Physical activity measurement and surveillance in adults: report of a scoping and planning meeting

27–28 November 2023
Physical activity measurement and surveillance in adults: report of a scoping and planning meeting

27–28 November 2023
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acknowledgements</td>
<td>iv</td>
</tr>
<tr>
<td>Background: Physical activity measurement and surveillance – developing global guidance</td>
<td>1</td>
</tr>
<tr>
<td>Objectives of the scoping meeting</td>
<td>5</td>
</tr>
<tr>
<td>Summary of presentations and discussions: day 1</td>
<td>7</td>
</tr>
<tr>
<td>Presentation 1: Current status of epidemiological evidence on physical activity (Dr Ulf Ekelund)</td>
<td>7</td>
</tr>
<tr>
<td>Presentation 2: Current use of devices in national surveillance (Dr Tessa Strain)</td>
<td>9</td>
</tr>
<tr>
<td>Summary of presentations and discussions: day 2</td>
<td>23</td>
</tr>
<tr>
<td>Presentation 3: Setting the scene (Dr Tessa Strain, Leanne Riley)</td>
<td>23</td>
</tr>
<tr>
<td>Presentation 4: Shortening GPAQ by collapsing moderate- and vigorous-intensity questions while maintaining a global trend estimate (Dr Tessa Strain, Dr Gretchen Stevens)</td>
<td>25</td>
</tr>
<tr>
<td>Presentation 5: GPAQ and measuring muscle strengthening (and balance) activities (Dr Tessa Strain)</td>
<td>27</td>
</tr>
<tr>
<td>Declaration of interests</td>
<td>31</td>
</tr>
<tr>
<td>References</td>
<td>32</td>
</tr>
<tr>
<td>Annex: List of participants</td>
<td>34</td>
</tr>
</tbody>
</table>
Acknowledgements

This meeting report was prepared by Dr Fiona Bull, Physical Activity Unit, World Health Organization (WHO) headquarters.

WHO extends its thanks to Dr Tessa Strain for the first draft, and to the delegates who attended this meeting for their review and additional contributions.

This meeting was supported by funding under the World Health Organization-State of Qatar-FIFA Partnership “Healthy FIFA World Cup Qatar 2022™ – Creating Legacy for Sport and Health”.
Background: Physical activity measurement and surveillance – developing global guidance

The global development policy known as Agenda 2030\(^1\) includes Sustainable Development Goal (SDG) Target 3.4 – a one-third reduction in premature mortality from noncommunicable diseases (NCDs) by 2030. Other key global health goals include the voluntary targets set for the prevention and management of NCDs by 2025, and the ambition of the Global Programme of Work of the World Health Organization (GPW13) to achieve the “Triple Billion” targets,\(^2\) which comprise:

- 1 billion more people benefitting from universal health coverage;
- 1 billion more people better protected from health emergencies;
- 1 billion more people enjoying better health and well-being.

In this policy landscape, the key physical activity policy documents and data that support the achievement of these goals are the:

- **Global action plan on physical activity** (GAPPA)(1), which sets out policy recommendations on how countries can increase population levels of physical activity;
- **Global status report on physical activity** (2), which presents global, regional and country progress on implementing GAPPA policy using a framework of indicators;
- **WHO guidelines on physical activity and sedentary behaviour** (3), which provide a set of recommendations on the amount and types of physical activity required for health benefits for different age groups;
- global target of a 15% (relative) reduction in insufficient physical activity by 2030 from a 2010 baseline (this extended the voluntary target of a 10% (relative) reduction by 2025);
- global estimates on levels of physical inactivity for adults (4) and adolescents (5).

---

2. See [https://www.who.int/data/triple-billion-dashboard](https://www.who.int/data/triple-billion-dashboard).
WHO-STEPS surveillance approach to monitor NCDs and their risk factors

To help countries monitor progress on reducing physical inactivity, WHO provides technical assistance and resources to implement the surveillance and monitoring of NCDs and their risk factors through the STEPwise approach (known as “STEPS”).

In brief, STEPS involves an approximately one-hour long, in-person interview which includes an interviewer-administered survey, and can and usually does include physical measurements (e.g. height and weight, blood pressure) as well as biochemical measures (e.g. fasting glucose). Some measures require individuals to fast and return to have the sample taken the next day.

STEPS includes the Global Physical Activity Questionnaire (GPAQ) to measure physical activity, which is used alongside measurement of other NCD risk factors (e.g. tobacco use, alcohol consumption, diet etc) and is recommended for use among adults aged 18 to 69 years. Some countries may include younger and older ages, and while there is interest in capturing data on the growing older adult population, this presents logistical and sampling challenges.

The GPAQ instrument (now in use in over 130 countries) was developed around 2006 and as such the instrument it is now due for review in terms of its feasibility and relevance for ongoing use in countries; its alignment with reporting on current physical activity guidelines; and its future applicability, particularly in context of developing digital technologies.

Beyond self-reporting – the potential of wearable devices to measure physical activity

While self-report methods are currently the most common for monitoring population levels of physical activity, they have well-known limitations, including the reliability of recall, response bias, and high respondent demands due to length and complexity of questions, as well as nuances within questions and response scales.

Because of this, the potential use of wearable (sensor) devices has received increasing attention in both research and to a lesser degree use in national health surveillance systems. There is however increasing use of wearable devices as the primary assessment method for physical activity in large-scale epidemiological studies on physical activity and health. These studies are now providing evidence on patterns of physical activity and health outcomes, revealing new and different associations in ways that were previously not possible to explore using self-report instruments.

---

3 See https://www.who.int/teams/noncommunicable-diseases/surveillance/systems-tools/steps#--text=The%20WHO%20STEPwise%20approach%20to%20NCD%20risk%20factors%20in%20countries
4 More details on STEPS sampling and protocols were presented and discussed on Day 2.
5 Wearable devices, in this context, refer to a variety of products known as ‘pedometers’ (step counters) and ‘accelerometers’ which are available and used for scientific research and for and by consumers. It is also recognized that ‘smartphones’ and linked ‘smartwatches’ have emerged over the last decade as another digital means to measure movement, heart rate and other physiologic measures through their in-built sensors, as well as location tracking.
The advantages of wearable devices are that they are able to continuously record movements over days or weeks, have been widely used in research among adults and youth, and can provide an assessment of physical activity and sedentary behaviours, as well as sleep.

**There is no global consensus on wearable digital devices as yet**

Despite the growth in the market for, and the use of, wearable digital devices, there is no global consensus on significant issues related to their use, such as, the variety and differences among devices and applications, and the impact that these differences can have on the output metric reported. This variability affects the consistency and comparability of results, and thereby hinders wider adoption in national health surveillance systems and the comparability of results and global knowledge transfer.

The lack of scientific agreement on key device characteristics, data processing and reporting protocols, combined with the absence of authoritative guidance to policy-makers and researchers, represent a major constraint to progress in the adoption and use of wearable devices in population health surveillance systems, and in the development of an optimal, fit-for-purpose wearable device to assess physical activity and sedentary behaviours.

There is therefore a need for WHO global guidance on this agenda – a role consistent with wider calls for WHO to support countries in strengthening digital health as outlined in the *WHO global strategy on digital health 2020–2025* (6), for which WHO serves as the Secretariat.

**Developing global guidance on the measurement and surveillance of physical activity**

Updating global guidance on the measurement and surveillance of physical activity was called for in the GAPPA and by the Guideline Development Group for the *WHO guidelines on physical activity and sedentary behaviour*. More recently, improving data information systems was identified as a priority action in the *Global status report on physical activity* to help countries measure outcomes and impact as a result of GAPPA implementation. As such, the development of updated guidance on physical activity measurement has been approved as a global technical product for the 2024–2025 biennium.

This report feeds back on the first of a series of planned technical meetings of scientific and technical specialists in measurement methods, wearable sensor technology and population surveillance of physical activity. This meeting supported WHO's efforts to scope the potential for using wearable digital devices in the pursuit of monitoring and surveillance of global physical activity in adults, and to review the use of GPAQ.

The report also sets out the objectives of this scoping meeting and summarizes the discussions and conclusions that were generated.
Objectives of the scoping meeting

Meeting objectives: day 1

1. To review evidence on the use of wearable digital devices in national surveillance systems to assess physical activity in adults and identify the technical and protocol issues that are arising.

2. To discuss the emerging new evidence on the dose-response relationships between physical activity, sedentary behaviours and clinical health outcomes from recent large epidemiological studies using wearable device-based measurements; and to identify the potential implications for future measurement and monitoring of physical activity in national surveillance systems, and the updating of global guidelines on physical activity and sedentary behaviour in 2030.

3. To develop recommendations on the technical and research tasks (and estimated timelines) required to support the development of global consensus on the instruments and protocols for use of wearable digital devices to measure physical activity in adult population surveillance systems.

4. To discuss and draft a programme, and identify potential key experts, for an expert meeting on physical activity measurement and surveillance in adult populations.

Meeting objectives: day 2

1. To review and discuss updating the GPAQ and draft preliminary recommendations on any potential changes to it.

2. To develop a draft plan of any field testing or analytical work required to support final decision-making on recommendations for changes (if any) to the GPAQ.
Summary of presentations and discussions: day 1

Following a presentation by Dr Fiona Bull and Leanne Riley on the current global policy context, current physical activity guidelines and the current approach to national surveillance of physical activity in adults (see content in the “Background” section of this meeting report), further presentations were made on the current status of epidemiological evidence around physical activity and on the current use of devices in national physical activity surveillance.

Presentation 1: Current status of epidemiological evidence on physical activity (Dr Ulf Ekelund)

Key learnings from epidemiological studies using self-report measures of physical activity:

- Evidence shows there is a marked reduction in the risk of mortality and NCDs among people who are active at levels lower than the current guidelines compared to those that do no exercise.
- Strong and consistent evidence shows that the shape of the dose-response is similar across outcomes – with the greatest benefits being had from doing some physical activity versus none, and maximal benefits from higher levels of activity.
- Findings show that there may be some variation in the benefits (strength of association) depending on the domain of physical activity – that is, occupational physical activity may not be (as) beneficial as recreational/other domains.
- Vigorous-intensity physical activity may be more beneficial than non-vigorous physical activity.
Key learnings from the growing body of evidence from epidemiological studies using device-based measurement of physical activity

- Devices have provided better quantification of light-intensity activity, shorter bouts of activity, and metrics of “step counts”, and this has introduced a new range of metrics to quantify physical activity levels.

- Much of the device-based epidemiological evidence comes from analysis of the United Kingdom Biobank dataset – multiple studies have been published but with a variety of different methods of sampling and analysis. These differences are unhelpful in developing a full understanding.

- Evidence from available device-based studies shows:
  - total physical activity (counts per minute or steps) is associated with reduced risk of mortality;
  - all intensities are associated with reduced risk of mortality;
  - a similar shape of dose-response as seen with self-reported methods but much higher risk reductions.

- Importantly, does the evidence suggest that the relative benefits of vigorous-intensity compared to moderate-intensity physical activity may not be in the “2:1 moderate-intensity physical activity to vigorous-intensity physical activity ratio (MPA:VPA)” which has been estimated in the past using self-report data and which is then used in data analysis, with, for example, approximately 3MET (or sometimes 3.5) and 6 MET for moderate and vigorous intensity, respectively?

- Evidence is emerging that vigorous-intensity physical activity may provide greater benefits than the “double” value of moderate-intensity physical activity.

- Device-based evidence provides an opportunity for better analyses of the interactions between physical activity and sedentary behaviour; results are revealing strong interactions between sedentary and moderate-to-vigorous intensity physical activity, with physical activity attenuating or maybe even offsetting the unfavourable associations between sedentary time and mortality.

What next for research using device-based measures?

- There is a need to understand comparability of data worn in different locations – waist, wrist and thigh-worn devices.

- More work is needed for wider range of health outcomes (to date most of the device data is on all-cause mortality and cardiovascular disease, with only some papers on site-specific cancers and even fewer on diabetes).

- More research is needed that allows for data triangulation to understand causality.
• More work is needed to understand the relative importance of different intensities of physical activity and whether the “total volume” of physical activity is most important.

• More work is needed to better identify and understand if there are age-specific associations.

Presentation 2:
Current use of devices in national surveillance (Dr Tessa Strain)

This presentation was supported by an accompanying background paper using data from a systematic review undertaken by Dr Hidde van der Ploeg and colleagues (currently under review for publication) that had been updated for this WHO meeting.

Key shared learnings were:

• Device-based measures of population levels of physical activity have been undertaken in 10 (inter)national surveillance systems, the majority of which are from high-income countries.

• There are wide differences in the characteristics of use between these 10 examples of physical activity surveillance using device-based measures. These differences (around 20) were grouped and summarized into four categories:
  1. Data collection
  2. Wear protocols
  3. Device properties
  4. Processing protocols

• Given the heterogeneity in use across these 10 (multi)country examples, and to support other countries going forward, there is a need for global standards and protocols on the use of wearable devices for population surveillance, particularly to enable the resulting data and estimates to be comparable.
Discussions

Use of wearable devices and experiences, and evidence to date

A group exercise consolidated the discussion into pros and cons of wearable and self-report instruments – see Table 1 and Table 2.

Table 1: Advantages and disadvantages of self-report instruments (and specifically GPAQ)

<table>
<thead>
<tr>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-report data are easy to integrate within other larger surveys</td>
<td>Can be annoying and/or difficult for participants because of length/number of questions, apparent repetition, complexity (e.g. due to potentially confusing domains, types, intensity)</td>
</tr>
<tr>
<td>Self-report instruments (including GPAQ) have usually been tested and reported levels of “validity” and “reliability” are available</td>
<td>Some debates around validity of self-report data</td>
</tr>
<tr>
<td>They are deemed relatively easy and cheap compared with potential costs (logistics and purchase) of wearable devices</td>
<td>Significant errors around reporting of minimums because questions can require a respondent to summarize (as with GPAQ) the “average” time spent doing physical activity (in minutes)</td>
</tr>
<tr>
<td></td>
<td>Errors due to social desirability bias, cognitive recall issues, comprehension difficulties</td>
</tr>
<tr>
<td>Self-report instruments can provide population trend data if the same instrument and methods are used in consecutive surveys</td>
<td>May have biases between subgroups that do not exist when using device-based measurement methods</td>
</tr>
<tr>
<td></td>
<td>Comparability between self-report questionnaires is not high – i.e. any changes or differences between question order or wording makes a difference to responses (including the inclusion or exclusion of minimum threshold of 10-min bouts (GPAQ removed this criteria in 2020 in response to new guidelines))</td>
</tr>
<tr>
<td></td>
<td>Self-report instruments with more questions generally lead to higher reported levels of activity</td>
</tr>
<tr>
<td></td>
<td>Mode of completion/delivery of instrument (in person, telephone or online) can affect response</td>
</tr>
<tr>
<td>Advantages</td>
<td>Disadvantages</td>
</tr>
<tr>
<td>------------</td>
<td>---------------</td>
</tr>
<tr>
<td>Self-reports assess the individuals' perceived level of physical activity intensity (i.e. is a relative measure rather than absolute/criteria defined)</td>
<td>Measures perception of activity/behaviour, not the actual behaviour</td>
</tr>
<tr>
<td>Self-reports can provide measures of domain-specific physical activity (i.e. a context for physical activity) which has policy relevance (i.e. transport, recreation/leisure, work/household)</td>
<td>Currently, most self-reports (including GPAQ) are limited to only assessing moderate- and vigorous-intensity physical activity</td>
</tr>
<tr>
<td>Self-report instruments could provide a measure of strength and balance physical activity if questions were added (but this has remained difficult)</td>
<td>Current instruments (including GAPQ) do not provide measure of light-intensity physical activity</td>
</tr>
<tr>
<td>Self-reports (and GPAQ) cannot collect a “step” output in minutes</td>
<td>Self-reports are not (easily) designed to capture short bouts of any type/intensity – a limitation particularly pertinent to an emerging body of evidence from wearable device studies showing steep beneficial dose-response curves of VPA and MVPA. As such, self-reports cannot assess adherence to the current physical activity guidelines as these are based on the concept that “every minute counts”</td>
</tr>
</tbody>
</table>

Table 2: Advantages and disadvantages of wearable devices for assessing physical activity

<table>
<thead>
<tr>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wearable devices measure the movement of an individual, not the perception and recall of movement by an individual</td>
<td>Wearable devices measure the movement of the device worn by an individual rather than perception of behaviour</td>
</tr>
<tr>
<td>Wearable device data are potentially more comparable between people than are questionnaires as they may eliminate some biases that arise in self-report instruments (including time recall and intensity recall)</td>
<td></td>
</tr>
<tr>
<td>Wearable devices can quantify physical activity on a small timescale (for example can capture bouts lasting from 5–10 seconds to just a few minutes)</td>
<td></td>
</tr>
<tr>
<td>Advantages</td>
<td>Disadvantages</td>
</tr>
<tr>
<td>--------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Wearable devices can assess intensity of physical activity, thus they align with reporting on MVPA, and would be able to assess light-intensity physical activity when defined</td>
<td>Some physical activity may not be well captured (e.g. cycling with wrist-worn devices)</td>
</tr>
<tr>
<td>Some devices and protocols can assess physical activity types and postures (e.g., walking, stair climbing, cycling, running, standing) – and this may improve with technological advances and machine-learning analysis</td>
<td>Cannot provide a measure of strength or balance activities (as yet - this may change in time)</td>
</tr>
<tr>
<td>Can provide an output metric in “steps”, which is considered an easier metric to communicate to the public and policy-makers. Also step data can be categorized as purposeful/incidental, for more refined transition into interventions, and data analysed to provide a measure of step intensity (cadence) (Note, there is no global “step count recommendation” as yet).</td>
<td>Currently, cannot provide a domain (context)-specific metric of physical activity as a standalone measure where these are deemed useful for policy (this may change over time and could be addressed in better algorithms and supported by questionnaire data)</td>
</tr>
<tr>
<td>Can assess the full 24-hour intensity spectrum of movement including sleep, permitting seamless application of compositional data-based analyses</td>
<td>No standard protocols for analysis, and analyses are complex (a specialist task)</td>
</tr>
<tr>
<td>Easy to integrate with another measure or survey</td>
<td>Devices need to be worn for multiple days and require a protocol for initiation, training, and device retrieval after a period of days (so adds complexity to a typical surveillance survey data-collection system)</td>
</tr>
<tr>
<td>Devices are getting cheaper – and could more so if demand and quality standards were established</td>
<td>Deemed expensive (financially and logistically) – compared to using a questionnaire</td>
</tr>
<tr>
<td>Devices may have battery-life issues (which may be solvable in time)</td>
<td>Devices may be lost/stolen/broken – current prices make this an unattractive additional cost to incur</td>
</tr>
<tr>
<td>Advantages</td>
<td>Disadvantages</td>
</tr>
<tr>
<td>---------------------------------------------------------------------------</td>
<td>------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>More types of sensors are now being added in to wearable devices – e.g. for heart-rate monitoring</td>
<td>More complex features in devices can make them more expensive and thus prohibitive for large-scale purchase and use in current population surveillance systems</td>
</tr>
<tr>
<td></td>
<td>Added complexity of devices with multiple sensors can also add complexity to the data processing and analysis (and require longer battery life, potential collection or recharging, and so more complex instructions to participants or logistics)</td>
</tr>
<tr>
<td>Many wearable devices are available, and many are well tested, and thus have publicly available results on validity and reliability</td>
<td>Heterogeneity between devices (functions and signals) and protocols (wear placement, duration etc.), make comparison difficult</td>
</tr>
<tr>
<td>Modern accelerometry methods are device agnostic, permitting use of different (e.g. best-value-for-money) devices in different countries/ settings</td>
<td>Significant data-storage and data transfer issues (which may be solvable in time)</td>
</tr>
<tr>
<td>Data collected and stored (raw) can be retrospectively analysed if national guidelines change and/or signal-processing algorithms improve – thus enabling the application of future analyses</td>
<td>No consensus on (1) standard output metrics for a wearable device AND (2) standard operating procedures for use and data analysis</td>
</tr>
<tr>
<td></td>
<td>Does not provide a comparable trend estimate with data derived using past self-report instruments</td>
</tr>
<tr>
<td>Device-based data can provide new insight and deeper understanding of the patterns of physical activity by different populations thanks to granular data over time</td>
<td></td>
</tr>
</tbody>
</table>
What needs to be measured and how can devices support this?

**Broad principles**

- The purpose of surveillance for WHO is to *measure guidelines*, monitor progress to targets, highlight subgroups at risk, and inform policies and interventions.
- At a minimum, we need to measure everything in the guidelines.
- Based on WHO needs, it is vital to remember the *importance of trend data*. Some elements must remain constant or have an overlap period to be able to allow for adjustment.
- *We may be moving towards more “personalised” guidelines* and potential future guidelines may offer multiple ways to achieve health benefits – i.e. “choose what suits you”.

**Step counts – as a more understandable metric for measuring level of physical activity**

- Step counts offer a way to link device data to current guidelines.
- The next physical activity guidelines in the United States of America (USA) may include step counts and the USA has formed the Steps for Health Collaborative (7) to produce relevant evidence. Other bodies are also publishing similar reviews (8–10).
- That said, using step counts is not always straightforward:
  - Defining a step is difficult. For example, one wearable device company is now on its sixth step count algorithm in 5 years. Polar, a brand of wearable devices, is using a “metabolic step” to include activities such as swimming/cross-country skiing into step counts.
  - There is a need to be aware of differences across devices – although most should be within ~500 steps/day. This could be considered to be tolerable between device accuracy.
  - Step count may be an inadequate measure of total activity for those who are active – i.e. more active people do more physical activities where “steps” are not detected, such as cycling or swimming.

**Where does light-intensity physical activity fit in?**

- There is a need to gather more evidence on the importance (associations) of light-intensity activity for health, and collecting light-intensity activity is possible with device-based measurement.
- There is still debate as to whether light-intensity activity is a proxy for total activity or an inverse of sedentary time.
What about sleep?

- WHO sleep recommendations exist only for children aged under 5 years. The remit for the WHO guideline update in 2020 did not include adding sleep for other age groups.

- Upcoming epidemiological evidence may allow for consideration of a recommendation for other age groups on sleep in the future.

Importance of vigorous-intensity physical activity and length of bout

- Device-based evidence may challenge the current understanding and guidelines on the recommended amount of vigorous-intensity physical activity – results show that shorter bouts of time may be more beneficial than previously thought. New insights are because device-based measures capture shorter bouts, which self-reports cannot.

- Prevalence of vigorous-intensity physical activity according to devices is much lower than that identified using self-report measures, due to questionnaires capturing the entire time-window during which a cluster of bouts of vigorous physical activity (including interruptions) takes place; devices capture only actual physical activity time.

- Measures of physical activity minutes (at light, moderate or vigorous intensity) is influenced by bout length, wear location, and how intensity thresholds are derived.

Daily or weekly guidelines

- Although many device studies present data as daily averages, it should not be expected that guidelines will change from weekly to daily indicators. Results from device studies still suggest that the “weekend warrior” pattern of physical activity is equally beneficial compared with physical activity evenly spread across the week – therefore guidelines should reflect this.

Strength and balance

- There was agreement that the guideline recommendation has not been included in assessments of physical activity; it has been overlooked, in part because of the complexity to measure it and no consensus on methods.

Measurements from wearables that go beyond accelerometry

- Wearables can measure other parameters, e.g. heart rate, blood pressure, eating patterns (via wrist accelerometry).
• Using heart-rate measurement has the potential to help distinguish physical activity intensity levels and non-wear time. But instruments with heart-rate sensors have additional complexity with the processing algorithms (and are more expensive due to the sensor).

• Difficulties were noted around the need for individual calibration to make sense of heart rate.

• Additional potential issues include wear location, accuracy on different skin tones, use in humid temperatures.

• New data and insights will come from the Fenland study (at the University of Cambridge, United Kingdom) that has used a chest-worn device combining an accelerometer and heart-rate sensor (11); and United Kingdom company Biobank have asked a sub-sample of 30 000 people to wear a chest patch to obtain cardiac data.

**Can wearable device measures complement self-report measures?**

• There was overall agreement that in the short- and mid-term, wearable devices can likely offer complementary measures to self-report measures.

• Until device-based measurement protocols exist, and devices are affordable and practical, many countries will still use self-report surveys to assess population levels of physical activity.

• It was suggested that WHO create a “toolkit” of measurement options including:
  – updated advice on continuing to measure physical activity by self-report only;
  – options for device-based assessment – with different options to suit the context and available resources (price of device as well as processing capabilities);
  – a questionnaire to “complement” the device-based measure because personal/social demographic and other data are still needed.

• It was recognized that it might take several years or more to introduce the use of wearable devices to countries and some may want to continue with self-report data collection.

• It is important to present wearable devices as complementary measures and not abandon or discredit the option of gathering self-report data.

• Research in the USA has been undertaken using the National Health and Nutrition Examination Survey (NHANES) dataset to harmonize GPAQ measures and device-based data (12).

• A questionnaire that complements wearable device measures should play to the strengths of self-report surveys (i.e. categorization of physical activity rather than quantification of physical activity), and could include questions on context of physical activity (e.g., domain/setting/location) and policy-relevant questions such as the type of physical activity – sport, cycling etc.
What is the best wear location?

- There was unanimous agreement that wrist-worn is the only feasible placement for global population surveillance as it is the most user-friendly across populations and achieves the best compliance (13).
- The non-dominant wrist is preferred as this minimizes (but does not eliminate) detection of non-activity movements of the dominant arm (e.g. chores, toothbrushing for example) but there may be other problems, e.g. wearing on non-dominant wrist might interfere with wearing a standard watch or another consumer wearable and affect adherence.
- There is recognition that other wear locations may be preferable for other research study purposes.

What is the best minimum analytical wear time?

- Wear time is a critical issue as it determines all the other logistics and protocols for use in health surveillance systems such as WHO-STEPS. And the logistics of using devices in surveillance studies are a key component of the total cost.
- The issue of reactivity (i.e. doing more activity because it is being measured by the device) by participants to being issued a device to measure physical activity was discussed, but overall it was deemed to be a minor issue when wear time is over multiple days, as usual behaviour returns relatively quickly.
- The NHANES dataset was cited, from which 7-day accelerometry was randomly sampled for different analytical wear times (14). At population level, one random day was sufficient for a stable estimate. Note – there may be issues for subgroups, and this is not the same as a 1-day wear protocol.
- The relatively small dataset of the USA’s “All of Us” research programme provides an opportunity to learn more about very long-term changes in behaviour (4.5 years of continuous monitoring using consumer wearables – which is perhaps unsuitable for surveillance).
- Research was cited to show that different wear times are needed for different study purposes (15).

Possibility of “1-day” data collection period

- A 1-day protocol was discussed as it presents the possibility of collecting the device within the data collection logistics of the WHO-STEPS protocol – i.e. starting to wear the device on the day of the survey/interview and returning the following day for biochemical measures.

6 See https://allofus.nih.gov/.
However, a 1-day protocol was deemed unlikely to provide a usable measure of physical activity because:

- it may not provide a full day of wear time due to the timings of the interviews etc. and therefore more wear days may be needed;
- the window of time/day would be atypical of usual patterns because of the activity of being involved in STEPS, and critically because of being asked to fast;
- data collected would not be random across the sample because logistics mean the day of data collection would not be within geographical clusters – this could possibly be overcome across sampling districts;
- data analysed and reported as “per day” were derived from participants wearing the device over multiple days and these days were randomly selected – meaning NHANES was not a one-day data sampling protocol.

Algorithms to classify intensity of physical activity

Overall there was agreement that defining vigorous intensity physical activity with devices is a key area to address as there is no consensus in scientific literature. In addition, what exists may not transfer and be applicable in all contexts as required by WHO to recommend in global surveillance guidance.

A good example of this is the lack of transferability from the United Kingdom Biobank to the China biobank context – this provided a reminder of the need for work to test global representation and validity.

Also noted before, there is also the need to test and validate across the age ranges of interest (e.g. 18–69 for WHO-STEPS surveillance).

From an industry perspective, manufacturers/product developers (e.g. of products such as Google Fitbit) work on a completely different scale in terms of the numbers of people (data points) and period of data collection available to them for analyses. In addition, they have extensive laboratory and field-based data collection (with multiple sensors and validation protocols) – and the use of machine learning.

Proposed options for WHO to consider moving forward include:

- investing in building and trying out “open access” algorithms;
- incentivizing consumer device manufacturers to release testing and validation data and results (and maybe their proprietary algorithms);
- collecting raw data alongside consumer algorithm results.

It was noted that cut points only use one feature of the signal to classify intensity levels. Machine learning methods have the potential to take into account more features and this has the potential to improve analysis.
Measuring physical activity domain and/or type of activity

For this discussion, type = specific activities; domain = context/purpose (e.g. work, household, travel, leisure).

- The activity domain is more important than the activity type for country policies.
- The WHO GPAQ domain analysis was mentioned, showing which domain provides the largest contribution to total physical activity across a set of countries (16).
- Activity-type recognition is possible from devices now and it is expected that the ability to assess and differentiate the type of physical activity will improve in future. Wi-Fi and GPS data can help.
- Assessing the domain of physical activity is harder than assessing type of activity for wearable devices as it is hard to know (accurately) where and why people do activities. For example, is a bout of activity work-related or just around the home? This means that data on work hours/location need to be known if a domain has to be attributed to the activity.
- As discussion of using triangulation and other data sources (GPS) unfolded, it was noted that this introduces additional data privacy issues as well as additional data collection and processing complexity.
- The use of diaries alongside devices is possible for determining work time. It is also worth looking at the research work of time-use epidemiologists.
- It was noted that using a combination of wearable devices and a self-report instrument is likely the most suitable method for the near future, and would allow for some domain (or context) differentiation (i.e. incidental physical activity vs leisure time exercise, or work versus travel or recreation). A modified self-report instrument capturing only the necessary data is needed (i.e. the self-report would not need to attempt to quantify the minutes but rather provide the context) and a modified GPAQ could serve this purpose.

Open or closed devices? (Impact of providing feedback to participant)

- Devices providing feedback are critical to participant engagement, and this does not necessarily need to be information related to data on their physical activity level – e.g. it could simply be just telling the wearer what time it is. In other words, there needs to be an incentive for wearing the device for multiple days for the main purpose (in this case the health surveillance system).
- Field experience from WHO-STEPS shows that anything that makes the device “more attractive” also makes them harder to get them returned and possibly increases the desire for them to be stolen. And until device costs are considerably lower, minimizing loss is very important.
- Another idea is to provide feedback to participants at a later date.
Use of smartphones and other sources of “big data”

- A question was posed on how passive data collection could be used in conjunction with traditional surveillance approaches (note: this is not necessarily limited to smartphones).

- An example of collaboration in Israel was shared whereby smartphone step data was linked to health records in a sample of around ~600,000 people. The results were promising and “passed basic sense checks”, and showed correlation with diabetes prevalence.

- Men “wear” phones more than women.

- A question was raised about the smartphone prevalence in low- and middle-income countries.

- Discussion of the application and potential of smartphones for assessing population physical activity is on the wider agenda of WHO programme of work but not seen as a very near-time solution – a reminder that the catalyst and first priority of the meeting was a review of the GPAQ self-report instrument and the scoping of current and future use of wearables.

Access to raw data

- Collecting raw data allows a return to the data for secondary analysis should guidelines change over time and makes it possible to retrospectively create metrics. Deemed essential to help “future proof” against changes.

- It was suggested that raw data is not necessarily always best unless you have a really good algorithm.

- Data storage is a concern, but it was suggested that it should not be an issue for consumer devices now as batteries and storage have increased capacity and are getting cheaper.

- There was discussion on how countries and WHO can reduce their reliance on external partners who are expert in processing (e.g. universities) to conduct data analysis, as this adds costs which in the long term are not sustainable.

- Open-source software is needed and does exist (see further discussion of this in the section on day 2).
Data sampling frequency

- Collecting data at a higher frequency than is necessary was cautioned against as this creates need for more storage for large data files.
- A minimum sampling frequency should be specified rather than an optimum.
- Discussion of 30 Hz being adequate for today’s algorithms but 50 Hz would be more “future proof”.

Device cost versus robustness

- To be practical for WHO to recommend their use to countries, device cost needs to be lower. A theoretical price of US$ 10–20 per unit was suggested, or preferably lower.
- While price of the device was a large cost it was also noted that the cost of the logistics and handling/processing of data is of concern and prohibitive for wider country use.
- There was consensus that devices need to be robustly designed to be fit for purpose and be used in WHO-STEPS and other country surveillance systems – they need to withstand being used multiple times and be designed to avoid damage and faults in diverse contexts of use (humidity, heat, cold etc.). Devices need to be waterproof.
- Allowing low cost to drive down to the “lowest common denominator” was cautioned against. Countries should be able to collect more data if they want – a suggestion that echoed the proposal of a tiered toolkit of different options for different budgets and contexts.
- There was general agreement that there are devices currently available that can provide data and fit this surveillance purpose, however the price is high (and notwithstanding other logistical problems) this limits scaling of use. One approach might be to find a current device that matches the desired specifications and discuss with industry how “cheaply” it could be made.
- There was also general agreement that industry needs to be involved in this discussion going forward.
- WHO is aware of this, and has a mechanism to meet and engage with the private sector and has already done so in past forums, starting with a smaller group of manufacturers. Deciding when to engage the industry more widely will be addressed by WHO as it follows the Framework for Engagement with Non-State Actors (FENSA) and other WHO processes.
Summary conclusions

- It was agreed that there was a clear role for wearable devices to be part of future physical activity surveillance systems.

- The idea emerged that WHO could provide guidance on a “toolkit” of different approaches in using devices to address the diversity of resources, contexts and capabilities in countries, and allow choice.

- WHO should provide guidance and protocols that are evidence based, affordable and feasible.

- There is a need for a minimum set of specifications for wearable devices for the purpose of population assessment of physical activity.

- There is need for more discussion on, and detailed knowledge of the constraints of integrating wearable devices into the current and foreseeable protocols and logistics of WHO-STEPS. There is also a need to agree on which physical activity outcomes are most important for physical activity surveillance systems to report and track, and then create and validate (open source) algorithms to track them.
Dr Fiona Bull opened day 2 by setting out the window of opportunity currently available to review and change the current GPAQ, and explaining that WHO is conducting a review and revision of the entire WHO-STEPS questionnaire. It was noted that any recommendations and developmental work on GPAQ changes would need to be finalized by July 2024.

**Presentation 3:**
**Setting the scene (Dr Tessa Strain, Leanne Riley)**

Summary data were presented on the current use of GPAQ worldwide, drawing from recent systematic collection of all physical activity data by WHO in 2022/23 (publication in review). The WHO-STEPS protocol was outlined in detail, including when and how GPAQ was typically administered in a country. The following issues have repeatedly been experienced and raised during field work:

- It takes a long time to administer GPAQ (there are too many questions).
- GPAQ questions feel repetitive (particularly because the sub-questions on moderate- and vigorous-intensity physical activity within domain are almost identical).
- Respondent confusion about what is moderate and vigorous intensity (note STEPS does include use of country-specific show cards illustrating relevant types of different intensity).
- Respondents are confused about housework and chores (more details on this confusion are required).
- The 10-minute threshold “was” confusing – this has been removed since WHO 2020 guidelines.
• Some respondents find it difficult to estimate time spent.
• Some confusion over sedentary activity versus sleep.
• Showcard examples are critical but difficult to contextualize.
• GPAQ does not capture some elements of the new WHO guidelines (2020), nor does it capture sleep – which is getting an increasing level of interest from countries, some of whom add their own question on it.

Summary details of WHO-STEPS logistics

• WHO headquarters technical team supports 10–15 countries per year to undertake WHO-STEPS, and currently often more as countries catch up on health surveillance since the COVID-19 pandemic.
• WHO-STEPS implementation in a country can take several months to plan, secure resources for, and train staff for data collection etc.
• Once started, WHO-STEPS data collection usually takes 2–3 months in a country, depending on context.
• Data collection tries to avoid rainy seasons and atypical periods (such as religious events).
• GPAQ is part of a 60-minute health survey interview that includes collecting demographic data.
• It is not known exactly how long GPAQ takes (although the collection done using tablets is time-stamped so it could be calculated).
• Countries can add extra questions to WHO-STEPS and to the physical activity section – for example on sleep, or muscle strength/balance, but there is no set WHO questions for these.
• The largest cost comes from difficulties with the logistics of using devices – specifically their retrieval from users given the WHO-STEPS data collection teams move from one area to another across the country. A larger number of devices is needed to reduce the time pressure of “collection and turnaround” for second/third use of devices.
• Other logistical issues include shipping costs and restrictions – shipping lithium batteries is burdensome.
• A rough calculation of the number of devices needed by WHO for a WHO-STEPS country survey is ~8000 devices for 15 countries, with the potential to lose 1600 devices if measuring ~80 000 participants per year. This is in range of a philanthropic donation.
Discussion

Clarifications on the removal of the “10 minute” threshold for bouts

- A question was raised around what is known about the effect of removing the “10-minute bout threshold” from the question in the GPAQ in WHO-STEPS data collection. This has not been investigated in WHO-STEPS data sets.
- A paper published in the USA (17) suggests little change in compliance with guidelines when not using the 10-minute bout threshold.
- Removing “10-minute-bout threshold” will have a huge impact on estimates of adherence to (meeting the) guidelines when using wearable devices – evidence shows that roughly 70% of adult Norwegians are active when counting “every minute” compared to only about 30–35% when using 10-minute bout criteria allowing for a 2-minute drop.
- Differences in prevalence-estimate comparisons vary but may be smaller with self-report instruments (likely explained by the fact that in self reports responders are not actually very effective at restricting their reporting to only those minutes of physical activity accrued in 10-minute bouts).
- It was also reported that this smaller difference was seen in work conducted in Sweden.
- Evidence is also available in another research paper (18).

Presentation 4:
Shortening GPAQ by collapsing moderate- and vigorous-intensity questions while maintaining a global trend estimate [Dr Tessa Strain, Dr Gretchen Stevens]

One proposal is to shorten and simplify GPAQ (and remove confusion and repetition) by asking about moderate- and vigorous-intensity activity as one question (currently these are two separate questions).

Results from an analysis of the impact of combining questions on moderate- and vigorous-intensity activity were presented using a WHO dataset of over 700 000 participants and their GPAQ responses. While the data set is not strictly “globally
representative" it nevertheless has global coverage based on the large number of countries included. Results showed that:

- ~30% individuals report “any” vigorous-intensity activity;
- moderate-intensity physical activity makes up the majority (average 77%) of total MET-minutes of MVPA per week;
- most people who meet the guidelines report >150 mins/week of moderate-intensity activity;
- only ~5% of individuals meet the guideline through some contribution of vigorous intensity activity.

A key issue (and concern) regarding the suggestion of combining moderate- and vigorous-intensity activity into one question would be the inability to assign different MET values (currently moderate intensity = 4 MET and vigorous intensity = 8 MET), and this would have implications for the ability of the surveillance system to respond to new evidence on the “value” (health benefit) of, for example, vigorous-intensity physical activity (see Day 1 discussion on emerging epidemiology that vigorous-intensity activity may have greater benefits and therefore require less time to accrue a health gain).

Results from analysis of a GPAQ data set presented to inform the group showed the following:

- On the assumption that the reported total duration of physical activity minutes would not change due to the different question structure (i.e., combined MVPA), and assigning 4 MET to all MVPA minutes, results showed this would, on average, reduce the prevalence estimates of meeting the guideline by around <1 percentage point.
- This could be adjusted to minimize the median absolute error further, although the assumption that reported durations would remain the same probably introduces greater error.

**Discussion**

- The discussion sought clarifications on methods and results that followed the presentation.
- Another potential suggested way to change GPAQ to address issues raised was the use of categorical duration responses.
- There was a suggestion to learn from the Eurobarometer use of this approach to assess the influence of this change.
Presentation 5: GPAQ and measuring muscle strengthening (and balance) activities (Dr Tessa Strain)

A brief overview was given on the absence of consensus on a self-report question to assess muscle strength and balance activities. The presentation cited a review of muscle strength and balance surveillance that indicated such surveillance is not widespread (19). The presentation highlighted two separate areas where differences arise:

1. Differences in guideline definitions of what counts as muscle strengthening activity.
2. Differences in how that is operationalized in surveillance – e.g. is muscle strengthening activity mutually exclusive from aerobic activity? Does it include everyday activities like carrying shopping?

The presentation provided examples of current measurement methods from the USA and Scotland, as well as examples of single-item questions that had been developed for specific uses in the United Kingdom of Great Britain and Northern Ireland but which had ultimately been unsatisfactory.

Discussion

- Use of grip strength as a measure was raised but it was noted that this is a physical (fitness) attribute not a behaviour. Also, it is not agreed that this is the best measure of whole-body strength (no consensus).
- It was noted that NHANES used standing grip strength as a measure to better reflect whole-body strength.
- There is a need to be careful about how to communicate guideline recommendations in ways that remain true to the evidence. For example, is there sufficient evidence of muscle strengthening in carrying shopping?
- Conclusion: there are more issues to consider in updating or changing GPAQ than can be addressed in the short window of time available for GPAQ changes. However, this should stay on the agenda for future.
GPAQ and sleep measurement

- It was suggested that consensus would be helpful on whether countries should ask a question on sleep – such a question is sometimes voluntarily asked (often in adolescent surveys).

- Devices would be best placed to give sleep duration and efficiency measures. Correlations of self-report and device-measured sleep are around ~0.3.

- It was noted that currently WHO has no global guidelines or recommendations on sleep in adults, and a process and time to create one is needed. It was noted that sleep is gaining wide interest across a large number of technical teams at WHO headquarters – so this is broadly on the horizon (subject to prioritization and funding).

Need for a complementary self-report instrument for use with a wearable device

These notes are a follow on to the day 1 discussion on this topic.

- The meeting was reminded that (at least) some socio-demographic data are needed alongside data from the wearable device, and so some form of survey is needed. Also, discussions on day 1 highlighted the potential need and opportunity for a self-report instrument (GPAQ) to collect relevant data on domain and type but not necessarily to attempt to quantify physical activity given that the device does this “better” than a survey.

- Therefore, it is foreseeable that a revised (“minimal”) questionnaire is needed to complement wearable devices to support having some domain analysis (and policy relevance) and it requires development – i.e. a “wearable questionnaire”.

- The group was reminded that day 1 had noted that there was a need for a GPAQ instrument for use by countries for the foreseeable future – at least 5 years. Therefore, in light of the emerging epidemiology and other gaps in GPAQ outputs and issues with GPAQ, there was a need to consider any potential changes by July 2024.

- The group also noted the need for planning and guidance for countries on switching or transitioning from self-report to wearable device measurement of physical activity, and the importance of an “overlap” period – this will also require some research to assist countries.
Day 2 afternoon – work group discussion forums

For 40 minutes the meeting participants divided into two small workshops to address two issues in greater depth:

1. Should any potential changes be made to the current GPAQ – and if so, what?
2. What is the way forward on using wearable devices in-country?

Group 1 – feedback and conclusions on GPAQ

- It is not feasible to develop and test any changes to the current GPAQ within the time period (by July 2024).
- Time, energy and resources would be better spent on developing the questionnaire(s) needed for use as a “complement” to wearable-device assessed physical activity rather than changes to the current GPAQ.
- A group should reconvene to discuss what indicators wearable devices will deliver and support a meeting to design and test a “complementary” questionnaire.

Group 2 – feedback on next steps to progress use of device-based measurement

The group chose to review and design a WHO-STEPS pilot study. Key features of the design attempted to address the logistical issues raised in the meeting, thus:

- The pilot study should be designed with a focus on testing feasibility and logistics (and as such could use any existing wearable device).
- The study design could include two intervention arms to test two different protocols for collection/return of devices – namely postal and 3rd party pick up.
- Other design features were to address the burden of data download and device set up for next use.
- A concern about the protocol remaining at 7 days of wear time with the hope of 1 day of wear time being possible was deemed not possible as per discussion on day 1 of the meeting.
- Experience of “pilot” work (in Malawi) by the WHO-STEPS team was outlined, as well as the purchase of 800 activity devices; it was also noted that work is currently underway in Georgia and Moldova, offering a learning opportunity. Delegates requested more information and learning on these activities to better inform next discussions.
Summary conclusions

- Wearable devices are feasible for population surveillance of physical activity, and this has been demonstrated in a set of 10 (inter)national surveillance systems to date.

- Adoption of wearables as a data collection approach by countries is feasible in the mid-term but not the short-term future. Issues of standard protocols and guidance as well as affordability need to be addressed.

- No changes should be made to the current GPAQ.

- Work should commence in 2024 to:
  - outline a minimum set of specifications for wearable devices;
  - engage industry and manufacturers in this agenda;
  - establish a set of required and desirable outcomes from wearable devices;
  - identify/develop (open source) algorithms for the outcomes;
  - develop and test a questionnaire suitable to complement the use of a wearable device within the WHO-STEPS surveillance protocol (where there is an existing health survey that may need additions or adaptation in relation to physical activity);
  - develop guidance on the data needed from a self-report instrument when there is no other health survey (which might be the context for some countries and situations where physical activity is being assessed outside or in addition to a national surveillance system);
  - undertake work plans with the idea of not limiting the possibilities of wearables and measurement of physical activity to only one option but rather to provide a choice and set of options for countries with different resource levels and needs;
  - keep sleep, light intensity, sedentary and muscle strengthening (and balance) on the agenda for further discussion within both self-report and wearable device agendas.
# Declaration of interests

<table>
<thead>
<tr>
<th>Name</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aidan Doherty</td>
<td>Research unit has received support from Novo Nordisk, GSK, SwissRe to conduct studies and purchase equipment. No conflict of interest identified.</td>
</tr>
<tr>
<td>Ulf Ekelund</td>
<td>No interests declared.</td>
</tr>
<tr>
<td>Maria Hagströmer</td>
<td>No interests declared.</td>
</tr>
<tr>
<td>Robert Harle</td>
<td>No interests declared.</td>
</tr>
<tr>
<td>Annemarie Koster</td>
<td>No interests declared.</td>
</tr>
<tr>
<td>I-Min Lee</td>
<td>No interests declared.</td>
</tr>
<tr>
<td>Rick Troiano</td>
<td>No interests declared.</td>
</tr>
<tr>
<td>Manos Stamatakis</td>
<td>No interests declared.</td>
</tr>
<tr>
<td>Hidde van der Ploeg</td>
<td>Research funding received from Statistics Netherlands. No conflict of interest identified.</td>
</tr>
</tbody>
</table>
References


Annex: List of participants

Meeting delegates (only those attending listed)

Professor Aiden Doherty, University of Oxford
Professor Ulf Ekelund, Norwegian School of Sport Sciences
Professor Maria Hagströmer, Karolinska Institute
Dr Robert Harle FITBIT, Google, University of Cambridge
Dr Annemarie Koster, Maastricht University
Professor I-Min Lee, Harvard Medical School
Dr Rick Troiano, External Advisor, DHSS, USA
Professor Manos Stamatakis, University of Sydney
Dr Hidde Van der Ploeg, Amsterdam University Medical Center

World Health Organization

Dr Fiona Bull, Physical Activity Unit, Department of Health Promotion
Dr Regina Guthold, Department of Maternal, Newborn, Child and Adolescent Health and Ageing
Dr Ruediger Krech, Department of Health Promotion
Dr Alain Labrique, Department of Digital Health and Innovation
Ms Leanne Riley, Surveillance, Measurement and Reporting Unit, Department of Noncommunicable Diseases
Dr Gretchen Stevens, (consultant) Surveillance, Measurement and Reporting Unit, Department of Noncommunicable Diseases
Dr Tessa Strain, (consultant) Physical Activity Unit, Department of Health Promotion
Ms Shelley Wallace, Physical Activity Unit, Department of Health Promotion
Dr Juana Willumsen, Physical Activity Unit, Department of Health Promotion
For further information, contact:
Department of Health Promotion
World Health Organization
20 avenue Appia
1211 Geneva 27
Switzerland
Website: https://www.who.int/