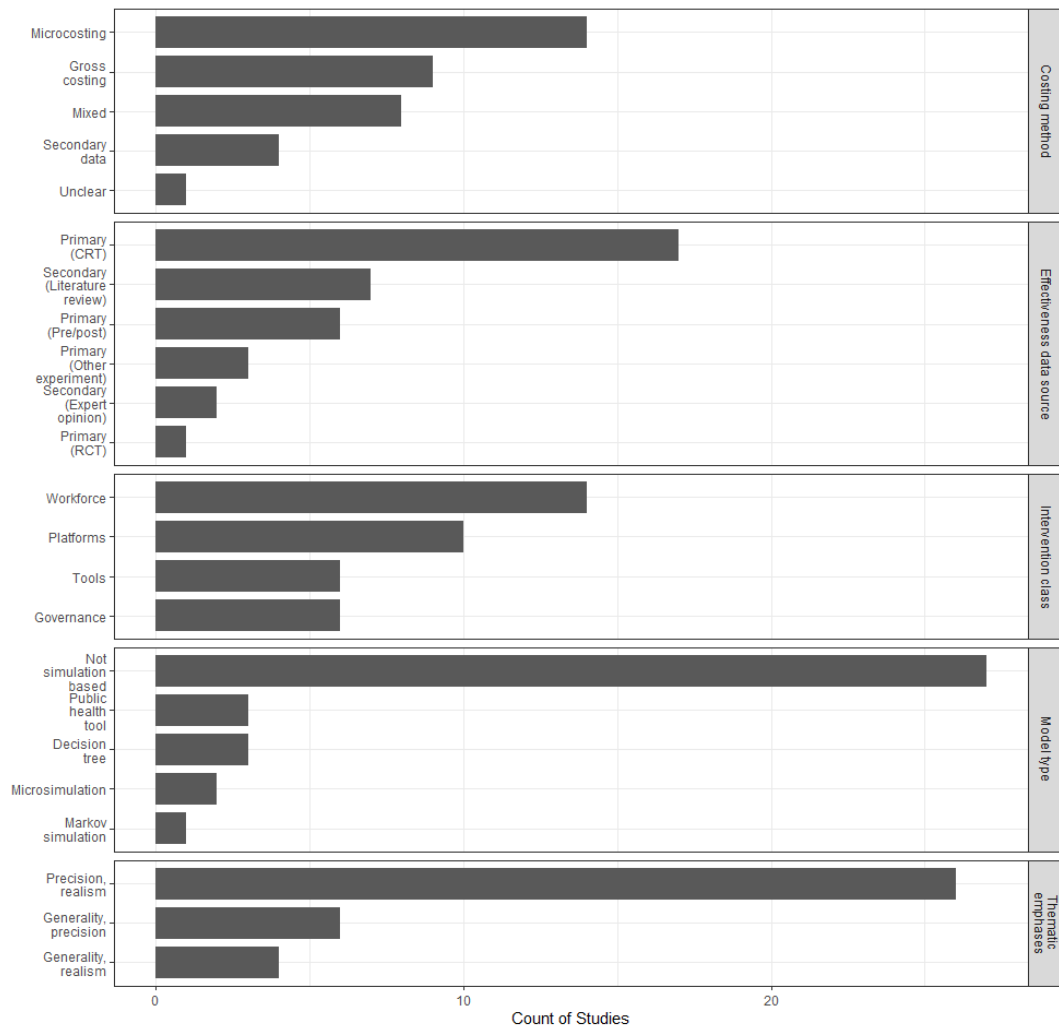
Figure 2 Summary characteristics of the included studies (n=36). CRT, Cluster Randomized Trial; RCT, Randomized Controlled Trial.

common source of this information across all included studies. Other outcomes reported included level of guideline adherence, episodes of illness averted, numbers of fully immunised children, lengths of hospital stay, and changes in productivity of both patients and providers. Only three studies exclusively reported outcomes that did not include disability or mortality, but a total of eight studies reported at least one such outcome.

Summary characteristics of studies on specific activities

We observed some differences in the studies we identified during the second stage of our literature review. Of the 16 studies included in this stage, 7 used disability-related outcomes and 8 used mortality-related outcomes. A substantial proportion of these studies used outcomes that were difficult to generalise across settings, such as length of hospital stay (eg, for EMR studies) or fully treated cases (eg, for task-shifting studies).

Compared with the studies identified in the first stage of literature review, which primarily used the health system perspective, a larger share (6 out of 16) of these studies also used the partial societal perspective (ie, including some costs from the payer and patient perspective). We also observed that more of these studies (14 of 16) used primary data for effectiveness estimates. These were generally cluster randomised trials, but preanalyses/postanalyses and open-label trials were also represented among these studies. Preanalyses/postanalyses used cost and health outcomes from before and after an intervention, while open-label trials gave participants the choice of which intervention they receive.

We observed a trend towards using similar methods within studies on the same topic. For example, among studies on task-shifting, all except one study used microcosting. However, we could not quantify the significance of any differences between the activity-specific articles and those we identified from the broader literature in the first stage of our search.

Study quality

Several of the 36 unique studies were of high quality and met virtually all the standards reviewed from the CHEERS checklist (n=8) (online supplemental appendix table S2). However, the perspectives used for costing and quality of

life calculations were partially explained in eight of the included studies. Also, 8 studies were unclear about the time horizon used; among the 21 studies that included preference-based measures, 9 did not fully detail the methods used for estimating the quality-of-life impacts of the interventions. Finally, four of the nine model-based studies provided succinct discussions of their models' structures or assumptions.

Included studies generally reported their results thoroughly. However, 13 did not explore the uncertainty around their findings. Only 10 studies included discussions about subgroup heterogeneity.

Holistic assessment of modelling choices

We assessed 26 of the 36 studies as focusing on precision and realism (table 1). Among those that included a focus on generalisability, we found that they were equally split between emphasising realism and precision. We primarily assessed studies as emphasising generalisability when they used simulation models or sensitivity analyses to extrapolate observed data or to explore uncertainty. Studies emphasising realism largely avoided basing any study inputs on simulation or mathematical formulas. Finally, we assessed studies as emphasising precision when they used inputs such as accounting records that tied the study results to specific settings.

DISCUSSION

Our goal for this study was to identify the methods used in economic evaluation studies pertaining to HSS activities and the strategies that researchers have used to cope with the challenges of conducting economic evaluations of HSS activities in LMICs. We identified a relatively small number of studies compared with the large body of CEA literature on disease-targeting interventions, although the settings studied were relatively diverse. Maternal and child health was the topical area of a large proportion of the studies included. Our literature review suggests that the most common strategy for conducting economic evaluations of HSS activities was to conduct analyses alongside a prospective trial rather than using simulation-based methods and secondary data, as is commonly found in the CEA literature. HSS activities generally produce a broad array of multifaceted impacts, whose effects are often quite context-specific. This creates challenges for the construction of simulation models, the identification of suitable inputs, as well as for generalisability. As such, we hypothesise that this choice of using inputs directly from prospective trials was a strategy that researchers used to avoid the difficulties associated with simulation methods in this context.

However, the conduct of economic analyses alongside prospective trials involves trade-offs. We used Levins's framework to identify these trade-offs and found that most included studies would sacrifice generalisability for precision and realism. This means that these studies' findings would likely be applicable only in specific settings and circumstances. As such, decision-makers from other settings would likely face substantial challenges, including considerations of uncertainty, in using these studies to strengthen their health systems (out of the specific context of these studies). Moreover, the designing, constructing and restructuring of health systems in LMICs is relatively more common compared with high-income countries. As such, the applicability of these relatively narrowly focused studies in reconfigured health systems—even in the same setting—might be rendered difficult.

In contrast, one-third of studies emphasised generalisability across settings and thus would have greater external applicability. These studies were more likely to rely on secondary data, such as literature review, for their inputs. They were also somewhat more likely to use simulation modelling methods, such as agent-based models, decision trees and Markov models. Several of these studies were grounded in results from trials in specific settings but used alternative inputs in sensitivity and scenario analyses to provide qualitative cost-effectiveness findings that would also be applicable in other settings.

Methodological limitations were common in the studies that we identified. There are many known challenges to conducting economic evaluations in LMICs.[25 26] For example, a large number of these studies that emphasised precision and realism did not conduct the types of sensitivity analyses that are informative for the interpretation of economic evaluation outcomes. Methods like one-way sensitivity analysis not only can point to the range of likely results, but also indicate what parameters are most influential in driving these results. By not including sensitivity analyses, many studies would neglect a potentially powerful source of insight about what data should be collected in future studies and how to effectively implement HSS activities.

Based on our findings, we offer a few suggestions for the conduct of future economic evaluations of HSS activities, both within a specific setting and beyond across environments.

First, we suggest using a broader array of outcomes to capture the distribution of impacts as well as non-health effects of HSS activities. Disability and mortality are clearly relevant to decision-makers, but financial risk protection and preventing medical impoverishment have also been acknowledged as major health system goals.[27–29] Only one of the included studies measured the consequences of HSS activities on financial protection.[30] Likewise, the distributional effects of HSS activities across socioeconomic groups were not considered, except in one study[30], even though HSS activities could potentially greatly benefit the poorest.

We also encourage researchers to consider emphasising both generalisability to larger populations within the same setting and transportability across populations. There are many barriers to conducting economic evaluations in LMICs including limited availability of data, and substantial variability in costs across settings.[31] These

difficulties are only compounded when attempting to evaluate HSS activities. For these reasons, a more generalisable approach may be desirable for local adaptations of economic evaluations, which could draw from the systematic literature review synthesised here and ensure that these economic evaluations are both tractable for researchers and useful to decision-makers.

Finally, we suggest developing novel analytic methods to capture the effects of HSS activities across sectors and time. Methodological innovations imported from fields such as operations research could give researchers insight into how HSS activities modify the dynamic relationships between different parts of the health system and beyond.[13 32 33] For example, HSS could be modelled as a set of interacting components that can result in non-linear improvements in health services delivery. These interactions can modify three attributes of the health system: economies of scale (size-based changes in efficiency), economies of scope (changes in efficiency brought about by the coproduction of related interventions) and the development of new platforms (novel channels for the delivery of services).[5] For instance, the conceptualisation of HSS activities as changing these three attributes might provide multipliers on the effects of disease-targeting interventions, thus allowing researchers to use elements of conventional CEA modelling to capture the effects of HSS.

While our suggestions outlined above may help conceptualise economic evaluations of HSS activities, a mature body of literature in this area would be ultimately responsive to the needs of decision-makers. Some research suggests that decision-makers have struggled to integrate the results of CEAs into their priority setting activities.[34 35] This may be due in part to a failure to adequately adapt CEA methods to the specific decision-making processes.[36] The refinement of methods for performing economic evaluation on HSS activities should therefore primarily depend on responding to decision-makers' needs (eg, see box 1 for the illustrative context of Ethiopia).

Our study was limited in several important ways. First, most importantly, we had a small sample of studies that covered a narrow range of topics. Second, the studies we identified may also reflect sponsorship and publication biases; for example, funding agencies may incentivise evaluations of individual projects, as opposed to modelling the comprehensive costs and benefits of a set of HSS activities implemented in different settings. This bias may also have manifested in the fact that most interventions examined were found to be highly cost-effective. Because of such potential biases, as well as uncertainties around the transferability of outcomes to different settings, we were unable to draw conclusions about the relative cost-effectiveness of different classes of HSS activity. Third, our choice of the CHEERS checklist over other guidelines for economic evaluations such as the International Decision Support Initiative's reference case may have brought our attention to certain methodological choices at the cost of others.[37] Finally, we were only able to search

the English-language literature, which may have excluded important contributions published in other languages (eg, French, Portuguese).

In conclusion, the existing literature on the cost-effectiveness of HSS has been primarily conducted alongside prospective trials, especially CRTs. Although this strategy can produce precise estimates for the specific setting in which they were conducted, it can limit the generalisability of the study's findings beyond to other settings. Existing methodologies offer ways of improving the relevance of this research. Methodological research could support this goal, such as by developing a list of best practices for evaluating the costs and benefits of HSS activities. However, the needs of decision-makers should ultimately drive this area of research. Future studies should be conducted to better characterise which features of economic evaluation methods can be best tailored for priority setting of HSS interventions.

> **Box 1 The potential use of economic evaluation to prioritise health system strengthening activities in Ethiopia**
>
> Concurrent with economic developments and sociodemographic changes, many low-income and middle-income countries have demonstrated a shift in disease burden toward non-communicable diseases. Such changes have created a demand for universal health coverage-type reforms that were accompanied by defining a national high priority health services package to be publicly financed and provided to citizens free of charge. As an example, Ethiopia, in 2019, revised its essential health services package (EHSP). The key principles used to prioritise health interventions were evidence on cost-effectiveness analysis (CEA) (interventions that maximise total population health), equity and financial risk protection benefits of the interventions. CEA was the key consideration in the prioritisation and ranking of interventions for inclusion in Ethiopia's EHSP. Effective and efficient implementation of the EHSP will now require substantial health system strengthening (HSS) around the health workforce, logistic management, health management information system, governance and other health system building blocks. Therefore, CEA of HSS activities could help prioritise the implementation and sequencing of the delivery of quality care in such setting.
> Source: Verguet *et al*.[38]

Author affiliations

[1]Department of Global Health and Population, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[2]Global Health Research and Consulting, Yaozhi, Yangzhou, Jiangsu, China
[3]Department of Management Science, University of Strathclyde, Glasgow, UK
[4]Addis Center for Ethics and Priority Setting, College of Health Sciences, Addis Ababa University, Addis Ababa, Ethiopia
[5]National Data Management Centre for Health, Ethiopian Public Health Institute, Addis Ababa, Ethiopia
[6]School of Public Health, University of Ghana, Accra, Ghana

Twitter Nathaniel Hendrix @ndhendrix and Stéphane Verguet @StephaneVerguet

the box text; NH drafted the first version of the paper; all authors provided critical review of the draft and suggestions for revision.

ORCID iDs
Nathaniel Hendrix http://orcid.org/0000-0001-8154-0276
Sarah Bolongaita http://orcid.org/0000-0003-0886-0454
Itamar Megiddo http://orcid.org/0000-0001-8391-6660
Solomon Tessema Memirie http://orcid.org/0000-0003-3806-2453
Stéphane Verguet http://orcid.org/0000-0003-4128-0849

REFERENCES

1 Kruk ME, Gage AD, Arsenault C, *et al*. High-quality health systems in the Sustainable Development Goals era: time for a revolution. *Lancet Glob Health* 2018;6:e1196–252.
2 Hafner T, Shiffman J. The emergence of global attention to health systems strengthening. *Health Policy Plan* 2013;28:41–50.
3 Chee G, Pielemeier N, Lion A, *et al*. Why differentiating between health system support and health system strengthening is needed. *Int J Health Plann Manage* 2013;28:85–94.
4 Adam T, de Savigny D. Systems thinking for strengthening health systems in LMICs: need for a paradigm shift. *Health Policy Plan* 2012;27 Suppl 4:iv1–3.
5 Hauck K, Morton A, Chalkidou K, *et al*. How can we evaluate the cost-effectiveness of health system strengthening? A typology and illustrations. *Soc Sci Med* 2019;220:141–9.
6 Macinko J, Starfield B, Erinosho T. The impact of primary healthcare on population health in low- and middle-income countries. *J Ambul Care Manage* 2009;32:150–71.
7 Naimoli JF, Saxena S, Hatt LE, *et al*. Health system strengthening: prospects and threats for its sustainability on the global health policy agenda. *Health Policy Plan* 2018;33:85–98.
8 Kraef C, Juma P, Kallestrup P, *et al*. The COVID-19 pandemic and non-communicable diseases-A wake-up call for primary health care system strengthening in sub-Saharan Africa. *J Prim Care Community Health* 2020;11:215013272094694.
9 Reich MR, Takemi K. G8 and strengthening of health systems: follow-up to the Toyako Summit. *Lancet* 2009;373:508–15.
10 Leatherman S, Ferris TG, Berwick D, *et al*. The role of quality improvement in strengthening health systems in developing countries. *Int J Qual Health Care* 2010;22:237–43.
11 Mills A. Health care systems in low- and middle-income countries. *N Engl J Med* 2014;370:552–7.
12 Bitton A, Ratcliffe HL, Veillard JH, *et al*. Primary health care as a foundation for strengthening health systems in low- and middle-income countries. *J Gen Intern Med* 2017;32:566–71.
13 Verguet S, Feldhaus I, Jiang Kwete X, *et al*. Health system modelling research: towards a whole-health-system perspective for identifying good value for money investments in health system strengthening. *BMJ Glob Health* 2019;4:e001311.
14 Adam T, Hsu J, de Savigny D, *et al*. Evaluating health systems strengthening interventions in low-income and middle-income countries: are we asking the right questions? *Health Policy Plan* 2012;27 Suppl 4:iv9–19.
15 Husereau D, Drummond M, Petrou S, *et al*. Consolidated Health Economic Evaluation Reporting Standards (CHEERS)--explanation and elaboration: a report of the ISPOR Health Economic Evaluation Publication Guidelines Good Reporting Practices Task Force. *Value Health* 2013;16:231–50.
16 Smith PC, Yip W. The economics of health system design. *Oxf Rev Econ Policy* 2016;32:21–40.
17 Institute for Health Metrics and Evaluation. Global burden of disease (GBD), 2019. Available: https://www.healthdata.org/gbd/2019
18 Airoldi M, Morton A. Adjusting life for quality or disability: stylistic difference or substantial dispute? *Health Econ* 2009;18:1237–47.
19 Levins R. The strategy of model building in population biology. *Am Sci* 1966;54:421–31.
20 Weisberg M. Forty years of 'The strategy': Levins on model building and idealization. *Biol Philos* 2007;21:623–45.
21 Fox MJ, Martorell R, van den Broek N, *et al*. Assumptions and methods in the Lives Saved Tool (LiST). *BMC Public Health* 2011;11 Suppl 3:I1.
22 Walker N, Tam Y, Friberg IK. Overview of the Lives Saved Tool (LiST). *BMC Public Health* 2013;13 Suppl 3:S1.
23 Raftery J. Costing in economic evaluation. *BMJ* 2000;320:1597.
24 Institute for Health Metrics and Evaluation. Global burden of disease (GBD) https://www.healthdata.org/gbd/2019
25 Griffiths UK, Legood R, Pitt C. Comparison of economic evaluation methods across low-income, middle-income and high-income countries: what are the differences and why?. *Health Economics* 2016;25:29–41.
26 Pitt C, Vassall A, Teerawattananon Y, *et al*. Foreword: health economic evaluations in low- and middle-income countries: methodological issues and challenges for priority setting. *Health Econ* 2016;25 Suppl 1:1–5.
27 Murray CJ, Frenk J. A framework for assessing the performance of health systems. *Bull World Health Organ* 2000;78:717–31.
28 Roberts M, Hsiao W, Berman P, *et al*, eds. *Getting Health Reform Right: A Guide to Improving Performance and Equity*. Oxford University Press, 2008.
29 Voorhoeve A, Ottersen T, Norheim OF. Making fair choices on the path to universal health coverage: a précis. *Health Econ Policy Law* 2016;11:71–7.
30 Nandi A, Colson AR, Verma A, *et al*. Health and economic benefits of scaling up a home-based neonatal care package in rural India: a modelling analysis. *Health Policy Plan* 2016;31:634–44.
31 Culyer AJ, Chalkidou K. Economic evaluation for health investments en route to universal health coverage: cost-benefit analysis or cost-effectiveness analysis? *Value Health* 2019;22:99–103.
32 Jbaily A, Feldhaus I, Bigelow B, *et al*. Toward health system strengthening in low- and middle-income countries: insights from mathematical modeling of drug supply chains. *BMC Health Serv Res* 2020;20:776.
33 Chang AY, Ogbuoji O, Atun R, *et al*. Dynamic modeling approaches to characterize the functioning of health systems: a systematic review of the literature. *Soc Sci Med* 2017;194:160–7.
34 Baltussen R, Jansen MP, Mikkelsen E, *et al*. Priority setting for universal health coverage: we need evidence-informed deliberative processes, not just more evidence on cost-effectiveness. *Int J Health Policy Manag* 2016;5:615–8.
35 Baltussen R, Marsh K, Thokala P, *et al*. Multicriteria decision analysis to support health technology assessment agencies: benefits, limitations, and the way forward. *Value Health* 2019;22:1283–8.
36 Leech AA, Kim DD, Cohen JT, *et al*. Use and misuse of cost-effectiveness analysis thresholds in low- and middle-income countries: trends in Cost-per-DALY studies. *Value Health* 2018;21:759–61.
37 Wilkinson T, Sculpher MJ, Claxton K, *et al*. The International decision support initiative reference case for economic evaluation: an aid to thought. *Value Health* 2016;19:921–8.
38 Verguet S, Hailu A, Eregata GT, *et al*. Toward universal health coverage in the post-COVID-19 era. *Nat Med* 2021;27:380–7.