Supplementary information

S.1 Data collection

Proximity sensors

Proximity data are measured with the `SocioPatterns` sensors that we here succinctly introduce, addressing the interested reader to [1] for a more detailed reference.

Their functioning is based on the emission of low-power and low-frequency signals. Participants are asked to wear the sensor on their chest, so that when they engage in a face-to-face interaction with another participant, the respective sensors can exchange packets of information with a frequency that does not exceed one packet per second. A contact is measured if, in the time-span of 20 seconds, two sensors exchange at least one packet, recording the unique identifier of the interacting sensor, the time at which the interaction occurred and the attenuation of the signal from the sender to the receiver. This attenuation is directly connected to the distance between the two sensors and can be used to filter only close proximity interactions.

Additionally, each sensor periodically records some status properties that allow the user to know whether it is functioning correctly or not. Among these, an accelerometer allows one to know if the sensor is moving (hence if it is worn) or not.

Collection procedure

The PHIRST study was a prospective household cohort study described previously [2, 3] In short, we enrolled a new cohort of households in 2016, 2017 and 2018 at two sites in South Africa (urban: Klerksdorp, North West and rural: Agincourt, Mpumalanga) and followed households up for 8 to 10 months. Consenting household members were visited twice weekly to collect nasopharyngeal specimens and self-reported symptom data to investigate the transmission of influenza respiratory syncytial virus (RSV) and Streptococcus pneumoniae. In the 2018 cohort, we deployed wearable proximity sensors for 10 – 14 days to all consenting household members to measure high-resolution household contact patterns during three periods of the year. Sensors were worn in PVC pouches on the chest or on a lanyard. Participants were requested to put the sensor on in the morning, keep it on the entire day (even when leaving the home), take it off at night and store it separately from other household member's sensors. Not all participants felt comfortable wearing sensors outside of the home and instead took sensors off when not at home. Participants were requested to complete a log to indicate the times the sensor was put on and taken off during the day. During the twice weekly visits to the household, study staff reminded participants to wear the sensors, monitored if all sensors were still working, and replaced batteries where sensors had stopped working. After at least a ten-day deployment, sensors were collected at the next routine household visit of study staff to the household and taken back to the study office where batteries were removed.

Data cleaning

The data cleaning procedure is described as follows:

1. All contacts measured by non-moving sensors are removed: this is to avoid including spurious contacts between sensors that are, for instance, kept in proximity inside a drawer.

2. Contacts are filtered according to their attenuation and only those with an attenuation of $-70\ dBm$ or less are kept. This threshold corresponds to an interaction between two sensors that are approximately at 2 meters, even if this is a context-dependent relation that depends on external parameters, such as, for instance, humidity.

3. All contacts happening before the official beginning of the deployment and after its end are removed. These contacts may exist, because sensors may be collected on different dates from the ones of the planned experiment, but they are removed because sensors' use may be non-systematic, hence unreliable. Moreover, the first and last day of measurement are removed as well. During these days, very intense activity patterns are typically observed due to the interaction with the people dispatching the sensors. Since this kind of interaction deviates from the standard conditions, it is not considered.

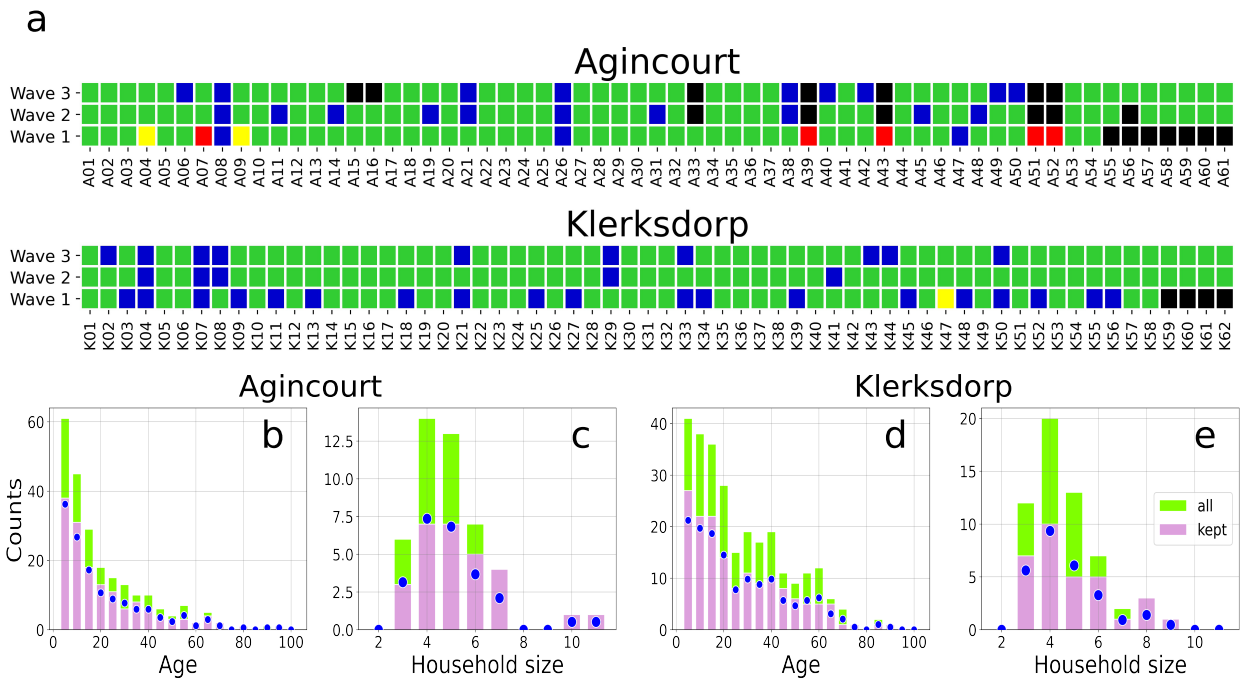

Figure S.1: **Raw data characteristics**. **a**: *data quality*. For each *(household-deployment)* we assign a color code: black indicates that the household did not participate; red that all household's sensors had data quality issues and did not provide valid measurements; blue that there are less than two days of measurement; yellow that a non circadian activity is observed; green none of the above. **b** and **d**: *age distribution in Agincourt and Klerksdorp, respectively*. The green bars are referred to the whole data-set, while the purple one only to the 60 households with valid measurements in all three deployments (see *a*). Blue dots are obtained multiplying the height of the green bars for the fraction of the included households. **c** and **e**: *household size distribution*. Legends and colors follow the description of *b* and *d*.

4. The data collected by the sensors contains information on the hardware identification code. A mapping relates this identifier with the individuals' anonymous identification code that allows us to relate contacts and meta-data. Errors at this stage make it impossible to relate contacts to people and results in the red dots shown in Figure S.1a.

5. As a minimal request, we impose that, after this cleaning procedure, a deployment can be considered valid only if it has two or more days of measurement. We found this to be a good trade-off between high quality data to work with and a still rather comprehensive inclusion principle. Household-deployment pairs that do not fulfill this condition are denoted in blue in Figure S.1a.

6. Finally, non-circadian activity patterns are identified. A high activity during the night was observed in only three cases (yellow dots of Figure S.1a) during the first deployment and are likely symptomatic of a misuse of the sensors happened at the beginning of the experiment.

Only the households in which all three deployments led to valid measurements (all green dots in Figure S.1a) were included in our study. Figures S.1b, c, d, e further show the age and household size histograms for the whole dataset against its cleaned version, showing that our inclusion principle did not affect either of the four distributions.

References

[1] Ciro Cattuto, Wouter Van den Broeck, Alain Barrat, Vittoria Colizza, Jean-François Pinton, and Alessandro Vespignani. Dynamics of person-to-person interactions from distributed rfid sensor networks. *PloS one*, 5(7):e11596, 2010.

[2] Cheryl Cohen, Meredith L McMorrow, Neil A Martinson, Kathleen Kahn, Florette K Treurnicht, Jocelyn Moyes, Thulisa Mkhencele, Orienka Hellferscee, Limakatso Lebina, Matebejane Moroe, et al. Cohort profile: A prospective household cohort study of influenza, respiratory syncytial virus and other respiratory pathogens community burden and transmission dynamics in south africa, 2016–2018. *Influenza and Other Respiratory Viruses*, 15(6):789–803, 2021.

[3] Cheryl Cohen, Jackie Kleynhans, Jocelyn Moyes, Meredith L McMorrow, Florette K Treurnicht, Orienka Hellferscee, Azwifarwi Mathunjwa, Anne von Gottberg, Nicole Wolter, Neil A Martinson, et al. Asymptomatic transmission and high community burden of seasonal influenza in an urban and a rural community in south africa, 2017–18 (phirst): a population cohort study. *The Lancet Global Health*, 9(6):e863–e874, 2021.