SYSTEMATIC REVIEW

A systematic review of the agreement of recall, home-based records, facility records, BCG scar, and serology for ascertaining vaccination status in low and middle-income countries [version 1; peer review: 1 approved, 1 approved with reservations]

Emily Dansereau¹, David Brown², Lena Stashko¹, M. Carolina Danovaro-Holliday¹

¹Strategic Information Group, Expanded Program on Immunization (EPI), Department of Immunization, Vaccines and Biologicals (IVB), World Health Organization, Geneva, Switzerland
²Brown Consulting Group Int’l LLC, Cornelius, NC, USA

Abstract

Background: Household survey data are frequently used to estimate vaccination coverage - a key indicator for monitoring and guiding immunization programs - in low and middle-income countries. Surveys typically rely on documented evidence from home-based records (HBR) and/or maternal recall to determine a child’s vaccination history, and may also include health facility sources, BCG scars, and/or serological data. However, there is no gold standard source for vaccination history and the accuracy of existing sources has been called into question.

Methods and Findings: We conducted a systematic review of peer-reviewed literature published January 1, 1957 through December 11, 2017 that compared vaccination status at the child-level from at least two sources of vaccination history. 27 articles met inclusion criteria. The percentage point difference in coverage estimates varied substantially when comparing caregiver recall to HBRs (median: +1, range: -43 to +17), to health facility records (median: +5, range: -29 to +34) and to serology (median: -20, range: -32 to +2). Ranges were also wide comparing HBRs to facility-based records (median: +17, range: -61 to +21) and to serology (median: +2, range: -38 to +36). Across 10 studies comparing recall to HBRs, Kappa values exceeded 0.60 in 45% of comparisons; across 7 studies comparing recall to facility-based records, Kappa never reached 0.60. Agreement varied depending on study setting, coverage level, antigen type, number of doses, and child age.

Conclusions: Recall and HBR provide relatively concordant vaccination histories in some settings, but both have poor agreement with facility-based records and serology. Long-term, improving clinical decision making and vaccination coverage estimates will depend on strengthening administrative systems and record keeping practices. Short-term, there must be greater

Open Peer Review

Reviewer Status

Invited Reviewers

version 1

published 21 Mar 2019

v1


recognition of imperfections across available vaccination history sources and explicit clarity regarding survey goals and the level of precision, potential biases, and associated resources needed to achieve these goals.

**Keywords**
immunization, vaccination, LMIC, coverage, survey, methodology, concordance, agreement, validity

**Corresponding author:** Emily Dansereau (Emily.Dansereau@gatesfoundation.org)

**Author roles:** **Dansereau E:** Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Visualization, Writing – Original Draft Preparation; **Brown D:** Conceptualization, Methodology, Writing – Review & Editing; **Stashko L:** Data Curation, Validation, Writing – Review & Editing; **Danovaro-Holliday MC:** Conceptualization, Funding Acquisition, Methodology, Supervision, Writing – Review & Editing

**Competing interests:** After completing this analysis and manuscript, ED was hired to work at the Gates Foundation.

**Grant information:** This work is supported by the Bill and Melinda Gates Foundation [OPP1055811; OPP1115427]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Copyright:** © 2019 Dansereau E et al. This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**How to cite this article:** Dansereau E, Brown D, Stashko L and Danovaro-Holliday MC. A systematic review of the agreement of recall, home-based records, facility records, BCG scar, and serology for ascertaining vaccination status in low and middle-income countries [version 1; peer review: 1 approved, 1 approved with reservations] Gates Open Research 2019, 3:923 (https://doi.org/10.12688/gatesopenres.12916.1)

**First published:** 21 Mar 2019, 3:923 (https://doi.org/10.12688/gatesopenres.12916.1)
Introduction

Vaccination coverage estimates are frequently used at the sub-national, national, and global levels to track performance, set priorities, make managerial and strategic decisions, and allocate funding for immunization programs. In some cases, vaccination coverage is continuously monitored through child-level registries, but these administrative sources are often unreliable, particularly in low and middle-income countries (LMIC). Therefore, LMICs frequently complement administrative recording and reporting data with vaccination coverage surveys, which typically rely on documented evidence in home-based records (HBR) and/or caregiver recall to ascertain a child’s vaccination history. In some cases, surveys also consult facility records, check for BCG scars, or analyze serological samples for evidence of immunity or prior vaccination. However, there is no single gold standard for validating whether a child has been vaccinated and the accuracy of these sources for informing coverage estimates remains uncertain.

Multiple factors can cause each vaccination history source to over- or under-estimate coverage. Caregivers may over-report recalled vaccination histories due to social desirability bias or be unable to recall which and how many vaccinations their children received, particularly as vaccination schedules become more complex. HBRs can be inaccurate if the record was not brought to every vaccination appointment or the provider made recording mistakes, including failing to record doses, recording doses that were not administered, or misrecording the vaccination date. Facility-based registries and records can be similarly incomplete. BCG vaccination typically leaves a characteristic scar as an indicator of vaccination; however 17 to 25% of vaccinated children may not develop a scar, independent of whether they develop immunity. Finally, while some consider serology the gold standard for measuring immunity to a disease, this differs conceptually from measuring receipt of a vaccine. Immunization and vaccination status can differ for multiple vaccine or host-related factors including natural infection, lack of immune response to a vaccine, waning immunity, or deactivation of vaccines due to exposure to extreme temperatures. Furthermore, some serological assays may misclassify true immunization status due to innate performance limitations. Nevertheless, serological information can inform vaccination coverage estimates, particularly when it is possible to rule out or distinguish natural infection (tetanus, hepatitis B) or in settings where a disease has been eliminated (measles, rubella, or polio).

A review conducted by Miles et al. synthesized the literature comparing vaccination history obtained from HBR and recall to health provider-based sources for 1975–2011. Compared to provider records, this review found that HBRs under-estimated coverage by a median of 13 percentage points (PP) (range: 61 PP lower to 1 PP higher), while recall over-estimated coverage by a median of 8 PP (range: 58 PP lower to 45 PP higher). The authors concluded that “household vaccination information may not be reliable, and should be interpreted with care.” A review of five studies reporting on validity of caregiver recall (three of the studies were also included in the review by Miles et al.) conducted by Modi and colleagues observed mixed evidence regarding the its usefulness compared to documented evidence of vaccination history in HBRs. Most importantly, however, only five of 45 articles in the Miles and associates’ review (and the two unique studies identified by Modi and colleagues) were conducted in LMICs. Given that immunization programmes located in LMICs are often the most reliant on survey data to help monitor programme performance and have the highest burden of vaccine-preventable diseases, the authors urged further research in these settings. Extending the inclusion criteria to include more sources of vaccination history and adding research from recent years provides a larger body of evidence from LMICs that should be analyzed. Furthermore, in a 2017 consultation by the World Health Organization (WHO), better understanding the reliability of recall was defined as one of the high research priorities around immunization.

We conducted a systematic review on the agreement between recall, HBR, health facility sources, BCG scars, and serological data in LMICs. We also investigated how agreement between these sources varies depending on factors including the type of vaccine, number of doses for a given vaccine, age of the child, and total doses in the country’s vaccination schedule.

Methods

Literature search

We searched Medline and EMBASE for peer-reviewed articles published from January 1, 1957 through December 11, 2017. The search was restricted to human-related publications and included all languages. We adapted the search terms from the Miles et al. review to include additional terms about serology, and restricted to articles with an immunization/vaccination term in the title. We verified that all articles analyzed in the Miles review were found by our search. Articles needed to contain at least one term from each of the following three categories:

- An immunization term in the title: immunization*, immunisation*, vaccin*;
- An agreement term in the title, abstract, MeSH terms or keywords: accuracy, bias, valid*, reliability, misclassification, error, overestimate*, underestimate*, concordance, agreement, sensitivity, specificity, predictive value, comparing*, compare*, comparison*, authentic*;

Reviews and meta-analyses were not eligible, but their reference lists were manually reviewed, as were the references of each eligible article. We consulted with vaccination experts, including researchers and partners who attended an April...
2017 WHO meeting on vaccination coverage surveys, to identify additional studies and unpublished analyses. The review protocol was created with feedback from experts.

The lead author screened all titles and abstracts, then reviewed the full text to confirm eligibility. Studies needed to meet several inclusion criteria. First, the review was restricted to LMIC, defined by the country’s World Bank income classification for the respective years in which the published studies were conducted. Second, studies needed to report on vaccines administered to children under 5 years of age. Third, eligible studies had to report and/or compare vaccination status at the child-level from at least two sources, including: recall, HBR, a facility-based source, serological data (see details below) or BCG scar. One article used records from a prospective study where mothers reported their children’s vaccinations on a weekly basis; those records were considered as health facility records. Serological studies were only included if the researcher could plausibly distinguish between immunity from vaccination and immunity from disease. This included tetanus, hepatitis B, and measles in non-measles endemic areas (as determined by the authors of each article). We excluded non population-based studies, including vaccine efficacy studies or studies among special populations such as pre-term infants.

Two researchers (ED and LS) independently extracted study meta data, measures of agreement, and findings on factors associated with agreement from each eligible study, using a pre-defined extraction template. Any discrepancies were discussed and reconciled between the two reviewers and the senior author.

**Analysis**

We extracted the following measures for each pair of vaccination history sources in each eligible paper: percentage points (PP) difference in coverage (point estimates only), concordance, kappa statistic, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) (Table 1). When papers did not explicitly report all measures, we attempted to calculate them using information provided in the papers. For example, if the paper reported a 2x2 table, we were able to calculate the desired measures of agreement, even if the author had not reported these in the paper. Sensitivity, specificity, PPV, and NPV require designating one source as the ‘gold standard’ or reference group; we used the same reference group(s) as chosen by the authors of each paper. However, we reiterate that in most settings there is no true gold standard for vaccination status to use as the reference. Therefore, these metrics should be interpreted as measures of agreement between

**Table 1a. 2x2 table comparing two sources of vaccination history, used to calculate measures of agreement.**

<table>
<thead>
<tr>
<th>Comparator source</th>
<th>+</th>
<th>-</th>
</tr>
</thead>
<tbody>
<tr>
<td>True positive</td>
<td></td>
<td></td>
</tr>
<tr>
<td>False positive</td>
<td></td>
<td></td>
</tr>
<tr>
<td>False negative</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 1b. Definitions of measures of agreement.**

<table>
<thead>
<tr>
<th>Measure</th>
<th>Definition</th>
<th>Calculation</th>
</tr>
</thead>
<tbody>
<tr>
<td>PP difference in coverage</td>
<td>Difference between coverage level estimated by the two sources</td>
<td>$\text{Coverage}<em>{\text{comparator}} - \text{Coverage}</em>{\text{reference}}$</td>
</tr>
<tr>
<td>Concordance</td>
<td>% of children with the same vaccination status from both sources</td>
<td>$\frac{\text{True Negative} \times \text{True Positive}}{\text{Total Children}}$</td>
</tr>
<tr>
<td>Kappa statistic</td>
<td>Measure of concordance that corrects for chance agreements. Interpretation: -0.2 = poor; 0.21-0.4 = fair; 0.41-0.6 = moderate; 0.61-0.8 = substantial; 0.81-1.0 = near perfect</td>
<td>$\frac{\text{Observed Agreement} - \text{Expected agreement}}{\text{1 - Expected agreement}}$</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>% of children vaccinated according to the reference source that are vaccinated according to the comparator source</td>
<td>$\frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$</td>
</tr>
<tr>
<td>Specificity</td>
<td>% of children unvaccinated according to the reference source that are unvaccinated according to the comparator source</td>
<td>$\frac{\text{True Negative}}{\text{True Negative} + \text{False Positive}}$</td>
</tr>
<tr>
<td>Positive predictive values</td>
<td>% of children vaccinated according to the comparator source who were vaccinated according to the reference source</td>
<td>$\frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$</td>
</tr>
<tr>
<td>Negative predictive values</td>
<td>% of children unvaccinated according to the comparator source who were unvaccinated according to the reference source</td>
<td>$\frac{\text{True Negative}}{\text{True Negative} + \text{False Negative}}$</td>
</tr>
</tbody>
</table>
two potentially flawed sources, as opposed to measures of validity compared to a gold standard.

For articles reporting on multiple countries or sub-regions within a country, we treated each geographic region as a separate study population.

For articles reporting on multiple age groups, we used the group closest to 12–23 months in the main analyses, and subsequently conducted a separate analysis of how agreement varied for different age groups within a given study.

Similarly, for articles reporting on multiple doses of the same antigen, we present the results for the most commonly reported dosages in the main analysis, and subsequently conducted a separate analysis of how agreement varied for different doses of the same antigen within a given study. The most common antigen-doses were: Bacille Calmette-Guerin (BCG), 1st dose Measles-Containing Vaccine (MCV1), 1st dose Oral Polio Vaccine (OPV1), and 1st and 3rd dose Diphtheria Tetanus Pertussis (DTP), including any DTP-containing combination vaccine. When reported, we also included summary measures for if the child was Up to Date (UTD) on vaccinations for their age, according to the definition used in the original study (with the limitation that that variation in age groups across studies could act as a confounder in the UTD metric).

Analyses were conducted using StataSE 15 and R version 3.3.1.

Results

Search results

The Medline and EMBASE searches identified a total of 4420 unique titles (Figure 1). 10 additional titles were identified by experts, and 2 were identified by manually reviewing references. This totaled to 4432 titles, of which 313 passed title and abstract screening and 27 were eligible for the study. Of these, 6 articles were published prior to 2000, 10 from 2000–2009, 8 from 2010–2017, and 3 were unpublished findings provided directly by researchers identified through the expert network (Table 2). One study contained information on two countries, and one presented results for three sub-national regions, resulting in a total of 30 study sites. 11 study sites were in the World Health Organization (WHO) African region, 5 in the Americas, 4 in the Eastern Mediterranean, 8 in South-East Asia and 2 in Western Pacific. 15 study sites reported on MCV, 14 on DTP, 10 on BCG, 2 on OPV, and 1 on pneumococcal conjugate vaccine (PCV). Three reported on measures of UTD.

Agreement of sources for all childhood vaccines assessed

Recall vs. HBR: Ten papers compared vaccination status based on recall to HBR (Table 3). The median percentage point difference in coverage estimated using the two was small (1 PP), but ranged from -43 to +17 PP. Recall-based coverage estimates were higher than those based on HBR for 12 of 18 data points, but were only over 10 percentage points higher in 3 cases (Figure 2). Median kappa (.55) and concordance (.88) between vaccination status based on recall and HBR were substantially higher than any other comparison, and kappa exceeded .60 ("substantial agreement") 45% of the time (Figure 3). PPV, sensitivity, NPV and specificity exceeded 80% in 94%, 81%, 56%, and 38% of cases, respectively.

Recall vs. Facility Records: Seven papers compared recall to health facility records. Recall-based coverage estimates were higher than those based on facility records in 9 of 14 comparisons, 5 of which exceeded +10 percentage points. The median PP

Figure 1. Article screening.
Table 2. Articles included in the systematic review.

<table>
<thead>
<tr>
<th>First Author</th>
<th>Published</th>
<th>Location</th>
<th>Survey period</th>
<th>Vaccines</th>
<th>Sources of vaccination data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aaby</td>
<td>1998</td>
<td>Guinea-Bissau</td>
<td>1998</td>
<td>MCV</td>
<td>Facility, recall</td>
</tr>
<tr>
<td>Adedire</td>
<td>2016</td>
<td>Nigeria</td>
<td>2013</td>
<td>UTD</td>
<td>HBR, recall</td>
</tr>
<tr>
<td>Colson</td>
<td>2015</td>
<td>Mexico, Nicaragua</td>
<td>2012 - 2013</td>
<td>MCV</td>
<td>HBR, serology</td>
</tr>
<tr>
<td>Dunem</td>
<td>2010</td>
<td>Angola</td>
<td>2005 - 2006</td>
<td>BCG</td>
<td>HBR, recall, scar</td>
</tr>
<tr>
<td>GAVI FCE</td>
<td>Unpublished</td>
<td>Uganda</td>
<td>2015</td>
<td>DTP, PCV</td>
<td>HBR, recall, HBR+recall, serology</td>
</tr>
<tr>
<td>GAVI FCE</td>
<td>Unpublished</td>
<td>Zambia</td>
<td>2015</td>
<td>DTP</td>
<td>HBR+recall, serology</td>
</tr>
<tr>
<td>Gareaballah</td>
<td>1989</td>
<td>Sudan</td>
<td>1998</td>
<td>MCV</td>
<td>HBR, recall</td>
</tr>
<tr>
<td>George</td>
<td>2017</td>
<td>India</td>
<td>2015</td>
<td>DTP</td>
<td>HBR, recall</td>
</tr>
<tr>
<td>Gong</td>
<td>Unpublished</td>
<td>Pakistan</td>
<td>2016</td>
<td>MCV</td>
<td>HBR, HBR+recall, serology</td>
</tr>
<tr>
<td>Hayford</td>
<td>2013</td>
<td>Bangladesh</td>
<td>2010 - 2011</td>
<td>BCG, DTP, MCV, OPV</td>
<td>Facility, HBR, recall, HBR+recall, serology</td>
</tr>
<tr>
<td>Jahn</td>
<td>2008</td>
<td>Malawi</td>
<td>2002 - 2004</td>
<td>BCG</td>
<td>HBR, scar</td>
</tr>
<tr>
<td>Liu</td>
<td>2017</td>
<td>China</td>
<td>2009 - 2015</td>
<td>MCV</td>
<td>Facility, recall</td>
</tr>
<tr>
<td>Luman</td>
<td>2009</td>
<td>N Mariana Islands</td>
<td>2005</td>
<td>UTD</td>
<td>Facility, HBR, recall, HBR+recall</td>
</tr>
<tr>
<td>Mast</td>
<td>2006</td>
<td>Uganda</td>
<td>Not given</td>
<td>DTP, MCV</td>
<td>HBR, recall</td>
</tr>
<tr>
<td>Murhekar</td>
<td>2017</td>
<td>India</td>
<td>2015</td>
<td>BCG, DTP, MCV, UTD</td>
<td>HBR, recall</td>
</tr>
<tr>
<td>Nanthavong</td>
<td>2015</td>
<td>Lao</td>
<td>2013</td>
<td>DTP</td>
<td>HBR, serology</td>
</tr>
<tr>
<td>Pereira</td>
<td>2001</td>
<td>Brazil</td>
<td>Not given</td>
<td>BCG</td>
<td>HBR, recall</td>
</tr>
<tr>
<td>Ramakrishnan</td>
<td>1999</td>
<td>India</td>
<td>Not given</td>
<td>BCG, DTP, MCV, OPV</td>
<td>Facility, recall</td>
</tr>
<tr>
<td>Ruiz-Gomez</td>
<td>2007</td>
<td>Mexico</td>
<td>1999 - 2000</td>
<td>MCV</td>
<td>HBR, serology</td>
</tr>
<tr>
<td>Selimuzzaman</td>
<td>2008</td>
<td>Bangladesh</td>
<td>Not given</td>
<td>MCV</td>
<td>HBR, recall</td>
</tr>
<tr>
<td>Sinno</td>
<td>2009</td>
<td>Lebanon</td>
<td>2003</td>
<td>UTD</td>
<td>Facility, recall</td>
</tr>
<tr>
<td>Srisaravananapanthan</td>
<td>2008</td>
<td>Sri Lanka</td>
<td>2006</td>
<td>BCG</td>
<td>HBR, scar</td>
</tr>
<tr>
<td>Tapia</td>
<td>2006</td>
<td>Mali</td>
<td>Not given</td>
<td>DTP</td>
<td>HBR+facility, serology</td>
</tr>
<tr>
<td>Travassos</td>
<td>2016</td>
<td>Ethiopia (3 regions)</td>
<td>2013</td>
<td>DTP</td>
<td>Facility, HBR, recall, serology</td>
</tr>
<tr>
<td>Ullah</td>
<td>2000</td>
<td>Bangladesh</td>
<td>Not given</td>
<td>BCG, MCV</td>
<td>Facility, recall</td>
</tr>
</tbody>
</table>

The difference was +5 PP. Median concordance was .78, and exceeded .80 for 29% of comparisons. Median kappa was .18, and never exceeded .60. Median sensitivity (.85) and PPV (.80) were higher than median specificity (.50) and NPV (.44).

**HBR vs. Facility Records:** Two papers compared HBR to facility records. Coverage estimates based on HBR were a median of 17 PP higher than those based on facility records, though the range was wide (-61 PP to +21 PP). Most measures of agreement were weak, including a median kappa of 0.00, specificity of 0.01 and NPV of 0.20. Concordance (median=0.77) never exceeded 0.80. Median sensitivity (0.95) and PPV (0.78) were relatively higher.

**Recall + HBR vs. Facility Records:** The same two studies that compared HBR to facility records also compared combined recall and HBR to facility records, with similar results as those noted above for the HBR vs facility records comparison.
Table 3. Summary measures of agreement for standard childhood vaccines and doses, including BCG, DTP3, MCV1, OPV1, PCV1, Yellow Fever (YF) and UTD.

<table>
<thead>
<tr>
<th></th>
<th>N articles</th>
<th>N data</th>
<th>PP diff in coverage est.</th>
<th>Kappa (minimum to maximum)</th>
<th>Sensitivity (minimum to maximum)</th>
<th>Specificity (minimum to maximum)</th>
<th>Concordance (minimum to maximum)</th>
<th>PPV (minimum to maximum)</th>
<th>NPV (minimum to maximum)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall vs. HBR</td>
<td>10</td>
<td>24</td>
<td>-43 to 17</td>
<td>0.55 (0.00 to 0.88)</td>
<td>0.95 (0.46 to 1.00)</td>
<td>0.73 (0.00 to 1.00)</td>
<td>0.88 (0.53 to 0.98)</td>
<td>0.93 (0.64 to 0.99)</td>
<td>0.83 (0.07 to 1.00)</td>
</tr>
<tr>
<td>Recall vs. HF</td>
<td>7</td>
<td>14</td>
<td>-29 to 34</td>
<td>0.18 (-0.01 to 0.57)</td>
<td>0.89 (0.51 to 1.00)</td>
<td>0.50 (0.00 to 0.76)</td>
<td>0.78 (0.50 to 0.94)</td>
<td>0.80 (0.49 to 0.99)</td>
<td>0.44 (0.20 to 0.86)</td>
</tr>
<tr>
<td>HRB vs. HF</td>
<td>2</td>
<td>5</td>
<td>-61 to 21</td>
<td>0.00 (-0.12 to 0.06)</td>
<td>0.95 (0.32 to 0.99)</td>
<td>0.01 (0.01 to 0.91)</td>
<td>0.77 (0.38 to 0.77)</td>
<td>0.78 (0.78 to 0.98)</td>
<td>0.20 (0.01 to 0.27)</td>
</tr>
<tr>
<td>HBR + recall vs. HF</td>
<td>2</td>
<td>5</td>
<td>-40 to 20</td>
<td>0.01 (-0.05 to 0.07)</td>
<td>0.94 (0.53 to 1.00)</td>
<td>0.05 (0.00 to 0.69)</td>
<td>0.77 (0.54 to 0.80)</td>
<td>0.80 (0.79 to 0.94)</td>
<td>0.17 (0.13 to 0.50)</td>
</tr>
<tr>
<td>Recall vs. serology</td>
<td>2</td>
<td>7</td>
<td>-32 to 2</td>
<td>0.26 (0.13 to 0.71)</td>
<td>0.23 (0.09 to 0.99)</td>
<td>0.90 (0.56 to 1.00)</td>
<td>0.73 (0.56 to 0.95)</td>
<td>0.95 (0.33 to 1.00)</td>
<td>0.79 (0.68 to 0.86)</td>
</tr>
<tr>
<td>HRB vs. serology</td>
<td>5</td>
<td>14</td>
<td>-38 to 36</td>
<td>0.21 (0.00 to 0.84)</td>
<td>0.91 (0.50 to 1.00)</td>
<td>0.44 (0.00 to 1.00)</td>
<td>0.79 (0.54 to 0.95)</td>
<td>0.93 (0.57 to 1.00)</td>
<td>0.52 (0.07 to 0.83)</td>
</tr>
<tr>
<td>HBR + recall vs. serology</td>
<td>3</td>
<td>4</td>
<td>-36 to 14</td>
<td>0.21 (0.02 to 0.48)</td>
<td>0.79 (0.60 to 0.91)</td>
<td>0.48 (0.38 to 0.65)</td>
<td>0.69 (0.60 to 0.88)</td>
<td>0.92 (0.69 to 0.96)</td>
<td>0.33 (0.05 to 0.70)</td>
</tr>
<tr>
<td>HF vs. serology</td>
<td>2</td>
<td>7</td>
<td>-3 to 4</td>
<td>0.05 (-0.09 to 0.23)</td>
<td>0.80 (0.62 to 0.93)</td>
<td>0.33 (0.04 to 0.60)</td>
<td>0.67 (0.60 to 0.88)</td>
<td>0.87 (0.71 to 0.94)</td>
<td>0.28 (0.03 to 0.40)</td>
</tr>
<tr>
<td>HF + HBR vs. serology</td>
<td>1</td>
<td>4</td>
<td>-6 to 20</td>
<td>0.00 (-0.1 to 0.00)</td>
<td>0.97 (0.98 to 1.00)</td>
<td>0.00 (0.00 to 0.00)</td>
<td>0.87 (0.74 to 1.00)</td>
<td>0.90 (0.79 to 1.00)</td>
<td>0.00 (0.00 to 0.00)</td>
</tr>
<tr>
<td>HBR vs. scar</td>
<td>3</td>
<td>3</td>
<td>-4 to 11</td>
<td>0.08 (0.00 to 0.31)</td>
<td>0.94 (0.85 to 1.00)</td>
<td>0.21 (0.00 to 0.54)</td>
<td>0.89 (0.67 to 0.93)</td>
<td>0.89 (0.74 to 0.98)</td>
<td>0.30 (0.25 to 0.36)</td>
</tr>
<tr>
<td>Recall vs. scar</td>
<td>1</td>
<td>1</td>
<td>2 (2 to 2)</td>
<td>0.43 (0.43 to 0.43)</td>
<td>0.93 (0.93 to 0.93)</td>
<td>0.48 (0.48 to 0.48)</td>
<td>0.86 (0.86 to 0.86)</td>
<td>0.91 (0.91 to 0.91)</td>
<td>0.54 (0.54 to 0.54)</td>
</tr>
</tbody>
</table>
Recall vs. Serology: Two papers including four study sites compared recall to serology. This included one article studying MCV1 vs. measles immunoglobulin G (IgG) and one article (with three study sites) studying pentavalent DTP-Hepatitis B (HepB)-Haemophilus influenzae type b (Hib) coverage compared to tetanus IgG and Hib polyribosylribitol phosphate (PRP) antibodies. In the pentavalent DTP-HepB-Hib study, recall consistently under-estimated compared to serology (range: -32 PP to -13 PP), while coverage estimates were similar in the MCV1 study (2 PP higher according to recall). Kappa showed substantial agreement in the measles study (0.71), and ranged from 0.13 to 0.65 in the pentavalent DTP-HepB-Hib study. NPV (median: 0.79, range: 0.68 to 0.86) and specificity (median: 0.90, range: 0.56 to 1.0) were high relative to other types of comparisons, while PPV (0.33 to 1.00) and sensitivity (0.09 to 0.99) varied widely.

HBR vs. Serology: Five papers including eight study sites compared HBR to serology. One study compared DTP to diphtheria and tetanus antibodies, one compared Pentavalent (with DTP as a proxy) to tetanus and Hib antibodies, and three compared to
measles antibodies. Coverage based on HBR was a median of 2 PP higher than serologically-confirmed coverage, but the difference ranged from -38 PP to +36 PP. Other measures of agreement also varied widely across the studies and antigens.

Recall + HBR vs. Serology: Three papers compared combined recall and HBR to serology, including two comparing DTP3 to tetanus antibodies and two comparing MCV1 to measles antibodies. Recall + HBR under-estimated DTP3 coverage in both cases (-15 to -36 PP). Recall + HBR over-estimated MCV1 coverage for one study (+14 PP) and under-estimated in the other (-4 PP). Kappa, sensitivity and NPV were higher in the MCV1 studies than the DTP3 studies.

Facility Records vs. Serology: Two papers containing four study sites compared facility records to serology, including a measles serum study in Bangladesh and a tetanus antibody study in Ethiopia. There was almost no difference in the population-level tetanus estimates for the three sites in Ethiopia (range: -1 to +4 PP) or the measles study in Bangladesh (-3 PP). Kappa was low (median: 0.05, range: -0.09 to 0.23). Sensitivity and PPV tended to be higher than specificity and NPV.

Facility Records + HBR vs. Serology: One paper compared tetanus serum and tetanus oral fluid to combined facility record and HBR information in Mali. In the 12–23 month-old group, it found that the Facility Record + HBR over-estimated coverage compared to the oral tetanus test by 14 PP, but under-estimated by 6 PP compared to the serum. Sensitivity and concordance was high for both, but the kappa and NPV were zero (or nearly zero).

BCG Scar studies: Four papers reported on BCG scars. Three compared HBR to BCG scars (with scars as the gold standard) and one compared recall to scars. HBR estimated 11 PP higher coverage than scars in one case and 4 PP lower in another, and kappa ranged from 0.00 to 0.31. Sensitivity was high (0.85 to 1.00), but specificity low (0.21 to 0.54). From the one data point available, recall estimated 2 PP higher coverage than scars, with high sensitivity (0.93) but lower specificity (0.48).

Factors associated with vaccination agreement between data sources

Variation by coverage level: When interpreting results, it is important to note that some measures of agreement are inherently affected by the level of vaccination coverage estimated by the reference source. According to mathematical principles, concordance tends to be lowest at 50% coverage and highest at the extremes; PPV increases with coverage; and NPV decreases with coverage. In contrast, kappa, sensitivity and specificity are not affected by vaccination coverage levels. These principles are visibly reflected when comparing agreement measures across studies and vaccines with different coverage levels (Figure 4). However, there is also confounding by factors such as the study setting, types of sources being compared, and type of vaccine. For example, in settings with >=75% coverage, very few data points report NPV above 0.5, with the exception of some comparing recall to HBR.

Variation by antigen: Four studies compared recall to HBR for multiple antigens. In all three cases where PP difference could be calculated, DTP3 coverage was underestimated (-45, -14, and -7 PP) more than any other vaccine or dose (Figure 5). While DTP3 also had the lowest concordance (and BCG the highest), this was explained in part by chance agreement, and no antigen had consistently higher or lower kappa.

Three studies compared recall to facility records for multiple antigens. Two of the studies included DTP3, and DTP3 had the lowest kappa in both (0.50 and 0.57).
Variation by number of doses: Figure 6 depicts data from five studies that reported on multiple doses of the same antigen, allowing us to analyze how agreement varies by dose. Lines connect points showing a different number of doses for the same antigen, type of comparison, and study site. In nearly all studies, the non-gold standard tends to over-estimate compared to the gold-standard for 1 dose, then come closer to the gold-standard value or even estimate lower coverage than the gold-standard at 2 and 3 doses. Kappa values decrease at higher doses in most studies, with the exception of a study comparing DTP from HBR to diphtheria and tetanus serology in Laos. Results are level or inconsistent for PPV and NPV across doses.

Variation by child age: Figure 7 shows the variation in agreement and recall between sources depending on the age of the child, using data from three of the previously described studies that stratified results for the same vaccine dose by age. Lines connect points showing different age groups for the same vaccine/dose and study site. In the Langsten study, the kappa of recall compared to HBR decreases with age. In the Tapia study, kappa for HBR or health facility record compared to serology decreases with age. In the Luman study, kappa for recall and/or HBR measuring UTD vaccination compared to facility records increase from 12–23 to 24–35 month-olds, but then decrease for 72–83 month-olds.

Variation by schedule complexity: It has been hypothesized that increasingly complex national vaccination schedules reflecting recommendations by WHO make it more difficult for caregivers to accurately recall their child’s vaccination history, particularly the number of doses received for multi-dose vaccines. We did not observe a clear, consistent relationship between the number of doses in the national vaccination schedule and the percentage point different in coverage estimates or the kappa statistic for recall as compared to HBR, facility records or serology (Figure 8) though there were relatively few studies.
available at periods of time when the national schedule recommended twelve or more vaccines.

Demographic and other factors associated with agreement:
Two studies analyzed factors associated with agreement. A study comparing recall to HBR in Costa Rica found that having more doses on the card (correlation coefficient: -0.61) and being an older child (correlation coefficient: -0.35) were associated with smaller error with a p-value<0.0001, while factors including community health worker visits, being recorded in health center
records, household size, maternal age and education and socio-economic status were not significant at the 0.0001 level (specific p-values were not provided)\(^5\). In India, a study comparing recall to ongoing prospective reporting found that agreement was higher for younger mothers (1.7 fold increase, \(p=0.03\))\(^7\). Other factors including “father’s age, sex of the child, place of dwelling, parity, mother’s education, family size, previous sibling status and mother’s occupation” were not significantly associated with agreement.

**Discussion**

Our study finds relatively good agreement between vaccination based on documented evidence in HBRs and that obtained from recall, but comparatively poor agreement versus facility-based records or serology in LMIC settings. Agreement varied substantially depending on the study setting, coverage level, type of antigen, number of doses, and child age.

These findings may be used to heighten awareness and inform discussions about the limitations of survey-based coverage estimates. Survey data have been treated as a ‘gold standard’ to validate or adjust administrative coverage sources, but this assumption may not always be appropriate\(^{46-48}\). Furthermore, countries with weak administrative systems for coverage estimation are often the same countries where card availability is low and surveys have to rely more on recall\(^7\). Those using survey-based vaccination coverage should carefully consider the quality of data underlying the estimates for their specific

---

**Figure 8.** Relationship between number of doses in national schedule and (a) percentage point difference in coverage; (b) kappa.
context(s). For example, current HBR availability has been found to vary considerably across Demographic and Health Surveys (DHS) conducted since 2010\textsuperscript{66}. Facility registries are also far more complete and accurate in some countries compared to others, and the ease to use them also varies depending on how they are organized (by date of birth, vs date of vaccination visit for example)\textsuperscript{91}. Additionally, while we did not observe that recall validity is changing over time, we believe this remains an open research question, including the influence of different factors including increasing national vaccination schedule complexity\textsuperscript{12} further complicated by decreasing fertility\textsuperscript{63} and changing patterns in maternal education\textsuperscript{65,65}. In order for decision makers to weigh these potential limitations, it is incumbent on those conducting surveys to be clear and thorough in the documentation of their work, including the limitations. Developing a standard template for vaccination coverage survey reports might further support this need for improved transparency.

We also believe additional steps can be taken during the survey design and data collection process to improve available information collected from respondent recall of child vaccination history. For example, DHS and UNICEF Multiple Indicator Cluster Surveys (MICS) currently require respondents to recall the number of doses the child has received for multi-dose vaccines (after obtaining an affirmative response that the child received the multi-dose vaccine). A response of “I don’t know” is most often not available in the standard response set. By requiring a numerical response (e.g., 0, 1, 2, 3 doses), even when the “true” response is “I don’t know”, respondents and enumerators are forced to undertake an ill-understood, unstandardized imputation processes in the field. The classification of “don’t know” responses has been shown to affect coverage estimates by nearly 20 percent points\textsuperscript{93}. Allowing “don’t know” responses would improve transparency around this important element of uncertainty and empower survey data users to impute in a more systematic way. Surveys might also explore collecting vaccination history from both caregiver recall (asked first of all respondents) and HBRs for all survey respondents, as done in some of the studies included in our review, in order to better assess recall validity among the subset with information from both sources and reveal the directionality and drivers of bias for that particular survey setting.

Despite their limitations and biases, surveys can and will continue to be an important source of information on vaccination programs. As emphasized in the recently updated WHO Survey Reference Manual, surveys will be most useful when they are designed to answer explicit questions\textsuperscript{1}. Clarity about the goals of a survey also gives context to the strengths and limitations of different ascertainment methods and whether additional precision and associated expenses are needed. For example, HBR and recall-based coverage estimates might be considered “good enough” for measuring global or national trends, even if they may over or under-estimate coverage or have poor child-level validity. However, the same data could be inappropriate for measuring achievement against results-based financing goals, as cautioned by the WHO’s Strategic Advisory Group of Experts on Immunization in 2011\textsuperscript{66}. Greater precision may also be needed to detect change in high-coverage settings\textsuperscript{57}. HBR and recall-based histories could also be problematic if a goal is to monitor equity across socioeconomic groups, as HBR availability and recall bias can vary by the same socioeconomic characteristics that are associated with vaccination coverage; more research is needed on this topic given the recent global emphasis on monitoring equity\textsuperscript{8,59}. Of course, survey objectives are often more complicated than the examples given here – a survey may have multiple goals or multiple stakeholders each with their own goals. National immunization programs and other survey implementers could benefit from additional WHO guidance about what type of survey design is most appropriate, if at all, given their specific objectives and available data conditions.

Particularly strong clarity about survey goals is needed to justify the added cost and effort of collecting serological samples, as well as to interpreting those findings\textsuperscript{7}. Across included studies, we find substantial discordance between serology and HBR or recall. This is expected given that serology measures something conceptually different than HBR and recall and reinforces that HBR and recall are poor proxies when a survey needs to measure immunization status, as opposed to vaccination status, of a population. Serology has an obvious added value when a decision should be based on population immunity, for example for disease elimination purposes\textsuperscript{13,60}. However, if the goal is to gather information on vaccination service utilization and dropout, a serosurvey might be difficult and time-consuming to implement and analyze, unnecessary and ultimately wasteful. As methods for collecting and analyzing serology become cheaper, easier and more accurate, researchers and public health officials should continue to explore potential applications, such as using serosurveys to trigger campaigns\textsuperscript{61}.

The intended use of a survey should also guide which specific vaccines are emphasized for analysis and reporting. DTP3 is frequently used as a standard indicator of immunization program performance\textsuperscript{62}. However, DTP3 recall (as compared to HBR and facility sources) is found to have lower concordance and under-estimate coverage by more percentage points than other vaccines in several studies. Therefore, survey users should consider examining other vaccines and doses if precise estimates are needed for decision-making. At the same time, DTP3 may be the most appropriate if the goals are oriented towards measuring delivery and retention in the routine immunization program, given that vaccines such as MCV are often delivered through campaigns in addition to routine immunization. However, the DTP retention metric or dropout (commonly calculated as the relative difference between DTP1 and DTP3 coverage) should still be interpreted with caution given our finding that bias may differ for the 3\textsuperscript{rd} versus 1\textsuperscript{st} dose.

Finally, the large inconsistencies between home and facility-based records when compared to each other, recall, and serology demonstrate inadequate information for health providers for determining which children have and have not been vaccinated. It is important to be aware that each of these sources is imperfect. Indeed, the primary purpose of these data sources is to serve frontline workers, rather than inform coverage
surveys. Without accurate and complete documentation of children’s vaccination histories, vaccinators will continue to miss opportunities to catch up unvaccinated children as well as waste resources re-vaccinating those who may already be protected. Such inefficiencies would likely be considered unacceptable in the private sector or other economic fields, and may be overcome using human centered design and other innovative approaches to optimize existing immunization programme resources.

Our study is subject to several limitations. First, although we believe our literature search to be comprehensive, it is always possible relevant studies were not identified. As a case in point, a similar yet distinct review of caregiver recall was published as this manuscript was being finalized. Second, the articles included in our review frequently reported data in inconsistent ways. We made every effort to ensure comparability across studies, but in some cases, we were missing necessary information about methodological or analytical details. For example, not all studies specified how they treated “don’t know” responses from respondents when asked about their child’s vaccination history and there were possible inconsistencies in how different authors counted the dose of polio recommended at birth (polio 0), when in the schedule. We also only focused on point estimates, thus, not taking into account sampling errors. Additionally, we expect there is special difficulty in differentiating vaccination received through routine delivery of vaccination versus campaign doses, including for MCV. As this issue was often not discussed by the source articles, it may not be well-addressed in our study. Most articles also did not document the phrasing of vaccination history recall questions; studying the best way to solicit recall, including the use of visual cues, is an area for future research. Some of these limitations may be addressed through further analysis of existing data, which the researchers approached as part of this review were agreeable to do.

In conclusion, while recall and HBR provide relatively concordant vaccination histories in some settings, both have poor agreement when compared to facility-based records and serology. In the long-term, improving clinical decision making for vaccination history recall questions; studying the best way to solicit recall, including the use of visual cues, is an area for future research. Some of these limitations may be addressed through further analysis of existing data, which the researchers approached as part of this review were agreeable to do.

Data availability
Underlying data
Open Science Framework: A systematic review of the agreement of recall, home-based records, facility records, BCG scar, and serology for ascertaining vaccination status in low and middle-income countries. https://doi.org/10.17605/OSF.IO/S5UBY

This project contains the following underlying data:
- Supplemental Table 1: List of all articles used in analysis.

Extended data
Open Science Framework: A systematic review of the agreement of recall, home-based records, facility records, BCG scar, and serology for ascertaining vaccination status in low and middle-income countries. https://doi.org/10.17605/OSF.IO/S5UBY

This project contains the following extended data:
- Search term syntax

Reporting guidelines
PRISMA checklist: https://doi.org/10.17605/OSF.IO/S5UBY

Data are available under the terms of the Creative Commons Zero “No rights reserved” data waiver (CC0 1.0 Public domain dedication).

Grant information
This work is supported by the Bill and Melinda Gates Foundation [OPP1055811; OPP1115427].

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements
We thank the authors of the studies that make up this review, the individuals involved in conducting the surveys, and the families who gave their time to respond to the survey questions. We would also like to acknowledge and thank the following individuals for their helpful feedback on the protocol and drafts of this manuscript: Elsy Dumit Bechara, Robin Biellik, Jethro Magwati Chakauya, Trevor Croft, Felicity Cutts, Melody Daly, Michael Edelstein, Marta Gacic-Dobo, Rebecca Grais, Kyla Hayford, Alan Hinman, Cao Lei, Nobubelo Ngandu, David Novillo Ortiz, Bo Pedersen, Bernardo Hernandez Prado, Tove Ryman, Heather Scobie, Messeret Eshetu Shibeshi, and Dana Thomson.

References
5. Demographic and Health Surveys (DHS): DHS Model Questionnaire - Phase 7.


Open Peer Review

Current Peer Review Status: ✔️ ❓

Version 1

Reviewer Report 24 April 2019

https://doi.org/10.21956/gatesopenres.14015.r27015

© 2019 Wiysonge C et al. This is an open access peer review report distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Charles Shey Wiysonge
Division of Health Systems and Public Health, Department of Global Health, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa

Olatunji Adetokunboh
Department of Global Health, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa

We commend Emily Dansereau and colleagues for conducting this timely systematic review on the agreement between recall, home-based records, health facility sources, BCG scars, and serological data in low and middle-income countries. While we applaud the efforts of the authors, we think the review could be improved, as follows:

1. The authors indicate that they prepared a protocol for the review but have not shared it. It is important to provide a link to the protocol. Was the protocol for the review registered in PROSPERO or similar platform? What sort of studies were expected to be included in the review? Was there any deviation from the protocol at the end of the research?

2. What informed the use of 1st January 1957 as the cut-off date for the search?

3. Given the increased research activity on immunization data, a search conducted up to December 2017 should be considered out of date for a paper submitted in March 2019. If the search is not updated, the authors should clearly identify this as a limitation.

4. Quality assessment of included studies is essential. We cannot make a sound conclusion without knowing the quality of the included studies. If this is not done, the authors should identify this as a limitation.

5. Are there particular reasons for not searching for grey literature?

6. In studies with multiple antigens, what informed the choice of antigen that was included in the analysis?
7. The authors should provide the complete search strategy used for one of the two databases, preferably Medline.

Are the rationale for, and objectives of, the Systematic Review clearly stated?  
Yes

Are sufficient details of the methods and analysis provided to allow replication by others?  
Partly

Is the statistical analysis and its interpretation appropriate?  
Yes

Are the conclusions drawn adequately supported by the results presented in the review?  
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Vaccine implementation research, evidence-based health care, epidemiology

We have read this submission. We believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Reviewer Report 09 April 2019

https://doi.org/10.21956/gatesopenres.14015.r27020

© 2019 Hanson C. This is an open access peer review report distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Celina M. Hanson


This is a great systematic review by Dansereau and colleagues examining the agreement of vaccination recall data, home-based records, health facility records, BCG scars and serological data in LMICs and I approve this article. This review is an updated and expanded version from one published by Miles et al. in 2013 and focuses on LMICs with some additional analyses. This article is timely given WHO's 2017 recommendation listing the reliability of vaccination recall a high research priority and given that the previous Miles et al. article only elicited 5 LMICs.

The article is written clearly and the methodology is also mostly clear and seems like it could be easily replicated. What was not entirely clear was the inclusion of grey literature as the article states the initial search for peer reviewed articles. However, manual reference searches and word-of-mouth experts provided means to include other articles; it was not entirely clear that grey literature was an inclusion criteria until reviewing the list of included articles. The article does list as a limitation that it may have missed some articles; however, one recommendation is the inclusion of specific antigen names in the
search terms as this may have elicited more articles.

Overall, this was a well-written and timely article with interesting and thorough analyses including the following variability factors: coverage level, antigen, number of doses, age of child, vaccine schedule complexity and demographic factors.

**Are the rationale for, and objectives of, the Systematic Review clearly stated?**
Yes

**Are sufficient details of the methods and analysis provided to allow replication by others?**
Yes

**Is the statistical analysis and its interpretation appropriate?**
Yes

**Are the conclusions drawn adequately supported by the results presented in the review?**
Yes

*Competing Interests:* No competing interests were disclosed.

*Reviewer Expertise:* Infectious diseases, epidemiology, and public health policies and implementation

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.