Using Proximity Sensors To Characterize Social Contact Patterns Relevant To The Transmission Of Respiratory Viruses In Schools And Households.

Thesis

How to cite:

Kiti, Moses Chapa (2019). Using Proximity Sensors To Characterize Social Contact Patterns Relevant To The Transmission Of Respiratory Viruses In Schools And Households. PhD thesis The Open University.

For guidance on citations see FAQs.

© 2019 The Author

https://creativecommons.org/licenses/by-nc-nd/4.0/

Version: Version of Record

Link(s) to article on publisher’s website:
http://dx.doi.org/doi:10.21954/ou.ro.00010ab6

Copyright and Moral Rights for the articles on this site are retained by the individual authors and/or other copyright owners. For more information on Open Research Online’s data policy on reuse of materials please consult the policies page.

oro.open.ac.uk
Using proximity sensors to characterize social contact patterns relevant to the transmission of respiratory viruses in schools and households.

Thesis presented for award of degree of Doctor of Philosophy

by

The Open University (UK)
Life and Bio-Molecular Sciences

Affiliated Research Centre
Kenya Medical Research Institute – Wellcome Trust Research Programme,
Kilifi, KENYA

Moses Chapa Kiti
BSc Geomatic Engineering and GIS (JKUAT, Kenya)
MSc Demography and Health (LSHTM, UK)

Submission date January 2019
Summary

This thesis reports the use of wearable sensors to obtain proximity data from students in schools and household residents in Kenya. Data on who-contacts-whom quantifies behaviour that underpins pathogen transmission usable in mathematical simulations to evaluate interventions such as vaccination.

A subset of students from two schools were recruited from rural and urban areas, including residents linked to some of the participating students. Data were collected on different dates for each site, but collection was hindered by low sensor numbers, administrative bottlenecks, student absenteeism, and resident out-migrations. This resulted in 4 study sites enrolling 595 (303 school) and 233 (136 school) participants in rural and urban locations, respectively. Each participant carried a proximity sensor over 7 days.

On average, each student was in contact with 38% (rural) and 31% (urban) other participating students. Contact number and duration were assortative by grade in the rural school, with considerable inter-grade mixing in the urban school. Weekends registered low interactions due to school closure, but mixing was reported by older students who attended school on Saturday.

In the households, children (5-14 years) in rural and adults (20-49 years) in urban sites interacted most with children under-5 years. The duration of interaction was higher within- versus between-households, suggesting that infections dependent on cumulative exposure time may have higher transmissibility within households. Observed high between-household network density suggests that these interactions may lead to substantial inter-household transmission of endemic infections.

Finally, a framework is proposed for an agent-based model that can be used in future to investigate the implications of observed intra- and inter-household interaction patterns on the transmission dynamics of respiratory syncytial virus (RSV). In the absence of a vaccine against RSV, the impact of alternative strategies to prevent infection in infants, such as vaccinating older children, can be safely investigated using computer simulations.
Dedication

This thesis is dedicated to all the teachers who have let me stand on their shoulders for the last 35 years. It does not stop now.
Acknowledgement

This starts from where my journey into being a PhD student begun: meeting Prof Kevin Marsh one idyllic afternoon in 2008 through two of my dear friends and mentors, Dr Martha Mwangome and Dr Benjamin Tsofa. I am forever indebted to you three for enabling me to “ng’alwa na matso” (open my eyes). Cheers to my two peers and great childhood friends, Dr James “Korona” Muye (deceased) and Dr Simon “Simo” Chengo, with whom we have shared over many years, drinks and endless conversation about moving the cheese. Korona, you are an inspiration, and you will forever be in my heart and thoughts.

That Dr Ciro Cattuto responded to my cold email one afternoon in 2010 has been one of the key highlights of my research work. Prof James Nokes gave me a carte blanche to engage with you, and it has led to fantastic work that we have done together since then. Through you, Ciro, I have met Drs Michele and Laura, and the warm team at ISI Foundation in Turin. James, from the time I was an intern in 2009, you have been confident enough to let me run with research ideas under your supervision. It has been a wonderful learning experience, and probably the three greatest lessons in life I have learnt from you, James, are: 1) always have a potential solution to problems, real or imaginary; 2) always dot the i’s and cross the t’s accurately, and 3) a researcher does not necessarily have to live a dull life (in relation to your family life, sailing trips and participation in Kilifi Gold Triathlons). Many thanks to the Wellcome Trust (UK) for providing my funding through James, Alessia for funding my trip to a conference in South Africa and introducing me to the wonderful team at FBK-Trento (grazie mille Piero and Stefano Merler), Ciro for giving me office space in Turin, and to KEMRI-WTRP (Kilifi) for hosting me for almost 10 years now.

I have to say a hearty thank you to my colleagues at KEMRI-WTRP, in particular the research team that I have worked with since 2009. I have had wonderful support from the Community Liaison Group. Special thanks to Hassan Alphan and Johnson Masha for helping make headway in Kilifi Township and Matsangoni schools and households. Many thanks to all students, staff and residents of both sites for patiently listening to me/us and agreeing to provide crucial data for our public health efforts. Shadrack, John and Grace you have been very instrumental in all my research projects. I am extremely
happy to see that you all have also progressed in your careers, as I have done. I also worked tirelessly with Monica, Micah, Fredrick and Chome in this project. Your dedication inspired me to lead the team well, and this ambitious project has been a tremendous success. I am forever indebted to you guys. Thank you very much Edward and Grieven for your assistance in data management, and not forgetting Mark for help with KHDSS data. Many thanks to my office mates – Irene, Joyce, James, Mike and John – for persevering through my “EUREKA” moments. To the rest of the VEC team (Patrick, George, Charles I, Charles II, Nelson, Tim and all lab and field staff), thank you for your camaraderie. Lastly, to my Rotary family in Kilifi: thank you for keeping me grounded, and for the opportunity to Serve above self.

I have had chance to travel to different places. Life would have been dull had I stayed in hotels. I have had wonderful hosts in 4 years: Piero and Stefano of FBK in Trento, Emanuele and Alessia of Bocconi University in Milano, Michele and Laura of ISI Foundation in Torino, Elena in Milan, and Chilalu and Korona (RIP) in Tarragona. By welcoming me to your lives and homes, you made me appreciate the importance of the phrase “casa mia è casa tua”. To my friends in Kenya who took time to meet and talk during my rare visits, thank you for your support and understanding. In particular Alex and Christine, and Simon and Evelyn. I truly cherish our friendship.

My family has been my greatest support pillar. Thank you, mom and dad, for ensuring that I get the best education in my formative years. Mom you once said “…wewe fuata kusoma, wacha sisi tutafute fees…”, which loosely translates to “…let us worry about the fees, make sure you excel in school…”). I hope I make you proud. Franco (and Halima) and Cza (and Joan), you have always been there for me through thick and thin and I am forever grateful. Thanks Chizi (and Nur) for your weekly calls, I know it is not only in return of my monthly visits to you when you were in high school :). Lastly to Anna Gonna aka Hawe Kadzo, my maternal grandmother, thank you for showering me with your blessings and always bringing good cheer to our family table.

And to GM, you know how we do!

Finally, thanks be to God. Thank you for listening to my many short and silent prayers.

Indeed, all things are possible through you.
Table of Contents

Summary .............................................................................................................................................. ii
Dedication ........................................................................................................................................... iii
Acknowledgement ............................................................................................................................ iv
List of tables ....................................................................................................................................... x
List of figures ...................................................................................................................................... xi
Abbreviations ...................................................................................................................................... xvii

1 Introduction ..................................................................................................................................... 1

1.1 Defining the question ......................................................................................................................... 1

1.1.1 Why is it important to understand the transmission of infections? ............................................ 1

1.1.2 Methods to measure social contact patterns ............................................................................. 2

1.1.3 The role of schools and households in transmission of respiratory infections .................... 5

1.2 Justification ....................................................................................................................................... 7

1.3 Research objectives ............................................................................................................................ 8

1.3.1 Specific objectives ...................................................................................................................... 8

1.4 Methods .......................................................................................................................................... 8

1.5 Declaration of author’s role .............................................................................................................. 9

1.6 Output ............................................................................................................................................ 10

1.7 Overview of the thesis ..................................................................................................................... 10

2 Literature review ............................................................................................................................... 11

2.1 Introduction ..................................................................................................................................... 11

2.2 Social mixing patterns in epidemiological models .......................................................................... 12

2.2.1 Social contact studies using self-report methods ................................................................... 12

2.2.2 Social contact studies using wireless proximity sensors ...................................................... 17

2.3 Social contact studies in Kilifi, Kenya .............................................................................................. 24

2.3.1 Collecting contact data using paper diaries ........................................................................... 24

2.3.2 Collecting social contact data using wireless proximity sensors ........................................... 25

2.3.3 Comparison between survey and sensor-based methods in Kenya ....................................... 26

2.4 How comparable are survey- and sensor-based methods? ............................................................. 27

2.4.1 Definition of a contact .............................................................................................................. 27

2.4.2 Acceptability of method .......................................................................................................... 28

2.4.3 Sampling strategies .................................................................................................................. 28
4.2 Baseline characteristics of participants from Matsangoni primary school ........................................ 68
  4.2.1 An overview of characteristics of the rural school ................................................................. 68
  4.2.2 Characteristics of the participating students ............................................................................. 69
4.3 Summary of the of number and duration of contact events ....................................................... 71
  4.3.1 Hourly count and mean contact events per day ...................................................................... 71
  4.3.2 Distribution of degree, number and duration of contact events ............................................ 74
4.4 Contact matrices of number and duration of contacts by grade .............................................. 78
4.5 Graphs of the contact network .................................................................................................... 79
4.6 Assessing the similarity of neighbourhood of the nodes ............................................................ 84
4.7 Discussion ...................................................................................................................................... 86
  4.7.1 Study design ............................................................................................................................. 86
  4.7.2 Network properties ................................................................................................................... 88
  4.7.3 Limitations ............................................................................................................................... 90
  4.7.4 Summary ................................................................................................................................. 91
5 Results from rural households ......................................................................................................... 92
  5.1 Introduction ................................................................................................................................... 92
  5.2 Baseline characteristics of participants from rural households .................................................. 93
  5.3 The distribution of number and duration of contact events .................................................... 96
    5.3.1 Hourly distribution of total number and mean contact events per day .............................. 96
    5.3.2 Distribution of degree, number and duration of contact events ........................................ 99
  5.4 Contact matrix of number and duration of contact events ....................................................... 102
  5.5 Contact activity distribution within and between households ................................................. 103
  5.6 Properties of the contact network in rural households ............................................................ 109
    5.7 Assessing the day-to-day stability of contacts ........................................................................ 115
  5.8 Discussion .................................................................................................................................... 116
    5.8.1 Recruitment of study participants ......................................................................................... 117
    5.8.2 Network properties ............................................................................................................... 118
    5.8.3 Limitations ........................................................................................................................... 119
6 Comparison of results from schools and households in rural and urban settings ....................... 121
  6.1 Introduction .................................................................................................................................... 121
    6.1.1 Data collection in the urban school ....................................................................................... 123
6.1.2 Demographic characteristics in the urban compared to rural households .................125
6.1.3 Comparing rural and urban contact event timelines in schools and households ..........126

6.2 Comparison of network properties between rural and urban schools and households ...........127
6.3 Comparison to random networks ..................................................................................131
6.4 Summary and discussion .................................................................................................132
  6.4.1 Recruitment in urban school and households ..............................................................132
  6.4.2 Participation rates and data abstraction ........................................................................134
  6.4.3 Contact patterns, number and duration of interactions ..............................................135

7 Discussion ............................................................................................................................137
  7.1 Purpose of the study .........................................................................................................137
  7.2 Summary of key findings .................................................................................................140
    7.2.1 Participation rates in schools and households ............................................................140
    7.2.2 Activity timelines, distribution of degree, number and duration of contacts .............141
  7.3 Importance of the findings ..............................................................................................143
    7.3.1 Mixing patterns at schools and households ...............................................................143
  7.4 Summary of limitations of the study ...............................................................................143
  7.5 Recommendations for further research .........................................................................146
    7.5.1 Participant recruitment, retention and improved data collection ...............................146
    7.5.2 Design of larger cluster-based studies .......................................................................146
    7.5.3 Epidemiological implications: formulation of a mathematical model against transmission of respiratory viruses .................................................................147

8 Conclusion ..........................................................................................................................149

References ...............................................................................................................................151
Appendices ...............................................................................................................................167
List of tables

Table 2.1. Characteristics of social contact patterns conducted in Sub-Saharan Africa. ................................................................. 14

Table 2.2. Summary of global studies that collected social contact data using wireless sensors. This shows the study country, proximity detection radius, setting, number of participants and days of study. ................................................................. 23

Table 2.3. Summary of key characteristics of paper diary and wireless proximity sensor studies. ................................................................. 29

Table 3.1. Definition of key terms used in this study. ................................................................. 45

Table 4.1. Baseline characteristics of participating students from the rural primary school. This table shows the grades, number of students by gender and the median age per grade. .................................................................................................................. 70

Table 4.2. Summary of statistical properties of the network in rural primary school. Number of nodes and edges, median degree, network density, clustering coefficient and mean contact duration in minutes by location, grade and day. ............... 78

Table 5.1. Baseline characteristics of rural households. This table shows the overall median degree and stratified by gender, age, day of the week, household size and location. .................................................................................................................. 96

Table 5.2. Summary of within and between household median degree by gender, age and household size. .................................................................................................................. 105

Table 5.3. Summary of properties of network within and between households. .. 109

Table 6.1. Participation rates in rural and urban schools and households in Kilifi. Participation rate is the proportion of individuals who carried sensors compared to the total number of sensors expected to be issued (350), while response rate is the proportion of sensors whose data were analysed, compared to the number of sensors actually issued. SR and SU are schools in rural and urban areas respectively. HR and HU are households in rural and urban areas, respectively. ........................................ 125

Table 6.2. Comparison of network properties in rural and urban schools and households................................................................................................. 128

Table 6.3. Properties of random networks in schools and households.................................. 132
List of figures

Figure 2.1. Social contact mixing patterns by age in Italy, Kenya and Zimbabwe. Panel (A) shows the overall mixing patterns by age. Panel (B) shows mixing patterns only in the home of the participant in Zimbabwe, rural and urban Kenya, respectively. These figures have been adapted from a presentation by Melegaro et al. (Population Association of America Annual Conference, 26-29 April 2018, Denver, USA)...........15

Figure 2.2. Different versions of wireless proximity sensors that have been used. In panel (A), v1 shows an RFID sensor used in the pilot study in rural Kilifi households [44], and v2 shows the wireless proximity sensor used in the current study. Panels (B) and (C) show how the sensors were worn. .................................................................21

Figure 3.1. Map of the study locations showing the rural and urban area. Panel (A) shows Kenya country map (grey) and Kilifi County shaded dark grey. Panel (B) shows the Kilifi County (grey) and the and the Kilifi Health and Demographic Surveillance Site in dark grey. The rural (orange) and urban (purple) sites are shown in panel (C), with the location of the school and health centre in each location shown. .........................41

Figure 3.2. Population pyramids of Matsangoni and Kilifi Township locations from the KHDSS data, round 26 of 2016. The proportion of males (blue) and females (red) is shown for 5-yearly age groups from age 0 to 85 years and above. ...............................42

Figure 3.3. Conceptualization of an interaction framework in schools and households. The school shows students (blue nodes) grouped into classes. The index student is marked X. Teachers are shown as red nodes. Interactions occur within and between classes. Household 1 hosts the index student. Household 2 represents a group of neighbouring households enclosed in a compound (dotted square boundary). Interactions can occur within or between households, between household residents and school-goers, and within and between students and teachers in various classes. ............43

Figure 3.4. Wireless proximity sensor used in this study. Panel (A) shows the sensor (left; diameter 30-mm) next to a 20-shilling coin. Panel (B) shows an adult wearing the sensor in a pouch hanging at the chest. Panel (C) shows a student wearing the sensor pinned to his school shirt. .................................................................46

Figure 3.5. Data collection schematic showing the sites and corresponding data collection date. SR and HR represent school and households in rural area, while SU and HU represents schools and households in urban areas.........................................................52

Figure 3.7. Activity timeline showing number of contact events across study days for raw rural household contact data. Panel (A) shows data between the start and end of collection, 18/10/2016 00:00:00 and 28/10/2016 23:59:59, respectively. The solid vertical lines are at midnight of each day. Panel (B) shows data between 22nd 00:00:00 and 24th 23:59:59 as highlighted in orange. The number of contacts were aggregated at 10-minute intervals.................................................................54

Figure 3.8. Activity timeline of Household 44. Row 1 shows charts of activity timelines for sensors (A – brown trace) and (B – green) representing data from residents of household 44. Chart (C) shows an overlay of activity timelines for both individuals. Spikes peaking at value 15 depict when a contact events were detected, and other values above 15 depict a sensor error (reboot). The grey boxes highlight 06:00:00 to 20:59:59
hours of each day, assumed to be when an individual was wearing a sensor. Each trace covers the same period from Oct 21st 00:00:00 to 27th 23:59:59.

Figure 3.9. Activity timeline for household 12. There were 6 residents in household 12 whose activities are shown by (A)-(F). Graph (G) is an overlay of the individual timelines for all 6 residents. The description of the graph details follows that of Figure 3.7.

Figure 3.10. Rescaled activity timelines for household 44, method 1. Panels (A) (brown) and panel (B) (green) represent data from two individuals in household 44. The black line represents smoothed data. The description of the graph details follows that of Figure 3.7.

Figure 3.11. Rescaled activity timelines for household 12, method 1. Panels (A)-(F) represent data from each of six individuals in household 12. The black line represents smoothed data.

Figure 3.12. Rescaled activity timelines for household 44, method 2. Panel (A) and (B) represent data from each of two individuals in household 12. The black line represents smoothed data.

Figure 3.13. Rescaled activity timelines for household 12, method 2. Panels (A) to (F) represent data from each of two individuals in household 12. The black line represents smoothed data.

Figure 3.14. Comparison of rescaled activity timelines for household 44. A and B represent data from two individuals in household 44. The black line represents smoothed data using method 1, while the red dotted line shows smoothed data using method 2.

Figure 3.15. Comparison of rescaled activity timelines for household 12. Panels (A) - (F) represent data from six individuals in household 12. The black line represents smoothed data using method 1, while the red dotted line shows smoothed data using method 2.

Figure 3.16. Schema demonstrating the definition of contact event and contact duration. One contact is a 20-second proximity event that occurs between individual i and j without a break. Here, there are 5 individuals. Each red horizontal line represents the start (left-hand side edge) and end (right-hand side edge) of a proximity event between i and j lasting s-secs. Individual 1 has zero contacts since the duration of contact is <20 seconds. Individual 2 has 1 contact event lasting exactly 20 secs. Individual 3 has three contact events cumulatively lasting 60-seconds, with a 20-sec interruption between the contact events. Individuals 4 and 5 have three and six contact events lasting 60- and 120-seconds, respectively.

Figure 4.1. Data collection flowchart in the rural primary school. This shows the total number of students in the school, number recruited, reasons for loss to follow up, and total number of sensors from students that were available for analysis.

Figure 4.2. Activity timeline of students from the rural primary school over 7 days. Total hourly number of contacts events are shown from midnight (0) to 11:59 pm (23). Each colored line represents a different day of the week. Dotted lines show weekend days. Faded lines depict discarded data. Friday (21st) and Thursday (27th) were the first and last days, respectively, of data collection.
Figure 4.3. Daily per capita activity timeline for students in the rural primary school. Average hourly number of contact events normalized by the number of unique sensors in proximity in each hour. Only data between 05:00 am and 06:00 pm are shown for each of the 7 days. Each colour represents a different day of the week, with dotted lined representing weekends. ................................................................. 73

Figure 4.4. Daily activity timeline of number of nodes and mean contact events in the rural primary school. The primary axis (red) shows the mean number of contact events per hour. The secondary axis (black) shows the number of nodes in proximity per hour. In the graph, time is highlighted by digits 5 and 18 representing 05:00 am and 06:00 pm, respectively, and a vertical grey line showing midday of each day. The grey vertical bars represent night time (19:00:00 to 04:59:59 hrs) of each day when data were discarded. There was a drop in the number of nodes on Monday and Tuesday as some students were sent home to collect school fees. ................................................................. 74

Figure 4.5. Statistics of the contact network showing the distribution of degree and number of contact events measured in the rural primary school. Panel (A) shows the probability degree distribution $P(K)$ of the contact network aggregated over 7 days. The black dotted line indicates the median degree, $<K> = 85.0$ ($CV^2 = 0.09$). Panel (B) shows the probability degree distribution for each day of the study. Panels (C) and (D) show the log-log probability distribution of the overall weight $w_{ij}$ and the daily probability distribution of the weights of the contact network. The weight $w_{ij}$ of an edge $i \rightarrow j$ represents the total number of contact events between $i$ and $j$. ................................................................. 75

Figure 4.6. Log-log probability density distribution ($P(dt)$) of contact durations (dt) measured over the experimental period (7 consecutive days) in the rural primary school. Panel (A) shows the overall distribution of contact event durations in minutes, highlighting interactions lasting 5 minutes, 30 minutes, 1 hour and 5 hours. The other panels show the probability distribution of contact duration by grade (B), gender (C) and day of the week (D). ................................................................. 77

Figure 4.7. Contact matrix showing the number and duration of contact events. This shows the total number (A) and duration (B) of contact events, respectively, that students of grade $i$ (column index) had with students of grade $j$ (row index) over 7 days. Panel (C) and (D) show the daily mean number and duration of contact events, respectively. Labels on the x and y axes report the grade of the individuals. Durations are reported in minutes. ................................................................. 79

Figure 4.8. Full network of aggregated contacts from the rural primary school. Each node represents a student (N=223). Each unweighted link represents the presence of at least one contact event between connected students aggregated over 7 days. Nodes are colour coded to represent the 9 grades (KG – grade 8). The graph is generated by the DrL force-directed graph generator provided in igraph (1.2.1) package available in R (3.4.3). The DrL algorithm emphasizes the presence of dense clusters as noticed in the interactions between students belonging to the same grade.5 ................................................................. 81

Figure 4.9. Daily networks of students from the rural primary school. Graphs were generated for each day of the week from Wednesday to Tuesday of the following week. The description of the nodes associated colours is similar to that of Figure 4.8. Here, the links represent at least one contact event aggregated per day. The DrL emphasizes the presence of dense graphs observed by grade but masks the same-grade links due to the overlapping of the nodes. ........................................................................................................... 82
Figure 4.10. Daily contact activity patterns by grade and selected days of the week. Only data for Friday-Sunday are shown in rows (A)-(C). The first column show the daily contact event timelines by grade, and the second column shows matrices of the contact event duration. Kindergarten (KG) and grade 8 activity timelines are highlighted, representing the variation in contact events in the youngest and oldest grades in the school.

Figure 4.11. The distribution of cosine similarity in the rural school. Panel (A) shows the normal cosine similarity, while panel (B) shows the cosine similarity in which the weights of the network are reshuffled among the edges (WR). The cosine similarity was calculated for each node of the full contact network (overall) and for each grade (KG, 1-8). Each distribution measured the cosine similarity of each node’s neighbourhood, for each pair of days of data collection. For each boxplot, the horizontal bar in the middle shows the median, the grey box shows the lower and upper bounds of the interquartile range (IQR), and the whiskers show 1.5 x IQR of the distributions.

Figure 5.1. Location of index-student households’ and school in rural setting. The index students were randomly selected from the list of participating students reported in Chapter 4. The household clusters are labelled 1-9, and the inset table shows the number of households (HH) and residents (Res) in each cluster. The lines show the main roads passing through the study area. The map also shows the location of the rural school (SR) and the main public health centre in Matsangoni location.

Figure 5.2. Data collection flow chart for rural households. This presents the number of residents located, sensors issued, reasons for loss to follow-up/discarding of data, and number of sensor-data analyzed. Attribute data include age, gender and household identifier.

Figure 5.3. Activity timeline of rural households over 7 days. Total hourly number of contacts are shown from midnight (0) to 11:59 pm (23). Each colored line represents a different day of the week. Dotted lines show weekend days. Faded lines depict discarded data. Friday (21st) and Thursday (27th) were the first and last days, respectively, of data collection.

Figure 5.4. Per capita daily activity timeline for Matsangoni households. Only data between 5 am and 6 pm are shown for each of the 7 days. Each colour represents a different day of the week. Contact events have been normalized by the number of participants per hour per day.

Figure 5.5. Daily activity timeline of mean contact events and number of nodes in the rural households. The primary axis (red) shows the mean number of contact events per hour. The secondary axis (black) shows the number of nodes in proximity per hour. In the graph, time is highlighted by digits 5 and 18 representing 05:00 am and 06:00 pm, respectively, and a vertical grey line showing midday of each day. The grey vertical bars represent night time (19:00:00 to 04:59:59 hrs) of each day when data were discarded.

Figure 5.6. Distribution of network degree and number of contact events in rural households. (A) Probability degree distribution P(K) of the contact network aggregated over 7 days. The black dotted line indicates the median degree, < K > = 11.0 (CV² = 0.4). (B) Probability degree distribution for each day of the study. (C) Log-log probability distribution of the weights of the aggregated contact network. (D) Log-log daily probability distribution of the weights of the aggregated contact network. The
weight $wij$ of an edge $i - j$ represents the total number of contact events between $i$ and $j$.

Figure 5.7. Log-log probability density distribution ($P(dt)$) of contact durations ($dt$) measured over the experimental period (7 consecutive days). Panel (A) shows the full distribution of contact durations, highlighting interactions lasting 5 minutes, 30 minutes, 1 hour and 3 hours. The other panels show distribution of contact duration by age in years (B), gender (C) and day of the week (D).

Figure 5.8. Contact matrix showing the number and duration of contact events. This shows the total number (A) and duration (B) of contact events, respectively, that residents of age group $i$ (column index) had with residents of age group $j$ (row index) over 7 days. Panel (C) and (D) show the daily mean number and duration of contact events, respectively. The number of residents per age group in panels (C) and (D) is given in brackets. Labels on the x and y axes report the age group of the individuals. Durations are reported in minutes.

Figure 5.9. Intra- and inter-household contact activity patterns in rural households. Daily contact events occurring only within (intra) members of the same household and exclusively between (inter) members of different households from 5 am to 7 pm. The grey background shows overall daily contacts. The grey vertical lines show 00:00:00 hours of each day.

Figure 5.10. The distribution of network degree within and between households. Panel (A) shows degree distribution within and between households, whereas (B), (C) and (D) shows distribution by gender, age group and household size, respectively, for within (brown) and between (grey) household.

Figure 5.11. Distribution of within and between household network degree, number and duration of contact events. Panel (A) shows the cumulative degree distribution of within and between household contacts, while (B) shows the probability density distribution ($P(dt)$) of contact durations ($dt$) measured over the experimental period (7 consecutive days) within and between households. Panel (C) and (D) show the daily average number of contacts that individuals of age $i$ (column index) had with individuals of age $j$ (row-index) within and between households, respectively. Panel (E) and (F) shows the daily average duration of contacts of an individual of age $i$ with individuals of age $j$. The labels of the x and y axes represent the age groups of the individuals, and the number in brackets is the number of people per age group. Duration of contacts are reported in seconds.

Figure 5.12. Within and between household mean contact matrices stratified by gender. Panels (A) and (B) show matrices within household for females (F) and males (M), respectively. Similarly, panel (C) and (D) show matrices for between household contacts for females and males, respectively, between households. Durations are shown in minutes.

Figure 5.13. Full networks of aggregated contacts from rural households over 7 days. Each node represents an individual. Each link represents an unweighted interaction between individuals who have at least one contact. Nodes in panels (A) are colour coded to represent the 9 household clusters linked to the index student, and the colours are similar to that on the map in Figure 5.1. Nodes in (B) are colour coded to represent age groups.
Figure 5.14. Intra-household networks of aggregated contacts from rural households over 7 days. The networks represent contacts that occur between individuals living in the same household. Nodes in panels (A) are colour coded to represent the 9 household clusters linked to the index student. Nodes in (B) are colour coded to represent age groups. .......................................................... 113

Figure 5.15. Inter-household networks of aggregated contacts from rural households over 7 days. The networks represent contacts that occur between individuals living in the different household. Nodes in panels (A) are colour coded to represent the 9 household clusters linked to the index student. Nodes in (B) are colour coded to represent age groups. .......................................................... 114

Figure 5.16. Distribution of cosine similarity in rural household clusters. Panel (A) shows the cosine similarity of the original network, while panel (B) shows the cosine similarity in which the weights of the network are reshuffled among the edges. The cosine similarity was calculated for each node of the full contact network (overall) and for each household cluster. Each distribution measured the cosine similarity of each node’s neighbourhood, for each pair of days of data collection. For each boxplot, the horizontal bar in the middle shows the median, the grey box shows the lower and upper bounds of the interquartile range (IQR), and the whiskers show 1.5 x IQR of the distributions. .......................................................... 115

Figure 6.1. Rural and urban locations of the KHDSS. Rural areas are shown in orange while urban areas are shown in purple, with study sites highlighted in darker shades. ........................................................................................................... 122

Figure 6.2. Comparing activity timelines in rural and urban schools and households. Figures in all panels represent hourly count of contact events divided by the number of people in contact per hour. Each coloured line represents a different day of the week, with dotted lines representing weekends. Only data between 05:00 am and 07:00 pm are reported. ........................................................................................................... 126

Figure 6.3. Comparison of distribution of degree, number and duration of contacts in rural and urban schools and households. Panel (A) shows boxplots of the distribution of degree. Panel (B) and (C) show the log-log distribution of number of contact events and contact duration, respectively. SR=Rural school, SU=Urban school, HR=Rural households, HU=Urban households. For each boxplot, the horizontal bar in the middle shows the median, the grey box shows the lower and upper bounds of the interquartile range (IQR), and the whiskers show 1.5 x IQR of the distributions...... 129

Figure 6.4. Household degree distribution by site. Panel (A) shows the degree distribution in rural and urban households by age of participants. Panel (B) shows the degree distribution stratified by household size................................................................. 130

Figure 6.5. Distribution of duration of interaction in various settings. Panels (A) and (C) represent rural school and households, respectively. Panels (B) and (D) represent urban school and households, respectively. Each colour in panels (A) and (B) depicts grade, and age-group in panels (C) and (D). A coloured link indicates interaction between the groups, and the thickness of the line represents the strength of interaction................................................................. 131
## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARI</td>
<td>Acute respiratory illness</td>
</tr>
<tr>
<td>CAST</td>
<td>Community Advice for Study Teams</td>
</tr>
<tr>
<td>CLG</td>
<td>Community Liaison Group</td>
</tr>
<tr>
<td>dB</td>
<td>Decibels</td>
</tr>
<tr>
<td>HiB</td>
<td><em>Haemophilus influenzae</em> type b</td>
</tr>
<tr>
<td>HR</td>
<td>Rural households</td>
</tr>
<tr>
<td>HU</td>
<td>Urban households</td>
</tr>
<tr>
<td>KEMRI-WTRP</td>
<td>Kenya Medical Research Institute-Wellcome Trust Research Programme</td>
</tr>
<tr>
<td>KG</td>
<td>Kindergarten</td>
</tr>
<tr>
<td>KHDSS</td>
<td>Kilifi Health and Demographic Surveillance System</td>
</tr>
<tr>
<td>LMIC</td>
<td>Low- and Middle-Income Countries</td>
</tr>
<tr>
<td>LRTI</td>
<td>Lower respiratory tract illness</td>
</tr>
<tr>
<td>mins</td>
<td>Minutes</td>
</tr>
<tr>
<td>PERCH</td>
<td>Pneumonia Etiology Research for Child Health</td>
</tr>
<tr>
<td>RSV</td>
<td>Respiratory syncytial virus</td>
</tr>
<tr>
<td>secs</td>
<td>Seconds</td>
</tr>
<tr>
<td>SMC</td>
<td>School management committee</td>
</tr>
<tr>
<td>SSA</td>
<td>Sub-Saharan Africa</td>
</tr>
<tr>
<td>SR</td>
<td>Rural school</td>
</tr>
<tr>
<td>SU</td>
<td>Urban school</td>
</tr>
<tr>
<td>USA</td>
<td>United States of America</td>
</tr>
<tr>
<td>US$</td>
<td>American dollars</td>
</tr>
</tbody>
</table>
1 Introduction

1.1 Defining the question

1.1.1 Why is it important to understand the transmission of infections?

Respiratory infectious diseases are a major cause of disease, hospitalisation, and death, and also pose significant social and economic burden to patients, their care-givers, and governments. According to the World Health Organization (WHO), lower respiratory tract infections (LRTI) caused the biggest burden of deaths due to communicable diseases\(^1\). Across all ages, LRTIs were the fourth leading cause of mortality, causing more than 3 million deaths globally in 2016. Pneumonia is a type of LRTI which affects the lungs and can be caused by bacteria, viruses or fungi. About 60% of deaths due to pneumonia are concentrated in 10 countries, 5 of which are in Sub-Saharan Africa (SSA) and the rest in southern Asia\(^2\). Populations at risk include children aged <5 years (74% of deaths occur in infancy), especially pre-term babies, those with pre-existing health conditions such as low-birth weight, congenital heart disease and immunosuppression [1,2]. There are various global efforts to understand the burden [3], pathways of transmission and pathogenesis of pneumonia, with more reduction in deaths anticipated from prevention of transmission using vaccines compared to curative efforts [4].


The transmission process of respiratory infections is still poorly understood. Essentially, some mechanistic realisation of a “close contact process” that puts an individual at risk of getting infected is required [5]. Transmission of some of the common airborne infections may occur if individuals are close enough to have a two-way conversation without shouting and without the presence of a physical barrier between them [6]. A variety of contact types relevant for transmission of respiratory pathogens exist, such as direct contact with large droplets of nasal secretions or fomites through touch, or indirect contact through inhalation of contaminated aerosols expelled by coughing and sneezing from infected hosts [7]. These contact types can be interactions within common spaces such as in households, schools or work places where individuals spend most of their time [8,9], and due to enhanced connectivity through transport modes that promote connectivity outside the ‘common spaces’ such as travel by road, air, sea or rail [10]. Knowing “who-contacts-whom” is a key principle in predictive modelling that is increasingly being used to pre-determine fundamental aspects of interventions such as who and when to vaccinate, or how many doses are required for community protection [11–14]. Broadly, two types of interventions against transmission of pathogens (or prevention of disease) exist: pharmaceutical such as vaccination to pre-immunize individuals [15], and non-pharmaceutical such as behavioural change to reduce contacts [4]. There exist vaccines against some of the pathogens transmitted via close contact such as pneumococcus, varicella zoster virus (VZV), *Haemophilus influenzae* type b virus (Hib), influenza, amongst others, but not for all e.g. respiratory syncytial virus (RSV). Strategies for intervention require an in-depth understanding of the epidemiology of each pathogen, particularly the transmission patterns in different population groups.

1.1.2 Methods to measure social contact patterns

*Using surveys and self-report questionnaires*

Over the last 15 years, several groups of investigators have attempted to quantify social mixing patterns in different regions and using different methods from the perspective of understanding respiratory infection transmission. Four of these methods: direct observations, using contact surveys (paper/web diaries, telephone interviews), proximity sensors and making inferences from secondary data (e.g. serological, demographic,
digital traces e.g. from internet use) have been discussed extensively in reviews by Read [7], Hoang (unpublished) [16] and Barrat [17]. In this thesis, the focus of discussion on methods is given to studies that used paper diaries as they have made the largest contribution to the field to date and provide a direct record of close contacts thought to be required for close contact transmission and the main comparator for the electronic proximity data that form the basis for this thesis. Hoang et al. [16] identified 64 published social contact surveys using paper diaries, out of which 95% targeted the general population and schools/universities. Only 4 studies recruited entire households, with 3 in upper-middle-income countries [18–20] and one in a lower-middle-income country [12]. Five community-based studies that recruited single participants from households have so far been conducted in SSA (South Africa [21], Zambia [22], Kenya [23], Zimbabwe [9] and Uganda [24]).

In surveys, individuals (referred to as nodes in a network) are requested to record social contacts that they make with other individuals either prospectively (record contacts as they occur) or retrospectively (report contacts the following day). Variations in these social contacts, regarded as proxies to a contagion process, can be explained by covariates such as age of participants and location of interactions, and weighted by the duration or frequency of interactions. The majority of studies used the retrospective design, with the biggest study to date having recorded contact patterns of 7,290 participants across 8 European countries (POLYMOD [25]). Mean contact rates per person per day varied across countries, but age-specific contact patterns were strikingly similar across all countries, with children having more assortative contacts than adults. Intergenerational contacts were also reported. These observations have also been reported to a higher degree, particularly within households, in various studies conducted in SSA [9,21,23], attributed to differences in demographic, socioeconomic and cultural factors. The majority of long-duration physical contacts mostly occurred at home or school, inferring that the risk of infection may vary due to location of the contact.

Finally, simulations of the initial phase of a hypothetical respiratory infection epidemic suggested that the incidence of infections is highly affected by the age-specific contact structure. In particular, school going children up to 19 years old would bear the biggest disease burden [9,25], and potentially act as introducers of infections to their younger siblings at home [26], or to adults due to intra-household mixing [24]. In general, these
studies provide data that are crucial in parameterizing transmission models, in particular, to support studies in high burden setting. Further questions that can be answered include the temporal variability of daily social contact patterns, the amount and complexity of data needed to capture representative contact patterns, and how the patterns may change due to forced social distancing as a result of illness, school or work closure [7].

Technological advances in collection of social contact data

Efforts to provide methods that improve social contact data collection have resulted in the advancement and use of automated methods of data collection. These include Bluetooth [27,28] and wearable proximity sensors [29–31] that communicate through radio frequencies (RF) in alternating cycles to transmit and receive signals. Bluetooth signals are detectable over distances ranging from 10-100 metres [32], suggesting that the coarseness of detection distance is unsuitable for diseases transmitted via close contact. A wireless proximity sensor, on the other hand, uses ultra-low-power radio signals that can be tuned to detect a similar device in its line of sight and within a clearly defined distance (e.g. 1.5 metres commonly used) [30]. This makes wearable sensors ideal for use in tightly knit populations such as schools [33–35], hospitals [31,36–38], workplace [30,39] and conferences [40] where individual behaviour may be of interest to epidemiological and social processes such as contagion and information sharing. The primary aim of the sensor platform has been to collect data from co-located individuals standing a few metres apart and facing each other, with the proximity detected suggestive of conversations or possible physical touch that can lead to a contagion process [30,35]. In addition, data collected include timestamps that can be used to automatically compute the frequency and duration of interactions, void of recall bias, to capture the heterogeneity of interactions and that can be used to model the transmission of infections [41–43].

The majority of sensor-based studies have reported a high participation rate (≥75%), suggesting that an unobtrusive way of data collection requiring minimal participant intervention may elicit better response rates compared to paper diaries [23]. This can be useful especially in settings with a high proportion of individuals unable to self-report their contacts [9,44]. Unlike surveys that have been able to recruit representative
samples within a country [45] or across several countries [25], saturation of study populations by sensors in order to define full networks rapidly encounters boundaries due to logistic, time and study cost constraints. The initial cost of purchase can be considered high in resource-poor settings, but the devices can be reused such as in global deployments by the SocioPatterns project\(^3\) thus reducing long-term costs. Unfortunately, sensors can only record proximity between similar co-located devices. Nonetheless, this property makes them relevant for use for a few individuals grouped in closed settings such as households [44,46], to hundreds of individuals in a highly dynamic settings such as museums and conferences [30,47] or health emergency simulations [48]. However, other non-trivial interactions with individuals not wearing sensors will not be recorded.

Sensors provide a rich source of dynamic temporal data, which even for partial networks can be used to investigate plausible characteristics of infection transmission on networks weighted by frequency and duration of interaction. Data collected from sensors are standardized and contain the sensor identifier (ID), measure of proximity and timestamps, thus making datasets easily comparable across various deployments [17]. In addition, statistical methods, supplemented by additional demographic assumptions, are available to infer the structure of the network at a population level by generating simulated interactions based on the gathered network properties [49]. In addition to understanding dynamic social interactions in various contexts, high resolution spatial and temporal data from sensors can be linked to microbiological data to elucidate pathways of transmission of infections through close contact [37]. Ultimately, the data can be used to inform mathematical models of control [50].

1.1.3 The role of schools and households in transmission of respiratory infections

Schools and households are locations where a high proportion of a population will spend most of their time and individual interactions are more frequent and intense (close proximity or physical) [9,12,21,25]. Generally, a household refers to a nuclear family living in the same residence (high-income countries) and may include extended families

\(^3\) https://www.sociopatterns.org
related by blood and marriage in low-income settings particularly in the rural areas (for example [51]). Primary and secondary schools host children, on average, with ages ranging from 4-14 to 15-18 years, respectively. The high intensity interactions at these settings may raise the likelihood of transmission of a range of respiratory infections [52–54]. For instance, children in particular are notable introducers and transmitters of respiratory infections to other household members [26,50,55], and to members of other households [24]. A few studies have demonstrated that infants spend more time with their mothers compared to other household members [46,56,57] intuitively suggesting that mothers may have a key role to play in the transmission of infection. Other studies also revealed that interactions between infants and non-household members were non-negligible but rarely captured [56], suggesting that only targeting same-household members to cocoon infants from infections may have a limited impact on transmission [56,57]. These highlight the importance of using empirical contact data in the design of intervention measures that confer both direct and indirect protection to the infant [21,58–60]. Measures such as school closure have been shown to be effective in reducing the magnitude of outbreaks by infections that spread via close contacts (e.g. influenza) [53,61], largely dependent on the transmissibility of the virus and the type of school closure (e.g. one class vs entire school [50]). As a side effect, school closure results in more heterogeneous interactions particularly with adults at home and with students from other schools [62]. This increases the potential of transmitting infections to students from other schools and working ages, thus suggesting that additional interventions such as vaccinating students from neighbouring schools and adults may need to be considered.

These interactions are still poorly understood especially among infants and within school-goers, more so in low-income countries. Contextual studies that may capture contact events in schools and households are thus suggested to provide empirical data that is needed in mathematical models that simulate transmission and assess the impact of various control measures. Currently, only two studies conducted in Kenya and Italy have recruited entire households to provide social contact data using wireless proximity sensors. In Kenya, participants were drawn from a rural coastal population in Kilifi. Household occupancy included more than two generations of related families resulting in large households (range 6 to 40 residents) and 54% of participants aged <14 years.
The Italian study recruited households of patients attending a paediatric hospital in Rome [46]. The household size ranged from 2 to 6 members, with a third of the participants aged <5 years old. Whereas children <5 years spent the majority of their time with older school-going siblings in Kilifi, children aged <5 years spent most of their time with their parents (mother) in Rome. Despite the small number of participants (N<100) in both studies, they illustrated differences in mixing patterns by age and due to family composition, suggestive of the different cultural roles played by siblings and parents in different settings. These also highlight the potential of different ages in the introduction and onward spread of infections [24,26], with direct implications in prevention of transmission such as through vaccination of mothers or older school-going siblings.

1.2 Justification

This thesis centres on gathering data on human social contacts in two key environments, the school and the household, using wireless proximity sensors. Schools and households are hubs of infection transmission making these locations potential targets in the prevention of transmission and control of infectious disease outbreaks. Individuals spend the majority of their time at home, in schools, and at work. Schools and households are sites of close and longer lasting interactions between individuals particularly in younger ages who are prone to infections. Primary schools in Kenya, attended by children aged between 5-14 years on average, bring together a large number of children at high densities who may be immunologically naïve to some respiratory infections. When infected, school-goers may transmit the infections to younger siblings and older individuals at home.

The focus of this thesis will be on the contact mechanisms that facilitate transmission of respiratory infections. In particular, data will be collected to compare the characteristic properties of close/proximal contacts in selected schools and households located in rural and urban settings of Kilifi. These data on social contact patterns will increase the understanding of properties of contact networks in low-and-middle income countries where very few data exist, and the burden of respiratory infections is still high. In addition, the data will be used to support modelling studies that will explore the
transmission of common respiratory viruses, such as RSV, and assess the impact of various control strategies such as vaccinations targeting specific groups by age or residence. An indirect benefit will be an assessment of the feasibility of using sensors in various settings in urban and rural areas in Kenya, generating recommendations that can be used in other low- and middle-income countries (LMIC) in future.

1.3 Research objectives

The general objective is to utilise radio frequency close proximity sensors to describe and understand the nature of human networks that have potential in the transmission of respiratory infectious diseases within a low resource population.

1.3.1 Specific objectives

(i) To collect data on close proximity interactions in schools and households in one rural and one urban area of coastal Kenya.

(ii) To characterize the number, duration, and temporal dynamics of social contact network properties, structures, and underlying determinants within the household and school settings.

1.4 Methods

This study was conducted in two locations within the Kilifi Health and Demographic Surveillance Site (KHDSS) in coastal Kenya. The study design was cross-sectional targeting two primary schools, one each from a rural and urban setting, and households linked to a subset of students attending the schools. From each selected school, 350 students were randomly selected and requested to participate. Nine index students from each school were then randomly selected from the list of participants. All the household residents of each index student were recruited into the study. Additionally, 3-5 neighbouring households of each index household were recruited. Proximity data were collected using wireless proximity sensors (WPS) to detect dyadic interactions between individuals. A contact occurred when sensors from two individuals separated by ≤1.5
metres registered an interaction, suggestive of a conversation or skin-to-skin touch such as a handshake. The sensors were worn on the chest area either using a lanyard hung around the neck or attached to the upper garment by a safety pin. Data were first collected at each school (not concurrently), then the related households also on different dates. Participants at schools and households wore the sensors for seven consecutive days during their normal activities, taking them off at night before sleeping. Data from each sensor were downloaded for analysis after each deployment.

Patterns of contact were quantified by the number of unique individuals contacted (degree), number of contacts (frequency of 20-second interval interactions), and the total time spent in contact between two individuals (in minutes). The heterogeneity of the contacts was assessed by stratifying by age, gender, setting (rural/urban) household size, and day of the week. A mathematical model in development will simulate the implications of various scenarios of contact structure (including those defined in this study) to investigate respiratory infection dynamics and control as future work arising from the thesis.

1.5 Declaration of author’s role

The studies described in this thesis were designed by the author and approved by his supervisors Prof James Nokes, Prof Alessia Melegaro, and Dr Ciro Cattuto. The author was responsible for the data collection, data analysis and write-up of the thesis. Six field staff assisted in the data collection procedures; field staff and field work were all recruited, trained and supervised by the author. Guided by the author, the database for storage of data collected during this project was created by Mr Edward Mundia and Mr Grieven Otieno (Epidemiology and Demography Department, KEMRI-WTRP). The author conducted the primary data analysis and was assisted by Dr Michele Tizzoni and Dr Laura Ozella (both of ISI Foundation, Turin, Italy) particularly in data interpretation. The author wrote the final thesis.
1.6 Output


(iii) Guest seminar series. 2017. ISI Foundation, Turin, Italy.


1.7 Overview of the thesis

This thesis is divided into 9 chapters. This first chapter gives an overview of the thesis. Chapter 2 presents the literature review, in which I review previous studies on social contact networks and their importance in understanding transmission (and control) of respiratory infections using mathematical models. Chapter 3 describes the study design and data collection procedures. Chapters 4 and 5 present the results from the rural school and household data, respectively. Chapter 6 presents a comparison of school and household data from rural and urban locations. Chapter 7 presents an overview of the formulation of a mathematical model to investigate transmission patterns of a respiratory virus and highlighting the impact of inter- and intra-household contact patterns. Finally, chapters 8 and 9 present a comprehensive discussion of the study and highlights the key points, while suggesting directions for future research in this field.
2 Literature review

2.1 Introduction

Human behaviour responsible for infection transmission, represented as social contacts or networks, is still poorly understood. This is despite the numerous studies undertaken to understand contact patterns across various geographic regions with different underlying demographic, economic, social and cultural characteristics. In principle, human social behaviour influences who we meet, where, when, the duration and form (e.g. conversation, physical) of these interactions, all of which influence infection transmission. This has direct implications on infections that are transmitted predominantly through non-sexual person-to-person interactions. The interface between social contact/network analysis and epidemiology focuses on understanding how interactions between individuals and groups of individuals may influence the transmission of respiratory or other communicable diseases through direct or indirect contact [63], or through the air via droplets or aerosols [64,65]. Transmission via direct contact involves transfer of pathogens through physical touch or skin-to-skin contact (such as a handshake, hug, or kiss), while indirect transmission involves coming into contact with pathogens deposited on the surface of objects (fomites).

To study and understand close social contact patterns relevant to the transmission of respiratory infections, several methods are in use. There are surveys (based on paper diaries, telephone interviews, focus group discussions) or automated methods (the use of wireless sensors embedded in portable devices) that will be discussed in the next two sub-sections. The basic tenet is to answer a customizable set of questions: who-contacts-
whom (name, age, gender), how many times, form, context (i.e. where the contacts occur), and duration (how long the contacts last). Extensions to these basic questions include whether other people were present (group contacts), health status of respondent, etc. The primary outcome from analysis of these data is the rate of contact that can be stratified mainly by age, or other covariates that affect it such as gender, day of week, and location of contacts.

2.2 Social mixing patterns in epidemiological models

2.2.1 Social contact studies using self-report methods

Conventionally, social contact patterns are self-reported by respondents who log individuals with whom they “contact” over a certain duration, or in other words, “who-contacts-whom”. Other metadata such as age, gender, relationship, number, duration and location of encounters, are reported to give possible explanations to the variations in observations. Data have been collected using various study designs. The respondents were trained to fill in their own questionnaire manually [13,66], or had a “shadow” fill in the questionnaire on their behalf especially for the very young or illiterate [23,67], or had an interviewer fill in the responses for the respondent [20]. Alternative approaches to estimate contact rates included models utilizing serological data [68], focus group discussions [69], and making inferences from time-use surveys [8] or demographic data [70–73]. Initial studies were exploratory and used small convenient samples (N<100) mainly selected from university settings involving students, staff and their families [6,66,74–76]. These studies were crucial in the development of paper-diary based data collection methods, such as evaluation of retrospective versus prospective study designs with reference to recall bias, or use of paper-diary versus web-based reporting to assess acceptability and ease of use [66]. These initial studies were considered highly successful, but largely non-generalisable to the rest of the population because they were conducted in institutions of higher learning where general understanding of study procedures was assumed to be high. The main recommendations were to conduct larger community-based studies in different settings and to involve all age groups, particularly children who experience the biggest burden of disease.
Studies have now been conducted across 24 countries in 5 continents: Europe [6,13,75–84,18,85–87,25,41,45,56,62,66,74], North America [20,88–91], Oceania [57,92,93], Asia [12,19,83,94–97], and Africa [9,21–23]. This was in response to the need to contextualize who-contacts-whom data for infectious disease modelling. In one of the pioneer studies conducted almost two decades ago on quantifying who-mixes-with-whom [6], a contact was defined as a conversation between two people facing each other (type I, non-physical), or an interaction involving skin-to-skin touch such as a kiss or hug (type II, physical). In both cases, co-location within a 2-metre distance was a must, without a physical barrier between the two individuals. Following this, all studies have adopted the same or some variation of this initial description of an at-risk contact. While non-physical contacts may enable the transmission of aerosolized pathogens such as influenza, direct physical contacts are relevant for infections such as RSV that spread via large droplets of infected saliva or mucous.

The largest survey to date, commonly referred to as POLYMOD [11], was conducted across 8 European countries and has been the reference point of most other paper-diary studies conducted. One participant per household was recruited through random digit dialling (Belgium, Italy, Luxemburg), face-to-face interviews (Great Britain, Denmark, Poland), or through demographic registers (Finland, The Netherlands). Participants collected data on age (or estimated age) and gender of contacts recorded only once per all encounter, as well as location, total duration and frequency of usual contacts with the contactee. Country level data were collected broadly across all ages, gender and geographic regions to be representative of the populations in each specific country. The country-specific results reported means ranging from 7.95 contacts per person per day (contact rates henceforth given in units of per person per day) in Germany to 19.77 in Italy, with the latter having a different family composition and anecdotally said to be more gregarious compared to other countries.

Table 2.1 presents the key characteristics of paper diaries conducted in SSA from 2010 to 2018. The table shows that the various studies had different contact definitions (Type I, Type II, or Type I and II). Studies number 4 and 5 were part of a project to assess the impact of demographic changes on infectious disease transmission (DECIDE⁴) in

⁴ http://www.dondena.unibocconi.it/wps/wcm/connect/Cdr/Centro_Dondena/Home/Research/DECIDE/
Zimbabwe and Kenya, respectively. Studies 4 and 5 were interviewer-led, with the former using a paper-based questionnaire and the latter a computer-based diary to record individual close-contact interactions over two continuous days. Study 5 was completed in 2016 and analysis is ongoing.

Four out of the 6 studies compared contact patterns between rural and urban areas with varying sample sizes. The studies reported averages as either mean or medians. In summary, different geographic regions exhibit variable contact patterns, and within country variations are also observed comparing rural and urban areas. In Zambia/South Africa (SA) and Kenya, the rural areas reported higher mean/median contact rates per person per day. Significantly higher rates were observed in urban Zimbabwe.

Table 2.1. Characteristics of social contact patterns conducted in Sub-Saharan Africa.

<table>
<thead>
<tr>
<th>#</th>
<th>Country</th>
<th>Year of data collection</th>
<th>Contact definition</th>
<th>Setting</th>
<th>N</th>
<th>Average number of contacts</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>South Africa (SA)</td>
<td>2010</td>
<td>Type I &amp; II</td>
<td>Rural</td>
<td>571</td>
<td>20.0 (13.0-29.0)²</td>
<td>[21]</td>
</tr>
<tr>
<td>2</td>
<td>Zambia &amp; SA</td>
<td>2011</td>
<td>Type I</td>
<td>Rural</td>
<td>296</td>
<td>5.6 (5.3-6.0)¹</td>
<td>[22]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Urban</td>
<td>3323</td>
<td>4.8 (4.5-5.2)³</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Kenya</td>
<td>2011-12</td>
<td>Type II</td>
<td>Rural</td>
<td>371</td>
<td>18.8 (17.5-20.1)¹</td>
<td>[23]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Urban</td>
<td>197</td>
<td>15.6 (16.6-18.9)¹</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Zimbabwe</td>
<td>2014</td>
<td>Type I &amp; II</td>
<td>Rural</td>
<td>1245</td>
<td>9.0 (6.0-14.0)²</td>
<td>[9]</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Urban</td>
<td>1714</td>
<td>10.0 (6.0-14.0)²</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Kenya</td>
<td>2015</td>
<td>Type I &amp; II</td>
<td>Rural</td>
<td>566</td>
<td>7.0 (0-25)³</td>
<td>Unpublished</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Urban</td>
<td>1714</td>
<td>7.3 (7.1-7.5)³</td>
<td></td>
</tr>
</tbody>
</table>

N = number of participants whose data were analyzed. Type I and II contacts are defined in the text. Average: ¹ = Mean (95% confidence interval), ² = Median (interquartile range, IQR), ³ = Median (range). The author of this thesis was the Principle Investigator (PI) of study #5 that was part of the DECIDE project conducted in rural Kilifi and urban areas (including slums) of Mombasa. Data analysis is complete, and a manuscript is in preparation for publication.

The images in Figure 2.1 show contact heat maps between individuals of different ages in Italy (from [25]), Kenya and Zimbabwe [9]. Panel (A) shows overall contact patterns for the three countries, while panel (B) shows mixing patterns for contact occurring at home only. In SSA, there was high assortativeness in contacts among school-age children (5-19 years), and the assortativity in contacts extended across a wider age range in Italy particularly in children and adults of working ages. Zimbabwe exhibits more interaction between young children and the elderly due to the presence of an
extended family structure. A high level of intergenerational mixing was observed at home (panel (B)) in rural Zimbabwe and Kenya suggestive of parents-children and grandparents-grandchildren, while urban Kenya was characterized by contact patterns within nuclear patterns. Parents-children contacts were also observed in Italy.

Figure 2.1. Social contact mixing patterns by age in Italy, Kenya and Zimbabwe. Panel (A) shows the overall mixing patterns by age. Panel (B) shows mixing patterns only in the home of the participant in Zimbabwe, rural and urban Kenya, respectively. These figures have been adapted from a presentation by Melegaro et al. (Population Association of America Annual Conference, 26-29\textsuperscript{th} April 2018, Denver, USA).

Differences in socio-demographic characteristics such as time-use [8,9], household composition [53], work size [98]) and spatial or community structure [99] have an influence on contact patterns, which may result in variability in transmission patterns and size of epidemics across regions. The contact matrices shown in Figure 2.1 have been used in models simulating the initial phase of an epidemic to compare the age-specific proportion of new infections between countries in SSA (Zimbabwe and Kenya) and Europe (Italy). Results from the simulations suggest that more than a third (35\%) of infections are expected to occur in individuals aged <15 years in SSA contexts. Around half this number (15\%) would occur in similar ages in the European context coupled
with a bigger burden of disease in individuals older than 35 years (50%) compared to SSA (30%). Infections in the older generation in SSA were attributed to the strong intergenerational contact patterns between the young and the elderly and attributed to the larger population of elderly in Europe. Back-and-forth interactions between places such as home, school, work and community may also influence the waves of transmission [100]. Events that change social behaviour such as being ill [101] and school closure/holiday [62,79,102] have been shown to reduce mixing rates by up to half and change child-child mixing to child-adult mixing, leading to changes in the age distribution of infections.

Utility and challenges of using paper diaries for data collection

Paper diaries have been used to generate data targeting small communities such as schools and households, entire countries, or spanning several countries to assess regional similarities and differences e.g. the Europe-wide POLYMOD [25], nationwide survey in the UK [45], and comparing rural and urban areas in SSA [9,23]. Understanding social contact structures in different regions and contexts is important for two reasons: firstly, to quantify the mixing patterns and understand the factors that influence differences within and between regions. These include demographic (age, household size and composition), socio-economic (rural vs urban areas) and cultural factors. Secondly, mixing patterns can be used to develop context-specific strategies for the prevention of a range of infectious diseases transmitted via close contact [21,58–60,103]. An example is the paper diary in rural Kenya [23] whose results have been used to investigate vaccination strategies against RSV [58,60], and assess or predict the impact of different strategies following the introduction of the pneumococcal conjugate vaccine in Kenya [59,103]. More recently is the use of contact pattern data from Uganda in a mathematical model to estimate the epidemic size and basic reproduction number for a hypothetical respiratory infection [24]. Results from the Ugandan study suggested that household structure and spatial mobility had a key role to play in understanding contextual contact patterns and the design of intervention measures against transmission of respiratory infections.

Several limitations regarding surveys have been discussed with some studies suggesting potential ways to make improvements on the study design processes. One of the key
challenges highlighted particularly is recall bias that may have resulted in underreporting of contacts [12,104,105]. Specifically, in retrospective studies, individuals generally recall longer-lasting encounters [106], or those that happened more recently (salience), especially those of importance to them, such as meeting a “best-friend” compared to a stranger. This generally results in an under-estimation of the number of contacts [106]. On the other hand, prospective diary keepers might be willing to keep the diary but may forget or deliberately choose not to fill in certain details leading to an underestimate of data being collected. Some surveys have also reported relatively low participation rates compared to data collected using sensors [107]. It is impractical to expect the very young to report their contacts [45], thus studies have resorted to using “shadows” or third-parties to record contacts on behalf of hard-to-reach ages [23,108]. A few studies have cited poor compliance [109,110], where participants did not fill in the diaries as requested or made some erroneous entries. Following suggestions from a rural Kenyan community, age-groups were represented as pictograms to minimize wrong reporting age categories. Perhaps one major shortcoming in resource-poor settings was short-term seasonal migration that made participant tracing difficult [23,111], suggesting that studies in some agricultural communities should be conducted during various seasons to assess if temporary migration will have an impact on contact patterns.

2.2.2 Social contact studies using wireless proximity sensors

Proximity sensors that use radio frequency (RF) signals embedded in portable devices have emerged as alternatives to self-reported data on social contacts [17]. The typical examples are Bluetooth enabled mobile phones and miniature Radio Frequency Identification (RFID) devices, both relying on a transmitter-receiver system of the RF signals to detect and record a similar device in proximity. With Bluetooth, the normal range of proximity detection of 10 - 100m radius [112] is arguably unsuitable for spread of infections requiring close or physical contact [32,112,113]. Nevertheless, Bluetooth signals can be important for investigating transmission of airborne infections such as tuberculosis and measles that do not require direct physical contact [114], or to make inferences on friends and friendship networks [27]. However, the mobile phones used in
epidemiological research such as the FluPhone\textsuperscript{5} project encountered some challenges in large-scale deployment. The application created to log user data were only compatible with specific phones that cost >100 GBP, which was considered expensive in 2010. Experimental studies could only be conducted on small groups or crowds hence reducing generalizability of the results. Running the application full time drained the phone battery very fast resulting in potential data gaps when the phone went off, and quick filling of internal storage particularly due to the large detection radius. Lastly, in remote areas of developing countries experiencing poor telephone network coverage, individual access to smart phones that run the customized software, and lack of electricity may hinder the deployment of data capture using Bluetooth systems. The biggest advantage is that the mobile phones are always carried by the owners and hence “do what the users do”, meaning that they can objectively capture the temporal dynamics of human activity over a specified period of time particularly if the devices are programmed to securely send data periodically to an external server over the internet. Applications developed for data collection can also be configured to collect location information via Global Positioning System sensors embedded in the phone for location tracking, as well as auxiliary data such as demographic and self-reported infection data [114].

Proximity sensors based on RFID technology (henceforth called RFID sensors) have been used to objectively measure human social interactions in different contexts (\textbf{Table 2.2}) and over one to several days of data collection in real time. The data collection infrastructure consisted of active wearable RFID sensors worn by participants, RFID readers installed in the experiment space, and a central computer. The RFID sensors had an on-board battery which powered an antenna that could transmit and receive a data packet in alternating cycles. This version of RFID sensors had no internal memory hence the readers detected incoming packets, encrypted the data packets and relayed the packets to the central server for archiving. These sensors had been pre-programmed to detect and record the presence of other sensors in its vicinity, a relation known as a “contact”. Further technical details of this distributed sensing platform are provided by Cattuto et al. [30]. These RFID sensors have been useful to understand proximity events that may lead to transmission of respiratory infections in settings such as work places,

\textsuperscript{5} \url{https://www.cl.cam.ac.uk/research/srg/netos/projects/archive/fluphone2/}
schools and conferences. For the RFID readers and central computer devices to communicate with the RFID sensors, the experiment space needed access to electricity to power the readers and computer. This was disadvantageous in settings where there is poor access to the power grid. In the first experiment in SSA [44], the RFID sensors were modified to have bigger rechargeable batteries, a water-proof plastic casing, and an internal flash memory to store data (Figure 2.2 (A)). This enabled the deployment of the sensors over several days in rural households, as well as assess the feasibility and acceptability of using the devices in remote settings.

Currently, the SocioPatterns projects uses more advanced wearable proximity sensors in experiments. Henceforth, the term wearable proximity sensors, or simply sensors, will be used to distinguish them from the RFID sensors. The main difference between them is the presence of an antenna (RFID, see Figure 1 in Cattuto et al. [30]) and a memory on board the wireless proximity sensors (henceforth referred to as sensors). The addition of an onboard memory ensures that data can be automatically collected and stored internally, making it easier to deploy the sensors in remote areas that have no electricity to support the reader and server infrastructure. The sensors operate in the 2.4 GHz ISM band of the RF spectrum. The total weight of the sensor inclusive of a lithium coin battery (CR2032) is <6 grams.

For the purpose of sensor-based studies, a contact event between two sensors is defined as the exchange of at least one data packet at the lowest power level between the two sensors in a 20-second time-window. A contact is considered broken if a 20-second window passes without the exchange of a data packet [30]. In this manner, the sensors enable the collection of face-to-face dyadic (between two or more sensors) time-varying proximity data with tuneable spatial and temporal granularity. This definition of a contact may give more leverage in determining the threshold of separation distance between individuals depending on pathogen-specific transmission requirements. Since timestamps for each contact event are recorded, it is possible to calculate the duration of contact events and the interval and between contacts.

Finally, sensors can be deployed simultaneously in large groups, for instance ranging from 25 to 575 individuals, as described by Cattuto et al. [30], and in entire schools [34]. Due to logistical challenges experienced during deployment, data collection can
ideally occur within a localised space such as school or hospital. However, household studies are now being explored [44,46]. Sensors are used to automatically and objectively collect unsupervised data on who-contacts-whom, how many times, and for how long, with the added advantage of potential seamless comparisons of network properties between different deployments. The proximity detected is suggestive of a close interaction which is relevant to the transmission of pathogens either by direct physical contact or indirectly (by aerosols) from coughing or sneezing [40]. It should be emphasized that the sensors to not detect physical touch between individuals rather their proximity. Individuals wear a sensor either attached to a lanyard around the neck so that it rests on the chest area, or it is pinned to the front of the blouse/ shirt as shown in Figure 2.2 (C), the latter method especially for younger children (in school). In this manner, only face-to-face proximity relations are detected.
Figure 2.2. Different versions of wireless proximity sensors that have been used. In panel (A), v1 shows an RFID sensor used in the pilot study in rural Kilifi households [44], and v2 shows the wireless proximity sensor used in the current study. Panels (B) and (C) show how the sensors were worn.

To date, wearable sensors have been used to collect data in households [44,46], schools [33–35,115,116], hospitals [31,36–38,117,118], among conference attendees [40,47,86], and in a museum [30]. These studies have been conducted mainly in Europe (N=13) and the USA (N=4), with only one published study from Africa [44] but with confirmed reports of similar longitudinal studies a in Kenyan school and South African community
(personal communication). All the studies in Europe and Africa used the same RFID sensing platform developed by the SocioPatterns project, while three out of four studies in the USA used “motes” that had different physical designs. However, the sensor characteristics in use were similar, with longer proximity detections ranging up to four metres. This suggest that the properties of the networks can be assessed using similar algorithms, and this is made increasingly available with open data available from the SocioPatterns website. Despite the different social contexts and number of people involved, a common property observed in all these studies is the heterogeneity of contact events and durations that are distributed in a manner that obeys a power-law. A power-law distribution means that the majority of contacts are brief or casual encounters, but long contacts are also observed [119]. This is suggestive of super-connectors, or individuals who have a high number of interactions Students have more contacts with their classmates than with those of other classes, and contacts are fewer among teachers as they spend most of their time in class [34].
Table 2.2. Summary of global studies that collected social contact data using wireless sensors. This shows the study country, proximity detection radius, setting, number of participants and days of study.

<table>
<thead>
<tr>
<th>Country</th>
<th>Year of publication</th>
<th>Proximity radius</th>
<th>Setting</th>
<th>N</th>
<th>Days</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>USA</td>
<td>2010</td>
<td>≤3 m</td>
<td>Primary school</td>
<td>788</td>
<td>1</td>
<td>[35]</td>
</tr>
<tr>
<td>Italy</td>
<td>2010</td>
<td>1 - 5 m</td>
<td>Conference &amp; museum</td>
<td>25-575</td>
<td>2-12</td>
<td>[30]</td>
</tr>
<tr>
<td>Italy</td>
<td>2011</td>
<td>≤1.5 m</td>
<td>Conference</td>
<td>100-14,000</td>
<td>3-90</td>
<td>[47]</td>
</tr>
<tr>
<td>Italy</td>
<td>2011</td>
<td>≤1.5 m</td>
<td>Hospital</td>
<td>119</td>
<td>7</td>
<td>[38]</td>
</tr>
<tr>
<td>France</td>
<td>2011</td>
<td>≤1.5 m</td>
<td>Primary school</td>
<td>242</td>
<td>2</td>
<td>[34]</td>
</tr>
<tr>
<td>France</td>
<td>2011</td>
<td>≤1.5 m</td>
<td>Conference</td>
<td>405</td>
<td>2</td>
<td>[47]</td>
</tr>
<tr>
<td>USA</td>
<td>2012</td>
<td>-</td>
<td>Hospital</td>
<td>-</td>
<td>7</td>
<td>[31]</td>
</tr>
<tr>
<td>USA</td>
<td>2013</td>
<td>≤1 m</td>
<td>Hospital</td>
<td>4732</td>
<td>365</td>
<td>[117]</td>
</tr>
<tr>
<td>France</td>
<td>2013</td>
<td>≤1.5 m</td>
<td>Hospital</td>
<td>75</td>
<td>4</td>
<td>[36]</td>
</tr>
<tr>
<td>France</td>
<td>2014</td>
<td>≤1.5 m</td>
<td>High school</td>
<td>298</td>
<td>11</td>
<td>[33]</td>
</tr>
<tr>
<td>USA</td>
<td>2014</td>
<td>≤2 m</td>
<td>High school</td>
<td>974</td>
<td>3</td>
<td>[33]</td>
</tr>
<tr>
<td>France</td>
<td>2015</td>
<td>≤1.5 m</td>
<td>Hospital</td>
<td>590</td>
<td>120</td>
<td>[118]</td>
</tr>
<tr>
<td>France</td>
<td>2015</td>
<td>≤1.5 m</td>
<td>Hospital</td>
<td>84</td>
<td>12</td>
<td>[37]</td>
</tr>
<tr>
<td>France</td>
<td>2015</td>
<td>≤1.5 m</td>
<td>High school</td>
<td>327</td>
<td>5</td>
<td>[116]</td>
</tr>
<tr>
<td>Germany</td>
<td>2016</td>
<td>≤1.5 m</td>
<td>Conference</td>
<td>74</td>
<td>1</td>
<td>[86]</td>
</tr>
<tr>
<td>Kenya</td>
<td>2016</td>
<td>≤1.5 m</td>
<td>Household</td>
<td>75</td>
<td>3</td>
<td>[44]</td>
</tr>
<tr>
<td>Italy</td>
<td>2018</td>
<td>≤1.5 m</td>
<td>Household</td>
<td>55</td>
<td>2 - 5</td>
<td>[46]</td>
</tr>
<tr>
<td>Italy</td>
<td>2018</td>
<td>≤1.5 m</td>
<td>Hospital</td>
<td>238</td>
<td>1</td>
<td>[48]</td>
</tr>
</tbody>
</table>

Ideally, one would want to saturate the entire population of interest with sensors to capture the full network of interactions that would be of epidemiological relevance [35]. Logistically, this proved to be very challenging in a multi-household setting. For example, in Kilifi each participant had to give individual consent, sensors had to be pre-programmed, participants needed to be trained on how to wear the sensors, and lastly the sensors had to be collected and data downloaded one sensor at a time. Data were only exchanged between two (or more) people wearing the sensors, meaning that there was only partial recognition of the complete network space. However, participation rates have been generally high (>90%) compared to paper diaries that sometimes report rates lower that 50% of the target sample size. Data has generally been collected from unique closed settings e.g. schools only or households only, essentially assuming that there are no interactions with individuals in other settings e.g. school-household-community. For instance, in modelling disease spread in schools, Salathé et al. [35] assumed an infection was introduced by one sick child, and that there was no interaction
with individuals outside the school. In reality, children continue to interact with other individuals and more so at home, thereby increasing the likelihood of getting an infection from other contexts. A similar argument can be made for contacts at hospitals, conferences and museums; it is expected that attendees will interact with other individuals outside these contexts.

Community use of wireless proximity sensors

As is common with the use of new tools in research, unforeseen challenges have been identified in the collection and statistical representation of electronically collected data. Simultaneous deployment in multiple settings, for example to include interactions in households plus schools, is mainly limited by available time and resources. It then becomes necessary to define representative populations in which data can be collected, which can be a daunting task. Even though automated data collection methods may provide a more objective way to collect data over several days compared to surveys, sensors will still capture a sub-set of the underlying network, and in some cases, result in a lot of “noise”. For instance, in the Kenyan study [44], it was noted that there were spikes in contacts recorded especially at night, suggesting that participants stored all the sensors in the same vicinity within range of the electronic signals. Arbitrary cut-off times were selected (8:00 pm - 6:00 am) but no further analysis was conducted to assess the effect on the distribution of the number and duration of contacts by dropping the night activity.

2.3 Social contact studies in Kilifi, Kenya

2.3.1 Collecting contact data using paper diaries

In Kilifi, coastal Kenya, two household surveys have been conducted between 2012 and 2015 to quantify who-contacts-whom, comparing mixing patterns between rural and urban areas, and across seasons (see Table 2.1). In the published survey listed as #3 [23], respondents of all ages were randomly selected across rural and semi-urban areas from a population based register of the KHDSS [51]. Individuals were requested to fill a text or pictorial paper diary reporting with whom they had a physical contact (e.g. touch
or embrace) in one day, and the age and gender of the contactee. Individuals were also given pre-programmed alarm wrist-watches that went off each hour to remind participants to make entries on the diary. The reported study (#3) conducted focus group discussions before the data collection and issued questionnaires after data collection to determine the views of respondents on the study methods. The second study listed (#5) collected data from a rural location within the KHDSS, and an urban suburb within Mombasa featuring both a middle-class residential area and a slum. This second study was an interviewer led survey that asked respondents to recall their contacts over two preceding days. The differences in age-specific contact patterns are displayed in Figure 2.1, with panel (B) highlighting the rural-urban differences in age-related contact patterns.

Overall daily contact rates per person per day reported in Kenya were higher than reported in South Africa [21] and Zimbabwe [9]. Contact rates were higher particularly in rural school-aged children (6-15 years old) compared to urban areas, attributable to differing larger households with bigger proportion of children. Several aspects of the Kilifi study #3 [23] made it novel: pre- and post-survey questionnaires to aid in the initial design and inquire on the ease of use of the diary, longitudinal survey to account for seasonality, the use of an alarm watch to remind participants to record contacts, the use of pictures to represent different ages rather than actual ages, the inclusion of “shadows” to record details on behalf of the illiterate (adults and <10 year olds), and comparison of rates between rural and urban areas. The biggest limitation was low participation rates (~50% participation, N=571) mainly due to lack of participant tracing, but this was comparable to other global studies.

2.3.2 Collecting social contact data using wireless proximity sensors

A feasibility study on acceptability and utility of using sensors within five households (N=75) was conducted in Kilifi over 3 days per household [44]. Of the 5 households, 3 collected data concurrently, making it possible to investigate both intra- and inter-household network patterns. The sensors deployed in the Kilifi pilot study were highly customised for use in the rural setting (see Figure 2.2 (B)). Each sensor was encased in a water-proof plastic case to prevent deliberate tampering by participants and exposure to external weather elements. It had an internal memory to store contact details, and a
battery that was large enough to enable deployment over several days of data collection. This was different from previous sensor deployments that relied on a network of RFID readers installed in specific locations to relay data to a central server for processing e.g. [30], meaning that proximity data could only be captured where the readers were located. The properties of the new sensors make them suitable for automatic collection of fine-grained temporal data that may highlight the dynamic nature of contact patterns in individuals.

This pioneer study in Africa revealed three key points. First, individuals were willing to carry the sensors for extended periods of time because the sensors were unobtrusive and did not require user or investigator intervention to collect data. Second, results suggested children spent more time in contact with other children at the household compared to other age groups, while adults appeared to act as bridges between households. Third, within household temporal contact patterns per day were stable across three days of observation, and contacts between individuals of different households were erratic. Residents aged ≥15 years were under-represented due to being away at school or at work. There were two main challenges in the feasibility study related to sensor data collection and participant recruitment. During analysis, unusual (likely erroneous) bursts of activity events emerged during night-time, suggesting that participants stored the sensors in close proximity before sleeping. For future studies, proper training of participants on sensor use and storage (e.g. store them separately when sleeping) coupled with collection of data over several days was identified as critical aspects to minimize these artefacts. In the study, there was only one infant present, but was not recruited. The parents to the infant raised safety concerns due to the relatively large size of the sensor and its pointed edges. The safety concerns have been minimized by a new sensor design that is smaller in size and round in shape, increasing portability and safety even to the very young [46]. These observations from the pilot study revealed important characteristics of study design that may be considered when conducting similar studies in other settings.

2.3.3 Comparison between survey and sensor-based methods in Kenya

In the survey and sensor studies conducted in Kilifi, extensive community-based focus group discussions (FGD) were conducted to assess the feasibility and acceptability of
using the different methods for contact data collection. The FGDs were conducted with participants of similar characteristics and were crucial in determining optimal ways of approaching the community for participation. For instance, for the diary study, two-thirds of the survey participants were illiterate and required a shadow to record their contacts, suspected to have led a behavioural change in participants, and a tendency of shadows reporting fewer contacts compared to self-reports [23]. The diaries had been modified to include pictograms representing 5 age categories, and these were preferred compared to more textual versions. While it was not possible to quantify the extent of behaviour change or reported contacts, these are expected to influence the overall transmission dynamics of infections in the models. Very few of the participants reported having not understood the diary-keeping procedures.

2.4 How comparable are survey- and sensor-based methods?

The interpretation and comparison of results from studies that use different definitions of a contact, study designs, approaches to data collection, and analysis methods require an in-depth understanding of the biases inherent in each approach. Yet, the challenges of collecting social contact data lie beyond controlling only for bias. A good approach should create a balance between the four items mentioned above.

2.4.1 Definition of a contact

The choice of the definition of a contact is primarily guided by the mode of transmission of the pathogen of interest. The majority of paper diary studies adopted a two-level definition of contacts: type I (conversation only) and type II (physical contact) which may be responsible for the transmission of aerosolised pathogens through coughing/sneezing, or transmission through touch, respectively. Sensors can be tuned to detect proximity when individuals are very close and facing each other. When considering proximal contacts that facilitate transmission of respiratory infections, distances of <3 metres have been used [35], down to less than 1.5 metres [30] which can mimic a very close conversation. With this definition, sensors do not record instances of
direct physical contacts, but suggestions have been made to generate body-area-networks using sensors that detect physical touch [38] to capture physical contacts between individuals.

2.4.2 Acceptability of method

Only a few sensor-based studies have reported on acceptability of different data collection methods e.g. [44], or discussed experiences during data collection e.g. [120]. While Kiti et al. deliberately held discussions with, and obtained viewpoints from, residents of the study area prior to sensor deployment [44], other studies in general report on perceptions of the scientists on the entire data collection exercise. One can however make inferences from the participation rates, for example, very high participation rates are reported in sensor-based studies compared to survey methods. This may not entirely be an indication of a method being unacceptable, but other reasons such as individuals being unavailable at place of recruitment, unwilling to give information on social behaviour [66], or a general lack of understanding of the study methods. Such limitations can be minimized by customizing the study tools, for example, pictograms in surveys for illiterate individuals [23], investigating respondent preferences prior to data collection [86], and extensive community engagement to address concerns of the potential participants [44].

2.4.3 Sampling strategies

Sensor studies have mainly targeted closed populations in environments that are easy to monitor and that minimize logistical constraints during deployments. Two studies [44,46] have demonstrated the logistical challenges in data collection from households. In these two sensor studies, individual households were given the sensors on separate days. Only in the Kenyan study was it possible to collect simultaneous data in 3 out of 5 households due to physical proximity of the households albeit with a one-day lag in sensor deployment between households. In all the other reported studies, sensors were issued on the same day since each participant was present almost at the same time in the schools, hospitals, conference or museum setting. This suggests that it is thus possible to target an entire set of individuals in a particular context and achieve high
participation rates but produce results that may be generalisable to the rest of the community.

Paper diary surveys have employed recruiting techniques that are different to sensor studies. Surveys have recruited participants through random digit dialling, physical visits to households of randomly selected individuals, or invitations through the internet. The biggest advantage of survey recruitment methods is that it may be possible to reach a larger set of spatially disaggregated respondents, such as the countrywide study in the Great Britain [45], or the multicounty European study [11], or spanning several locations of rural and urban areas [19,23]. This results in studies that can be generalised within regions or countries.

Table 2.3. Summary of key characteristics of paper diary and wireless proximity sensor studies.

<table>
<thead>
<tr>
<th></th>
<th>Diary studies</th>
<th>Wireless proximity sensors</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Study design</strong></td>
<td>Prospective or retrospective</td>
<td>Prospective data collection</td>
</tr>
<tr>
<td></td>
<td>Paper diaries, web-based interfaces, or telephone</td>
<td></td>
</tr>
<tr>
<td></td>
<td>interviews.</td>
<td></td>
</tr>
<tr>
<td><strong>Recall bias</strong></td>
<td>Respondents may forget to fill diary, record all contacts, or deliberately omit certain contacts.</td>
<td>Only devices within range are logged.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Contacts with individuals not wearing tags are not logged.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Logs may be interrupted by battery failure or faulty devices.</td>
</tr>
<tr>
<td><strong>Resolution</strong></td>
<td>Data mainly collected over one or two consecutive days, or two days separated by a time interval.</td>
<td>Data can be collected continuously and over longer periods. Limits imposed by storage size, battery life and participant consent.</td>
</tr>
<tr>
<td>(spatial and temporal)</td>
<td></td>
<td>Generates a large and complex dataset that poses challenges in analysis and interpretation.</td>
</tr>
<tr>
<td>Aspect</td>
<td>Description</td>
<td>Description</td>
</tr>
<tr>
<td>---------------------</td>
<td>----------------------------------------------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Obtrusiveness</td>
<td>Can be cumbersome due to size, thus may pose challenges in understanding how to fill out in countries with low literacy rates. For easier recall, one might have to carry diary or extra sheet to note down contacts.</td>
<td>Current devices are small (diameter 30-mm) and light (&lt;20 grams) Can be easily integrated into normal clothing e.g. lanyard around neck.</td>
</tr>
<tr>
<td>Ease of deployment</td>
<td>Respondent must fill in the diary manually or respond to questions. Special groups e.g. infants and illiterates need a shadow/assistant to fill in diary Possibility of misinterpretation due to legibility and transcription errors (paper diaries).</td>
<td>Data collected automatically by sensor; no need for user intervention to collect data. Might not be appropriate for young children despite small size. Data automatically downloaded from sensors and ready to use for analysis. Process is tedious for large deployments.</td>
</tr>
<tr>
<td>Scalability</td>
<td>It is possible to get data from people in different geographic settings at the same time e.g. telephone interviews. There is no restriction, unless by study design, on number of participants or the contacts that can be listed in a diary. In most studies, only one person per household is selected to participate.</td>
<td>May be logistically impossible to collect data simultaneously since physical presence is needed. The number of expected nodes is limited to the number of sensors deployed per study. Possible to include all members e.g. in a household, school etc in a study. Limited by number of sensors available.</td>
</tr>
<tr>
<td>Cost</td>
<td>Low cost for developing and printing. Costs increases with study size. Initial cost of design and calibration is high. However, devices can be reused, thus long-term cost can be low. RFID sensors are cheaper than other electronic devices e.g. mobile phones used in Bluetooth data collection.</td>
<td>Response rates are low and may necessitate selecting</td>
</tr>
<tr>
<td>Compliance/Response</td>
<td>Response rates are generally high.</td>
<td></td>
</tr>
</tbody>
</table>
sample sizes to have ample power.

Compliance is subjective

Compliance requires that all nodes in the network participate to get the complete picture, which may not always happen e.g. refusal, absenteeism, etc

(a) Recall: the ability to accurately remember all contacts during the study period. (b) Spatial and temporal resolution: ability to generate dynamic contact data throughout the study period in different locations. (c) Obtrusiveness: the ease of using data collection method without arousing interest of other people, hence leading to e.g. change of behaviour. (d) Ease of deployment: illiteracy and use of shadows, self-reporting or automated data collection. (e) Scalability: the ease with which method can be replicated in different populations. (f) Cost: how expensive the devices are to design and produce en masse. (g) Compliance/Response: how many agree to participate and accurately keep the diary as instructed.

2.5 Using contact patterns in models of disease transmission and control

2.5.1 Mechanism of transmission of respiratory infections

The risk of transmission of respiratory pathogens is influenced by several factors including the interplay between the age-related immunity of the host (i.e. susceptibility), characteristics of the pathogen (e.g. its infectiousness), and the social mixing behaviour of the host which brings susceptible and infectious hosts into contact enabling transmission [5,121]. In infectious disease epidemiology, social mixing behaviour can be summarised by an age-specific square contact matrix $C$ with $i$ age groups. Each element $c_{ij}$ of the matrix represents the number of contacts that age group $i$ makes with age group $j$. Age-specific transmission parameters of a pathogen can then be estimated by multiplying matrix $C$ by a proportionality factor $q$ that measures the infectivity of the pathogen under consideration. The next generation matrix estimated by $N = (n_{ij}) = (q \ c_{ij})$ gives the transmission rates between age groups and characterizes the variability of the risk of infection in the age groups. This transmission hypothesis is referred to as the “social contact hypothesis” [122]. The largest eigenvalue of $N$ gives the basic reproduction number $R_o$, which is the average number of successful infections by an infectious individual in an entirely naïve population [123]. Social contact surveys can thus be used to estimate the relative changes in $R_o$ due to heterogeneities in contact
patterns caused by confounders such as location (e.g. rural vs urban) and setting (e.g. school vs household).

2.5.2 Incorporating transmission patterns in mathematical models of respiratory infections

Following a logical step-wise manner, humans and the parameters that describe their real-life behaviour can be reduced to equations that provide a representation of the real world, i.e. a model [5,124]. The best studied population model for an infectious disease comprises of three compartments, susceptible-infected-recovered (SIR), and an individual can exist in only one compartment at one point in time. Individuals move from one compartment to another based on transition rates. Additional compartmental structures can be used to define different epidemiological characteristics of the disease (e.g. incorporating loss of immunity, or multiple infection states) and demographic or social structures of relevance to disease transmission (e.g. age structure, school and households) and depending on information available to parameterize the model [5]. The basic assumption is of a force of infection acting on susceptible individuals the magnitude of which is determined by the rates of effective contacts (i.e. $q c_{ij}$ – see earlier) and the prevalence of infectious contacts (i.e. proportion infected). In early models the population mixed at random with equal probability of contacting any other individual as in an ideal gas mixture. This was then extended to incorporate heterogeneities of importance in transmission, particularly associated with social grouping e.g. age classes, urban-rural [5].

Empirical social contact structures have successfully incorporated age-specific mixing patterns in models studying transmission and control of several respiratory infections such as influenza [13,122], RSV [58,60], pertussis [125], pneumococcal [59] and Mycoplasma pneumonia [126]. Other studies have demonstrated that a better description of age-dependent serological data for different airborne infections can be attained by using data from contact surveys compared to theoretical contact matrices in use [121,122,127]. In addition, including data on both frequency and duration of contacts would improve the predictive power of disease transmission models [41,42,75,128].
There are several reasons why it has become more appealing to conduct *in silico* studies that focus on understanding the transmission of respiratory infections and the assessment of intervention strategies. For instance, it is difficult and expensive to conduct natural experiments *in vivo*, and ethical requirements have to be satisfied before human studies can be conducted [129]. Mathematical models offer intuitive and tractable tools to study the potential health and cost-benefit impact of introduction of vaccines [60,130,131], to assess the long-term impact of vaccines [59,103], in identifying key vaccine-target populations [54], and assessing contextual intervention measures [50,79,132]. Finally, the utility of contact profiles can be strongly demonstrated by the results of four recent compartmental models that assessed the potential impacts of different vaccine regimes in a low resource setting [58–60,103]. All these age-stratified compartmental models were structurally different but relied on contact data from rural coastal Kenya [23]. Similarly, contact data from the European POLYMOD study has been applied in several transmission models and assessment of intervention measures against various pathogens [42,53,133,134]. These contact studies highlighted the difference in mixing patterns in different populations and the utility of contextual data from resource poor settings.

2.5.3 Using individual based models to study transmission pathways and control of respiratory infections

Individual based models (IBM) are a different way of studying transmission dynamics of infections compared to compartmental models. In IBMs, individuals are explicitly represented, with each individual having a set of epidemiologically relevant characteristics that are tracked over time [129]. This inclusion of interacting social (e.g. contact profile, membership in a group such as a household) and biological (e.g. age, serology profile) characteristics of individuals results in a model that may be better programmed to fit the complex dynamics of disease transmission. The challenge of this is that it requires highly detailed spatial and temporal epidemic and demographic data for validation, and the results obtained may be too specific to be generalised to a larger population [124]. Furthermore, IBMs are more complex to code and computationally more intensive [135]. There has been a growing interest in investigating the transmission of infections in multiple smaller but interconnected subpopulations, mainly
households, schools, workplaces and the community. Studies strongly suggest that households and schools are the hubs of infection spread [26,52,136], and that transmission between the contexts may be linked. A recent review of studies published from 2006-2015 [135] identified two studies from Africa that used IBMs to investigate transmission dynamics of measles [137] and polio [138] in Nigeria. Two other studies on potential vaccination strategies in the absence of a vaccine against RSV in Kenya [131] and the effectiveness of targeted vaccination strategies against Ebola in Sierra Leone [139] have since been published. This demonstrates the rising interest in using IBMs to assess the potential impact of different interventions against endemic (e.g. RSV, no vaccine) and emerging (e.g. ebola, vaccine available) pathogens in resource poor settings. In particular, the study in Kenya incorporated demographic data from the Demographic and Health Survey (DHS) and household transmission dynamics of RSV [26] from the Kilifi Health and Demographic Surveillance System [51]. The public availability of detailed spatial and temporal population data from the DHS [140] and HDSS in low and middle income countries [141] coupled with epidemiological data (e.g. [26]) has helped fuel the interest and practical applications of IBMs in Africa.

Of importance in a social network are two elements: heterogeneity, evidenced by individuals interacting with a subset of population based on socio-demographic characteristics such as age; and clustering or community structure, whereby two or more individuals in a group are more likely to have an interaction than individuals selected at random [98,129,142]. When preferential or assortative mixing is demonstrated, individual based models may provide a more intuitive, complete and accurate way of describing complex social contact structures using simple graph theories [126]. In a network, all susceptible individuals linked to infective individuals have a non-uniform risk of acquiring infection given a contact, potentially leading to different rates of disease transmission in diverse contexts. Based on this, different control strategies can be implemented for different contexts, such as contact tracing for treatment [139,143], non-random vaccinations [144] such as targeting of “super-spreaders” to break the chain of transmission, social distancing such as school or class closure [50], and improvement of hygiene (e.g. hand washing) [108] and standards of living (e.g. adequate ventilation systems) [4]. Older school going children potentially infect younger siblings and caregivers (parents) at high rates at home [26,131,145,146]. Increased number and
duration of close and physical interactions at school and home translates into higher potential for getting infected. Previous studies have suggested that such locations are important hubs for infection and thus good targets for both pharmaceutical [147] and non-pharmaceutical [4,148] intervention measures. A knowledge of the burden of disease coupled with an understanding of the mechanisms through which pneumonia-causing pathogens are transmitted may provide a strong basis for introducing contextual intervention measures against transmission of disease.

2.6 Pneumonia: using contact patterns to understand transmission and control measures

2.6.1 The global burden of disease due to pneumonia

Globally, infectious diseases are the biggest cause of morbidity and mortality across all ages. In 2013, infectious diseases were estimated to have caused 3.3 million (51.8%) out of 9.3 million deaths of children under 5 years [149]. Developing countries bear the biggest brunt of the deaths, with countries in SSA and southern Asia contributing roughly half (3.1 million) and a third (2.0 million) of the deaths in under-5s, respectively. Of the infectious diseases, respiratory diseases have been singled out as the largest cause of death [150]. The Forum for International Respiratory Societies6 highlighted five conditions that significantly contribute to the global burden of respiratory diseases: chronic obstructive pulmonary disease (COPD), asthma, acute respiratory infections (ARI), tuberculosis (TB) and lung cancer. Worldwide, acute lower respiratory infections (ALRI) cause up to 4 million deaths per annum, with the majority of these deaths occurring in infants and young children aged <5 years mainly in South Asia and SSA [149,151]. Pneumonia is a form of acute respiratory infection of the lungs that can be caused by bacteria, viruses or fungi. In 2015, approximately 16% of all global deaths of children <5 years were attributed to pneumonia7, making it the biggest childhood killer. Streptococcus pneumoniae and Haemophilus influenzae type B (Hib) are the two most common causes of bacterial pneumonia [152], while respiratory

---

6 https://www.firsnet.org/
syncytial virus (RSV) is a common cause of viral pneumonia and a major cause of hospitalizations in children under 5 years [153].

2.6.2 The impact of a vaccine against pneumococcal infections

All is not bleak though especially for bacterial causes of pneumonia. By the end of 2017, the Gavi Vaccine Alliance (GAVI) had supported the introduction of pneumococcal vaccines in 60 countries across three continents, resulting in vaccination of more than 143 million children\(^8\). This has led to a considerable reduction (51%) in the proportion of global deaths attributable to ALRIs [152]. The proportion of surviving infants who received the third dose of PCV vaccine between 2008 and 2016 has increased by up to 30% globally, with more than 70% of infants in Eastern and Southern Africa surviving after getting the three doses\(^9\). In Kenya, PCV10 was introduced in routine childhood immunization in 2011, leading to a 68% overall decline in invasive pneumococcal disease (IPD) and pneumonia in both vaccinated and unvaccinated children <5 years by 2016 [154]. Similar reductions have been observed in other countries where PCV13 was introduced such as the USA (64%) and Britain (46% in children <2 years) [155,156]. This demonstrates the importance of the pneumococcal vaccine against the carriage and transmission of pneumococcal bacterial in both middle- and high-income countries, as well as other unquantified measures such as better hygiene and improved access to health care. However, 35 of the countries in the world, and 9 (16%) of the 54 African countries have not introduced PCV in their national immunization programme. With the reduction in cases of pneumonia due to bacterial pathogens, infections due to respiratory viruses may become more prominent. As such, more efforts are still needed to accelerate the decline of deaths attributed to pneumonia globally and within countries so as to achieve the Sustainable Development Goal 3 of ensuring the “good health and well-being of children” [157].

---

\(^8\) [https://www.gavi.org/support/nvs/pneumococcal/](https://www.gavi.org/support/nvs/pneumococcal/)

2.6.3 The epidemiology of RSV pneumonia

In 2015, RSV was globally responsible for an estimated 33.1 million ALRI disease episodes in children aged <5 years, out of which 9.7% were severe cases that resulted to hospitalization, and from which ~2% died [153]. An estimated 90% of RSV-ALRI infections in children aged <5 years occurred in developing countries, with hospital admission rates highest in the first half-year of life. In Kilifi, RSV was identified as the major aetiological agent accounting for >30% (N=759) of all severe hospitalised pneumonias in children aged 1 day to 12 years in 2007 [158]. When comparing inpatient to outpatient cases in Kilifi, RSV was associated with more severe disease resulting to admission at the Kilifi hospital [159].

RSV is ubiquitous and the disease exhibits seasonal patterns and variations in different geographic regions [160], with epidemics starting from the southern to the northern hemisphere from March and June and September to December, respectively. The virus is transmitted through direct or indirect contact with oral or nasal secretions from coughing or sneezing. Following paediatric primary infection and recovery from RSV disease, individuals remain partially susceptible to reinfections throughout life [161], with infants experiencing up to three separate episodes of infections [162,163].

Importantly, the global burden of disease due to RSV is increasingly being recognized [164,165]. Unfortunately, there is no licensed vaccine against RSV. Following the discovery of RSV in the 1960s, a formalin-inactivated vaccine (FI-RSV) to confer protection against RSV disease was tested in infants. Unfortunately, the FI-RSV vaccine resulted in enhanced disease in children, leading to 80% hospitalization and two deaths [166]. Since then, 3 Phase III trials in older adults and infants have been unsuccessful probably due to inadequate study design and implementation, as well as a limited understanding of the development of immunity to infection [165]. Currently, 19 out of 45 candidate vaccines are in phase 1 and 2 trials on various ages: 6 for the elderly aged ≥60 years, 2 for pregnant women, 8 paediatric (see Appendix A). Six vaccines are in Phase II trials, and one maternal vaccine has been in Phase III clinical trials since 2015. In high-risk infants, a passive immunotherapy, Palivizumab (Synagis, MedImmune), is available for short-term prophylactic use against severe RSV disease.
However, Palivizumab is expensive and only recommended for use in high-risk infants i.e. those born prematurely or with exacerbating comorbidities [167].

In the development of vaccines against RSV disease and transmission, four strategies are considered. These are (1) vaccinating infants <6 months old for direct protection, (2) vaccinating children >6 months old for direct protection and prevention of transmission to the other ages, (3) vaccinating pregnant women to boost protective maternal antibodies in new-borns, and (4) vaccinating the elderly [165,168]. Increasingly, mathematical modelling (in silico) studies are used to investigate different vaccination strategies that can be introduced in different groups [58,60], and also to assess the potential maximal health benefits at minimal cost due to introducing different vaccination regimes [54]. In this light, a consultative meeting involving a group of investigators looking at mathematical modelling of RSV intervention programmes was held at the University of Warwick in 2017 [169]. Out of the five main recommendations from the meeting, one key theme was to enhance the understanding of epidemiological factors that influence the transmission of RSV; in particular an improved understanding of social contact patterns focusing on early childhood, school going children and the elderly.
3 Methods

3.1 Study design

3.1.1 Selection of the rural and urban settings within Kilifi County, Kenya

A description of Kilifi County, Kenya

From a general perspective, this study was conducted in Kilifi county, Kenya\(^\text{10}\) (Figure 3.1 (A)). Kilifi county (henceforth referred to as Kilifi) is located approximately 426 kilometres from the capital city, Nairobi. Kilifi is one of the four counties that borders the Indian Ocean on the eastern side of Kenya and covers an area of 12,609 square kilometres (km\(^2\)) (the other three counties are Tana River, Kwale and Mombasa). According to the last census in 2009, Kilifi county had a population of 1,109,735 residents (2.9% of Kenya’s total population), with about two-thirds living in the rural areas. Kilifi has a population density of 88 people/km\(^2\). More than two-thirds (70%) of the residents live in poverty (less than US$1.90 per day\(^\text{11}\)), making it one of the counties with the highest poverty rates. The mean household size in Kilifi was higher than the national mean (5.6 vs 4.4, respectively), with about a third of the population living households with >7 people (national average 19%). Agriculture is the predominant economic activity, with other individuals engaging in fishing, charcoal-making, mining

---

\(^{10}\) Administratively, Kenya is divided into 47 counties, with subsequently smaller sub-divisions into districts, divisions, locations and sub-locations.

and seasonal tourism activities, and a small proportion (18%) being formally employed with a regular monthly salary. The people living in Kilifi are multi-cultural and are comprised of several coastal tribes, with influence from other Kenyans and people of Arab, Indian and European descent [170]. Within Kilifi county lies the Kilifi Health and Demographic Surveillance System (KHDSS). The KHDSS consists of 15 locations covering an area of 891-km² (7% of Kilifi) and is considered as the main catchment area of paediatric inpatients at the Kilifi County Hospital. These 15 locations have been under surveillance since April 2002 onwards, and all household demographic information (births, pregnancy events, deaths, migrations), and geographic location details have been linked to clinical surveillance data from the Kilifi County Hospital [51]. The KHDSS thus contains a rich dataset of geo-located individuals under continuous health and demographic surveillance grouped into households. For community-based studies that recruit participants from within the KHDSS, appropriate datasets can be used to locate participants for recruitment.

Selection of study locations within the KHDSS

For this cross-sectional study two locations, Matsangoni and Kilifi Township (Figure 3.1 (C)), were purposively selected from the KHDSS. These sites are along the Mombasa-Malindi highway and were selected to represent rural (R) and urban (U) settings, respectively. The rural site has been the source of data for, among others, studies investigating the transmission of respiratory infections [26] and was the pilot site for an earlier version of sensors [44] used for contact detection. In addition, both sites have participated in a cross-sectional diary-based study on contact patterns [23]. Further, each site has a public health centre, with Matsangoni Health Centre being located approximately 30 kilometres from the Kilifi County Hospital.
Figure 3.1. Map of the study locations showing the rural and urban area. Panel (A) shows Kenya country map (grey) and Kilifi County shaded dark grey. Panel (B) shows the Kilifi County (grey) and the Kilifi Health and Demographic Surveillance Site in dark grey. The rural (orange) and urban (purple) sites are shown in panel (C), with the location of the school and health centre in each location shown.

There are differences in demographic patterns, with about 3 times the number of people living in Kilifi Township compared to Matsangoni (51,937 vs 15,556 residents, 2016 KHDSS data). There are also evident differences in the distribution of population by age and gender (Figure 3.2), particularly for ages 15 years and above. The proportion of males in Matsangoni location aged 15-50 years shows a marked decline mainly attributed to rural-urban migration in search of education and employment opportunities in urban areas within Kilifi and surrounding towns [51]. Kilifi Township is dominated by individuals living in modern (stone) houses with cemented floors predominantly in their own compound, or in apartments with shared compounds. Most times, individuals living within the same compound occupy rental houses but are not related. In contrast, rural houses are predominantly mud-walled with palm-leaf-thatch roof. Several related families (by blood and marriage) live in the same compound with shared facilities such as outdoor kitchen and toilets. Within the urban area, adults of working age are formally
employed in offices, engage in mining of coral rock, or fishing along the creek. In the rural area, individuals practise subsistence farming, operate small businesses and practise fishing. Typical of urban centres in Kenya, Kilifi Township is multi-ethnic while Matsangoni is dominated by one of the local tribes found along the coast (Giriama).

**Figure 3.2. Population pyramids of Matsangoni and Kilifi Township locations from the KHDSS data, round 26 of 2016.** The proportion of males (blue) and females (red) is shown for 5-yearly age groups from age 0 to 85 years and above.

The theoretical framework of study design

An ideal study design would have been to saturate the whole community with sensors to investigate the complete network of interactions. Capturing a full network would enable the description of interactions whether or not they resulted in transmission of a pathogen resulting in an outbreak. In such a case, one would expect to observe the presence of interactions within and between individuals in different settings (schools, households, workplaces, places of worship, transport hubs, etc) and characterization of network properties such as strength and duration of interactions. However, several considerations have to be made as discussed in a review by Eames et al. [171], while putting in perspective the infectious disease in question and the research interests of the investigators. Assuming that most transmission occurs within and between schools and households, the potential interaction networks can be visualised as in Figure 3.3. In
practice the approach taken for this study was constrained by resource limitations – specifically the number of proximity sensors available.

![Figure 3.3. Conceptualization of an interaction framework in schools and households.](image)

The school shows students (blue nodes) grouped into classes. The index student is marked X. Teachers are shown as red nodes. Interactions occur within and between classes. Household 1 hosts the index student. Household 2 represents a group of neighbouring households enclosed in a compound (dotted square boundary). Interactions can occur within or between households, between household residents and school-goers, and within and between students and teachers in various classes.

3.1.2 Selection of schools

Two schools, one each from the selected rural and urban locations, agreed to participate in the study. These were Matsangoni Primary School and Kilimo Primary School from the rural and urban settings, respectively, shown on the map in Figure 3.1. The selection of the schools, based on inclusion criteria below, was guided by discussions with the Kilifi County Education Office. Structural differences such as school and class composition and size were expected, which in turn would affect the network structure [34,172]. Records from the Kilifi County Education Office estimated the total number of students per school in Kilifi as 600-900 (personal communication). Each school was
further stratified into preschool (kindergarten, KG), lower primary (grade 1-4) and upper primary (grade 5-8) strata. In Kenya, students in these grades are generally within the ages 3-5 years, 6-9 years and 10-15 years, respectively. The different strategies for selection of students for each school were devised as explained below.

Criteria for enrolling a school

(i) School included primary and early childhood development (ECD) or kindergarten.

(ii) Approval from the County Education Office (responsible for all the administrative matters regarding education in the region. All engagement with the school required express permission from the officer in charge).

(iii) Approval from Headteacher and school’s Board of Management.

3.1.3 Selection of households

For each of the three strata at the school (preschool, lower and upper primary), 3 index students were selected randomly and followed to their households, giving an initial 9 index households per setting. For each of these 9 households, an additional 3 to 5 (minimum-maximum) neighbouring households were recruited into the study to give an expected maximum of 495 and 405 household participants in rural and urban setting, respectively (average household size is 11 and 9 in rural and urban areas, respectively. Unpublished KHDSS data). The expected sample size was considered based on the number of sensors available to conduct the experiment per setting, with provision for lost or damaged sensors that would need replacement.

Criteria for enrolling a household

(i) Assent from head of household of index student or of neighbouring household.

Criteria for exclusion of a household

(i) More than a third of members refusing to participate
### Table 3.1. Definition of key terms used in this study.

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Household</td>
<td>A group of individuals eating from the same kitchen and referring to one of these individuals as the head.</td>
</tr>
<tr>
<td>Building unit</td>
<td>A dwelling in which individuals live. One or more building units can form a household.</td>
</tr>
<tr>
<td>Index household</td>
<td>The household in which an index student selected from the school resides.</td>
</tr>
<tr>
<td>Neighbouring household</td>
<td>One that shares a common fence/ boundary with the index household (rural/ urban), or several households that are co-located in a compound owned by one individual e.g. rental houses (urban only).</td>
</tr>
<tr>
<td>Compound</td>
<td>The area enclosing a household or several related or non-related households whose boundaries defined by law or mutual agreement between neighbours.</td>
</tr>
</tbody>
</table>

#### 3.2 Data collection infrastructure

Proximity data were collected using wearable proximity sensors (See Figure 3.4) developed by the SocioPatterns collaboration (a European consortium of institutions and investigators focused on social dynamics, refer to [www.sociopatterns.org](http://www.sociopatterns.org) for more details) based on the opensource OpenBeacon open hardware design ([https://www.openbeacon.org](https://www.openbeacon.org)).

The sensors exchanged ultra-low power radio packets in a peer-to-peer fashion by transmitting and scanning their neighbourhood for packets sent by nearby sensors. Sensors in proximity exchanged a maximum of 1 data packet per second and could store up to 5 million contact events in the on-board memory, corresponding to over 1000-hours of continuous data collection. This exchange of low-power radio-packets due to the spatial proximity of the individuals wearing the sensors was used as a proxy of a face-to-face interaction [30,47]. To estimate how close individuals were, the attenuation of the signals with distance was computed as the difference between the received and transmitted power. Proximity between individuals was asserted when the median attenuation over a given time interval exceeded a specified attenuation threshold (in dB). The duration of time intervals over which the time-varying graph was computed
(‘time slice duration’) defined the temporal resolution of the contact network. This resolution could be set freely between 1 second and any integer multiple of a second. Previous sensor deployment in humans suggested that 20s can be considered as a reasonable duration to describe “the fastest social interactions”, such as a quick conversation or handshake, in a social gathering [30].

During the data collection, some simple measures were put in place to minimize data loss resulting from not carrying or tampering with the sensors. For instance, one class representative was appointed to ensure that each student participating in the study wore the sensor as expected. Whenever possible, the head of the household was also asked to ensure that each household member wore the tag correctly each morning and stored them separately after taking them off. These actions were not expected to be a major role that would have affected the normal class/ household routine, but they were difficult for the research team to monitor directly.

Figure 3.4. Wireless proximity sensor used in this study. Panel (A) shows the sensor (left; diameter 30-mm) next to a 20-shilling coin. Panel (B) shows an adult wearing the sensor in a pouch hanging at the chest. Panel (C) shows a student wearing the sensor pinned to his school shirt.
3.3 Ethical considerations

Ethical approval was sought from Scientific and Ethical Review Unit (SERU, Kenya) and Biomedical and Scientific Research Ethics Committee (BSREC, University of Warwick). See Appendix B and Appendix C, respectively, for the approval letters.

The following guidelines were observed:

(i) “First, do no harm”. An earlier version of the wireless proximity sensors were used in a previous study in Kilifi that involved piloting the use of sensors in the community, understanding community concerns and learning best practice methods for deployment [44], as well as in several other referenced studies. More recently, modified sensors were used in a household-based study in Italy that suggested the appropriateness of the sensors for use in all ages [46], unlike the Kilifi study that omitted one infant due to the size of the sensors. There are no known risks posed by the low power frequency signals emitted by the sensors. The sensors were first inserted in a polythene zip-lock bag, which was then inserted in a pouch made of cloth. The zip-lock bag was to prevent the sensor coming in contact with water and dust in the environment. For each school, the pouch was made from cloth that had a similar colour to the school’s shirt/blouse. Different colored pouches were made for the rest of the participants, and each was allowed to choose a preferred colour. These measures were taken to ensure personal safety, to minimize device loss by theft or misplacement (e.g. participants exchanging sensors by mistake), or data loss through tampering by the participants.

(ii) Benefits to participants:

   a. Parents who attend the engagement sessions at the school were refunded travel expenses. This did not exceed USD 1.5 (~KES 200.00). Participating households also benefited from health talks that focused on prevention of communicable diseases such as pneumonia and diarrhoea. This included talks on importance of washings hands (before and after visiting toilets, before handling food, before handling infants, etc), use of handkerchiefs or
disposable tissues when sneezing or coughing, and a demonstration of proper hand washing techniques. Each household received two bars of hand-washing soap at the end of data collection.

b. At the end of the study period at schools, each student received a stationery pack consisting of a mathematical set, 2 writing books, 2 pens and a ruler. The school was given several bars of soap and hand washing stations (a bucket with a tap) to be used by the students. The research team also organized an elaborate open day with various activities at the end of the school terms. The author presented a summary of the research progress, including participation rates, challenges encountered, and a message of appreciation for the cooperation from all involved. The research team was present to respond to further questions from the meeting attendees. In addition, members of staff from KEMRI helped to organize simple explanations of activities undertaken by the research institute and their benefits to the health of the community. A personal hygiene session was conducted, with demonstrations to students and parents on how to correctly wash their hands and dispose rubbish off. At the households, smaller discussions were held with the families, and each family received a hygiene pack comprising soap and hand washing stations. Generally, these benefits and discussions were appreciated by the research communities.

(iii) Consent and assent forms were back-translated from English to two local languages, Swahili and Giriama. Participants were free to choose the language in which they wanted the information presented. Sample English consent forms are found in Supplementary Information file 2.

(iv) Participants were free to leave the study or to request the withdrawal of their data at any time and for whatever reason without explanation and without penalty.
3.4 Recruitment procedures and community engagement in schools and households

3.4.1 Enrolment of students from the rural school

Matsangoni Primary School in the rural setting (SR) had a full school register with 907 students. The register contained details of all the students enrolled in the year and the classes they belonged to. From the school register, 350 students were randomly selected. Due to the high number of students per grade, parents were invited to attend a meeting with the research team on different days depending on the grade of their child. At each meeting, an explanation of the overall study objectives was explained by the author, and the parents were then separated into smaller groups for in-depth explanation of study procedures and expectations from the parents by well-trained fieldworkers. Attendees could ask questions on the research objectives and procedures, and the responses were compiled into a frequently asked questions (FAQ) booklet that was given to the participants in subsequent meetings (Appendix D). Following this extensive session that lasted about 2 hours, parents who agreed to let their children participate in the study signed a consent form (Appendix E). On a later date and during normal school breaks, students whose parents had given consent were also informed of the study procedures; only those who gave verbal assent (<10 years) and written consent (>10 years) could participate.

3.4.2 Enrolment of students from the urban school

The first school (name withheld) that was approached delayed in giving a positive response to the research team despite several meetings with the school administration, school management committee and parent-class representatives. This was anecdotally attributed to various reasons such as a feeling of expectation (what is the direct benefit to the student versus benefit to the community), parents lacking time to attend meetings at the school, and a general apathy to research activities due to misinformation about research. Further engagement with the school was stopped, and a letter of withdrawal was presented to the school administration and the SMC.
Following the first unsuccessful attempt to recruit an urban school, the KEMRI-WTRP (henceforth known as KEMRI) Community Assistance for Study Team (CAST\textsuperscript{12}) suggested a different community engagement strategy to recruit a second school (SU). The school management committee and parent representatives of the urban school were first invited to an in-depth tour of the KEMRI facility. The teams visited the laboratories and offices to understand the research activities undertaken by the research institute. On the same day, a meeting was held with the school team to discuss the intended study. At the school, individual letters addressed to parents were issued through the students inviting them to a meeting at the school co-chaired by the parent representative for each grade. The parents who showed up for the meeting received further information on the study, and those who agreed to allow their children to participate signed a consent form for parental approval.

Following written approval by the headteacher and the PTA for each school, the research team addressed all the students during one regular morning assembly to ensure they were aware of the intended research activity. Potential participants were identified through the school register and letters of invitation issued to their parents. For consenting, parents of selected students were either (a) requested to come to school for a group consenting exercise, or (b) followed up to their household for consent by matching KHDSS household records with those of the students. Only students whose parents gave consent were asked for individual assent to participate. Teachers and other staff also provided individual consent. To minimize disruptions to the normal school routine, engagement sessions with the students were arranged during their normal breaks, such as class recess, lunchtime or sports breaks.

3.4.3 Enrolment of household residents

A list of all index household residents and the selected neighbours was prepared from the KHDSS database. Using these household lists, all residents physically present at the house during the recruitment phase were enrolled. To simplify the recruitment process, all residents present during a household visit were requested to meet with the research

\textsuperscript{12} A CAST is formed by the KEMRI Community Liaison Group (CLG). The task of CAST is to advise researchers on best practices for community engagement, such as approaches to participant recruitment, developing key messages, etc.
team during one session for a group discussion. Scheduled appointments for return visits were made for any resident not present at the initial home visit due to other activities (e.g. school, work, travel, etc).

Following consent from all parties, participants were informed of the data collection procedures at least three days prior to the data collection start date. Background socio-demographic data (gender, date of birth, household identifier, etc) were available from the KHDSS database [51].

3.4.4 Summary of enrolment criteria for individual subjects from within school or within household.

(i) School or household member provides written consent (teachers, adult, caregiver) or assent (child aged <10 years old)
(ii) Member of a household which contains an index student or of a household immediately neighbouring a house of an index student.

3.4.5 Data collection timeline

The timelines for data collection in the different settings (rural and urban schools and households) are shown in Error! Reference source not found.. Data in the rural setting were collected between September and October 2016, and in the urban setting between February and March 2017. Here, SR and HR represent school and households in the rural setting, respectively, and SU and HU represents the school and households in urban setting, respectively. Gaps between S and H periods were to allow data collection, download and verification. From November 2016 to early January 2017, all primary schools were closed for holiday and to allow students in grade 8 to sit their national exams. For each setting, data were collected over a 10- to 12-day period. This allowed a maximum of three days of recruitment and giving of sensors, seven full days of data collection, and a few days after data collection to allow retrieval of sensors from participants.
3.5 Data cleaning

The sensor firmware and data cleaning procedures have been developed as part of the custom SocioPatterns software developed for this and other studies. The publicly available version of the SocioPatterns tag firmware is a branch of the OpenBeacon firmware that has been developed, tested and verified by the ISI Foundation and the SocioPatterns project (https://github.com/sociopatterns/openbeacon-ng).

The cleaning of data for each site was conducted in several steps:

(i) cleaning the main participant file data (metadata) which contains the demographic details of participants. The household residents were divided into five age classes to reflect the Kenyan education system and social grouping: (0-4), [5-14], [15-19], [20-49] and [50+] years. Household sizes were based on quartiles: 1, 2-4, 5-9, >10.

(ii) cleaning the contact dataset.
   a. identify and select the period of study - based on participant recruitment and number of study days
   b. observe micro-activity timelines, i.e.
      • all data over the 7 days. Take note of patterns, trends, etc.
      • check global daily patterns and trends
   c. observe macro-activity timelines, i.e.
      • check daily patterns and trends by age group
- check daily patterns and trends by household
- check daily patterns and trends by individual

(iii) eliminate spurious contact data
   a. check micro-patterns for randomly selected individuals
   b. assess algorithms to smooth out the data and remove spurious events

3.5.1 Step 1: Assessing contact event timeline using data from Matsangoni households (HR)

The following gives a demonstration of the cleaning process, using data from Matsangoni households as the example. The first step was to plot the overall activity timeline of each dataset. For Matsangoni, data were collected between 19th and 29th October and the activity timeline over all 10 days is shown in Figure 3.6. Data from the first two and the last day were discarded to eliminate data collected during initial issuing and collection of sensors. Data thus reported cover 7 days from 21/10/2016 00:00:00 hrs to 27/10/2018 23:59:59 hrs. Contact event data were then aggregated into 10-minute intervals and plotted against the time of day.

From the plot of raw data (Figure 3.6), it is possible to observe the regular temporal troughs and peaks for each day (panel (A)). A keen look at the daily patterns (panel (B)) reveals that activity patterns are occurring in an opposite direction to those expected. That is, contacts are very low during the day, and rapidly increase in the evening from about 18:00:00 hours to peak a few hours later and remain relatively constant, and then drop drastically from around 04:00:00 hours the next morning and remain low throughout the day with observable peaks at various times during the day. In reality, we would expect contact activity to peak during the day and be low at night (almost zero if participants have no interactions or store their sensors separately when sleeping).
Figure 3.6. Activity timeline showing number of contact events across study days for raw rural household contact data. Panel (A) shows data between the start and end of collection, 18/10/2016 00:00:00 and 28/10/2016 23:59:59, respectively. The solid vertical lines are at midnight of each day. Panel (B) shows data between 22nd 00:00:00 and 24th 23:59:59 as highlighted in orange. The number of contacts were aggregated at 10-minute intervals.

This observation from the raw data had two possible explanations: first is that individuals had short bursts of interaction separated by short intervals (spurious contact events), or that the sensors periodically experienced continuous episodes of interaction. In the two reported household based studies in Kilifi [44] and Italy [46], night-time data (from 8:00 pm-7:00 am) were discarded due to an unexpected high volume of contacts during this time. Rather than discard the “night-time” data in this current study, methods of cleaning the data to extract a dataset that would be representative of the expected temporal activities of individuals living in households were explored, as described below.
In this exercise, the y-axis accelerometer data were important as it defined the position of the sensor relative to zero, where a value of zero refers to when the sensor is completely flat facing upwards. A sensor (individual) was deemed to be in correct use if the y-values were $0 < y \leq 1$ or $-1 =< y < 0$ and a flag value of 16, meaning that y lay between -1 and 1 but not equal to zero (zero signified a sensor at rest, such as when placed flat on a table). For instance, increased activity was observed late at night for continuous periods in some sensors. Individuals were supposed not to wear the sensors when asleep and to store the sensors separately when not being worn, suggesting that activity indicator would be equal to zero. However, if sensors were all placed together in close proximity at the same location then they would still continuously record contacts.

Two different methods were used to separately identify activity timelines, as described below in method 2 and 3. For illustrative purposes, data from only two households are shown for both tests: household 44 (n=2) and household 12 (n=6). The household identifier was automatically generated by grouping all residents sharing a building unit identifier.

### 3.5.2 Step 2: Assessing day and night contact event timelines

In the second step, for each sensor a dummy dataset to define blocks of day and night was created. The dummy data assumed that daytime lasted from 06:00:00 to 20:59:59 hours, and night time from 21:00:00 to 05:59:59 hours, for each of the 7 study days. The day and night periods were theoretical and for visualization purpose but guided by observing individual activity timeline data to estimate when sensors were collecting contact event data. The dummy data were graphed over individual activity timelines of residents of 8 (out of 44) randomly selected households.

Results for household 44 are shown in Figure 3.7. Panels (A) and (B) represent the activity of two unique individuals over 7 days of data collection. The grey boxes represent times from 6:00 am – 9:00 pm for each day, when we assume an individual should be actively interacting with other study participants. Each coloured line represented accelerometer data values per household resident throughout each of the 7 days. Sensor (A) was collecting proximity data during the day for the first two days. On
the night of October 23-24\textsuperscript{th}, there was heightened activity recorded by the sensor that diminished for the rest of the days. During the afternoon of October 24\textsuperscript{th} the data suggests that the sensor experienced a malfunction and remained dormant for almost one day. The sensor restarted in the early morning of 26\textsuperscript{th} and continued to collect data till the end of the data collection period. For sensor (B), data were collected consistently during the daytime hours but only for the first two days. It was not clear what caused the sensor to stop collecting data on 22\textsuperscript{nd} October. For the rest of the sensors, interpretation of the graphs was done in a similar manner.

![Image](image.jpg)

**Figure 3.7. Activity timeline of Household 44.** Row 1 shows charts of activity timelines for sensors (A – brown trace) and (B – green) representing data from residents of household 44. Chart (C) shows an overlay of activity timelines for both individuals. Spikes peaking at value 15 depict when a contact events were detected, and other values above 15 depict a sensor error (reboot). The grey boxes highlight 06:00:00 to 20:59:59 hours of each day, assumed to be when an individual was wearing a sensor. Each trace covers the same period from Oct 21\textsuperscript{st} 00:00:00 to 27\textsuperscript{th} 23:59:59.

Household 12, shown in **Figure 3.8,** was a 6-member household, with 5 out of 6 members having regular data collection over the entire data collection period. It is not known why sensor (C) had no data collected. Looking at the combined graph (G), there is clearly less activity recorded by the sensors throughout most of the night times. Interestingly, on the nights of October 22\textsuperscript{nd} and 24\textsuperscript{th}, a lot of activity was observed for majority of the sensors. Given that October 24\textsuperscript{th} was a Saturday, it was easy to assume
that the household members would experience enhanced contact activity either within or out of the household.

**Figure 3.8. Activity timeline for household 12.** There were 6 residents in household 12 whose activities are shown by (A)-(F). Graph (G) is an overlay of the individual timelines for all 6 residents. The description of the graph details follows that of Figure 3.7.

3.5.3 Step 3: Assessing performance of thresholds to filter contact event data

In the third step, an inbuilt rolling mean algorithm available in the *Pandas* package [173] (Python v2.7) was implemented using various parameters. In this experiment, a window of size $k$ was defined, and the average ($\bar{k}$) of $k$ consecutive time points was calculated. For each run, if $\bar{k}$ was below a threshold specified a priori, then the data were discarded.

Several thresholds were assessed:
A threshold of 0.4 with a sliding window of 4. This means that the sensor was expected to have been collecting data for more than 40% of the time for each hour, corresponding to at least 24 minutes per hour. Figure 3.9 shows the rescaled activity timelines same individuals from household 44 (A and B) in brown and green, respectively, while Figure 3.10 shows the rescaled activity timelines same individuals from household 12. The black line overlay depicts the results from the rolling mean algorithm.

Figure 3.9. Rescaled activity timelines for household 44, method 1. Panels (A) (brown) and panel (B) (green) represent data from two individuals in household 44. The black line represents smoothed data. The description of the graph details follows that of Figure 3.7.
Figure 3.10. Rescaled activity timelines for household 12, method 1. Panels (A)-(F) represent data from each of six individuals in household 12. The black line represents smoothed data.

(ii) A threshold of 0.75 with a with sliding windows of 4. This means that the sensor was expected to have been collecting data for more than 40 minutes per hour. **Figure 3.11** shows the rescaled activity timelines same individuals from household 44 (A and B) in brown and green, respectively. **Figure 3.12** shows the scaled activity timelines for the 6 individuals in household 12. The black line overlay depicts the results from the rolling mean algorithm.
Figure 3.11. Rescaled activity timelines for household 44, method 2. Panel (A) and (B) represent data from each of two individuals in household 12. The black line represents smoothed data.
Figure 3.12. Rescaled activity timelines for household 12, method 2. Panels (A) to (F) represent data from each of two individuals in household 12. The black line represents smoothed data.

3.5.4 Step 4: Selecting the best-fitting data filter

The last step was thus to select the threshold specification that omitted periods that suggested spurious contacts particularly during the night times when we did not expect to see contacts since individuals were expected to be asleep and the sensors stored separately. The rationale was that when at rest or at night, one would expect these times to exhibit long durations of inactivity. These two rolling-mean methods were visually compared against each as shown in Figure 3.13 and Figure 3.14, whereby the solid black and dotted red lines represented data from methods 2 and 3, respectively.

From Figure 3.13 and Figure 3.14, it was observed that method 3 omits majority of the spikes compared to method 2. These omitted spikes were considered as spurious data because they represented periods of contact event outbursts outside of the specified
threshold. This was observed across all the other 18 households selected for this exercise. Thus, method 3 was chosen as the data filter of choice for all participants.

Figure 3.13. Comparison of rescaled activity timelines for household 44. A and B represent data from two individuals in household 44. The black line represents smoothed data using method 1, while the red dotted line shows smoothed data using method 2.
3.6 Data analysis

Patterns of contact between participants were analysed by statistical distributions describing: a) the number of contacts in households and schools, b) the duration of the contacts, c) the cumulative time spent in contact, and d) the temporal evolution of the networks. Heterogeneity of the contacts and their statistical distributions were assessed across five key variables: age (0-4, 5-14, 15-19, 20-49, >50 years)), gender (female and male), temporality (hourly, daily and weekly), grade (kindergarten, grades 1-8), and setting (rural/ urban). Analysis was conducted using Python (Python Data Analysis library, Pandas) and custom and non-public data processing software by the SocioPatterns project (data cleaning and management), R (statistical analysis and
network visualisation), Gephi (network visualization) and QGIS (cartography). Network data analysis and visualization were aggregated at the school and household level with nodes representing students and household residents, respectively. Links between two individuals $i$ and $j$ in contact were weighted by the cumulative duration of interaction between them. Temporal data were aggregated into time windows of 10 minutes, hourly, daily and over the entire duration of the study (7 days).

3.6.1 Definition of outcomes

A contact event was considered detected if one sensor recorded a radio packet from another sensor, and the incoming radio power was higher than a given attenuation threshold calibrated to corresponded to about 1.5 metres of separation distance [44].

![Figure 3.15. Schema demonstrating the definition of contact event and contact duration.](image)

One contact is a 20-second proximity event that occurs between individual $i$ and $j$ without a break. Here, there are 5 individuals. Each red horizontal line represents the start (left-hand side edge) and end (right-hand side edge) of a proximity event between $i$ and $j$ lasting $s$-seconds. Individual 1 has zero contacts since the duration of contact is <20 seconds. Individual 2 has 1 contact event lasting exactly 20 secs. Individual 3 has three contact events cumulatively lasting 60-seconds, with a 20-sec interruption between the contact events. Individuals 4 and 5 have three and six contact events lasting 60- and 120-seconds, respectively.
The primary outcome of interest was the median degree \( < k > \) and corresponding interquartile range (IQR, 25th and 75th percentiles). For a contact network, the following quantities, similar to a household study in Kilifi [44], were defined:

(i) The degree \( k_i \) of a node \( i \) is the number of other nodes to which it was linked during the time window.

(ii) A contact event is a continuous set of 20-second interactions between two tags without a 20-second break.

(iii) The weight \( n_{ij} \) of an edge between nodes \( i \) and \( j \) was the number of contact events recorded between these individuals during the time window (see Figure 3.15 for illustration). The mean number of contact events was computed as the sum of the individual contact events divided by the number of node-pairs in proximity, \( n, \frac{\sum n_{ij}}{n} \).

(iv) The weight \( w_{ij} \) of an edge between nodes \( i \) and \( j \) was the total duration of contact events recorded between these individuals during a given time window (see Figure 3.15 for illustration). The mean contact duration was computed as the sum of individual contact duration divided by the number of node-pairs in proximity, \( n, \frac{\sum w_{ij}}{n} \).

(v) The network density is the ratio of the number of observed edges formed in a network to the maximum number of expected edges [46].

(vi) The clustering coefficient measures the cohesiveness of local groups of nodes by calculating the probability of two different contacts of individual \( i \) also contacting each other.

(vii) For two consecutive day-pairs e.g. day 1 - day 2, day 2 - day 3 etc, the cosine similarity was defined as an individual’s tendency to have repeated contacts with the same individual, taking into account the duration of contact (weight) \( w_{ij,1} \) and \( w_{ij,2} \) on the edge \( i \leftrightarrow j \) measured at time \( t_1 \) and \( t_2 \) [34], calculated as:

\[
\text{sim}(i) = \frac{\sum_j (w_{ij,1})(w_{ij,2})}{\sqrt{\sum_j (w_{ij,1}^2) \sqrt{\sum_j w_{ij,2}^2}}} \tag{1}
\]
Essentially, the cosine similarity calculates the changes in the neighbourhood of each node in each pair of daily networks, suggestive of whether a node $i$ was in contact with and spent the same amount of time with the same nodes for each successive day pair. The cumulative cosine similarity over the 7-day study period was calculated as:

$$\frac{\sum_{i=1}^{n} \text{sim}(i)}{n-1}$$

where $n = 7$.

Cosine similarity takes values ranging from 0 and 1, with lower values suggesting that neighbouring edges are not the same at time $t_1$ and $t_2$, while accounting for time spent in contact. In order to assess the magnitude of the cosine similarities, these values were compared to a null model. The null model randomly swapped the weights of the networks among the edges but did not change the topology of the network.

### 3.6.2 Data visualisation methods

It was expected that the values of the contact events and duration of interaction would vary over a large range. To best visualize the probability distribution of the number of contact events or contact event duration of a set of individuals (e.g. grouped by age), log-log plots displaying data on both axes using the logarithmic scale were used. The continuous data were divided into 50 bins as in previous studies [44,46]. To visualize the networks, the Distributed Recursive Graph Layout (DrL) and Fruchterman-Reingold (F-R) force directed algorithms available in the igraph package in R were used. In a force-directed algorithm, the attractive forces act upon the edges while repulsive forces act between nodes. The F-R algorithm minimizes edge crossing and node overlap thus distributing nodes evenly in the visualization frame while ensuring that the lengths of edges are similar [174]. Due to this, individuals will be clustered together as the density of the links among them increases. The DrL algorithm aims primarily to minimize the overlap of large clusters, or in other words, to emphasize dense clusters [175].
4 Results from the rural school

4.1 Introduction

This chapter presents results of data collection from the rural primary school. The specific objectives of the rural school study were:

(i) To collect data from a subset of students from Matsangoni primary school in Kilifi, Kenya.
(ii) To quantify the degree, number and duration of contact events by grade, gender, and day of the week.
(iii) To assess the dynamic variation of the school network by time of day and day of the week.

Social proximity contacts were collected using wireless proximity sensors over 7 continuous days. Individuals were instructed to wear the sensors each day on waking up before starting their normal business of the day (e.g. attending school) and take them off in the evening before going to sleep or engaging in chores that involved bodily contact with water such as taking a bath. The sensors were worn at chest level, thus only able to detect close face-to-face interactions between students wearing sensors.

The baseline characteristics of the school are presented giving an overview of the school structure and distribution of students by grade, gender and age. Results of the
distribution of number and duration of contact events are presented as probability density functions and matrices. These results are further stratified by grade, gender and day of the week to assess the effects of these covariates on the distributions. Network diagrams of the interactions are drawn, showing the overall connectivity, with nodes representing a student (each coloured by their grade) and edges weighted by the total (and daily) duration of interaction between two nodes. The properties of these networks (degree distribution, density, clustering coefficient and cosine similarity) were then compared to random networks having the same number of nodes and link density.

4.2 Baseline characteristics of participants from Matsangoni primary school

4.2.1 An overview of characteristics of the rural school

Matsangoni Primary School (SR) had 907 students (51% of which were female) in April 2016. The school was divided into 11 grades ranging from the youngest in 3 grades being kindergarten (KG1-3), and grades 1 to 8. In the analysis, KG1-3 are presented as one grade (KG) due to the small number of students per grade. There was also one class with special need students who did not participate in the study due to ethical constraints. For grades 1-8, there were two classes per grade: Blue (B) and Green (G). There were 27 physical rooms in total in four blocks, with separate lower and upper primary blocks (19), and the rest including the staff room, computer room and school stores. In the results, all three kindergarten grades were combined and presented as one grade due to the small number of participants per class. Class size ranged from 16 (KG1) to 60 (class 7) students, and the largest grade had 134 students in total (grade 7). There were 13 teachers and 3 ancillary members of staff (1 watchman and 2 cooks). The school had three terms in a year, and this study was conducted in term three that ran from 05/09/2016 to 08/11/2016. School opened early and students arrived in school from 7:00 am each day. Normal learning begun at 8:00 am to end at 3:00 pm, with two breaks at 10.00 - 10.30 am and 12:00 noon – 2:00 pm (lunch), and students were free to go home in the evening from 4:00 pm. Between 3:00 and 4:00 pm,
students attended a session of physical exercises, where they participated in coordinated or uncoordinated outdoor games such as football.

4.2.2 Characteristics of the participating students

A sample of 350 students (46% of total) were randomly selected from the school register proportional to class size and gender distribution. This number of students was dictated by the number of sensors available for the study. Out of the 350 selected students, 42 (15%) were unavailable for consent due to various reasons presented in Figure 4.1 (e.g. transfer to another school, not present in school at time of consent), while 5 refused to participate in the study (reasons not collected). Only 7 teachers participated in the study (2 refusals, 1 dropout, 3 absent during recruitment days). However, only data from students are presented. Only data between Wednesday 28/09/2016 to Thursday 04/10/2016 of the following week were analyzed.

Data were retrieved from 299/303 student sensors (4 sensors lost). After cleaning as described in section 3.5 and matching to participant metadata (date of birth, gender, grade), complete data were available for 223 individuals. A summary of distribution of participants by grade and gender (52.5% female) is shown in Table 4.1. The median age for students in SR was 13.4 years, with the oldest participant being 26.5 years old in grade 5, and the youngest in KG1 aged 3.9 years. The median ages of the students increased with grade from KG to class 8 albeit with overlap in the ranges by about 2 years. Grade 7 had the highest number of participants (n=40) and the lowest number in grade 1 (n=16).
Figure 4.1. Data collection flowchart in the rural primary school. This shows the total number of students in the school, number recruited, reasons for loss to follow up, and total number of sensors from students that were available for analysis.

Table 4.1. Baseline characteristics of participating students from the rural primary school. This table shows the grades, number of students by gender and the median age per grade.

<table>
<thead>
<tr>
<th>Grade</th>
<th>Number of students</th>
<th>Female</th>
<th>Male</th>
<th>Median age years (IQR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>KG(^1)</td>
<td>21</td>
<td>12</td>
<td>9</td>
<td>8.5 (6.7-9.3)</td>
</tr>
<tr>
<td>1</td>
<td>16</td>
<td>5</td>
<td>11</td>
<td>9.4 (9.0-9.7)</td>
</tr>
<tr>
<td>2</td>
<td>23</td>
<td>13</td>
<td>10</td>
<td>10.8 (9.9-11.8)</td>
</tr>
<tr>
<td>3</td>
<td>19</td>
<td>10</td>
<td>9</td>
<td>11.8 (10.4-12.6)</td>
</tr>
<tr>
<td>4</td>
<td>26</td>
<td>14</td>
<td>12</td>
<td>12.9 (11.8-14.0)</td>
</tr>
<tr>
<td>5</td>
<td>26</td>
<td>14</td>
<td>12</td>
<td>14.2 (13.2-16.0)</td>
</tr>
<tr>
<td>6</td>
<td>29</td>
<td>16</td>
<td>13</td>
<td>15.0 (14.0-16.2)</td>
</tr>
<tr>
<td>7</td>
<td>38</td>
<td>23</td>
<td>15</td>
<td>16.3 (14.8-17.4)</td>
</tr>
<tr>
<td>8</td>
<td>25</td>
<td>10</td>
<td>15</td>
<td>16.8 (16.2-18.1)</td>
</tr>
<tr>
<td>Overall</td>
<td>223</td>
<td>117</td>
<td>106</td>
<td>13.4 (10.8-16.2)</td>
</tr>
</tbody>
</table>

\(^1\) KG = combined kindergarten grades 1-3. IQR interquartile range.
4.3 Summary of the number and duration of contact events

4.3.1 Hourly count and mean contact events per day

The number of contacts events aggregated per hour from 0 to 23 (representing midnight to 11:00 pm, respectively) for each day of the week is shown in Figure 4.2. Each colour represents a different day of the week, ranging from Wednesday 28/09/2016 to Thursday 04/10/2016 of the following week. The dotted lines represent weekends (orange = Saturday, black = Sunday). The faded sections on the graph from 0 (midnight) to 5 (5:00 am) and 18 to 23 (6:00 pm - 11:59 pm) were relatively constant. This suggested that either students had contacts out of school (most probably early morning or early evening before and after school) and that there were participants living in the same households (siblings or relatives such as cousins) who had contacts out of school or placed their sensors together overnight, the latter leading to a continuous exchange of data packets even when the students were not physically present. These data were hence omitted from further analysis.

A total of 422,173 contacts events were recorded, with the majority (355,155 (84%)) of these contact events occurring among students belonging to the same class. The total hourly number of contacts exhibit a circadian rhythm during school days (Monday-Friday). On all weekdays, three peaks are observed at 7:00 am, 11:00 am and 2:00 pm corresponding to students arriving at school in the morning, interactions during mid-morning break from class, and return to school after lunch, respectively. Troughs are observed between noon and 2:00 pm corresponding to lunch break. Contacts reduce drastically from 3:00 pm corresponding to the end of the school session for the day and remain relatively constant from 6pm. Fewer interactions are recorded during the weekend compared to weekdays, with Sunday recording the least number of interactions between students. On Saturday, there are delayed contact activity peaks compared to weekdays and very few interactions are recorded after 2:00 pm on both weekend days.
Figure 4.2. Activity timeline of students from the rural primary school over 7 days. Total hourly number of contacts events are shown from midnight (0) to 11:59 pm (23). Each colored line represents a different day of the week. Dotted lines show weekend days. Faded lines depict discarded data. Friday (21st) and Thursday (27th) were the first and last days, respectively, of data collection.

To compute the mean contact events per hour, the total number of contact events per hour (shown in Figure 4.2) were divided by the number of sensor-pairs in proximity at each hour. Figure 4.3 shows that a distinct peak activity on all days was observed at 5:00 am particularly over both weekend days, followed by a steep drop in contact events between 5:00 am and 6:00 am. The description of the axes for Figure 4.3 follows that of Figure 4.2.
Figure 4.3. Daily per capita activity timeline for students in the rural primary school. Average hourly number of contact events normalized by the number of unique sensors in proximity in each hour. Only data between 05:00 am and 06:00 pm are shown for each of the 7 days. Each colour represents a different day of the week, with dotted lined representing weekends.

To investigate the reason for the spike in mean contact events at 5:00 am, the mean number of contact events and number of nodes in proximity at each hour were graphed as shown in Figure 4.4. and Table 4.2 shows the daily summary of total sensors that were recorded. These two show both daily and hourly fluctuations in the number of detected nodes. A general drop is observed from the first day to the last. Over the weekend, a number of nodes were not detected, and this could have been due to students not wearing the sensors, or students not coming into contact with another student with a sensor. On Monday and Tuesday, a section of the students was sent home to collect school fees, including some of the participants. Between 05:00 and 05:59 am of the first day of the study, there were about 10 sensors in proximity having an average of 34 contacts. On the rest of the days, the number of sensors in proximity within the same time interval ranged from 18-36, with higher mean contact events recorded over the weekend days compared to weekdays. This suggests that a small fraction of students share a household and store their tags together overnight. It will be counterintuitive to suggest that the students leave the sensors in school, since students report to school from 7:00 am as observed by the peaks that occurred only on the weekdays. On
Saturday and Sunday morning, few sensors (27 and 18, respectively) were in proximity but recorded the highest number of contact events. Considering that this happened at 5:00 am on a non-school day, it is suggestive of sensors either being stored close together till early morning, or that students living in the same household wake up earlier than during the weekdays and e.g. do communal chores at home. Weekend evenings displayed an increase in the number of nodes in proximity but a decline in the mean number of contact events. This is suggestive of short or casual interactions between individuals. From Figure 4.4 and, it is evident that the number of nodes fluctuates during the day and by day of the week.

Figure 4.4. Daily activity timeline of number of nodes and mean contact events in the rural primary school. The primary axis (red) shows the mean number of contact events per hour. The secondary axis (black) shows the number of nodes in proximity per hour. In the graph, time is highlighted by digits 5 and 18 representing 05:00 am and 06:00 pm, respectively, and a vertical grey line showing midday of each day. The grey vertical bars represent night time (19:00:00 to 04:59:59 hrs) of each day when data were discarded. There was a drop in the number of nodes on Monday and Tuesday as some students were sent home to collect school fees.

4.3.2 Distribution of degree, number and duration of contact events

Panels (A) and (C) of Figure 4.5 show the distribution of overall degree aggregated over the 7 days and weight (number of contacts, log-log scale), respectively, while panels (B) and (D) show the degree and weight (log-log scale) distribution by day of week, respectively. The overall median degree over the 7-day period (panel (A) dotted line) was 85.0 (IQR 66.5-104.0, CV²=0.08). Daily values ranged from 25.0 (Tuesday) to 43.0 (Friday) during the week, with dramatic decline over the weekend (median 5.0
[IQR 2.0-15.0] and 3.0 [IQR 1.0-6.0] for Saturday and Sunday, respectively). The decline in median degree over the weekend was due to official school closure; however, students in grades 7-8 attend extra classes on Saturday thus partly explaining the high contact events in the morning to midday hours and the broad degree IQR.

The broadest degree distributions were observed on Thursday and Friday (panel B), but there was no major difference observed in the shape of the distribution over all week days. Overall, there was no evidence of difference in the median weekly degree values by gender (female = 89.0 [IQR 70.0, 101.0], male = 83.5 [63.2, 105.0], Wilcoxon Rank Sum test p=0.47). Median degree value was smallest in KG students (63.0), increased with grade to peak at grade 4 (104.0), then declined with increasing grade to 77.6 at grade 8.

Figure 4.5. Statistics of the contact network showing the distribution of degree and number of contact events measured in the rural primary school. Panel (A) shows the probability degree distribution $P(K)$ of the contact network aggregated over 7 days. The black dotted line indicates the median degree, $<K> = 85.0$ ($CV^2 = 0.09$). Panel (B) shows the probability degree distribution for each day of the study. Panels (C) and
(D) show the log-log probability distribution of the overall weight $w_{ij}$ and the daily probability distribution of the weights of the contact network. The weight $w_{ij}$ of an edge $i \rightarrow j$ represents the total number of contact events between $i$ and $j$.

The probability density distribution of person-to-person contact durations is shown in **Figure 4.6** Panel (A) shows distribution for all students over the entire study period, and stratified by grade (B), gender (C) and day of the week (D). The overall daily mean contact duration between different grades was 3.2 minutes ($CV^2=9.1$). The majority of contact events (62%) were of short duration lasting less than 5 minutes. These would represent casual contacts such as brief interactions during lessons within students of the same grade, or interactions between students of different classes or grades. The distribution of contact duration showed a similar trend across all grades, with longer mean contact durations experienced by KG students (8.1 mins) and class 1 students (8.4 mins) and the shortest by grade 8 (5.7 mins). The longest mean duration of interaction was 18.7 minutes for grade 4. It was also interesting to note the longer duration of interaction on Saturday compared to all the other days.
Figure 4.6. Log-log probability density distribution ($P(dt)$) of contact durations ($dt$) measured over the experimental period (7 consecutive days) in the rural primary school. Panel (A) shows the overall distribution of contact event durations in minutes, highlighting interactions lasting 5 minutes, 30 minutes, 1 hour and 5 hours. The other panels show the probability distribution of contact duration by grade (B), gender (C) and day of the week (D).

A summary of the properties of the networks is provided in Table 4.2. This table shows the number of nodes and median degree (i.e., contacted individuals), number of edges (links between students) and network density, the clustering coefficient, and the mean contact duration in minutes by different covariates.
Table 4.2. Summary of statistical properties of the network in rural primary school. Number of nodes and edges, median degree, network density, clustering coefficient and mean contact duration in minutes by location, grade and day.

<table>
<thead>
<tr>
<th></th>
<th>Nodes</th>
<th>Median degree (IQR)</th>
<th>Edges</th>
<th>Density</th>
<th>Clustering coefficient</th>
<th>Mean contact duration (mins)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>223</td>
<td>85.0 (66.5, 104.0)</td>
<td>9,564</td>
<td>0.386</td>
<td>0.532</td>
<td>22.7</td>
</tr>
<tr>
<td><strong>Grade</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KG</td>
<td>21</td>
<td>83.5 (66.8, 101.0)</td>
<td>1,032</td>
<td>0.059</td>
<td>0.716</td>
<td>12.5</td>
</tr>
<tr>
<td>1</td>
<td>16</td>
<td>82.0 (69.0, 93.0)</td>
<td>1,068</td>
<td>0.053</td>
<td>0.792</td>
<td>12.3</td>
</tr>
<tr>
<td>2</td>
<td>23</td>
<td>89.0 (68.5, 112.0)</td>
<td>1,368</td>
<td>0.067</td>
<td>0.744</td>
<td>12.3</td>
</tr>
<tr>
<td>3</td>
<td>19</td>
<td>104.0 (93.5, 127.0)</td>
<td>1,288</td>
<td>0.059</td>
<td>0.845</td>
<td>17.8</td>
</tr>
<tr>
<td>4</td>
<td>26</td>
<td>96.0 (72.2, 105.0)</td>
<td>1,995</td>
<td>0.085</td>
<td>0.876</td>
<td>18.7</td>
</tr>
<tr>
<td>5</td>
<td>26</td>
<td>89.0 (72.0, 106.0)</td>
<td>1,711</td>
<td>0.071</td>
<td>0.830</td>
<td>10.2</td>
</tr>
<tr>
<td>6</td>
<td>29</td>
<td>82.5 (73.0, 103.0)</td>
<td>1,725</td>
<td>0.079</td>
<td>0.826</td>
<td>5.7</td>
</tr>
<tr>
<td>7</td>
<td>38</td>
<td>81.0 (58.0, 93.0)</td>
<td>2,200</td>
<td>0.092</td>
<td>0.812</td>
<td>14.3</td>
</tr>
<tr>
<td>8</td>
<td>25</td>
<td>63.0 (55.0, 78.0)</td>
<td>1,318</td>
<td>0.062</td>
<td>0.786</td>
<td>10.3</td>
</tr>
<tr>
<td><strong>Day</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wed</td>
<td>221</td>
<td>31.0 (25.0, 40.0)</td>
<td>3,568</td>
<td>0.147</td>
<td>0.410</td>
<td>3.3</td>
</tr>
<tr>
<td>Thur</td>
<td>220</td>
<td>38.5 (31.0, 49.3)</td>
<td>4,424</td>
<td>0.184</td>
<td>0.449</td>
<td>2.3</td>
</tr>
<tr>
<td>Fri</td>
<td>217</td>
<td>43.0 (30.0, 53.0)</td>
<td>4,415</td>
<td>0.188</td>
<td>0.424</td>
<td>1.7</td>
</tr>
<tr>
<td>Sat</td>
<td>178</td>
<td>5.0 (2.0, 15.0)</td>
<td>739</td>
<td>0.047</td>
<td>0.400</td>
<td>4.7</td>
</tr>
<tr>
<td>Sun</td>
<td>166</td>
<td>3.0 (1.0, 6.0)</td>
<td>411</td>
<td>0.030</td>
<td>0.349</td>
<td>2.8</td>
</tr>
<tr>
<td>Mon</td>
<td>212</td>
<td>30.0 (20.0, 39.0)</td>
<td>3,119</td>
<td>0.139</td>
<td>0.393</td>
<td>2.0</td>
</tr>
<tr>
<td>Tue</td>
<td>211</td>
<td>25.0 (16.0, 33.0)</td>
<td>2,629</td>
<td>0.119</td>
<td>0.411</td>
<td>3.0</td>
</tr>
<tr>
<td><strong>Location</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within</td>
<td>223</td>
<td>24.0 (20.0, 26.0)</td>
<td>2,693</td>
<td>0.109</td>
<td>0.964</td>
<td>45.7</td>
</tr>
<tr>
<td>Between</td>
<td>223</td>
<td>61.0 (42.0, 79.0)</td>
<td>6,871</td>
<td>0.278</td>
<td>0.362</td>
<td>5.5</td>
</tr>
</tbody>
</table>

4.4 Contact matrices of number and duration of contacts by grade

The matrices showing the total (A and B) and mean (C and D) number of contact events and the total duration of contacts (in minutes) occurring between grades over the 7 days are shown in Figure 4.7. Panels (A) and (B) show that the total number and duration of contacts are symmetric, with higher number and longer duration of contacts recorded among people of the same grade. Panels (C) and (D) suggests that both mean contact event numbers and duration show the same feature of progressive decline in number/duration as the grade gap increases, but do not appear to vary in a systematic way by grade. The older a student gets by grade, the less time they spend with students.
in lower grades. For instance, grade 8 students rarely interact with kindergarten students.

Figure 4.7. Contact matrix showing the number and duration of contact events. This shows the total number (A) and duration (B) of contact events, respectively, that students of grade $i$ (column index) had with students of grade $j$ (row index) over 7 days. Panel (C) and (D) show the daily mean number and duration of contact events, respectively. Labels on the $x$ and $y$ axes report the grade of the individuals. Durations are reported in minutes.

4.5 Graphs of the contact network

Graphs were generated to depict the networks of the school: the main one showing the entire school interactions over all 7 days of the study (Figure 4.8), and further stratified by day of the week (Figure 4.9). The overall network reveals that the graph is not fully connected (a fully connected network would have had a density of 1 meaning that each node was connected to all other nodes). Clustering by grade is observed as similar
coloured nodes are grouped together, particularly grades 4-8, and as highlighted by high clustering coefficients in Table 4.2. Of note, grades 7 and 8 are more isolated from the full network and there are very few direct links to lower grades (KG to grade 3).

When stratified by the day of the week (Figure 4.9), weekdays show different topologies when compared to weekends. From Wednesday to Friday (day 1 to 3), grade-related clustering was also observed, but grade 8 students were consistently disjointed from the network. On Saturday and Sunday (day 4-5), the number of nodes and links reduces considerably (Table 4.2) as fewer students come into contact, resulting in several isolated nodes. The structure also reveals more inter-grade interactions as students meet in the community outside the school. Following the weekend, clustering by grade is once again observed in panels 6 and 7 (Monday and Tuesday), with a few unlinked nodes observed in the periphery of the graph resulting in an artificial dense structure of the main graph. On these two days, a reasonable explanation would be that the students appearing in the periphery were absent from school due to non-payment of school fees.
Figure 4.8. Full network of aggregated contacts from the rural primary school. Each node represents a student (N=223). Each unweighted link represents the presence of at least one contact event between connected students aggregated over 7 days. Nodes are colour coded to represent the 9 grades (KG – grade 8). The graph is generated by the DrL force-directed graph generator provided in igraph (1.2.1) package available in R (3.4.3). The DrL algorithm emphasizes the presence of dense clusters as noticed in the interactions between students belonging to the same grade.
Figure 4.9. Daily networks of students from the rural primary school. Graphs were generated for each day of the week from Wednesday to Tuesday of the following week. The description of the nodes associated colours is similar to that of Figure 4.8. Here, the links represent at least one contact event aggregated per day. The $DrL$ emphasizes the presence of dense graphs observed by grade but masks the same-grade links due to the overlapping of the nodes.
Having observed the marked differences in activity timelines (Figure 4.3) and network topology (Figure 4.9) between weekdays and weekends, activity timelines and matrices of duration of contact for each day of the week were generated. In Figure 4.10 the distribution of contact events and matrices of contact event duration by grade for Friday are shown representing other weekdays (panel A). Panels (B) and (C) show the same distributions for Saturday and Sunday. Contact event timelines for KG and grade 8 are highlighted in the graph to show the variation between the youngest and oldest grades, respectively. Contact events generally reduce over the weekend compared to weekday and lowest contact events are recorded on Sunday, with different activity levels displayed by students of the various grades. Grade 8 students tend to have lower activity levels particularly during normal class sessions. This can be explained by the classroom orientation: KG students normally sit in round tables facing each other thus providing opportunity for a higher number of interactions that are longer, while all grades from 1 to 8 sit three per desk with all facing forward. Students in grades 7-8 attend school on Saturday leading to a higher number and duration of contact events compared to KG students from morning to 1:00 pm (panel (B)). Also, on Saturday, more mixing between students of different grades was registered, suggesting that the propensity to meet students of other grades out of school is higher over the weekend compared to weekday. Interestingly, in the afternoon of on both week end days, very few contact events are reported.
Figure 4.10. Daily contact activity patterns by grade and selected days of the week. Only data for Friday-Sunday are shown in rows (A)-(C). The first column show the daily contact event timelines by grade, and the second column shows matrices of the contact event duration. Kindergarten (KG) and grade 8 activity timelines are highlighted, representing the variation in contact events in the youngest and oldest grades in the school.

4.6 Assessing the similarity of neighbourhood of the nodes

To assess if each student $i$ was linked to the same students in each daily pair of networks (e.g. day 1-day 2, day 2-day 3, etc) and spent an equal fraction of time in their neighbourhood, the cosine similarity was calculated for each grade and is shown in Figure 4.11. The overall median cosine similarity was 0.180, and ranged from 0.070 (grade 8) to 0.271 (grade 2), shown in panel (A). The mean network density varied...
between 0.053 (grade 2) and 0.092 (grade 7). The values of the cosine similarity closer to zero suggest that individual stability of contact patterns was very low, and that the durations of time spent in proximity with different students was not similar from day to day. In order to assess the magnitude of the cosine similarities, these values were compared to a null model (panel (B)). The null model did not change the topology of the network but reshuffled the weights of the networks among the edges. The overall median cosine similarity value on the null model was 0.016, and corresponding median values by grade were also smaller compared to the original network and varied between 0.007 (grade 8) and 0.028 (grade 2).

**Figure 4.11. The distribution of cosine similarity in the rural school.** Panel (A) shows the normal cosine similarity, while panel (B) shows the cosine similarity in which the weights of the network are reshuffled among the edges (WR). The cosine similarity was calculated for each node of the full contact network (overall) and for each grade (KG, 1-8). Each distribution measured the cosine similarity of each node’s neighbourhood, for each pair of days of data collection. For each boxplot, the horizontal bar in the middle shows the median, the grey box shows the lower and upper bounds of the interquartile range (IQR), and the whiskers show 1.5 x IQR of the distributions.
4.7 Discussion

This chapter has reported the results of proximity data arising from interactions between students in a rural primary school in Kilifi, coastal Kenya. The study aimed to recruit 350 students (out of 907) in Matsangoni Primary School, the sample size being limited by the number of sensors available for the study. Previous studies have either recruited the entire school (students and staff) [34,35,115] or a section of the school [33,116]. Students (and staff) were given wireless proximity sensors to detect and record contact events over 7 continuous days. Data from teachers and other staff were discarded from the analysis. The analysis revealed details of networks between students and provided key information on the structure and heterogeneity of contacts between students in different grades and the nature of repeat contacts over several days. This thesis reports the first study in SSA that used wireless proximity sensors to collect data on social networks of students in a primary school. Data were collected from a selection of students over one week, including both week end days. The sensors in this study enabled the collection of data even from the youngest school children without direct supervision and over several days, compared to a previous paper diary study over one day that required young students to have two shadows, one at home and at school, to record their contacts [23].

4.7.1 Study design

Wireless proximity sensors provide an alternative to paper diaries for collecting proximity data from groups of individuals particularly in closed settings. The sensors used ultra-low power radio frequencies to automatically detect and record another sensor facing it that was within a pre-defined threshold. When distributed and worn on the chest area by a group of individuals, a “contact event” occurred when the sensors sustained an exchange of data packets (containing the sensor ID, timestamp, transmission power) for 20 seconds without a 20-second break. This definition of a contact is different from that used in conventional paper diaries whereby the latter defined a contact as a conversation or direct touch between two co-located individuals. Similar sensors based on the SocioPatterns project have been used in schools in Europe [33,34,116] to understand the dynamic properties of contacts between students and
model the impact of intervention measures such as school or grade closure against transmission of respiratory infections such as influenza [50]. From these studies, some further important questions were raised, particularly on the study design e.g. increasing the diversity of geographic contexts from which data are collected, understanding the role of selection of participants, the duration of collection of contact data, and the analysis and presentation of results from complex data sets [17].

Sensor based studies in schools have had various study designs, either recruiting the entire school [34,35,115] or highlighting contact patterns in specific groups of students over one or several days [176], or repeated in different years [33]. These studies managed to recruit >80% of the school population (including staff), but none discussed any challenges encountered during participant recruitment and data collection, and what measures (if any) were taken to generate interest and ensure high participation and retention rates. Only one study reported on sensor design issues that addressed the technical challenges experienced during data collection such as node reboots and time-synchronization of the sensors [120]. The sample size for this current study was limited by the number of sensors available for deployment, and the selection of additional students to replace the refusals or those not available was hampered by the time available for the data collection. In the Kenyan primary school calendar, the third term runs between September and November. Students from KG to grade 7 proceed for their holiday break in the first week of November to allow the annual grade 8 students to sit for their national school exams that starts in mid-November for one week. During this time, no visitors are allowed in the schools. There was thus a narrow window-period in which data could be collected. In retrospect, a bigger number of students should have been recruited preferably during term 2 to give sufficient recruitment time that would cater for expected dropouts and lost sensors, damaged sensors leading to unreadable/corrupt data, and discarding data due to inconsistencies. Nevertheless, a compromise has to be made between the number of expected recruits and duration of the study to minimize the logistical challenges discussed and the possibility of students requiring a refresher training on study procedures before issuing the sensors to them. Other studies have reported minimal data loss probably due to mishandling of sensors used, or censoring of time to eliminate suspected spurious data [44,46].
Following community recommendations before the start of the present studies, students were given sensors encased in polythene zip-lock bags to prevent direct contact with water. The sensors were then inserted in a cloth pouch of similar colour to the school uniform to minimize unnecessary attention from fellow students and other parties. The cloth pouch was then sealed to prevent physical tampering and was equipped with a fastener to attach it to the front of a student’s shirt/blouse. During collection, there was no evidence of deliberate attempts to access the sensor in the pouches. However, in a few sensors that were pinned to the school shirts, it was noted that the polythene cover had been pierced severally by the pin fastener, explaining why some tags had evidence of contact with water. Future studies in schools may consider alternatives to polythene sheathing, such as customized covers made using appropriate cheap technology e.g. 3-D printed sheaths, or reusable silicon covers, that may be more durable and weather-proof compared to the polythene and cloth pouches. In all the other studies reported, the sensors were inserted in pouches similar to conference badges and had minimal reports of tampering. A heat-sealed conference badge could also be a potential solution to reduce costs considerably.

4.7.2 Network properties

An assessment of the degree, number and duration of contacts within this rural primary school setting revealed 3 key things. The first is that in general, students mixed preferentially with students within their grade. This was as expected because students within the same grade shared a classroom and had the same school schedule. There was a decreasing trend in number of contact events and duration as grade difference increased (off-diagonal). This was also revealed by the network graphs that exhibited grade-specific clusters, with very few links between the older (grade 6-8) and younger (KG-grade 3) students. This hierarchical contact mixing structure was also observed in a primary school in France [34] with students highly likely to have repeat contacts with the same students from one day to the next. In the current school study, the cosine similarity values between day pairs were low, suggesting that the proportion of repeat contacts was low.

The second key observation is that the median degree was higher during the weekdays (33.5 different individuals per day) compared to weekend (3.5). Daily weekday median
degree ranged from 25.0 (Tuesday) to 43.0 (Friday) different individuals per day, and 3.0 (Sunday) to 5.0 (Saturday) over the weekend. The higher median degree reported on Friday could be due to the congregation of all students at the general assembly in the morning. A similar high degree would have been expected on the following Monday due to the morning assembly. However, the degree reported for Monday was lower because some students were requested to go back home to collect fees owed to the school. Students going home for this reason happens a week or two before the school-based exams are due, thus this disruption is not common. The very low degree values reported over the weekend are not unexpected since students were not constrained within the confines of their classrooms and school but were able to meet across the community – hence the dilution effect. In addition, students in grade 6-8 attend classes for half a day over the weekend, explaining the high number of contact events and duration for these grades (Figure 4.10 panel (B)). If we assume weekends to be a short holiday period between school attendance, then a similar effect of reduced contacts over the weekend compared to weekdays was demonstrated in a study investigating contact patterns during school term and holiday period in the UK [62]. The number of unique people met reduced by half during the holidays compared to school term. Contact events were assortative by grade during the term and students interacted with older people over the holidays, particularly those over 50 years. In France, a primary school study revealed lower mean degree values than reported in this study, also with relatively small differences between the two days of the study [34]. However, no data were available over the weekends, but it is expected that the number of distinct individuals met would fall considerably. Finally, majority (6 out of 8) of the countries in the POLYMOD study revealed a 12 - 26% reduction in the basic reproduction ($R_0$) number when comparing weekdays to weekend [53]. This reduction in $R_0$ suggests decreased infections within schools, coupled with an increase in infections in older adults due to family-based mixing at home.

The third key observation from this work was that overall daily number and duration of contact events vary by grade as shown in the contact matrices presented in Figure 4.7 and further by day of the week as shown in Figure 4.10. An illustration is given for the youngest (KG) and oldest grade (8), whereby KG students exhibit higher activity patterns because their writing spaces are arranged such that they face each other during
Kindergarten students (and class one, not shown) do not attend classes in the afternoon, thus explaining the drop in their activity from midday. Older students normally sit two-to-three per study bench, and face forwards towards the teacher. This explains why their activity timelines drop after they enter their classes and become seated at 8 am and rise again during breaks when they leave their desks and are free to interact face-to-face at 11 am and over lunch time, also giving credibility to the sensors on the detection of face-to-face contacts. Opposite to what was observed with the KG students, all students from grade 2-8 return to school in the afternoon, and grade 8 students stay in school until 5:00 pm. On Saturday, grade 6-8 students attend extra classes at school for half a day explaining the dense clusters that are formed in the weekend networks for these grades. These patterns depict the normal school day for public (state-run) schools in Kenya. In addition to highlighting the activity patterns in various grades, these results also suggest that the majority of school children had a high level of adherence to carrying the sensors throughout the week, with data discrepancies brought about by sensors coming into contact with water or firmware malfunctions that led to multiple rebooting of the sensors.

4.7.3 Limitations

There are a few caveats that have to be put in perspective for the interpretation of the results and generalization. The sensors only recorded contact events between students wearing them. However, the sensors were worn by the students continuously from the time they woke up in the morning to when they went to sleep in the evening. There was no way of physically verifying this by observing students during the study period, but data from the analyzed sensors revealed realistic activity patterns that can be attributed to the school schedules. Further, the expected sample size was limited by the number of sensors available for deployment. Even though data were available for a quarter of the entire school population, the study has revealed important temporal patterns of contact events in a rural primary school in Kenya that can be use in mathematical models to understand dynamics of infections spread via close contact and potential ways of mitigating transmission.

The data provided by the sensors were a record of proximity events between students rather than an actual physical contact. This suggests that the data may only be
appropriate for studying respiratory infections transmitted through secretions from coughing or sneezing in the physical proximity of another individual. However, the advantage of this data is that one can vary the proximity threshold to mimic very close interactions (but not skin-to-skin contacts) to interactions between individuals standing a few metres apart. This makes it highly customizable for a range of respiratory infections such as RSV (requiring inoculation with large droplets that do not travel far following sneezing) or influenza that can be transmitted via aerosols.

Generalization of these results to other schools within rural Kilifi is possible but should be viewed with caution for rural schools outside Kilifi. For public primary schools in Kenya, the school schedule is the same nationally, and the temporal patterns would be expected to be similar in different regions. However, there are private boarding schools (single or mixed gender) where the students stay in school for the entire term. For these boarding schools, transmission events would be confined to the school population with teachers living out of the school acting as potential bridges to either import or export infections. To verify this assertion, studies targeting schools with different schedules, spatial and classroom compositions should be considered in future.

4.7.4 Summary

This chapter presents a unique dataset from a rural primary school in SSA that increases our understanding of mixing patterns in schools. The results are suggestive of homogeneous mixing within students of the same grades, and non-negligible mixing between different grades. Data collected over the weekend revealed preferential mixing among older students due to school attendance on Saturday and high mobility, albeit at lower levels that during the week. The statistical properties are similar to those observed in earlier school studies, suggestive of a few individuals with many and long-duration contacts who can be considered super-spreaders in case of an infectious disease outbreak. These data can be used to assess different micro-simulation transmission models (e.g. comparing school vs grade closure) and suggest the most efficient and effective disease prevention strategies.
5 Results from rural households

5.1 Introduction

In the previous chapter, we reported on the results of a rural primary school. In this chapter, we move to the household setting of a subset of children from the rural school. Ideally, investigating the relationship between the school and household mixing patterns would have required simultaneous deployment of the sensors at both settings. Such data would have, to some extent, enabled the exploration of contacts between students attending the rural school and their residents of households selected to participate in the study. The simultaneous data collection, or alternatively collection of data from a higher number of households, was hampered by the number of sensors available for the study and logistical constraints as previously discussed in Chapter 4. By linking households to students attending a primary school in a rural area, this chapter reports on the use of an innovative study design to collect data on social proximity patterns that are important to understand transmission of respiratory infectious diseases at the household setting.

The specific objectives of the study in rural households were:

(i) To collect social contact data from a subset of household residents linked to students from a primary school in Matsangoni location.
(ii) To quantify the degree, number and duration of contact events based on intra- and inter-household mixing, stratified by age, household size and day of the week.

(iii) To assess the dynamic variation of the intra- and inter-household networks by time of day and day of the week.

5.2 Baseline characteristics of participants from rural households

From the 303 students in the rural school who carried a sensor, nine index students were randomly selected. The number of index students selected was increased by 3, from 9 to 12, to cater for potential non-response and refusal to consent by household members. The unique identifier for each student, as assigned in the KHDSS register, was used to match the student to his/her household (known as the index household) and identify the members of his/her household. Each index household was further linked to 3-5 neighbouring households (defined in Table 3.1) to form a cluster. At the end of the follow up, 9 household clusters were available to participate in the study (Figure 5.1). Of the 3 that were unavailable for consent, 2 had been bereaved hence we could not approach them to get consent, and in the third household the parents to the student were not living together and the mother could not give full consent without the father’s assent.

Following extensive community engagement sessions, data were collected between 18th and 29th October 2016. Data for the first three days and last two days were excluded from the analysis. Some household residents were not present to take the sensors on the first two days, and some sensors were collected up to two days after the proposed end date of the study period (27th). In this analysis, data included are from 21st to 27th, a total of 7 continuous days. Out of 328 residents listed in the KHDSS register, 292 (66 index) members were present and received tags for data collection (summary shown in Figure 5.2).
Figure 5.1. Location of index-student households’ and school in rural setting. The index students were randomly selected from the list of participating students reported in Chapter 4. The household clusters are labelled 1-9, and the inset table shows the number of households (HH) and residents (Res) in each cluster. The lines show the main roads passing through the study area. The map also shows the location of the rural school (SR) and the main public health centre in Matsangoni location.

Of the 292 tags issued, data were available for 180 (62%, 111 female) of the residents living in 40 households as summarised in Table 5.1. The median age was 14.5 years (IQR 9.1-40.1) with one infant (0.8 months) and the oldest individual aged 92.5 years. Household size ranged from 1 to 10 members. The gender distribution was similar in all age groups apart from age 0-4 years (twice as many females than males) and 20-49 years (five times as many females as males). Residents were listed as not present if they were not available for the entire duration of data collection (7 days). Sensors were listed as corrupt if it was impossible to extract data from them, whether due to software malfunction or direct physical damage to the sensors (such as contact with moisture). Sensors miss attribute data (such as age, gender, household residence, etc) if they cannot be linked to a file with demographic metadata, suggesting the possibility of signals being picked from other tags in proximity but not participating in the study e.g.
during issuing of tags. Data collection was considered inconsistent if there were less than 6 consecutive night periods, either due to no data collection or continuous data collection if tags are placed in close proximity throughout the study.

**Figure 5.2. Data collection flow chart for rural households.** This presents the number of residents located, sensors issued, reasons for loss to follow-up/discarding of data, and number of sensor-data analyzed. Attribute data include age, gender and household identifier.
Table 5.1. **Baseline characteristics of rural households.** This table shows the overall median degree and stratified by gender, age, day of the week, household size and location.

<table>
<thead>
<tr>
<th></th>
<th>Number of nodes</th>
<th>Median Degree (IQR)</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Overall</strong></td>
<td>180</td>
<td>11.0 (7.0, 18.0)</td>
<td>-</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>111</td>
<td>10.0 (6.5, 15.0)</td>
<td>0.0181</td>
</tr>
<tr>
<td>Male</td>
<td>69</td>
<td>13.0 (8.0-22.0)</td>
<td></td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-4</td>
<td>18</td>
<td>8.5 (6.3-14.0)</td>
<td>&lt;0.0012</td>
</tr>
<tr>
<td>5-14</td>
<td>69</td>
<td>19.0 (12.0-27.0)</td>
<td></td>
</tr>
<tr>
<td>15-19</td>
<td>27</td>
<td>13.0 (10.5-19.5)</td>
<td></td>
</tr>
<tr>
<td>20-49</td>
<td>31</td>
<td>8.0 (6.5-12.0)</td>
<td></td>
</tr>
<tr>
<td>50+</td>
<td>35</td>
<td>6.0 (5.0-10.5)</td>
<td></td>
</tr>
<tr>
<td><strong>Day of the week</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Friday</td>
<td>172</td>
<td>6.0 (4.0, 9.3)</td>
<td>0.0052</td>
</tr>
<tr>
<td>Saturday</td>
<td>163</td>
<td>5.0 (3.0, 8.0)</td>
<td></td>
</tr>
<tr>
<td>Sunday</td>
<td>158</td>
<td>5.0 (3.0, 7.0)</td>
<td></td>
</tr>
<tr>
<td>Monday</td>
<td>163</td>
<td>5.0 (3.0, 7.5)</td>
<td></td>
</tr>
<tr>
<td>Tuesday</td>
<td>154</td>
<td>5.0 (3.0, 7.0)</td>
<td></td>
</tr>
<tr>
<td>Wednesday</td>
<td>143</td>
<td>5.0 (2.0, 7.0)</td>
<td></td>
</tr>
<tr>
<td>Thursday</td>
<td>146</td>
<td>5.0 (5.0, 8.8)</td>
<td></td>
</tr>
<tr>
<td><strong>Household size</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>6</td>
<td>10.5 (1.5-25.5)</td>
<td>0.4112</td>
</tr>
<tr>
<td>2-4</td>
<td>51</td>
<td>11.0 (7.5-15.5)</td>
<td></td>
</tr>
<tr>
<td>5-6</td>
<td>47</td>
<td>9.0 (6.5-18.5)</td>
<td></td>
</tr>
<tr>
<td>7-9</td>
<td>46</td>
<td>12.0 (9.0-20.0)</td>
<td></td>
</tr>
<tr>
<td>10-11</td>
<td>30</td>
<td>11.5 (8.0-17.5)</td>
<td></td>
</tr>
<tr>
<td><strong>Location</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within</td>
<td>173</td>
<td>5.0 (3.0-6.0)</td>
<td>&lt;0.0011</td>
</tr>
<tr>
<td>Between</td>
<td>163</td>
<td>8.0 (4.0-14.5)</td>
<td></td>
</tr>
</tbody>
</table>

1 = Wilcoxon Rank Sum test; 2 = Kruskal-Wallis Rank Sum test

### 5.3 The distribution of number and duration of contact events

#### 5.3.1 Hourly distribution of total number and mean contact events per day

**Figure 5.3** shows the hourly aggregated total number of contacts from midnight (0) to 11 pm (23) for each day of the week from Friday 21\(^{\text{st}}\) to Thursday 27\(^{\text{th}}\) October 2016. Each colour represents a different day of the week, and weekend days are represented by dotted lines. Throughout the week, a consistent circadian activity was displayed, with a sharp rise in total contacts that started from 03:00 am to peak at 05:00 am. During the weekdays, the total number of contact events decline sharply to reach lowest
number between 08:00 and 09:00 am, with higher number of contact events recorded up to 03:00 pm during the weekend. Contacts remain relatively constant or rise gradually to peak again between noon and 02:00 pm. Apart from Thursday, contacts decline markedly between 3:00 pm and 5:00 pm and reached the lowest levels in the evening to night time. The highest number of contact events were recorded on the early morning of day 1 (Friday 21st) and evening of day 7 (Thursday 27th). Reduced number of contacts are recorded on Sunday compared to Saturday between 9:00 am and 3:00 pm, and then both trajectories follow a dipping pattern as for the rest of the days of the week.

![Activity timeline of rural households over 7 days](image)

**Figure 5.3. Activity timeline of rural households over 7 days.** Total hourly number of contacts are shown from midnight (0) to 11:59 pm (23). Each colored line represents a different day of the week. Dotted lines show weekend days. Faded lines depict discarded data. Friday (21st) and Thursday (27th) were the first and last days, respectively, of data collection.

Between 6:00 pm and 2:00 am of each day, there are low numbers of contact events reported apart from Friday (day 1) and Thursday (day 7). The low number of contact events in the evening and night are as expected and is suggestive of few sensors in contact, with a few recording contact events throughout the night. On each day between 3:00 am and 5:00 am (Figure 5.3), a rise in the number of contact events was observed, reasonably reflecting the start of the day. To conform to the analysis in the rural school,
only daily contact events recorded between 05:00 and 18:59 hours were considered in further analysis.

A total of 171,522 contact events were recorded in 7 days. Of these, 129,101 (75%) contacts were recorded between members of the same household, and 43,421 contact events between members of different households. Figure 5.4 shows the daily mean number of contact events (total number of hourly contact events divided by number of sensors in proximity per hour) between 05:00 am and 07:00 pm. This figure demonstrates the very high number of contacts recorded at 05:00 am and the steep drop up to around 08:00 am, coupled with an increase in the number of nodes in proximity shown in Figure 5.5. In addition, Figure 5.5 suggests that there is a decline in the number of nodes in proximity very early in the morning (05:00 am) from 80 in day 1 to 26 in the last day. In addition to suggesting normal morning activities at the household, this decline could also suggest a tendency of participants to store sensors separately in the course of the study, as anticipated. During the rest of the day, the average number of contacts remains relatively constant and following the same trend daily, with gradual dips observed from 03:00 pm and a rise from 05:00 pm. The number of participants and median degree values by gender, age, household size and location of contacts are summarized in Table 5.1.

Figure 5.4. Per capita daily activity timeline for Matsangoni households. Only data between 5 am and 6 pm are shown for each of the 7 days. Each colour represents a
different day of the week. Contact events have been normalized by the number of participants per hour per day.

Figure 5.5. Daily activity timeline of mean contact events and number of nodes in the rural households. The primary axis (red) shows the mean number of contact events per hour. The secondary axis (black) shows the number of nodes in proximity per hour. In the graph, time is highlighted by digits 5 and 18 representing 05:00 am and 06:00 pm, respectively, and a vertical grey line showing midday of each day. The grey vertical bars represent night time (19:00:00 to 04:59:59 hrs) of each day when data were discarded.

5.3.2 Distribution of degree, number and duration of contact events

The overall network degree distribution of rural households is shown in Figure 5.6. Figure 5.7 panel (A), while panel (B) shows degree distribution stratified by day of the week. Panel (C) of Figure 5.7 shows distribution of weight (number of contact events), and further stratified by day of the week (D). Overall, the median degree was 11.0 (IQR 7.8-18.0). There is no evidence that the degree distributions vary by day of the week, with degree decreasing marginally over the weekend. There was evidence of difference in median degree values by gender, age and whether the participants shared a household or not (location). A male participant was, on average, connected to slightly more participants than a female participant (13.0 vs 10.0, respectively, Wilcoxon rank sum test p=0.016) despite males being half the number of females. The median degree rose sharply from toddlers (0-4 y, median=8.5) and peaked at school going children (5-14 y, median degree 19.0), and declined to the lowest value of 6.0 for the elderly (>50 y). Overall, the median degree values increased slightly by household size to peak at size 5-6 (13.0) and remained relatively constant after. In general, there was a negative
correlation between degree values and age (-0.43), and minimal effect due to increasing household size (0.16). Figure 5.6 (C) shows that there is a high probability of having very few contacts events; however, a high number of contact events still occur (almost $10^4$) but they are of very low probability ($P(dt) = 10^{-6}$).

Figure 5.6. Distribution of network degree and number of contact events in rural households. (A) Probability degree distribution $P(K)$ of the contact network aggregated over 7 days. The black dotted line indicates the median degree, $<K> = 11.0$ ($CV^2 = 0.4$). (B) Probability degree distribution for each day of the study. (C) Log-log probability distribution of the weights of the aggregated contact network. (D) Log-log daily probability distribution of the weights of the aggregated contact network. The weight $w_{ij}$ of an edge $i - j$ represents the total number of contact events between $i$ and $j$.

Figure 5.7 displays the overall probability distribution of person-to-person contact duration (A), and stratified by age group (B), gender (C) and day of the week (D). The average daily contact duration across all contact events is 8.6 minutes, with 60% of all contacts exceeding 5 minutes (the squared coefficient of variation, $CV^2$, of the full
contact duration distribution is $= 5.7$). The distribution by gender and day of week exhibits similar characteristics per strata, but women spent twice as much time in contact with other participants compared to males (10.9 vs 5.4 minutes, respectively, Kolmogorov-Smirnov test $p=0.01$). There was evidence to suggest a difference in the distribution of contact durations by age. Children aged 5-14 years, who constitute more than half (53%) of the total sample, recorded the longest contacts per person per day lasting 8.3 minutes (Figure 5.7 (B)), whereas adult contacts (20-49 y) were the shortest lasting only 1.7 minutes per day.

![Figure 5.7. Log-log probability density distribution $(P(dt))$ of contact durations $(dt)$ measured over the experimental period (7 consecutive days). Panel (A) shows the full distribution of contact durations, highlighting interactions lasting 5 minutes, 30 minutes, 1 hour and 3 hours. The other panels show distribution of contact duration by age in years (B), gender (C) and day of the week (D).](image-url)
5.4 Contact matrix of number and duration of contact events

Contact matrices were generated as shown in Figure 5.8. These show the total (A and B) and mean (C and D) daily number and duration of contact events by age group, respectively. The mean daily contacts are the total number and duration of contact events by age group scaled by the number of participants per age group and number of study days.

Figure 5.8. Contact matrix showing the number and duration of contact events. This shows the total number (A) and duration (B) of contact events, respectively, that residents of age group $i$ (column index) had with residents of age group $j$ (row index) over 7 days. Panel (C) and (D) show the daily mean number and duration of contact events, respectively. The number of residents per age group in panels (C) and (D) is given in brackets. Labels on the $x$ and $y$ axes report the age group of the individuals. Durations are reported in minutes.
Panels (A) and (B) highlight the symmetric nature of total number and duration of contact events due to the reciprocate nature of contacts as captured by the sensors, which is not observed in paper diary studies. When scaled by the number of participants per age group and number of study days, the matrices become asymmetrical due to the sizes of the age groups. The contact matrices were characterised by an L-shape with offsets at the 5-14 and 20-49-year-old participants. There is little emphasis on assortative mixing by age group, with children aged 5-14 years displaying the highest assortativity with regards to the total number (A) and duration (B) of contacts.

When scaled by the number of participants per age-group, children aged 0-4 years had the highest number of contact events (C) over the longest duration (D) with 5-14-year olds and also spent a considerable amount of time with 20-49 years-olds. This represents time that the very young spent with older school going children (presumably siblings and other relatives) and their caregivers (parents). The elderly (aged >50 years) spent majority of their time with primary school children (5-14).

### 5.5 Contact activity distribution within and between households

Data were stratified according to whether contacts were intra-household (within, i.e. contacts occurred between members of the same household), or inter-household (between members of different households). Figure 5.9 shows the daily distribution of intra- and inter-household contact activity patterns between 05:00 am and 06:59 pm, against a background of total contact events. In the morning hours, the number of contact events within household residents is relatively higher compared to between household residents, whereas during the day and in the evening the between number of contact events increases. On average, an individual spent 16.3 minutes ($CV^2=2.6$) with a member of the same household, compared to 3.5 minutes ($CV^2=13.9$) with a member of a different household (Kolmogorov-Smirnov $p<0.001$). In addition, the majority (82%) of contacts within households lasted more than 5 minutes, compared to 44% between residents of different households. Thus, contact events within households lasted 5 times longer compared to contact events between people of different households.
**Figure 5.9. Intra- and inter-household contact activity patterns in rural households.** Daily contact events occurring only within (intra) members of the same household and exclusively between (inter) members of different households from 5 am to 7 pm. The grey background shows overall daily contacts. The grey vertical lines show 00:00:00 hours of each day.

A summary of the distribution of median degree values for intra- and inter-household interactions is shown in **Table 5.2**. There was strong evidence that the median degree of contacts by individuals living in different households (inter-household contacts) was higher than that of individuals sharing a household (intra-household contacts) (8.0 vs 5.0, respectively, Kruskal-Wallis p<0.01, **Figure 5.10** panel (A)). Differences in median degree were also observed by gender in two ways: first, males had almost two times the number of contacts as women when interacting with individuals not sharing their households (10.0 vs 6.5, respectively, Wilcox rank sum test p=0.006) and second, males interacted with more individuals from other households compared to those within their households as shown in **Figure 5.10** panel (B) (10.0 Vs 5.0, p<0.0001).

Variations were also observed by age group and household size. Whereas the intra-household distribution of median degree by age group remained relatively constant by age (**Table 5.2** and **Figure 5.10** panel (C)), there was a considerable increase in inter-household contacts especially in children aged 5-19 years (school going). In addition, for contacts occurring within households, there was a linear increase (Spearman correlation coefficient = 0.87) in the median degree with increasing household size (**Figure 5.10** panel (D)).
Table 5.2. Summary of within and between household median degree by gender, age and household size.

<table>
<thead>
<tr>
<th></th>
<th>Within</th>
<th></th>
<th>Between</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of nodes</td>
<td>Median degree (25%, 75%)</td>
<td>Number of nodes</td>
<td>Median degree (25%, 75%)</td>
</tr>
<tr>
<td><strong>Gender</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>112</td>
<td>5.0 (3.0, 6.0)</td>
<td>108</td>
<td>6.5 (3.0, 15.0)</td>
</tr>
<tr>
<td>Male</td>
<td>68</td>
<td>5.0 (3.0, 7.0)</td>
<td>61</td>
<td>10.0 (5.0, 23.0)</td>
</tr>
<tr>
<td><strong>Age</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0-4</td>
<td>20</td>
<td>5.0 (4.0, 6.0)</td>
<td>17</td>
<td>5.0 (3.0, 10.0)</td>
</tr>
<tr>
<td>5-14</td>
<td>70</td>
<td>5.0 (4.0, 7.0)</td>
<td>72</td>
<td>15.0 (7.0, 21.0)</td>
</tr>
<tr>
<td>15-19</td>
<td>27</td>
<td>4.0 (3.0, 6.0)</td>
<td>26</td>
<td>11.0 (6.0, 17.5)</td>
</tr>
<tr>
<td>20-49</td>
<td>30</td>
<td>4.0 (3.0, 6.0)</td>
<td>27</td>
<td>4.0 (2.0, 8.0)</td>
</tr>
<tr>
<td>50+</td>
<td>33</td>
<td>4.0 (2.0, 6.0)</td>
<td>27</td>
<td>4.0 (2.0, 7.0)</td>
</tr>
<tr>
<td><strong>Household size</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>-</td>
<td>-</td>
<td>6</td>
<td>12.0 (1.5, 26.2)</td>
</tr>
<tr>
<td>2-4</td>
<td>43</td>
<td>2.0 (1.0, 3.0)</td>
<td>50</td>
<td>8.5 (5.3, 14.8)</td>
</tr>
<tr>
<td>5-6</td>
<td>58</td>
<td>4.0 (4.0, 5.0)</td>
<td>56</td>
<td>8.0 (3.5, 17.0)</td>
</tr>
<tr>
<td>7-9</td>
<td>37</td>
<td>6.0 (6.0, 7.0)</td>
<td>37</td>
<td>6.0 (3.0, 15.0)</td>
</tr>
<tr>
<td>10-11</td>
<td>42</td>
<td>8.0 (6.0, 9.0)</td>
<td>21</td>
<td>11.0 (7.0, 17.0)</td>
</tr>
</tbody>
</table>
Figure 5.10. The distribution of network degree within and between households. Panel (A) shows degree distribution within and between households, whereas (B), (C) and (D) shows distribution by gender, age group and household size, respectively, for within (brown) and between (grey) household.

The distribution of degree (A) and contact events (B) within and between households for the whole study period is shown in Figure 5.11. This figure also shows the mean number of contact events within (C) and between (D) households, and the mean duration of contact events within (E) and between (F) households, respectively, for the entire study period. The number and duration of contacts are mainly driven by individuals living in the same households particularly by school going children aged 5-14 years.
Figure 5.11. Distribution of within and between household network degree, number and duration of contact events. Panel (A) shows the cumulative degree distribution of within and between household contacts, while (B) shows the probability density distribution \( P(dt) \) of contact durations \( dt \) measured over the experimental period (7 consecutive days) within and between households. Panel (C) and (D) show the daily average number of contacts that individuals of age \( i \) (column index) had with individuals of age \( j \) (row-index) within and between households, respectively. Panel (E) and (F) shows the daily average duration of contacts of an individual of age \( i \) with individuals of age \( j \). The labels of the x and y axes represent the age groups of the individuals, and the number in brackets is the number of people per age group. Duration of contacts are reported in seconds.
School-going children (5-14 years) exhibit the highest assortative interactions, exhibiting the highest number and duration of contact events among children of the same ages. These students also exhibit high number of interactions of long durations with younger children and the elderly individuals. Interestingly, children have the least average number of with those aged 20-49 years within households and tend to interact less with elderly individuals not living in the same households.

Having observed a marked difference in contacts by gender within and between households, the contact duration matrices were further stratified by gender as shown in Figure 5.12.

**Figure 5.12.** Within and between household mean contact matrices stratified by gender. Panels (A) and (B) show matrices within household for females (F) and males (M), respectively. Similarly, panel (C) and (D) show matrices for between household contacts for females and males, respectively, between households. Durations are shown in minutes.
In Figure 5.12, panels (A) and (C) show female participants (x-axis) mixing with either gender, while panels (B) and (D) show mixing preferences for male participants. Within households (panel A), female children aged 0-4 y spent the majority of their time with school-going children (67 minutes) and adults (40 minutes). Female residents aged 5-14 years as well as adults spend most of their time with children aged 0-4 years. Across both genders, the majority of the time is spent with primary school going children (5-14). Within households, male toddlers (0-4 y) spent half as much time with school-going children as female toddlers. Notably for both genders, there was very little assortative interaction between toddlers. When considering inter-household contacts, male residents aged 15-49 years had no interactions with other adult residents (panel B).

5.6 Properties of the contact network in rural households

The structure of the fully aggregated network and stratified by within and between households was investigated both statistically and visually using a network diagram. Table 5.3 summarises the number of nodes and links for each network, clustering coefficient, density and distribution of contact duration.

Table 5.3. Summary of properties of network within and between households.

<table>
<thead>
<tr>
<th></th>
<th>Nodes</th>
<th>Edges</th>
<th>Clustering coefficient</th>
<th>Network density</th>
<th>Proximity duration (minutes)</th>
<th>CV²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall</td>
<td>180</td>
<td>1,235</td>
<td>0.558</td>
<td>0.077</td>
<td>8.6</td>
<td>5.7</td>
</tr>
<tr>
<td>Within</td>
<td>173</td>
<td>402</td>
<td>0.834</td>
<td>0.027</td>
<td>15.7</td>
<td>2.7</td>
</tr>
<tr>
<td>Between</td>
<td>163</td>
<td>828</td>
<td>0.248</td>
<td>0.063</td>
<td>3.6</td>
<td>14.5</td>
</tr>
</tbody>
</table>

1 Mean daily duration of interaction (minutes)

While there were 180 individuals over the entire duration study, only 173 and 163 distinct individuals were recorded for intra- and inter-household interactions, respectively. This suggests that there are 7 individuals who did not record contacts within their households, and 17 individuals who did not interact with residents of other households. Overall, the clustering coefficient was 0.556, with the clustering coefficient being about four-times higher within households (0.834) compared to between
households (0.248). The higher clustering of contacts in within-household contacts is as expected due to the restricted environment in this setting. Network density from the full network suggests that only 7% (1,235) of the expected links (if each person was to contact another person once, the fully connected network would have approximately 19,200 edges) were formed. Conversely, the intra-household network density was three times higher than the inter-household network (0.063 vs 0.027, respectively) due to the possibility of interacting with a larger number of individuals outside the household. On average, within household contacts lasted five times as long as between household contacts with higher variability observed in the latter (CV^2=14.5). More than 80% of contacts occurring within the household last for more than 5 minutes compared to 45% for individuals not sharing a household.

The series of figures from Figure 5.13 to Figure 5.15 represents networks of interaction for all participants, intra-household interactions and inter-household interactions, respectively. The position of the nodes is the same in panels (A) and (B), but different in each subsequently numbered figure.

Panel (A) of Figure 5.13 shows that nodes of the same colour clustered highly and had dark-coloured links joining them. There are also links between different households, but these were of a lighter shade. Nodes with single links to another node of similar colour are observed on the periphery of the graph, while different coloured nodes with strong links are drawn towards the middle of the graph (panel (A)). This suggests that individuals of the same household spent more time in contact compared to individuals of different households. When coloured by their age group (panel (B)), this revealed that individuals aged 6-19 years cluster towards the middle of the graph and have more links compared to children and individuals older than 20 years.

Figure 5.14 panel (A) shows nodes grouped together only if they had contacts with members of the same cluster. Some clusters are split into more than one group of different sizes (for example households 3, 4, 6 and 8 have 3, 3, 3, and 2 groups, respectively), representing distinct families (building units) in the cluster. In most groups, an individual has a contact with everyone. Panel (B) shows intra-connected nodes colour coded by age. This showed that households are of different compositions, with most containing both children and adults. In the bigger households, while mixing
seemed heterogenous, nodes of same colour (same age-group) clustered together, in particular age 6-14 years highlighting their propensity to spend time together.

From Figure 5.15, inter-household contacts were driven mainly by individuals living in households in the same cluster. Inter-household contacts are abundant and are driven by school going children as demonstrated by the same-coloured nodes that cluster in the middle.
Figure 5.13. Full networks of aggregated contacts from rural households over 7 days. Each node represents an individual. Each link represents an unweighted interaction between individuals who have at least one contact. Nodes in panels (A) are colour coded to represent the 9 household clusters linked to the index student, and the colours are similar to that on the map in Figure 5.1. Nodes in (B) are colour coded to represent age groups.
Figure 5.14. Intra-household networks of aggregated contacts from rural households over 7 days. The networks represent contacts that occur between individuals living in the same household. Nodes in panels (A) are colour coded to represent the 9 household clusters linked to the index student. Nodes in (B) are colour coded to represent age groups.
Figure 5.15. Inter-household networks of aggregated contacts from rural households over 7 days. The networks represent contacts that occur between individuals living in the different household. Nodes in panels (A) are colour coded to represent the 9 household clusters linked to the index student. Nodes in (B) are colour coded to represent age groups.
5.7 Assessing the day-to-day stability of contacts

Similar to the rural school, the changes in the contacts of each household member were assessed by calculating the cosine similarities between the neighbourhoods of each member in each daily pair of networks. These distributions are shown in Figure 5.16 aggregated for all household clusters (overall) and by each cluster (numbered 1-9) for the original network (panel (A)) and on a network with reshuffled links (panel (B)).

![Figure 5.16. Distribution of cosine similarity in rural household clusters.](image)

Panel (A) shows the cosine similarity of the original network, while panel (B) shows the cosine similarity in which the weights of the network are reshuffled among the edges. The cosine similarity was calculated for each node of the full contact network (overall) and for each household cluster. Each distribution measured the cosine similarity of each node’s neighbourhood, for each pair of days of data collection. For each boxplot, the horizontal bar in the middle shows the median, the grey box shows the lower and upper bounds of the interquartile range (IQR), and the whiskers show $1.5 \times IQR$ of the distributions.

The overall median similarity was 0.723 (IQR 0.71) and ranged from 0.637 to 0.925 per household cluster. Values for the null model were lower (overall median 0.171, IQR
0.661) and also lower for each household cluster when compared to the original network.

## 5.8 Discussion

This chapter reports on mixing patterns in clusters of households from a rural location, Matsangoni, located in coastal Kenya. The index households participating in this study were identified from a subset of 9 students who participated in the school-based study reported in **Chapter 4**. Entire neighbouring households (defined in **Table 3.1**) to the index households were also recruited. The primary aim of this school-household linkage was to assess the nature of interactions of families linked to students from a primary school and within a community of neighbours. Generally, individuals spend most of their time at home [9,12], and the intimate (non-sexual) interactions that occur at home have been linked to a higher propensity to transmit infections to individuals across all ages [26,131,177]. In addition, asymptomatic but infectious individuals requiring no hospitalization may expose other household members and caregivers to possibility of acquiring infection [101]. To date, there have been only two studies using proximity sensors to understand the nature of contacts within households [46] and between individuals living in different households [44]. The former study was conducted in Italian households (N=55) assessing interactions between infants and other household members, while the latter was a pilot study in Kilifi (N=75) to assess the feasibility and acceptability of using sensors to collect data on mixing patterns in households. Prior to this current study, Matsangoni location was one of the sites of a self-report paper diary study that recruited individuals from households [23]. This current study additionally reports on intra- (contact events occurring among members of the same household) and inter-household (contact events occurring between members of different households) interactions that may be relevant understanding the transmission of respiratory infections via close contact.
5.8.1 Recruitment of study participants

This current study in rural households recruited 292 (out of an expected 328) residents drawn from 40 households aggregated into 9 clusters. The 36 individuals unavailable to participate were absent from the households mainly due to temporary outmigration for work and school. This unavailability of participants was expected and has also been reported from a paper survey in the same setting [23]. From the 292 participants, 4 sensors were lost while the rest (108) were either corrupted/ damaged or did not have participant demographic details, the latter being due to inconsistencies in the KHDSS personal data. Data for analysis were thus available for 180 (62%) participants. The methods put in place to encourage participation and minimize physical loss of or damage to sensors, such as extensive community engagement, enclosing sensors in a waterproof bags and camouflaging the pouches in which sensors were inserted, followed recommendations from a pilot study conducted earlier in the same setting [44].

To avoid future loss of data, periodic physical checks can be conducted on the sensors to ensure that they are working as expected. Additional baseline data collection forms should also be provided to study teams to verify and update demographic data from participants particularly those not listed in the KHDSS database.

One unexpected challenge encountered was that two of the households approached to participate were bereaved making it impossible to recruit the households. Social events such as funerals that host big crowds of relatives and friends for a short period are important in the transmission of highly infectious pathogens such as ebola [178]. It would be considered insensitive and unethical to recruit individuals into research activities as they grieve, and methods such as retrospective contact tracing have been shown to be effective in identifying individuals who may be infected before implementing control procedures such as ring vaccination [139]. This further suggests that future studies in different societies should be mindful of the social and cultural aspects of the participants during recruitment to minimize friction between the potential participants and research team.

Lastly, all night time data collected between 19:00 and 04:59 hours were discarded in toto, as was done in two previous household studies [44,46]. Household data revealed unusually high number of contacts during the night that can only be explained by
continuous data collection due to a number of the sensors being stored in proximity. Aggregating the data as shown in Figure 5.3 masks any true individual proximity events and makes it impossible to filter out prolonged contact events that are not representative of actual contact events. Though challenging in rural populations due to high illiteracy levels [23], future studies can consider asking participants to keep a time-diary of when they woke up and went to sleep. These start and end times of data collection can be used to filter out any additional data collected, thus depicting proximity events more accurately as was done in Italian households [46]. This can be done using retrospective questionnaires by research teams to avoid the need of shadows. Despite these losses to follow up, data were available for more than half of the expected number of participants. For the first major study in a rural setting in SSA this was a major achievement that has revealed important information on mixing patterns within and between households.

5.8.2 Network properties

This study reveals important network properties about household interaction patterns, particularly highlighting differences in intra- and inter-household mixing patterns only in individuals wearing the sensors. The median degree reported in this study was lower than that reported in social contact studies from the same area [23,44] using paper diaries and proximity sensors, presumably because only individuals with sensors were recorded as contacts. The degree (number of unique people met per person), number and duration of contact events were dominated by children <14 years. There were many interactions between children aged <5 years and those older than them, particularly school-going children (5-14 years) and adults (20-49 years). The relevance of this information has been demonstrated by the role played by older siblings (school going aged 5-14 years) in the introduction of a range of respiratory infections to infants in the household [26,136]. Intergenerational mixing was frequent at the household with primary school-going children dominating the frequency and duration of interactions. The contact patterns arising from this household study can be used to parameterize mathematical models that investigate different vaccinating strategies without the apparent danger of adverse effects in clinical trials involving infants.
One major addition to knowledge from this study is the interactions among people sharing a household, and between those living in different households. Within-household interactions were low and limited by the number of people per household but were dominated by high frequency and longer duration contact events particularly between children <5 years and 5-14 years, and also with 20-49-year olds. This study also shows that within households, girls aged 5-14 years and female adults (presumably mothers) had a non-negligible role to play particularly in time spent with children <5 years. This is suggestive of the gender roles played by young girls in rural areas of coastal Kenya, where they are responsible for primary childcare.

This study also assessed the longitudinal stability of contacts by assessing the temporal distribution of contact events for each day (05:00-18:59 hours), and the daily amount of time spent with other participants. The results suggested that individuals within rural household clusters interacted with and spent a similar amount of time with other individuals from day-to-day. The distribution of the median cosine similarity values were similar to those of a pilot study conducted in the same site [44]. However, similarity values in rural households in Kenya were relatively low when compared to a study conducted in Italian households [46]. Rather surprising was the very low mean network density in the rural households (0.077) compared to Italian households (~0.835) [46]. This has to be carefully interpreted particularly with reference to the context of the study design. The Italian study recruited entire households that were independent of each other, and there was no report of any inter-household interactions [46]. Households in rural coastal Kenya are composed of several related extended families sharing a compound, and the study also recruited neighbouring households with whom index households would generally interact. This suggests that proximity data from rural households can be collected over a few days to give a snapshot of general contact patterns. However, the effects of seasonal social (e.g. school, agricultural practices) and economic (e.g. tourism) patterns particularly in the coastal areas of Kenya have to be accounted for by collecting data within the different seasons.

5.8.3 Limitations

There are some key limitations of the present study. These limitations are similar to those reported in section 4.7.3, but this time related to households rather than schools.
As such, only additional limitations not discussed in the previous chapter are highlighted. National and regional mixing patterns are expected to vary due to differences in demographic, social, cultural and economic profiles of study locations [73]. This current study was limited to several household clusters in one location, whereby the community predominantly practise farming and fishing. Some residents practise temporary seasonal migration from their households, suggesting that repeat data collection during different seasons may be necessary to assess similarities and differences in the nature of contact events, if any. However, data collection over a 7-day period showed remarkable stability, suggesting that collection of data over even shorter periods may give robust data for use in models. In addition, data were only collected by people wearing the sensors, meaning that all encounters with non-participants were not reported. In future studies, location awareness would be an added advantage to assess daily individual trajectories to understand where individuals spend most of their time, and with who.

Participant recruitment in the rural households required the field team to schedule appointments with potential participants very early in the morning (before 7:00 am) or early evening (4:00 pm) or over the weekends. Recruitment was made more difficult in some instances when married women required the “permission” of their husbands to participate in this research (and any KEMRI-related research in general). More often than not, the husbands were in nearby urban towns in search of employment and would only visit their household over the weekend. This delayed the recruitment process but ensured that all consent was given which increased the number of participants. This is an important factor to consider in future when conducting studies in rural parts of Kenya.
6 Comparison of results from schools and households in rural and urban settings

6.1 Introduction

Countries in the developing world have distinct rural and urban areas. Of the 15 locations in the KHDSS [51], 3 locations including Kilifi Township are considered urban, while the other 12 including Matsangoni are rural (Figure 6.1). These rural and urban areas are characterised by different demographic (e.g. Figure 3.2 showing population pyramids for one rural and one urban area), socioeconomic and cultural factors. In rural areas, households are composed of a cluster of 2-6 contiguous building units, each accommodating a nuclear family. Up to three generations have been observed to live in one household in rural areas [131]. In urban areas one or several building units located in the same compound may be owned by an individual or institution, thus defining a household. The overall sex-ratio (number of males per 100 females) is the lower in the rural compared to urban sites (87 vs 89 males per 100 females, respectively), suggestive of male rural-urban out-migration within the KHDSS. When stratified by age, the sex ratio between the ages of 15-24 years is lower in the urban compared to rural, anecdotally attributable to temporary out-migration for education and out-migration to more urban areas such as Mombasa in search of employment.
Figure 6.1. Rural and urban locations of the KHDSS. Rural areas are shown in orange while urban areas are shown in purple, with study sites highlighted in darker shades.

Synthetic (proxy) age-related contact mixing matrices generated from KHDSS household occupancy data revealed rural-urban differences in the age-related mixing patterns [58]. In general, age-assortative mixing was observed in the synthetic matrices within all the administrative locations with strong off-diagonal mixing indicative of intergenerational interactions. A paper-diary study revealed that Kilifi Township (urban) was characterized by strong mixing among adults aged 20-30 years, while mixing in Matsangoni (and other rural areas) exhibited preferential mixing particularly in school going children aged 5-14 years and parent-child interactions [23]. Similar age-assortative mixing among school-going children has also been observed in African
countries (Zimbabwe [9] and Uganda [179]), but was different from mixing patterns in Europe that was assortative even among the elderly [25]. These studies highlight that there are different mixing patterns within regions in a country and between different countries, suggesting that interventions against infections transmitted via close contact may need to factor the contextual heterogeneities for effective control.

The results presented in this chapter introduces new data from an urban school (SU) and set of households (HU). The description follows that presented in Chapters 4 and 5 which presented results from a rural school and subset of households linked to students in the school. This chapter also compares the statistical properties of the networks in the rural and urban settings and discusses the implications of the results to public health.

The objectives for this chapter were:

(i) To collect social contact data from students in an urban school.
(ii) To collect social contact data from a subset of households linked to students from objective (i).
(iii) To quantify and assess the dynamic variation of the degree, number, and duration of contact events in the rural school and households.
(iv) To assess the differences in statistical properties of the networks in the rural and urban schools and households.

6.1.1 Data collection in the urban school

The first school (name withheld) that was approached in the urban setting declined to participate. This decline was categorised as a “silent refusal” since the initial discussions with the school management committee were protracted and lasted for more than two months until the school closed in November 2016 as per the national primary school calendar. From the KHDSS handbook, a silent refusal is when a participant does not give a direct “no” to participation but exhibits deliberate reluctant behaviour to either give consent or not.

Kilimo Primary School (SU) was then selected and approached in January 2017. The reception from the school management committee was swift and cordial. Following further discussions with the SMC and representatives of parents, these two teams were
invited to the KEMRI research laboratories for an open day session with pre-arranged visits to the labs and presentation from a few scientists on the range of social, clinical and laboratory research ongoing in KEMRI. The aim of this was to boost the confidence of the school teams so as to endorse the research activities at the school. Following this meeting, all parents were invited to the school by the headteacher through letters delivered by the students. Only students whose parents attended the meetings at the school and gave written consent were allowed to participate in the study, and the opt-in selection process led to lower participation rates compared to the rural school.

The urban school had 792 students. The school structure was similar to that of the rural school. That is, it was divided into kindergarten 1-3, and 2 physical classes for each of grades 1-8. The age range was 4.2 - 17.3 years (median 10.5 years) with a gradual increase in older grades. The daily schedule was also the same as the rural school, with classes officially starting at 8:00 am, and two breaks between 10:00 - 10.30 am and 12 noon – 2:00 pm. Teaching ended at 03.00 pm and students were free to go home from 05:00 pm, with a one-hour period for school games/ physical exercise in between. In both schools, students in grade 7 and 8 attended half-day school sessions on Saturday from 8:00 am to midday.

In the urban school, 130 (16%) students out of the entire school (N=792) were recruited into the study using an opt-in approach. The majority of the parents were not available at meetings held at the school to give consent for their children to participate in the study. Most parents said that they were unable to leave work to attend the meetings at the school. Following informal discussions with the parents who attended the meetings but still declined consent for their children, majority cited lack of understanding of the work that KEMRI does in general and were thus skeptical to allow their children to participate.

Recruited students carried sensors in term 1 of 2017 from 19\textsuperscript{th} to 25\textsuperscript{th} February. From the 130 sensors issued, there were 4 withdrawals, 8 damaged sensors due to contact with water, and 4 lost sensors. Out of the 114 sensors available, data from only 67 sensors were analyzed. The rest had no matching student details (17), had inconsistent (15) or corrupted (9) data, or had no data at all (6).
6.1.2 Demographic characteristics in the urban compared to rural households

From the urban school, 9 index households linked to 9 students were identified by simple random sampling. Only 5 opted-into the household-based data collection phase (HU), with the 4 politely declining to participate. This resulted in 5 clusters with 10 distinct households, compared to 40 households in 9 clusters in the rural site. The number of household residents per cluster size in the urban area ranged from 2 to 21 (2 clusters with 2 residents each, 1 cluster size 3, 1 cluster size 6, and 1 cluster size 21). The largest cluster in urban households was composed of several nuclear families residing in a rented block of houses. The median household size was 5.0 (IQR 2.0-8.5) in the urban site and 6.0 (4.0-8.0) in the rural site. The median age in years among participants was higher in the rural compared to urban site (14.5 (IQR 9.1, 40.1) vs 11.2 (IQR 6.5, 26.2), respectively), suggesting younger but smaller households with fewer grandparents in the urban area. From the KHDSS enumeration round 27 (2016) household data, the proportion of children <14 years was 61% (N=15,856) in the rural and 52% (N=51,937) in the urban area, compared to 48% (n=180) and 68% (n=34) in the study population. This shows that in both sites, children aged <14 years compose almost half of the population within the households as expected. The participation rates for schools and households in rural and urban site are summarized in Table 6.1.

Table 6.1. Participation rates in rural and urban schools and households in Kilifi. Participation rate is the proportion of individuals who carried sensors compared to the total number of sensors expected to be issued (350), while response rate is the proportion of sensors whose data were analysed, compared to the number of sensors actually issued. SR and SU are schools in rural and urban areas respectively. HR and HU are households in rural and urban areas, respectively.

<table>
<thead>
<tr>
<th>Site</th>
<th>Number of students/residents (A)</th>
<th>Number of sensors issued (B)</th>
<th>Number of sensors analyzed (C)</th>
<th>Participation rate (B/350) %</th>
<th>Response rate (C/B) %</th>
</tr>
</thead>
<tbody>
<tr>
<td>SR</td>
<td>907</td>
<td>303</td>
<td>223</td>
<td>86.6</td>
<td>73.6</td>
</tr>
<tr>
<td>SU</td>
<td>792</td>
<td>136</td>
<td>67</td>
<td>38.9</td>
<td>49.2</td>
</tr>
<tr>
<td>HR</td>
<td>328</td>
<td>292</td>
<td>180</td>
<td>83.4</td>
<td>61.6</td>
</tr>
<tr>
<td>HU</td>
<td>125</td>
<td>97</td>
<td>34</td>
<td>27.2</td>
<td>35.1</td>
</tr>
</tbody>
</table>

1 The number of sensors available for data collection.
6.1.3 Comparing rural and urban contact event timelines in schools and households

Daily contact activity timelines were compared between schools and between households in rural and urban areas (Figure 6.2). Only the rural school (panel (A)) showed slight temporal variation in daily mean number of contact events recorded per hour with marked dips during the lunch break and post-5 pm when students returned home from school. Interactions in the urban school (panel (C)) showed no variations on all weekdays and weekends, apart from Saturday when higher interactions were recorded with a major peak occurring in the afternoon. The daily timelines for rural households also exhibited a constant trend from 7:00 am to 4:00 pm with major peaks observed consistently on all days at 5:00 am. There is no difference in the distribution throughout weekdays and weekends (see also Figure 5.7 (D)).

Figure 6.2. Comparing activity timelines in rural and urban schools and households. Figures in all panels represent hourly count of contact events divided by the number of people in contact per hour. Each coloured line represents a different day of the week, with dotted lines representing weekends. Only data between 05:00 am and 07:00 pm are reported.
Conversely in the urban households, from early morning to mid-afternoon (3:00 pm), there is no characteristic patterns that can be observed in the data for all days, suggesting that contact patterns may be different across all days. However, this remains inconclusive especially due to the small sample size. Many peaks are observed on each day, and these interactions seem to stabilize in the afternoon from 3:00 pm and remain constant till 6:00 pm.

6.2 Comparison of network properties between rural and urban schools and households

The network properties of the rural and urban schools and households have been summarized in Table 6.2 and Figure 6.3. Panel (A) of Figure 6.3 displays the distribution of the median (IQR) degree values of the full networks over the entire duration of study. The median degree was higher in the rural compared to urban school (85.0 vs 21.0, respectively) as a consequence of the sampling. The differences in median degree in schools may be explained by the sample size since the rural school had 3 times the number of participants compared to the urban school. The median degree was twice as much in the rural compared to the urban households (11.0 vs 5.5, respectively). The total number of contact events was almost similar in both urban and rural schools and households. However, there were differences observed in school and household based mean contact duration. Contact events generally lasted longer in households compared to schools, with the urban setting exhibiting higher mean contact duration in both schools and households. Clustering was higher in the rural compared to urban school, but higher in urban compared to rural households. Strangely, rural households exhibited the lowest network density indicative of the very low proportion of links that form between individuals compared to the total links expected.
Table 6.2. Comparison of network properties in rural and urban schools and households.

<table>
<thead>
<tr>
<th></th>
<th>Number of nodes</th>
<th>Median Degree (IQR)*</th>
<th>Number of Edges</th>
<th>Number of contact events</th>
<th>Mean clustering</th>
<th>Mean density</th>
<th>Mean contact duration in minutes (CV²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SR</td>
<td>223</td>
<td>85 (66-104)</td>
<td>9,564</td>
<td>28,440</td>
<td>0.532</td>
<td>0.386</td>
<td>3.2 (9.1)</td>
</tr>
<tr>
<td>SU</td>
<td>67</td>
<td>21 (17-25)</td>
<td>711</td>
<td>28,439</td>
<td>0.469</td>
<td>0.322</td>
<td>4.0 (11.3)</td>
</tr>
<tr>
<td>HR</td>
<td>180</td>
<td>11 (7-18)</td>
<td>1,230</td>
<td>30,240</td>
<td>0.556</td>
<td>0.077</td>
<td>8.6 (5.7)</td>
</tr>
<tr>
<td>HU</td>
<td>34</td>
<td>5 (2-9)</td>
<td>102</td>
<td>28,440</td>
<td>0.650</td>
<td>0.182</td>
<td>16.3 (9.3)</td>
</tr>
</tbody>
</table>

*Median degree values presented with no decimals.

In Figure 6.3, panels (B) and (C) display the log-log distribution of probability of observing number of contact events of contacts of a certain duration, respectively. Panel (B) suggests that the distribution of the number of contacts depends on the sites. That is, the distribution follows a similar trend in rural and urban schools, and the trend is slightly different when considering the corresponding households. The distribution of short contacts (< 5 minutes) shows a similar trend across all four sites (panel (C)). In the rural and urban schools, a third of the contact events lasted more than 5 minutes.
(38% and 35%, respectively), while more than two-thirds of contact events (61% and 73%, respectively) lasted more than 5 minutes in households.

Figure 6.3. Comparison of distribution of degree, number and duration of contacts in rural and urban schools and households. Panel (A) shows boxplots of the distribution of degree. Panel (B) and (C) show the log-log distribution of number of contact events and contact duration, respectively. SR=Rural school, SU=Urban school, HR=Rural households, HU=Urban households. For each boxplot, the horizontal bar in the middle shows the median, the grey box shows the lower and upper bounds of the interquartile range (IQR), and the whiskers show 1.5 x IQR of the distributions.

Figure 6.4 displays the distribution of degree by age (A) and household size (B) comparing rural and urban areas, and Appendices F and G highlight the differences in the number of participants by age and household size distribution of number of contacts. Whereas the median degree peaked in children aged 0-4 years (10.0, IQR 8.3-11.0) in urban households, it was highest in school going children in the rural area (19.0, IQR 12.0-27.0). At the household level (B), median degree values did not vary by household size. In contrast, the values showed an increasing trend in urban households.
Figure 6.4. Household degree distribution by site. Panel (A) shows the degree distribution in rural and urban households by age of participants. Panel (B) shows the degree distribution stratified by household size.

Network plots for urban households were not presented because it was not possible to decipher any clear pattern due to the small cluster sizes and total number of residents (NUHH = 34). Similarly, the network diagram for the urban school is not presented. Instead, circle plots displaying the magnitude of the distribution of interactions within and between grades in schools ((A) and (B)), and within and between age groups in households ((C) and (D)) are shown in Figure 6.5. These distributions have been normalised by the number of people per age group. Each colour represents a different school grade (or household age group), the size of each colour band is proportional to the group size, and the links within and between groups are weighted by the duration of interaction. Panels (A) and (C) have been discussed extensively in Chapters 4 and 5. While these data were only collected within two locations of the KHDSS [51], the plots at households and schools show evidence of heterogeneity between settings worth mentioning. In the rural school (panel (A)), interaction is highly assortative within the same grades. Most of the heterogeneous contacts occur in the lower grades (KG-grade 4), and there are thin links (minimal interaction times) between lower grades and upper grades (5-8). In the urban school, apart from the almost exclusive interaction within grade 8 students, there is considerable cross-grade interaction. Similar to the rural school, these inter-grade interactions are more pronounced within the lower primary.
Different patterns also emerge at the household level. In the rural area the majority of the interaction time is spent between the 0-4- and 5-14-year age groups, while interactions between 0-4 and 20-49 take precedence in the urban area. Children aged <14 years also spend a considerable amount of time with the elderly (50+ years) in the rural compared to urban households.

### 6.3 Comparison to random networks

Random networks were generated based on the same number of nodes and network density for each of the settings. The expected median degree values and expected clustering coefficient for random networks were calculated at the schools and
households with corresponding equal number of nodes and edges (Table 6.3). The median degree values in the random networks of rural school and households were significantly lower than observed, but there was no difference between observed and expected values in the urban areas. In all the settings, the clustering coefficient in the random networks was lower than was seen in the original network. This suggests that even the strong heterogeneities observed due to age (household) and grade (school) were not sufficient to explain the observed social structures.

Table 6.3. Properties of random networks in schools and households.

<table>
<thead>
<tr>
<th></th>
<th>Nodes</th>
<th>Median degree</th>
<th>K-S test P-value*</th>
<th>Clustering coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>SR</td>
<td>223</td>
<td>86.0 (81.0, 91.0)</td>
<td>&lt;0.001</td>
<td>0.387↓</td>
</tr>
<tr>
<td>SU</td>
<td>67</td>
<td>21.0 (18.0, 24.0)</td>
<td>0.414</td>
<td>0.320↓</td>
</tr>
<tr>
<td>HR</td>
<td>180</td>
<td>15.0 (12.0, 17.0)</td>
<td>&lt;0.001</td>
<td>0.080↓</td>
</tr>
<tr>
<td>HU</td>
<td>34</td>
<td>5.0 (4.0, 6.0)</td>
<td>0.085</td>
<td>0.138↓</td>
</tr>
</tbody>
</table>

*Kolmogorov-Smirnov test on difference in distribution of degree values, comparing original network and random network. ↓ indicates that the value in the random network was lower than the original network.

6.4 Summary and discussion

6.4.1 Recruitment in urban school and households

The first school that was approached in the urban area declined to participate in the study. In retrospect, there are two reasons that potentially resulted in the delay of getting a favourable response from the school management committee (SMC) of this school. First, it took a long time to set up a meeting with the SMC due to conflicting work schedules of the members. Being in an urban area, most of the members were in full employment and they expressed difficulty in getting time off work to attend the engagement meetings with the research staff. Eventually, the meeting was held on the day that the school was closing for the August holiday. The research team was invited to the school but could not engage the SMC and parents directly since the school had internal ceremonies to commemorate the closing of the school. Second, there was even less time for formal engagement with the school administration and SMC in the last term of the school calendar (runs between September and November). The last term is
shorter than the first two since the grade 8 students sit for their national exam in late October. During this time, non-academic activities are not allowed in all primary schools, and all students between kindergarten and grade 7 break from school at least two weeks before the start of the national exams. This evidently posed a challenge to the continuation of the engagement with the school and a decision had to be made to withdraw from the school despite the massive efforts that had been taken to enrol the school. This “refusal” scenario could be replicated in other government-funded (public) schools in Kenya since all follow the same administrative procedures.

A few recommendations can be implemented to reduce the amount of time spent in recruitment at schools for similar studies. Firstly, research studies targeting schools in Kenya should aim to conduct activities within the first two terms because they are longer and not subject to compulsory administrative interruptions e.g. due to national exams. However, this may minimize the ability of long-term studies to assess differences in contact patterns due to school schedules such as weekends and holidays, which have been shown to reduce the number of assortative contacts within school children with subsequent implications in the transmission of infections [62,80]. Secondly, the research staff for this current study organized a science open-day held at the school at the end of the study, inviting parents, students, staff and the neighbouring community. At this event, various scientists from the KEMRI were present to discuss various research activities going on at the institute. This was both informative and to give the scientist a chance to respond to concerns from the community regarding health research. In future, the best approach would be to engage the school in discussions related to a study either early in the first or third term. This would give ample time for discussions between researchers and potential participants before recruitment, particularly because recruitment of minors (<18-year-old) requires initial parental/guardian consent. Similar sentiments of “bringing the research scientists closer to the community” were expressed in the rural school. By encouraging local British secondary school students to interact with primary school students, researchers enriched the experiences and knowledge of the students leading to high participation rates [180] particularly among the younger students.

This narrative is important because several lessons can be drawn from the school engagement in the urban location with relation to potential future studies in urban
settings. First, due to their higher education levels, formal employment and more access to various forms of media, parents were more aware of research activities within Kilifi (KEMRI) and globally. At times, misconceptions towards research (for instance, questions arose on the link between electromagnetic signals from portable devices and cancer) formed the basis of refusal to participate in research. This misconception, in addition to several other recurrent themes in both the rural and urban area, led to the development of a frequently-asked-questions (FAQ) booklet that was presented to the research participants (Appendix D).

6.4.2 Participation rates and data abstraction

Different recruitment strategies for participants were employed in the rural and urban schools. With the availability of a complete school register from the rural school, it was possible to randomly select participants proportional to grade and gender in line with the number of sensors available for data collection. However, parents were free to opt-into the research activity in the urban school, with most of the parents not attending recruitment meetings due to work obligations. This resulted in fewer participants from the urban compared to the rural school. In addition, there was generally a reduction in the amount of data analysed due to data quality checks described in the Chapter 3 (Methods). With reference to the school size, participation in the rural and urban areas was 25% and 8%, respectively, and less than 1% within the households. When considering only the number of participants expected guided by the sensors available, data were more than 50% apart from the urban households. With network studies, the aim would be to capture a sample of individuals that would be representative of populations while considering the cost, logistics and time implications.

Previous studies have not described in-depth any data quality check issues, apart from two that reported censoring data particularly during the night time described [44,46]. Here in this study, a comprehensive description of data cleaning methods has been provided. The cleaning process involved assessing individual data files for inconsistent data, for example, continuous data collection overnight suggestive of sensors placed together during storage. Despite this, there was still a large volume of data that was discarded between 07:00 pm and 04:59 am. As such, there is no evidence of significant contact activity at night, which would have resulted in higher mean duration of contact
events particularly in the households. Despite this, the number of participants in the urban setting was close to that reported in previous household studies that used similar methods of data collection [44,46]. Even with this small number of residents participating, this study reports an important and novel study that describes the mixing patterns in schools and households particularly in SSA.

6.4.3 Contact patterns, number and duration of interactions

Due to the skewed distribution of degree, this study has reported the median and corresponding interquartile range (25th and 75th percentiles) values. However, previous studies conducted in SSA have reported mean values. A paper diary study spanning rural and urban areas in the KHDSS reported an overall mean of 17.7 contacts (degree), with rural areas reporting significantly higher rates compared to urban areas (18.8 vs 15.6, p=0.002) [23]. The opposite observation was observed in Zimbabwe, where peri-urban participants reported more contacts that their rural counterparts (11.6 vs 10.8, p<0.001) [9]. Using wireless proximity sensors in a pilot study, a mean degree = 15.3 was reported [44]. In the rural households of this current study, the degree values presented here are robust, despite being lower than previously reported. The sensors only record interactions with other individuals who have sensors, meaning that all interactions with other community members not participating in the study were not captured. The only solution out of this when using sensors is to saturate the entire community, which will be practically challenging due to recruitment procedures and issuing of sensors to participants. For instance, for this relatively small study involving 9 household clusters in a rural area, it took 6 research staff working from 7:00 am – 6:00 pm at least 3 days to recruit all individuals from the households, and a similar amount of time to collect all the sensors at the end of the study. Similarly, for the schools, the research team could only engage the students when out-of-class, that is early morning before class, during morning and lunch break, and in the evening after class for only about one hour. This presents challenges that are probably not unique to the rural coastal setting because all public primary schools follow the same school schedule.

A high number of contacts were recorded between toddlers (0 to 5 years) and adults (20-49 years) in the urban areas, and consequently more time spent in proximity
between these two groups (Figure 6.5 panel (D)). This was also observed in households in Italy, where 70% of the contacts were between infants and other family members. In peri-urban areas of Zimbabwe, high inter-generational contacts between parents and children were also observed. However, this is different to what was observed in the rural households in Kilifi, where toddlers spent predominantly more time with older school-going children (siblings and other relatives within the household). The subtle differences on with-whom individuals spend their time may have an impact on transmission of respiratory infections. Transmission patterns from a household study suggested that the majority of infection in infants are conferred by older school-going children sharing a household with the infected infant [26]. This has also been alluded to through a mathematical model investigating transmission of RSV in a network of households and schools in rural Kilifi [131]. However, given the differences in mixing by age within rural and urban areas, it can be argued that urban adults may play a more significant role in introducing infections into the household.

Despite the within school and household differences in demographic patterns reported here, there are similarities in the distribution of the number of contact events and duration of interaction. From this study, there is no evidence that there were differences between those who participated and those who did not. However, there is a likelihood that this exists, and the generalisation can only be assessed by recruiting more schools (and households) from a bigger geographical region. Despite the same data collection methods used in both settings, the interpretation of the school-school and household-household differences in degree is greatly hampered by the small sample sizes. In other words, the studies did not have sample sizes that were large enough to conclusively say that there was a difference between the schools, and between the households.
7 Discussion

7.1 Purpose of the study

It has been demonstrated that the transmission of respiratory infectious diseases depends on, among other factors, the mixing patterns in the population. Several factors are still not clearly known, such as the form of these interactions in various environments (e.g. households versus schools), in different groups (e.g. working adults versus school children), in different settings (e.g. in rural versus urban areas) and regions (developed versus developing countries), and how long an interaction should last for transmission to occur. Previous studies to quantify mixing patterns in different populations have mainly collected data on who-contacts-whom using egocentric paper diaries as presented in a recent comprehensive review [16]. Other indirect methods that exist include simulations using demographic data [71–73], and estimations from time-use data [8]. Each method has its own level of intricacy, but the biggest challenge with the most frequently used method, self-reported paper diaries, is recall bias. In settings with low education levels, inability to self-report required the use of third parties to record the contacts thus increasing the cost and complexity of the study while providing potentially biased results [23]. With paper diaries, individuals of various ages have been requested to report on whom they had a contact with, for how long, context, and the frequency of the contacts. In the majority of the cases the relevant contact was that seen as sufficient enough to enable the transmission of a pathogen that may cause a respiratory infection, such as influenza, pertussis, tuberculosis, or RSV. A contact was thus originally defined as either a direct (involving physical touch) or indirect (such as a
two-way conversation between individuals in the same space) [6], with some variations between studies. Over one or repeated days, individuals prospectively or retrospectively fill in a paper or electronic diary (or respond to an interviewer). Contact pattern data are largely applied in models that aim to understand transmission of various pathogens in different contexts, with a view to model various ways of control and prevention, such as vaccination (random vs targeted) and social distancing measures (shutting down transport hubs, school and work closure, etc). Several of these contextual models do exist. Unfortunately, there are very few published studies from SSA (e.g. [21,58–60,131]) providing contextual evidence on the need for and utility of contact pattern data in the fight against contagion processes particularly through vaccination.

Data collection using wireless proximity sensors is becoming popular in human social network studies. Since 2010 to date, several studies have been conducted in various settings: schools [33–35,115,116], hospitals [31,36–38,117,118], social and work settings [30,40,86], households [44,46], and more recently at a simulation of a health emergency evacuation procedure [48]. Proximity sensors provide an automated way of logging even the shortest face-to-face encounters between two individuals, providing crucial data of potential contagion with major applications in health. However, there still remains a paucity of contact data from low- and middle-income countries where the disease burden is high and such data is needed to design tailor-made interventions. In particular, there is almost no data on school and household mixing patterns in these settings, where most infections occur. To date, only one pilot study has reported the feasibility of use and acceptability of proximity sensors in 5 low-income households in Kilifi, coastal Kenya [44]. Focus group discussions with the residents before the pilot study suggested that they would be amenable to carry the sensors for several days of data collection. Compared to a previous diary study in a similar population that was wrought with reporting challenges due to low education levels of participants [23], the pilot sensor study experienced high participation rates (75% vs 54% for diary study) and residents reported minimal challenges with using the sensors continuously over three days. Two studies in Europe that compared sensor-based methods to paper diaries [86] and a web-based survey [115] reported that participants were more at ease using sensors compared to survey methods. These studies also reported underreporting of contacts especially those of short durations (<5 mins), suggesting that longer contacts were
easier to recall. The interplay between time of exposure (and frequency of contacts) has been shown to be important in simulations investigating transmission of parvovirus-B19 (and varicella) [42] and influenza [37]. At the household setting, two key advantages of the proximity sensors compared to the paper diaries were: participation from all individuals within a household including individuals of all ages, and the automated detection of proximity events among residents sharing a household and between residents of different households. The pilot study in Kilifi [44] provided proximity data at very high temporal (contacts aggregated at 20-second intervals) and spatial (<1.5-meter separation distance between individuals) resolutions, while also generating a favorable framework for community engagement for future studies. Only one infant was excluded in the pilot study due to concerns about the size and shape of the proximity sensor, but every other individual reported being comfortable to carry the sensors for extended periods of time in future.

To our knowledge, this thesis reports the first large-scale study in an African setting that presents finely-grained data on proximal (face-to-face) interactions between students in schools and linked households. Further, it compared patterns of contact between rural and urban locations. A unique component was to study patterns both within and between neighbouring households. This study had two main objectives: the first was to characterize and compare the contact network patterns in rural and urban schools and households in coastal setting of Kenya. The descriptive results from the observed mixing patterns are important in the modelling of the transmission of various pathogens that are spread via close contact. Thus, the second objective was to develop the framework of a mathematical model parameterized by the ensuing contact structure to aid in understanding the transmission of a respiratory virus within and between households in a rural setting. In addition, this study provided an opportunity to develop a comprehensive data cleaning and analysis pipeline, particularly one that can be used to eliminate “noise” in data. Finally, this study was able to explore social and logistical constraints on the use of wearable proximity sensors in rural and urban schools and households in a developing-world setting.
7.2 Summary of key findings

7.2.1 Participation rates in schools and households

There have been discussions on what proportion of individuals should be sampled for social contact studies to be representative of a population, or for how long such studies should be conducted. When investigating disease outbreaks in networks, it is important to investigate all contact events whether or not they have been involved in the transmission of pathogens [17,181]. Previous studies collecting data on networks of epidemiological relevance have been bound in space, time and scope (see Table 2.2 for a summary of studies using proximity sensors). That is, studies have focused on structured semi-closed groups such as schools, hospitals and conferences, collecting data over various periods ranging from one day to several distinct periods in one year. Compared to diary-based data collection, studies using wireless proximity sensors in closed populations such as hospitals and schools have generally reported participation rates in excess of 90% [86,115]. Household-based studies in Kenya and Italy have reported data extraction rates of 75% (data analyzed/ number of participants) [44,46], with individuals indicating that they would be willing to participate in similar studies in future attributed to the lack of need to self-report contacts.

This study has extensively discussed some challenges that were encountered by the research study team during the data collection. Recruitment has generally been observed to be easier in rural compared to urban areas [19,23]. Individuals in urban areas are highly mobile and are more likely to travel further for work, reducing the chances of involving them in research due to absenteeism at home. Previous studies in primary and secondary schools have recruited selected classes [33,116] for convenience or entire schools [34,35,115] due to availability of sensors. For this current study, only 350 sensors were available for the study hence limiting the sample size from the onset. The reasons for non-participation differed greatly between the two schools in this study: absenteeism from school (mainly due to non-payment of school fees) was the biggest reason for non-availability of students in the rural school, while parent and consequently student reluctance to participate in the study were the main reasons for non-participation in the urban school. In urban households, there were different challenges, some of which were unsurmountable. For instance, outmigration was a key challenge and the
KHDSS data available at the time of study (enumeration round 27 of 2016) was outdated particularly in the urban area, and some listed residents could not be traced. In urban areas some residents were not available to give consent due to work schedules, and others cited the necessity of spousal consent before participating particularly in the rural areas. Lastly, in two households, the research team encountered one visually impaired person and a couple with hearing impairment, but all were not recruited. It was not possible for the research team to ensure complete understanding of the study procedures by the last three participants. All these suggest that in different settings, there could be numerous valid reasons for non-participation that future studies have to consider during the study design and data collection phase [171].

Following pre-processing, there was great variation in the proportion of sensors with data for processing, with higher values in the rural compared to urban setting. The reason for the low rates was the elimination of nodes with data that exhibited unusual patterns, such as enhanced activity during the night, as described in section 3.5. While these rates may seem low, this study bears in mind that previous studies in schools in Europe and the USA mostly targeted all the students per school or class including the staff. This current study was limited from the start by the number of sensors available to collect data per setting, hence we could only recruit individuals based on the number of sensors available.

7.2.2 Activity timelines, distribution of degree, number and duration of contacts

The versatility of the wireless proximity sensors makes it possible to deploy them repeatedly across different settings, and to directly compare their statistical distributions [17,30]. The definition of a contact in this study was predefined by a threshold that corresponded to close face-to-face interactions and is similar to what has been used in previous studies using the same sensors. Since sensors also capture date-time data at the second level, it is also possible to accurately compute the number and duration of interactions both at individual and group level.

Results from this study reveal that that despite the low number of participants in the urban settings compared to rural, the probability distributions of number and duration of
contacts are strikingly similar Figure 6.3. The daily activity timelines by setting also revealed circadian rhythms, with schools revealing patterns that conform to the school schedule while households are characterized by seemingly unchanging interaction patterns. Schools revealed sustained interactions during the day with dips in number of interactions during lunch hours and in the evening after school, and lower contact events (and degree) recorded over the weekends. One of the earliest studies harnessed similar wireless proximity sensors to collect data in different social contexts (office and scientific conference) with 25 to 575 individuals participants [30]. The statistical signatures from the different contexts displayed remarkable self-similarity and identified a few super-connected individuals. Data from wireless studies reported in Table 2.2 suggested that regardless of the setting, most contacts are of short durations, but there exist contacts with non-negligible probability of long durations (described as heavy-tailed distributions). If the transmission probability of a disease depends on the duration of interaction between two individuals, then contact events of shorter durations may correspond to a small transmission probability while longer interactions could have a more important role in disease dynamics [33,42].

The mean duration of interactions reported in this study were longer than those reported previously in schools and households in European countries and the USA. In principle, this current study assumed that all contacts occurred at home or in school, which may not be so particularly for older children and adults who are considered to be highly mobile. Rural school patterns show daily reduction in contact events over lunch hour and lowest levels in the evening when the school session ended. It would have been informative to have data on location of contact and amount of time spent at each location to potentially gauge the risk of transmission given various settings. Suggestions on how this data would have been collected are through self-reporting or responding to questionnaires such as in Zimbabwe [9] or using portable Global Positioning Systems (GPS) devices e.g. Vazquez-Prokopec et al. [182] to track the movements of students (and household residents) throughout the day. In addition, the considerable reduction in number of contacts over the weekends in schools suggests that there exist differences between school and non-school days. If holidays are considered as extended weekend periods, the data suggests that there would be a reduction in school-based contacts, and a corresponding increase in household-based contacts especially with older individuals.
7.3 Importance of the findings

7.3.1 Mixing patterns at schools and households

Evidence from schools and household transmission studies highlight the importance of children as potential conduits of infections in the community [26,131,136]. However, no study has collected mixing pattern data linking the two settings, that is, data on interactions between students in schools and students and family members at home. This current study, in part, collected data from households linked to students attending a school in the vicinity of the households. Even though the data were not collected concurrently, i.e. simultaneously at school and households per setting, or simultaneously in schools and households of rural and urban settings, the study still offers invaluable insights into the statistical similarities (distribution of contact number and duration) and differences (distribution of degree in schools and households per site). These data suggest that formulation of contextual mathematical models may need to explicitly account for the heterogeneity in degree of contacts due to setting (schools and households) and demographic factors (age and household size). This has been proposed by including the framework of a mathematical model that will incorporate details of mixing within and between households. The flexibility of the model allows details of mixing in other settings such as schools and the general community to be added in future.

7.4 Summary of limitations of the study

The transmission process of respiratory infections is still poorly understood. Transmission is highly dependent on contact patterns, but also on other phenomena such as disease-specific transmission probability and individual immunity. It should be emphasized that the proximity sensors do not record an actual physical contact that is considered as the putative at-risk event, but rather mutual proximity of individuals wearing the sensors. In other words, proximity sensors cannot distinguish physical and conversational contacts – and hence cannot dissect the different patterns of transmission with certainty dependent on contact type. One proof-of-concept study attempting to link
transmission to contact events has been conducted in a hospital setting in France [37]. This hospital-based study simultaneously collected detailed contact pattern data using sensors and nasopharyngeal samples from doctors and nurses attending to geriatric patients admitted to the hospital. By tracking individuals contact patterns and influenza infectious status, the study demonstrated that infectious healthcare workers were a potential source of infection for patients and vice versa. However, generalizability was limited by the small number of patients and lack of consideration of other contextual factors that influence transmission such as individual and microbial characteristics.

Larger studies can be conducted perhaps in controlled settings such as kindergartens where mixing rates are high while targeting more abundant a respiratory pathogen such as human rhinovirus that circulates all year round [159,183] compared to e.g. flu and RSV that are seasonal and may thus be missed.

One major advantage of proximity sensors is that they record a measure of proximity between individuals, the number and duration of contacts at fine temporal resolutions. Future studies can investigate the impact of altering the cut-off threshold of proximity on the number and duration of contacts, and its impact the degree distribution and hence disease transmission in models. The quality of the data is not easy to judge. Numerous yet unexpected contact events were continuously recorded from early evening (~7:00 pm), through the night and up to early morning (~5:00 am). An algorithm was first designed to filter out excess contact events by taking the mean of subsequent contact events. Further, data between 7:00 pm and 4:59 am were discarded due to non-resolution of the excess contact events particularly during the night time. The only plausible explanation for these excess contact events is that individuals stored the sensors together, as was seen in a previous study in the same setting [44]. Future studies should consider requesting individuals to record the times when the sensors were actually worn (or removed) from the body, or use active or passive surveillance such as body cameras or body-area-networks [34]. This would enable direct elimination of noise accrued due to excess data collection and potentially establish the fraction of actual physical contacts.

Because we had a limited number of sensors insufficient for all the studies to run simultaneously, we had to collect the data in waves, starting at the schools and linking households to a subset of participating students. This means that a synchronous view of
the entire system was not observed; however, all important periodicities and heterogeneities were probably captured, and this data is invaluable in modelling studies. This limitation might be removed in future studies by using more sensors or focusing on smaller communities. There was a completeness or scale trade-off, this study has established that state-of-the-art data collection tools can be utilised in larger future studies.

Bias may have been introduced by the selection of participants, particularly in the urban school. Students, through their parents, opted into the research as compared to a random selection from the school register in the rural site, or inviting the entire school to participate [34,35,115]. One way of resolving this bias would have been to compare contact patterns from these studies to another source of proximity contacts, such as was done using paper diaries or web-based surveys. However, these were studies conducted in high schools with older students [115,116] and scientific conferences [86] where the general level of understanding may be considered high and the participants were able to access the internet freely. It would be practically impossible to introduce web-surveys to the rural population of the current study due to lack of personal computers at home, and that it would also increase the complexity of data collection particularly in the rural areas with relatively higher levels of illiteracy.

One of the key advantages of using automated sensors is the ability to accurately identify and record close social contacts. However, due to the threshold constraints exercised in the analysis (detection of contacts within approximately 1.5 metres), it is possible that contacts occurring at group/social events may have been missed such as during class breaks or out-of-school social activities (e.g. religious activities, weddings, funerals, etc) that are prevalent in this community particularly over weekends. Early studies in schools and a museum [30,35] used a different cut-off threshold (3 metres) so as to capture the effects of group structures in communities. Future analysis can vary the power thresholds to assess the dynamic nature of formation and dissolution of cliques and its potential impact of spreading of infections.
7.5 Recommendations for further research

7.5.1 Participant recruitment, retention and improved data collection

In addition to the availability of data on contact patterns in different settings over several days, this study has demonstrated the challenges of recruiting participants in rural and urban areas of Kenya. Given the similarity of school composition and settings within Kenya, it is possible that studies in other regions may encounter similar challenges particularly in urban areas. From this experience, it is important to involve the parents early when conducting school studies. Community engagement is expensive and takes time thus ample time should be set aside in the design of the future studies, particularly considering school schedules. It may also be useful to explore other ways of reaching out to potential participants such as through phone calls, digital media such as websites with full information on study procedures, or visits to work places as appropriate. At the end of the data collection period for this study, all parents regardless of participation, were invited to the school for an open day to pass on key public health messages (such as hand washing techniques) and demonstrate some research activities undertaken by KEMRI. This activity, together with a demonstration of the data collected by this study and its intended applications, reduced the skepticism of the parents to the just concluded work. If replicated in future before and after studies, such open health days are expected to improve participation rates.

To minimize the amount of noise observed in the data due to continuous logging of contact events, sensors can be switched on or off during issuance to or collection from participants, respectively. Other strategies to filter out noise during nighttime include daily recording of the time that each sensor was worn and put down and have static sensors in frequently visited locations in the vicinity of the study area.

7.5.2 Design of larger cluster-based studies

The study reported in this thesis purposively selected two schools, one each in a rural and urban area, and clusters of households related by blood/marriage and neighbours with shared boundaries. A better understanding of social networks in different areas may be assessed by saturating multiple community clusters with sensors over a period
of time. In order to generate a bigger sample size and to account for within-country (e.g. Kenya) heterogeneity due to demographic, socio-economic and cultural reasons, cluster-based studies can be designed. This can adapt a two-stage sampling design guided by data collected during National Housing Surveys by the Kenya National Bureau of Statistics\textsuperscript{13}. A representative sample of households can be extracted from the National Sample Survey and Evaluation Program V (NASSEP V) frame which covers rural and urban areas of all 44 counties (administrative units) in Kenya. A two-stage cluster sampling design can be adapted, with the first stage selecting a number of clusters and the second stage selecting households as the primary sampling units proportional to the size of the clusters. Even with this design, limitations such as the number of sensors available for deployment and logistical challenges during recruitment need to be considered as discussed earlier.

7.5.3 Epidemiological implications: formulation of a mathematical model against transmission of respiratory viruses

The motivation to collect data on social contact patterns in different settings was the need to understand the transmission of a range of respiratory infections, in particular RSV. RSV causes seasonal ALRI epidemics globally \cite{160}, and disease is exacerbated in in preterm infants, patients with congenital defects such as heart disease as well as (elderly) individuals who are frail or immunocompromised \cite{2}. Severe disease leads to hospitalization which causes a huge social, economic and health burden to families and national governments \cite{184,185}, especially those living in low-and-middle income countries that have poor access to quality health care. Research efforts are underway to develop and test candidate vaccines to prevent disease particularly in infants. In anticipation of a publicly-available vaccine in 5-10 years, more effort is needed to improve the design of vaccination strategies that will, for instance, directly protect the infant or indirectly prevent transmission to infants through a cocooning effect achieved by vaccinating older individuals.

\textsuperscript{13} \url{http://statistics.knbs.or.ke/nada/index.php/catalog/83/related_materials}
This study has defined empirical networks for schools and households in a rural and urban setting in coastal Kenya. Here, the framework of a mathematical model to simulate transmission mimicking RSV epidemics in Kilifi is proposed. An individual based model can account for the age-specific RSV serological profile and contact patterns occurring within and between residents of the same or different households, respectively. During initialization of the model, household residents will be characterized by their household identifier, age, contact degree of participants, and with temporary protection against RSV based on an age-specific serological profile [186]. If an individual $i$ within the network is selected randomly and infected, the ensuing epidemic can be monitored by counting the number of secondary cases infected by the primary case. During the simulation, the probability of transmission can vary depending on two things: firstly, the serological profile determining the individual history of infection, and secondly, whether the primary case and its neighbourhood share a household or not, given by $j$. This will make it possible to characterize 1) the expected number of secondary cases that each infector $i$ will generate in household $j$ ($R_i^j$), hence capturing variation between individuals and household membership, and 2) the distribution of secondary cases due to $i$ ($R_i$) that estimates the total number of infections at the end of the epidemic. In this manner, it will be possible to compute the age-specific proportion of infections due to transmission within or between households determined, in part, by the contact degree of each individual. This will enable the quantification of proportion of infections arising from same-household members or from the community, with a view to design and assess alternative targeted vaccination strategies such as individual, age-related or household based vaccination.

The main limitation of such a model will be the generalizability of the outcomes due to the specific nature of the contact patterns from Kilifi schools and households. However, the model can importantly act as a first step that can lead into local or regional models that incorporate contextual data on social and demographic data. In addition, the framework of this model can be used to assess control strategies against RSV (and other respiratory infections) such as vaccine administration targeting individuals such as the very young or children of school-going age can be simulated due to the explicit representation of each individual in the households.
8 Conclusion

There is a growing emphasis on understanding human social networks through time and in different geographical regions in order to elucidate context-specific mechanisms of transmission and hence implementation of interventions that factor these heterogeneities. It is becoming increasingly clear that there are many different combinations of contacts, such as contacts involving conversation only or including touch, which need to be understood in order to better simulate what happens especially at the onset of epidemics. Understanding these variations in proximal contacts is an important step in improving models that simulate transmission patterns of respiratory infections, provide more robust data to predict epidemic trends, and generate invaluable input in mathematical models assessing prevention and control measures against transmission.

In general, the current study complements previous studies in households [44,46], schools [33–35,115,116], hospitals [31,36–38,117,118], among conference attendees [40,47,86], and in highly dynamic scenarios such as a museum [30] and simulation of a health emergency evacuation [48]. Data from this study conducted in coastal Kenya has provided additional data on social contact patterns in rural and urban schools and households in a developing country setting, arguably making available useful data that did not exist to this magnitude. The results show that there are significant differences in the number of contacts based on location (rural vs urban) and setting (schools vs households). The study has demonstrated the strong assortative nature of contacts in children aged <14 years and inter-generational mixing at the households. This confirms that an assumption of homogeneous mixing in populations may misinform
mathematical models against transmission and control of respiratory infections. Specifically, the number of contact patterns in schools were invariable during the school days, but reduced significantly over the weekends, whereas household contact patterns remain similar over one week. Age of household members and grade of students were the biggest driver of differences in contact patterns in households and schools, respectively. In households, the number and duration of contact events remained statistically similar over seven days. However, school data revealed significant decline in contact activity patterns during the weekend compared to school days. This suggests that data collected over one or two days in schools, as previously done using paper diaries, may not sufficiently capture the heterogeneity in age-distribution of contacts events particularly over weekends (or holidays). Nonetheless, it is important to collect additional data from schools, households and other contexts in other geographic locations and at different periods of the year to validate the results presented in this study.

This thesis also briefly provides potential future direction for similar sensor-based studies. In order to get representative samples from multiple sites, two-stage sampling has been proposed with households as the primary sampling unit. In addition, the results of degree distributions in households obtained from this study can be used to parameterize a mathematical model seeking to quantify the contribution of inter- and intra-household contacts on the transmission of RSV and other related respiratory infections. In the absence of a vaccine against RSV coupled with challenges in designing a vaccine suitable for infants, other prevention and control measures such as vaccinating older ages particularly school-going children should be investigated. Finally, there is little empirical evidence to support the assertion of social contacts in determining the transmission patterns of respiratory infections. Studies that simultaneously collect both complex contact network and microbiological data from participants can be used in future to model contact patterns and the risk of transmission of disease.
References


23. Kiti MC, Kinyanjui TM, Koech DC, Munywoki PK, Medley GF, Nokes DJ.


58. Kinyanjui TM, House TA, Kiti MC, Cane PA, Nokes DJ, Medley GF. Vaccine
doi:10.1371/journal.pone.0138018


doi:10.1093/infdis/141.1.98


82. Willem L, Van Kerckhove K, Chao DL, Hens N, Beutels P. A nice day for an


140. The DHS Program Website. [Internet]. 2018 [cited 3 Sep 2018]. Available: www.dhsprogram.com


145. Okiro EA, Nokes DJ. Transmission Dynamics of Respiratory Syncytial Virus within the Household and in the Community. Open University. 2007.


183. Munywoki PK, Koech DC, Agoti CN, Cane PA, Medley GF, Nokes DJ. Continuous Invasion by Respiratory Viruses Observed in Rural Households During a Respiratory Syncytial Virus Seasonal Outbreak in Coastal Kenya. Clin Infect Dis. 2018; doi:10.1093/cid/ciy313


Appendices

A. A snapshot of RSV candidate vaccine trials

This is the first approval from SERU. Subsequent approvals (up to December 2018) are available.
C. Ethical approval letter (BSREC, UK)

PRIVATE
Prof D J Nokes
Life Sciences
University of Warwick
Coventry
CV4 7AL

20th January 2019

Dear Professor Nokes,

Study Title and BSREC Reference: Using proximity and location tracking methods to define specific contact patterns relevant to the transmission of respiratory viruses
REGO-2016-1738

Thank you for submitting your revision to the above-named study to the University of Warwick’s Biomedical and Scientific Research Ethics sub-Committee for approval.

I am pleased to confirm that approval is granted and that your study may commence.

In undertaking your study, you are required to comply with the University of Warwick’s Research Data Management Policy, details of which may be found on the Research and Impact Services’ webpages, under “Codes of Practice & Policies” > “Research Data Management Policy”, at: http://www2.warwick.ac.uk/services/impact/researchdata/management/policies/researchdata_management_policy

You are also required to comply with the University of Warwick’s Information Classification and Handling Procedure, details of which may be found on the University’s Governance webpages, under “Governance” > “Information Security” > “Information Classification and Handling Procedure”, at: http://www2.warwick.ac.uk/services/governance/informationsecurity/handling

Investigators should familiarise themselves with the classifications of information defined therein, and the requirements for the storage and transportation of information within the different classifications:

Information Classifications:
http://www2.warwick.ac.uk/services/gov/informationsecurity/handling/classifications

Handling Electronic Information:
http://www2.warwick.ac.uk/services/gov/informationsecurity/handling/electronic/

Handling Paper or other media
http://www2.warwick.ac.uk/services/gov/informationsecurity/handling/paper/

Please also be aware that BSREC grants ethical approval for studies. The seeking and obtaining of all other necessary approvals is the responsibility of the investigator.
D. Frequently asked questions

This documents some of the frequently asked questions pertaining to the study.

i) How does the sensor work? What information does the sensor collect?

The sensor uses low power radio frequencies to detect when two or more sensors come close (≤1.5 metres apart) to each other. When a sensor comes close to another one, it records: the serial number of the other sensor(s), the time of the interaction, and the distance between the two sensors. This will enable us to tell who you meet, how close you were and how long the interaction was.

ii) Are the signals from these devices harmful to health?

No, these signals are not harmful to your health. Research done elsewhere has not shown any harmful effect of these signals even when you carry the devices for long periods of time.

iii) How should I carry the sensor?

The participant will decide whether to carry the sensor in a pouch/plastic case. The sensor should be carried in the front chest area (Figure 4) with the antennae always facing outwards. If the antennae faces the individual carrying it, no data will be collected.

iv) Do I have to carry my sensor all the time?

Yes. Please carry the sensor at all reasonable times during the entire duration of the study.

v) What happens if I lose my sensor?

If you lose your sensor, we will not have access to your data. Kindly try not to lose the sensor.

vi) What happens if I exchange my sensor with someone else?

If you exchange your sensor with another individual, it will give us wrong data. This will have a negative impact on how we interpret the information after analysis. Each sensor is to be carried only by the person to whom it is assigned.

vii) How can I ensure that I do not change my sensor with someone else?

We will try to give each individual a sensor with a special way of identifying it, such as a pouch with a different colour to the rest. However, we also encourage you to, for example, make a special mark on the outside cover so that you can remember your sensor.

viii) What happens if I meet people who have not worn a sensor?

If you meet someone without a sensor, no data will be collected. They may ask you questions about the sensor. Feel free to answer the questions as per the information we have given you. If this is not possible, then kindly give them our study number and we will call and respond to their questions.

ix) What happens if I move out of the study area?

The sensor will still work even if you move out of the study area for one or several days, but will only collect data when you come close to another individual with a sensor. In case you
will be away from the study area for a period longer than the study period, we will either not
give you the sensor, or we will arrange to pick the sensor before you leave.

x) What happens if the sensor comes into contact with water?

The sensor has been sealed in a waterproof plastic pouch; hence we do not expect it to get into
direct contact with water. Do not attempt to remove the sensor from the pouch at any time. The
tracker is waterproof; however, do not submerge it in water deliberately.

xi) If you have already conducted a similar study in Matsangoni, why are you doing it again?

The first study was smaller and involved only five homesteads and 100 individuals. We were
looking at how the community would react to using new technology for data collection, and if
the sensors actually collected data that we can use in our research. Now, we are expanding the
study to involve students and members of several homesteads to see how interactions happen in
different locations and times.
E. Consent forms

1. INFORMATION SHEET AND CONSENT FORM FOR INVOLVING HOUSEHOLD IN RESEARCH

STUDY TITLE: Using proximity and location tracking methods to define social contact patterns relevant to the transmission of respiratory viruses.

LAY TITLE: Understanding how people interact at school and at home and how it can lead to spread of infections.

<table>
<thead>
<tr>
<th>Institution</th>
<th>Investigators</th>
</tr>
</thead>
<tbody>
<tr>
<td>KEMRI-WTRP</td>
<td>Moses C. Kiti, James Nokes</td>
</tr>
<tr>
<td>Bocconi University</td>
<td>Alessia Melegaro</td>
</tr>
<tr>
<td>ISI Foundation</td>
<td>Ciro Cattuto</td>
</tr>
</tbody>
</table>

Who is carrying out this study and what is this study about?

This study is being carried out by KEMRI in collaboration with Bocconi University and Institute for Scientific Interchange (ISI) Foundation based in Italy. KEMRI is a government organization that carries out medical research to find better ways of preventing and treating illness in the future for everybody’s benefit.

KEMRI is currently doing research to learn more about how people mix at the home and in school and how this can affect the spreading of diseases such as pneumonia. The study will include all household members in a household who consent to participate.

Why do you want to talk to me and what does it involve?

As the head of the household, we would like to seek your permission to include your household in the study. Even if you give us permission to proceed, we will then discuss the study with all adults, parents of children and young people to seek their own consent because we would like them to have a chance to make their own decision on whether or not to participate in research.

We will discuss with you first about the study using a separate [consent] form and seek your consent, and then discuss with the other household members.

We are therefore asking for involvement of your household in the study, and for your permission to talk to members of your household.

What will happen if I refuse to participate?

All participation in research is voluntary. You are free to decide if you want your household to take part or not. If you do agree you can change your mind at any time.
without any consequences. If you refuse, then the household will not be recruited into the study.

**What if I have any questions?**

You are free to ask me any question about this research. If you have any further questions about the study, you are free to contact the research team using the contacts below:

**Mr Moses Chapa Kiti**, KEMRI Wellcome Trust Research Programme, P.O. Box 230, Kilifi. Telephone: 0702 221591 or 0722 203417, 0733 522063, 041 7522063

**If you want to ask someone independent anything about this research please contact:**

**Community Liaison Manager**, KEMRI – Wellcome Trust, P. O. Box 230, Kilifi. Telephone: 0723 342 780 or 041 7522 063

And

**The Secretary**, KEMRI Scientific and Ethics Review Unit (SERU), P. O. BOX 54840-00200, Nairobi, Tel number: 020 272 2541 Mobile: 0722 205 901 or 0733 400 003

[The following section is recommended, and should be signed by person undertaking informed consent.]

I have sought permission from the head/member of the household of ________________ [household name] to involve the household in the study. S/he has been given opportunity to ask questions which have been answered satisfactorily.

**Designee/investigator’s signature**

Nameitials: ____________________________ Date ____________

THE HOUSEHOLD HEAD SHOULD NOW BE GIVEN A SIGNED COPY TO KEEP
2. INFORMATION SHEET AND INFORMED CONSENT FOR ADULTS (≥18 years).

STUDY TITLE: Using proximity and location tracking methods to define social contact patterns relevant to the transmission of respiratory viruses.

LAY TITLE: Understanding how people interact at school and at home and how it can lead to spread of infections.

<table>
<thead>
<tr>
<th>Institution</th>
<th>Investigators</th>
</tr>
</thead>
<tbody>
<tr>
<td>KEMRI-WTRP</td>
<td>Moses C. Kiti, James Nokes</td>
</tr>
<tr>
<td>Bocconi University</td>
<td>Alessia Melegaro</td>
</tr>
<tr>
<td>ISI Foundation</td>
<td>Ciro Cattuto</td>
</tr>
</tbody>
</table>

Who is carrying out this study?

This study is being carried out by KEMRI in collaboration with Bocconi University and Institute for Scientific Interchange (ISI) Foundation based in Italy. KEMRI is a government organisation that carries out medical research to find better ways of preventing and treating illness in the future for everybody’s benefit.

What is this research study trying to find out and why are you being asked to participate?

KEMRI is currently doing research to learn more about how people mix at home or at school and how this can affect the spreading of diseases such as pneumonia. In this study we aim to find out how many people you meet, what their ages are, how much time you spend with them, where you meet, and what work they do. To do this, we will request you to carry a device that will detect when you come face-to-face and close enough to talk with someone else carrying a similar device, known as a sensor [show sensor to participant]. This will record the people that you have met, and for how long you remain close to the other person;

The study will involve 2 schools (1 in Kilifi Township, 1 in Matsangoni), and 90 households (45 in Kilifi Township, 45 in Matsangoni). From each school, 9 index students will be selected at random. You are being requested to participate because:

[If at the index household, select option (i); If at the neighbour, select option (ii); If school teacher, select (iii)]:

i) your child, [mention name], is a student at [mention name] Primary School

ii) your household shares a fence with [mention name of index household head] and you are thus considered their neighbour
iii) you are a member of staff at [mention name] Primary School

We are asking all members of this household/ all teachers at the school to participate in the study. We have asked for permission from the household head/ the headmaster to talk to each household member/ teacher, and now would like to tell you what this study is about.

If you decide to take part in this research, the following things will happen:

i) You will be given a pouch containing the sensor and the logger, and we will train you appropriately on how to use these.

ii) The sensor will be carried for 7 days

Are there any advantages of taking part?

Taking care of your health and wellbeing is important to us. There will be minimal disruption to the normal routine at school/ home during the initial engagement and training. We will schedule appointments with you at convenient times, such as early morning before you leave for your daily chores/ before you start your lessons or in between normal school breaks.

What will you gain from participating in this research study? How will you benefit from taking part in this study?

By taking part in this study, you will have helped us understand more the issues being studied and the knowledge gained can hopefully help other people in the future.

[For household residents at home]

There will be no individual benefit to adults at home. However, all school-going children in the household will be given a storybook that highlights how to prevent the spread of respiratory infections such as pneumonia. The research team will also spend time giving health talks at your household, including benefits of proper hygiene measures such as hand washing which may prevent the spread of respiratory and other infections. Your household will then receive two bars of hand washing soap.

[For teachers at school]

The research team will spend time giving health talks to you and the students, including benefits of proper hygiene measures such as hand washing which may prevent the spread of respiratory and other infections. Each class in the school will receive two bars of hand washing soap.

What happens to the data?

All individual names will be removed from the data collected and will be replaced by numbers to ensure that data can only be linked to the participants by people closely
concerned with the research. Most of the research analysis will be done here in Kilifi, and some will be conducted at the Bocconi University and the ISI Foundation in Italy.

In future, information collected or generated during this study may be stored and used to support new research by other researchers in Kenya and other countries on other health problems. Any access to the data by people outside the investigators and specific collaborators will require permissions from the principle investigator and a designated specialist subgroup of KEMRI-CGMRC CSC.

Who will have access to the information I give?

The information collected from this study will be stored in securely locked cabinets and password protected computers. This information will only be shared with people who are concerned with the research. The information will be summarised, and all the names of the participants will be removed from the documents. This study information may be used for future work; the information will only be provided after a national independent committee checks and agrees that you will not be affected in any way.

What will happen if I refuse to participate?

All participation in research is voluntary, and you are free to decide whether you want to take part or not. If you agree now, you can change your mind and stop participating in the future and no one will be upset with you. This will not affect your health care now or in the future.

What if you have any questions?

You are free to discuss your decision about taking part in this study with other people and you can ask to be given time to go and discuss this with them.

You are free to ask questions to any of the staff at any time. You can also contact the research team using these contacts:

Mr Moses Chapa Kiti, KEMRI Wellcome Trust Research Programme, P. O. Box 230, Kilifi. Telephone: 0702 221591 or 0722 203417, 0733 522063, 041 7522063

If you want to ask someone who is not related to this research about this work, please contact:

Community Liaison Manager, KEMRI Wellcome Trust Research Programme, P.O. Box 230, Kilifi. Telephone: 0723 342 780 or 041 7522 063

And

The Secretary, KEMRI/Scientific and Ethics Review Unit, P. O. BOX 54840-00200, Nairobi, Tel number: 020 272 2541 Mobile: 0722 205 901 or 0733 400 003
I, ______________________________ (name of Participant), have had the research explained to me. I have understood all that has been read/explained and had my questions answered satisfactorily.

Please insert the boxes below or add others where relevant

☐ Yes  please tick  I agree to take part in this research

☐ Yes  please tick  I agree to data being stored and used for future research

☐ Yes  please tick  I agree to this data being shared with other researchers.

I understand that I can change my mind at any time and it will not affect me in any way.

Subject’s signature: ___________________________  Date: _____________

Subject’s name: ___________________________  Time: _____________

(Please print name)

Only necessary if the participant cannot read:

I* attest that the information concerning this research was accurately explained to and apparently understood by the subject and that informed consent was freely given by the participant.

Witness’ signature: ___________________________  Date _____________

Witness’ name: ___________________________  Time _____________

(Please print name)

* A witness is a person who is independent from the trial or a member of staff who was not involved in gaining the consent.

Thumbprint of the subject as named above if they cannot write:

________________________________________________________________________

________________________________________________________________________

I have followed the study SOP to obtain consent from the [participant]. S/he apparently understood the nature and the purpose of the study and consents to participation in the study. S/he has been given opportunity to ask questions which have been answered satisfactorily.
Designee/investigator’s signature: ______________________ Date __________
Designee/investigator’s name: ______________________ Time __________
(Please print name)

THE PARTICIPANT SHOULD NOW BE GIVEN A SIGNED COPY TO KEEP
3. INFORMATION SHEET AND INFORMED CONSENT FOR PARENTS/GUARDIANS (participant ≤18 years).

STUDY TITLE: Using proximity and location tracking methods to define social contact patterns relevant to the transmission of respiratory viruses.

LAY TITLE: Understanding how people interact at school and at home and how it can lead to spread of infections.

<table>
<thead>
<tr>
<th>Institution</th>
<th>Investigators</th>
</tr>
</thead>
<tbody>
<tr>
<td>KEMRI-WTRP</td>
<td>Moses C. Kiti, James Nokes</td>
</tr>
<tr>
<td>Bocconi University</td>
<td>Alessia Melegaro</td>
</tr>
<tr>
<td>ISI Foundation</td>
<td>Ciro Cattuto</td>
</tr>
</tbody>
</table>

Who is carrying out this study?

This study is being carried out by KEMRI in collaboration with Bocconi University and Institute for Scientific Interchange (ISI) Foundation based in Italy. KEMRI is a government organisation that carries out medical research to find better ways of preventing and treating illness in the future for everybody’s benefit.

What is this research study trying to find out and why are you/your child being asked to participate?

KEMRI is currently doing research to learn more about how people mix at home or at school and how this can affect the spreading of diseases such as pneumonia. In this study we aim to find out how many people your child meets, what their ages are, how much time your child spends with them, and where they meet. To do this, we will request your child to carry a device that will detect when your child come face-to-face and close enough to talk with someone else carrying a similar device, known as a sensor [show sensor to participant]. This will record the people that your child has met, and for how long your child remains close to the other person;

The study will involve 2 schools: 1 in Kilifi Township (urban area) and 1 in Matsangoni (rural area), and 90 households (45 in Kilifi Township, 45 in Matsangoni). From each school, 9 index students will be selected at random. Your child has been selected because:

[If at the index household, select option (a);
If at the neighbour, select option (b)]:

(a) [mention name] is a student at [mention name] Primary School
(b) your household shares a fence with [mention name of index household head] and you are thus considered their neighbour

We are asking each parent/guardian in this household to allow their child/children to participate in the study.

If you agree to your child participating, the following things will happen:

(i) Your child will be given a pouch containing the sensor. We will train your child on how to use the sensor as appropriate.

Are there any advantages of taking part?

Taking care of your child’s health and wellbeing is important to us. There will be minimal disruption to the normal routine at school/home during the initial engagement and training. We will schedule appointments with you and your child at convenient times, such as early morning before you/your child leaves for his/her daily chores, or in between normal school breaks. When necessary, we will invite parents of selected students to the school for further discussions at suitable times. When this happens, we will reimburse your actual travel costs.

What will you gain from participating in this research study? How will your child benefit from taking part in this study?

By taking part in this study, your child will have helped us understand more of the issues being studied and the knowledge gained can hopefully help other children in the future.

All school-going children in the household will be given a storybook that highlights how to prevent the spread of respiratory infections such as pneumonia. The research team will also spend time giving health talks at your household, including benefits of proper hygiene measures such as hand washing which may prevent the spread of respiratory and other infections. Your household will then receive two bars of hand washing soap.

What happens to the data?

All individual names will be removed from the data collected and will be replaced by numbers to ensure that data can only be linked to the participants by people closely concerned with the research. Most of the research analysis will be done here in Kilifi, and some will be conducted at the Bocconi University and the ISI Foundation in Italy.

In future, information collected or generated during this study may be stored and used to support new research by other researchers in Kenya and other countries on other health problems. Any access to the data by people outside the investigators and specific collaborators will require permissions from the principle investigator and a designated specialist subgroup of KEMRI-CGMRC CSC.
Who will have access to the information I give?

The information collected from this study will be stored in securely locked cabinets and password protected computers. This information will only be shared with people who are concerned with the research. The information will be summarised, and all the names of the participants will be removed from the documents. This study information may be used for future work; the information will only be provided after a national independent committee checks and agrees that you will not be affected in any way.

What will happen if I refuse to participate?

All participation in research is voluntary, and you are free to decide whether you want your child to take part or not. If you agree now, you/your child can change your mind and stop participating in the future and no one will be upset with you. This will not affect you/your child’s health care now or in the future.

What if you/your child has any questions?

You/your child are free to discuss your decision about taking part in this study with other people and you can ask to be given time to go and discuss this with them.

You/your child are free to ask questions to any of the staff at any time. You/your child can also contact the research team using these contacts:

Mr Moses Chapa Kiti, KEMRI Wellcome Trust Research Programme, P. O. Box 230, Kilifi. Telephone: 0702 221591 or 0722 203417, 0733 522063, 041 7522063

If you/your child wants to ask someone who is not related to this research about this work please contact:

Community Liaison Manager, KEMRI Wellcome Trust Research Programme, P.O. Box 230, Kilifi. Telephone: 0723 342 780 or 041 7522 063

And

The Secretary, KEMRI/Scientific and Ethics Review Unit, P. O. BOX 54840-00200, Nairobi, Tel number: 020 272 2541 Mobile: 0722 205 901 or 0733 400 003

I, ___________________________ [name of parent], have had the research explained to me. I have understood all that has been read/explained and had my questions answered satisfactorily.

Please insert the boxes below or add others where relevant

☐ Yes please tick I agree to my child taking part in this research

☐ Yes please tick I agree to data being stored and used for future research
□ Yes  please tick  I agree to this data being shared with other researchers.

I understand that I can change my mind at any time and it will not affect me/ my child in any way.

I have followed the study SOP to obtain consent from the parent/guardian. S/he apparently understood the nature and the purpose of the study and consents to participation in the study. S/he has been given opportunity to ask questions which have been answered satisfactorily.

Designee/investigator’s signature: __________________________ Date ____________

Designee/investigator’s name: __________________________ Time ____________

Please print name)

THE PARTICIPANT SHOULD NOW BE GIVEN A SIGNED COPY TO KEEP
4. MINORS ASSENT DOCUMENT (participants aged >13 years and <18 years).

STUDY TITLE: Using proximity and location tracking methods to define social contact patterns relevant to the transmission of respiratory viruses.

LAY TITLE: Understanding how people interact at school and at home and how it can lead to spread of infections.

<table>
<thead>
<tr>
<th>Institution</th>
<th>Investigators</th>
</tr>
</thead>
<tbody>
<tr>
<td>KEMRI-WTRP</td>
<td>Moses C. Kiti, Irene Adema, James Nokes</td>
</tr>
<tr>
<td>Bocconi University</td>
<td>Alessia Melegaro</td>
</tr>
<tr>
<td>ISI Foundation</td>
<td>Ciro Cattuto</td>
</tr>
</tbody>
</table>

Who is carrying out this study?

This study is being carried out by KEMRI in collaboration with Bocconi University and ISI Foundation based in Italy. KEMRI is a government organisation that carries out medical research to find better ways of preventing and treating illness in the future for everybody’s benefit.

What is this research study trying to find out and why are you being asked to participate?

KEMRI is currently doing research to learn more about how people mix at home or at school and how this can affect the spreading of diseases such as pneumonia. In this study we aim to find out how many people you meet, what their ages are, how much time you spend with them, where you meet, and what work they do. To do this, we will request you to carry a device that will detect when you come face-to-face and close enough to talk with someone else carrying a similar device, known as a sensor. This will record the people that you have met, and for how long you remain close to each other.

[If child is at school, select option (a).]

If at household or not attending school, select option (b)]

(a) We are asking children of primary school in [Matsangoni location] or [Kilifi Township] to participate in the study. This study will involve 500 students from your school. We are asking if you would participate in this study.

(b) We are asking all children of this household to participate in the study.

What will happen if I refuse to participate?

We have spoken to your headteacher and parent about this research and they are aware that we are talking to you. All participation in research is voluntary, and you are free to decide whether you want to take part or not. If you agree now, you can change your
If you decide to take part in this research, the following things will happen:

(i) You will be given a pouch containing a sensor

(ii) The sensor will be carried for 7 days

Are there any advantages of taking part?

Taking care of your health and wellbeing is important to us. There will be minimal disruption to the normal routine at school/home during the initial engagement and training. We will visit schools during normal breaks (e.g. lunch time, sports break) or schedule appointments with you or your parents/guardians at convenient times.

What will you gain from participating in this research study? /How will you benefit from taking part in this study?

By taking part in this study, you will have helped us understand more the issues being studied and the knowledge gained can hopefully help other children in the future.

You will receive a stationery pack containing schoolbooks and writing material. You will also be given a storybook that highlights how you can prevent the spread of respiratory infections such as pneumonia. The research team will also spend time giving health talks at your school and household, including benefits of proper hygiene measures such as hand washing which may prevent the spread of respiratory and other infections. Each class or household will then receive two bars of hand washing soap.

What happens to the data?

All individual names will be removed from the data collected and will be replaced by numbers to ensure that data can only be linked to the participants by people closely concerned with the research. Most of the research analysis will be done here in Kilifi, and some will be conducted at the Bocconi University and the ISI Foundation in Italy.

In future, information collected or generated during this study may be stored and used to support new research by other researchers in Kenya and other countries on other health problems. Any access to the data by people outside the investigators and specific collaborators will require permissions from the principle investigator and a designated specialist subgroup of KEMRI-CGMRC CSC.

Who will have access to the information I give?

The information collected from this study will be stored in securely locked cabinets and password protected computers. This information will only be shared with people who are concerned with the research. The information will be summarised, and all the names of the participants will be removed from the documents. This study information may be
used for future work; the information will only be provided after a national independent committee checks and agrees that you will not be affected in any way.

**What if I have any questions?**

You are free to discuss your decision about taking part in this study with your parent/guardian or other people and you can ask to be given time to go and discuss this with them.

You are free to ask questions to any of the staff at any time. You can also contact the research team using these contacts:

**Mr Moses Chapa Kiti**, KEMRI Wellcome Trust Research Programme, P. O. Box 230, Kilifi. Telephone: 0702 221591 or 0722 203417, 0733 522063, 041 7522063

If you want to ask someone who is not related to this research about this work please contact:

**Community Liaison Manager**, KEMRI Wellcome Trust Research Programme, P.O. Box 230, Kilifi. Telephone: 0723 342 780 or 041 7522 063

And

**The Secretary**, KEMRI/Scientific and Ethics Review Unit, P. O. BOX 54840-00200, Nairobi, Tel number: 020 272 2541 Mobile: 0722 205 901 or 0733 400 003

--------------------------------------------------

I have followed the study SOP to obtain consent from the [participant]. S/he apparently understood the nature and the purpose of the study and consents to participation in the study. S/he has been given opportunity to ask questions which have been answered satisfactorily.

Designee/investigator’s signature: ___________________________ Date ___________

Designee/investigator’s name: ___________________________ Time ___________

(Please print name)

**THE PARTICIPANT SHOULD NOW BE GIVEN A SIGNED COPY TO KEEP**
5. PERMISSION FOR INVOLVING SCHOOL IN RESEARCH

STUDY TITLE: Using proximity and location tracking methods to define social contact patterns relevant to the transmission of respiratory viruses.

LAY TITLE: Understanding how people interact at school and at home and how it can lead to spread of infections.

<table>
<thead>
<tr>
<th>Institution</th>
<th>Investigators</th>
</tr>
</thead>
<tbody>
<tr>
<td>KEMRI-WTRP</td>
<td>Moses C. Kiti, James Nokes</td>
</tr>
<tr>
<td>Bocconi University</td>
<td>Alessia Melegaro</td>
</tr>
<tr>
<td>ISI Foundation</td>
<td>Ciro Cattuto</td>
</tr>
</tbody>
</table>

Who is carrying out this study and what is this study about?

This study is being carried out by KEMRI in collaboration with Bocconi University and ISI Foundation based in Italy. KEMRI is a government organization that carries out medical research to find better ways of preventing and treating illness in the future for everybody’s benefit.

KEMRI is currently doing research to learn more about how people mix at the home and in school and how this can affect the spreading of diseases such as pneumonia. In this study we aim to find out how many people each selected student and staff member of your school meets, what their ages are, where they met, and how much time they spent together.

Why do you want to talk to me and what does it involve?

As the Head of the school, we would like to seek your permission to include your school in the study. Even if you give us permission to proceed, we will then discuss the study with all students and their parents to seek their own consent because we would like them to have a chance to make their own decision on whether or not to participate in research.

We are therefore asking for involvement of your school in the study, and for your permission to talk to students and staff of your school.

What will it involve for my school?

In order to select to students who will participate in this research, we request you to provide us with the students register.

From the school register, we will randomly select half of the students from each grade (e.g. grade 1, 2…8) to participate in this research. We will also request all teachers and staff to participate.
On a day and time to be agreed upon, we will visit the school to explain to the students and teachers about the study. We will contact the parents/guardians of all the selected students at an appropriate location (e.g. school or home) to get parental/guardian consent.

Students who give assent and also have parental consent will carry a sensor enclosed in a pouch. This will be carried around the neck for 7 days. The first day for all students will be selected at random by the Headteacher so that data can be collected within the same period. The same will be done for staff who give consent.

Participants aged 5 years and above will also carry a logger for two days. The Headteacher will select these days at random.

When collecting the loggers after the study, participants will be asked a set of questions detailing how they spent their time, where, with whom, and for how long.

We will also select 9 other students (3 each from kindergarten, class 1-4 and 5-8) whom we will follow up to their households.

**What will happen if I refuse to participate?**

All participation in research is voluntary. You are free to decide if you want your school to take part or not. If you do agree you can change your mind at any time without any consequences. If you refuse, then the school will not be recruited into the study.

**What if I have any questions?**

You are free to ask me any question about this research. If you have any further questions about the study, you are free to contact the research team using the contacts below:

**Mr Moses Chapa Kitii**, KEMRI Wellcome Trust Research Programme, P.O. Box 230, Kilifi. Telephone: 0702 221591 or 0722 203417, 0733 522063, 041 7522063

If you want to ask someone independent anything about this research, please contact:

**Community Liaison Manager**, KEMRI – Wellcome Trust, P. O. Box 230, Kilifi. Telephone: 0723 342 780 or 041 7522 063

And

**The Secretary**, KEMRI Scientific and Ethics Review Unit (SERU), P. O. BOX 54840-00200, Nairobi, Tel number: 020 272 2541 Mobile: 0722 205 901 or 0733 400 003

[Following section is recommended, and should be signed by person undertaking informed consent.]

I have sought permission from the Headteacher of __________________________ Primary School to involve the school in the study. S/he has been given opportunity to ask questions which have been answered satisfactorily.
These other approvals may include, but are not limited to:

1. Any necessary agreements, approvals, or permissions required in order to comply with the University of Warwick's Financial Regulations and Procedures.
2. Any necessary approval or permission required in order to comply with the University of Warwick's Quality Management System and Standard Operating Procedures for the governance, acquisition, storage, use, and disposal of human samples for research.
3. All relevant University, Faculty, and Divisional/Departmental approvals, if an employee or student of the University of Warwick.
4. Approval from the applicant's academic supervisor and course module leader (as appropriate), if a student of the University of Warwick.
5. NHS Trust R&D Management Approval for research studies undertaken in NHS Trusts.
6. NHS Trust Clinical Audit Approval for clinical audit studies undertaken in NHS Trusts.
7. Approval from Departmental or Divisional Heads, as required under local procedures, within Health and Social Care organisations hosting the study.
8. Local ethical approval for studies undertaken overseas, or in other HE institutions in the UK.
9. Approval from Heads (or delegates thereof) of UK Medical Schools, for studies involving medical students as participants.
10. Permission from Warwick Medical School to enrol medical students or medical student data for research of evaluation purposes.
11. NHS Trust Caldicott Guardian Approval, for studies where identifiable data is being transferred outside of the direct clinical care team. Individual NHS Trust procedures vary in their implementation of Caldicott guidance, and local guidance must be sought.
12. Any other approval required by the institution hosting the study, or by the applicant's employer.

There is no requirement to supply documentary evidence of any of the above to BSREC, but applicants should hold such evidence in their Study Mission File for University of Warwick auditing and monitoring purposes. You may be required to supply evidence of any necessary approvals to other University functions, e.g. The Finance Office, Research & Impact Services (RUS), or your Department/School.

May I take this opportunity to wish you success with your study, and to remind you that any Substantial Amendments to your study require approval from BSREC before they may be implemented.

Yours sincerely,

[Signature]

Professor Scott Welch
Chair
Biomedical and Scientific Research Ethics Sub-Committee

Biomedical and Scientific Research Ethics Sub-Committee
A010 Medical School Building
Warwick Medical School
Coventry, CV4 7AL
T: 024 76-522707
E: BSREC@warwick.ac.uk

http://www2.warwick.ac.uk/faculty/smed/researchethics/committees/bsrec/
### F. Demographic distribution of participants in the rural and urban households

<table>
<thead>
<tr>
<th>Gender</th>
<th>Number of participants</th>
<th>Rural</th>
<th>Urban</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>111</td>
<td></td>
<td>17</td>
</tr>
<tr>
<td>Male</td>
<td>69</td>
<td></td>
<td>17</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Age group</th>
<th>Number of participants</th>
<th>Median age</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-4</td>
<td>18</td>
<td>8.4 (6.9-9.2)</td>
</tr>
<tr>
<td>5-14</td>
<td>69</td>
<td>10.4 (9.4-11.6)</td>
</tr>
<tr>
<td>15-19</td>
<td>27</td>
<td>11.7 (10.2-12.5)</td>
</tr>
<tr>
<td>20-49</td>
<td>31</td>
<td>13.0 (12.0-14.3)</td>
</tr>
<tr>
<td>50+</td>
<td>35</td>
<td>15.0 (13.9-16.0)</td>
</tr>
</tbody>
</table>

### G. Demographic distribution of participants in the rural and urban schools

<table>
<thead>
<tr>
<th>Grade</th>
<th>Rural</th>
<th>Median age</th>
<th>Urban</th>
<th>Median age</th>
</tr>
</thead>
<tbody>
<tr>
<td>KG</td>
<td>21</td>
<td>8.4 (6.9-9.2)</td>
<td>5</td>
<td>4.7 (4.3, 4.9)</td>
</tr>
<tr>
<td>1</td>
<td>16</td>
<td>9.5 (8.9-9.9)</td>
<td>5</td>
<td>7.2 (6.4, 7.8)</td>
</tr>
<tr>
<td>2</td>
<td>23</td>
<td>10.4 (9.4-11.6)</td>
<td>6</td>
<td>7.8 (7.4, 8.3)</td>
</tr>
<tr>
<td>3</td>
<td>19</td>
<td>11.7 (10.2-12.5)</td>
<td>4</td>
<td>9.5 (8.7, 10.0)</td>
</tr>
<tr>
<td>4</td>
<td>26</td>
<td>13.0 (12.0-14.3)</td>
<td>12</td>
<td>9.4 (9.0, 9.7)</td>
</tr>
<tr>
<td>5</td>
<td>26</td>
<td>13.8 (12.8-14.6)</td>
<td>4</td>
<td>10.5 (10.1, 10.9)</td>
</tr>
<tr>
<td>6</td>
<td>29</td>
<td>15.0 (13.9-16.0)</td>
<td>10</td>
<td>11.8 (10.9, 12.5)</td>
</tr>
<tr>
<td>7</td>
<td>38</td>
<td>16.5 (15.0-17.6)</td>
<td>11</td>
<td>12.3 (12.1, 13.1)</td>
</tr>
<tr>
<td>8</td>
<td>25</td>
<td>17.2 (16.2-18.2)</td>
<td>10</td>
<td>14.4 (14.0, 14.7)</td>
</tr>
<tr>
<td>Overall</td>
<td>223</td>
<td>13.5 (13.0, 13.1)</td>
<td>65</td>
<td>10.5 (9.7, 11.2)</td>
</tr>
</tbody>
</table>