Online Appendix

Does Precise Case Information Limit Precautionary Behavior? Evidence from COVID-19 in Singapore

Aljoscha Janssen* Matthew Shapiro†

June 18, 2020

Contents

1 Data Manipulation **2**

1.1 Location Data Description 2

1.2 Data Cleaning 3

1.3 Geography Data 6

1.4 Generating Outcome Data 8

2 Additional Analysis **9**

2.1 Full Regression Models 9

2.2 Excluding Dates 11

2.3 Definition of Local Cases 11

*Singapore Management University and IFN, ajanssen@smu.edu.sg, School of Economics, Singapore
†Singapore Management University, mshapiro@smu.edu.sg, School of Economics, Singapore

1 Data Manipulation

1.1 Location Data Description

Lifesight provided our principal dataset, granular location information for individuals over long time periods. Each observation in this main dataset is an individual ID, unique to the phone; a timestamp; and a lat-long coordinate for that person at that time. Though they provided data for January through March 2020 with parts of 2019, our main analysis uses the 2020 data through March 17$^{\text{th}}$, the end of our period of study. Table 1 reports raw observation counts — after dropping duplicates — by month.

Table 1: **Basic Monthly Summary**

Time Period	Observations	ID-Day Count	Unique IDs
Jan 2019	311,021,056	6,909,862	1,295,054
Feb 2019	209,078,692	4,222,157	873,720
Oct 2019	195,046,083	8,789,171	1,894,819
Nov 2019	626,863	64,894	64,881
Dec 2019	171,326,775	7,131,877	1,722,650
Jan 2020	405,059,907	11,180,595	2,312,115
Feb 2020	512,084,761	12,502,090	2,492,687
Mar 2020	642,199,102	10,292,256	1,934,968

Note: We include the summary for data for months in which we have partial data, most extreme November 2019, though our analysis only includes the 2020 data for which we have complete information.

Our dataset is not balanced; individuals do not appear in our coverage period daily nor do they provide locations on a consistent basis. In our empirical analysis we assume that any data missing for an individual is not a function of their behavior. That is, the data available for an individual should be representative, if not completely, of their movement within the day. Table 2 provides a deeper look at observations available by ID (person), the first section on unique people per day for a given month and the second section on the number of observations per person-day. The first section makes clear that ID counts alone are not sufficient as measures of activity volume. There is significant variance both across months and within months in terms of people identified in the dataset. The second demonstrates

the wide variance in observations per person-day; the data are highly skewed right, though there is not much variance in the number of times we see a person, conditional on that individual. This latter point supports our usage of models with individual fixed effects to capture inherent differences in observation frequency across individuals as well as focusing on day-level outcomes rather than putting significant weight on individual observations.

Table 2: **Summary Statistics on IDs**

	Unique IDs by Day		Daily Observations by ID		
Time Period	Average	Variance	Average	Median	Avg Deviation
Jan 2019	222,865	55,455	45.02	6	0.019
Feb 2019	150,778	31,214	49.52	6	−0.043
Oct 2019	283,492	101,562	22.19	2	−0.022
Nov 2019	2,163	11,823	9.66	1	−0.609
Dec 2019	230,042	151,142	24.02	2	−0.077
Jan 2020	360,639	74,252	36.23	3	−0.016
Feb 2020	431,047	66,220	40.97	3	0.016
Mar 2020	331,960	89,307	62.41	6	0.077

Note: The average "deviation" in daily observations by ID (person) is constructed by first calculating for each individual how many observations they produce per day deviates from their average (standard score) and averaging this score across all individuals for the month.

1.2 Data Cleaning

We take several passes at cleaning the raw dataset and remove observations that fit into one of several categories in the following order:

[1] Inaccurate longitude-latitude data

[2] People linked to devices with unrealistic travel behavior

Category 1 observations arise from issues with how the data is collected. One issue is that in the event that the precise location of a person cannot be determined, they might be "resolved" to a particular lat-long based on a guessed IP address. In the data this manifests as an unrealistic number of people in the exact same long-lat coordinates. For long-lat where we observe more than 500 unique perople per week, we remove all observations associated

with the long-lat for that particular week. Additionally, GPS data is occasionally collected with accuracy at a such low resolution it is useless for our purposes. We hence remove specific observations where the horizontal accuracy of a GPS reading is higher than 250m.

For category 2 observations we carry out simple operations on the data to calculate an individual's travel distance and speed between successive pings. Travel distances, like speeds, are not directly observable in the data. To calculate distance we take the aerial distance between two successive pings for an individual. To calculate speed between two pings, we take the distance and divide it by the elapsed time. This speed calculated is only a lower bound on speed and hence appropriate to check for unrealstically fast travel speeds.

What we classify as category 2 observations include people with excessive travel — more than 100 km per day — or unrealistic travel speeds, which we calculate as the distance traveled over the time elapsed between two location observations — over 140 kph, which is far greater than the highest speed limit within Singapore of 90 kph — in which case we remove observations for that person-day. This category also includes people with insufficient movement — individuals who do not change their long-lat position for the duration they are in the sample — in which case we remove that person entirely.

Table 3 illustrates how much data we drop through this procedure. Generally, about a third of observations are dropped per month, but the number of devices, or people, kept in the sample drops to a quarter. Most of these devices are dropped because they exhibited little movement of the life we observe them in the sample. Unfortunately, we cannot distinguish between collection errors or genuine lack of movement. Given the high bar we set for removing these observations — essentially no movement every single day — it seems reasonable that in the worst case that the data collection is correct for some of these people, they do not represent a large fraction of the populace. Recall that we have data prior to knowledge of the pandemic so these true cases would not be individuals exhibiting extreme isolation. This change is more obvious in Table 4 where the average daily observations per device increases significantly for the months of our study.

Finally, as described in the main text, part of the analysis is conducted on a subsample

Table 3: **Summary of Basic Cleaned Data**

Time Period	No Conditions			Cleaned Data		
	Observations	ID-Day Count	Unique IDs	Observations	ID-Day Count	Unique IDs
Jan 2019	311,021,056	6,909,862	1,295,054	210,125,557	4,431,761	493,989
Feb 2019	209,078,692	4,222,157	873,720	130,457,919	2,836,238	390,246
Oct 2019	195,046,083	8,789,171	1,894,819	141,109,196	2,990,717	490,593
Nov 2019	626,863	64,894	64,881	429,868	15,937	15,927
Dec 2019	171,326,775	7,131,877	1,722,650	118,353,840	2,285,094	411,184
Jan 2020	405,059,907	11,180,595	2,312,115	287,106,021	4,140,000	546,178
Feb 2020	512,084,761	12,502,090	2,492,687	329,464,674	4,762,227	569,803
Mar 2020	642,199,102	10,292,256	1,934,968	413,326,755	4,616,688	457,482

Table 4: **ID Statistics for Cleaned Data**

Time Period	No Conditions				Cleaned Data			
	Unique Daily IDs		Daily Observations by ID		Unique Daily IDs		Daily Observations by ID	
	Average	Variance	Average	Avg Deviation	Average	Variance	Average	Avg Deviation
Jan 2019	222,865	55,455	45.02	0.019	142,931	25,847	47.42	−0.009
Feb 2019	150,778	31,214	49.52	−0.043	101,283	17,156	46.00	−0.096
Oct 2019	283,492	101,562	22.19	−0.022	96,454	48,280	47.19	0.106
Nov 2019	2,163	11,823	9.66	−0.609	637	3,173	26.97	−0.666
Dec 2019	230,042	151,142	24.02	−0.077	73,706	62,134	51.80	−0.138
Jan 2020	360,639	74,252	36.23	−0.016	133,538	40,266	69.35	0.005
Feb 2020	431,047	66,220	40.97	0.016	164,172	41,679	69.20	0.030
Mar 2020	331,960	89,307	62.41	0.077	148,891	28,568	89.55	0.034

Note: The average "deviation" in daily observations by ID (person) is constructed by first calculating for each individual how many observations they produce per day deviates from their average (standard score) and averaging this score across all individuals for the month.

of the data for which we have estimates of a person's home. Lifesight created these estimates based on regular pings from a person's phone during off hours, like the early morning or late evening. Table 5 recreates the first two panels of the same table from the main paper using this slice of the data. The additional restriction maintains about 80 to 90% of the original sample, though on about two-thirds of individuals. While we have no reason to believe the home estimation is selects individuals in a way that is biased, individuals who ping more will be more likely to have a home estimate. This is evident from the higher observations per day in this subsample. We also highlight that the principle travel statistics we share in Panel B are roughly the same across the two subsamples. Note that in Section 2.3 of this appendix we additionally test whether results from the main analysis are robust to this home definition or subsample use; we find that they are robust.

Table 5: **Data Summary**

	Jan 2020	Feb 2020	Mar 2020
Panel A: Cell Phone Data			
Person-Day Count	3,425,997	3,818,295	2,244,850
Unique People	332,256	362,912	264,570
Avg Obs Per Person-Day	78.04	81.71	104.84
	(131.96)	(157.51)	(145.62)
Panel B: Travel Statistics			
Avg KM Traveled Per Day	20.00	14.29	16.95
	(25.90)	(22.86)	(24.70)
Avg % Staying Home	22.85	27.81	26.43
	(0.18)	(0.15)	(0.10)
Avg Areas Visited Per Day	2.95	2.09	2.81
	(2.92)	(1.97)	(2.76)

Note 1: Data for March 2020 only covers through the 17^{th}, the end of our period of study. The standard deviation for select averages are presented in parentheses.
Note 2: Panels A and B use the a subsample of the data with home estimates available.

1.3 Geography Data

These notes cover the process of linking Lifesight longitude-latitude data to specific places in Singapore, typically known as reverse geocoding. Reverse geocoding is easily possible by mapping services like Google but is prohibitive for over hundreds of millions of observations, and does not produce results necessarily useful for our analysis. The same concerns apply to open source solutions like Nominatim through Open Street Maps.

The methodology presented here instead uses more standard point-in-polygon identification using place and location data from Open Street Maps. The data provided lists building and land area common names and use classifications. We use the latter to determine if particular areas are of a commercial, industrial, retail, residential, etc. nature.

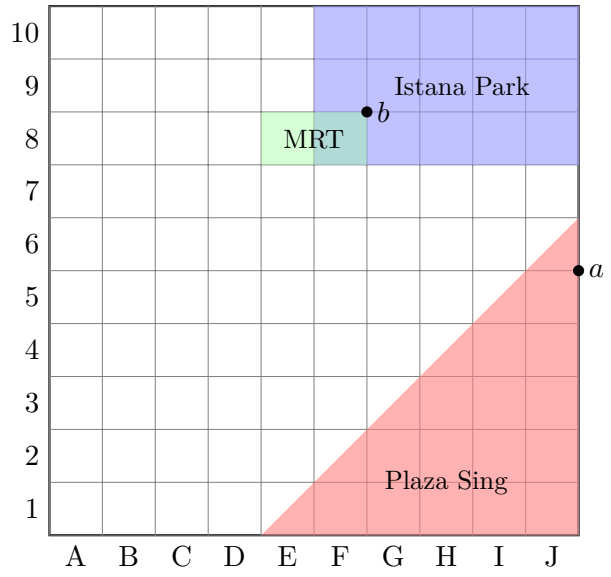
An issue that is not resolved with this methodology is that these places can overlap. In dense cities like Singapore there are many multi-use buildings, and horizontal GPS coordinates alone are not sufficient to identify what part of the building the individual uses. An additional trade off is that buildings and land-use records do not form a complete cover of

Singapore.

Figure 1 depicts a highly stylized map of Singapore with some common challenges. The Istana Park, Plaza Singapura, and a mass rapid transit (MRT) station are included along with two sample location points, denoted a and b. Notice the Istana and MRT overlap in one square. White space is meant to depict areas without any specific place information available.

A simple geocoding operation would indicate location a is in Plaza Singapura. Location b cannot be assigned to a particular place, as it is contained within both Istana Park and the MRT station. In the main analysis of this paper we are somewhat agnostic in the specific locations of a and b in that we do not try to claim if b was in the park or the MRT station. As we describe in the next subsection, if the person is ever in a land area with a particular classification type, we claim that person visited that classification type that type. Location a would be tagged as retail. Location b would be tagged as both transit and recretational, though we do not use either of the latter classifications in the main analysis of the paper.

Figure 1: **Sample Geography**



1.4 Generating Outcome Data

Our outcome variable for individual i at time t with a home location in subzone j of area k are gathered in the set a_{it}. They include travel distance in meters ($TravelDist_{ijkt}$); a dummy which takes the value one if i stays within the subzone of their home ($StayHome_{ijkt}$); a dummy which takes the value one if i visits an area with an industrial-, commercial-, or retail-use classification ($IndRetCom_{ijkt}$); a dummy which takes the value one if i visits an with a residential-use classification ($Residential_{ijkt}$) outside the own home. In the following we provide a brief description on the construction of the summary variables:

1. $TravelDist_{ijkt}$: For each individual and day we order the device signals according to time. We proceed to calculate the aerial distance. Finally, $TravelDist_{ijkt}$ is the sum of these distances during a day.

2. $StayHome_{ijkt}$: For each individual and day we consider if two conditions are met. First, the individual needs to send signals from only one subzone. Secondly, the signals have to be within the subzone of their home. If both conditions are fulfilled, $StayHome_{ijkt}$ takes the value one.

3. $IndRetCom_{ijkt}$: For each individual and day we consider if an individual has sent at least one signal during the day from within a land class (see Section 1.3 of the Appendix) that is defined as industrial, retail or commercial.

4. $Residential_{ijkt}$: For each individual and day we consider if an individual has sent at least one signal during the day from within a land class (see Section 1.3 of the Appendix) that is defined as residential. To avoid that we simply count individuals when they are at home, we exclude observations that are extremely close to their home location. Specifically we exclude observations that have a long-lat coordinate identical to their home's up to the fourth decimal place.

2 Additional Analysis

2.1 Full Regression Models

In the following section we present additional regression evidence for model 1, the outward travel regressions, of the main paper. In detail, each of the following tables presents one outcome variable of the set a_{it}: travel distance in meters ($TravelDist_{ijkt}$); a dummy which takes the value one if i stays within the subzone of their home ($StayHome_{ijkt}$); a dummy which takes the value one if i visits an area with an industrial-, commercial-, or retail-use classification ($IndRetCom_{ijkt}$); a dummy which takes the value one if i visits an with a residential-use classification ($Residential_{ijkt}$) outside the own home. For each outcome variable the alternative specifications in each table drop fixed effects and run a naive pooled regression in specification (1), include only date fixed effect in specification (2), only individual fixed effects in specification (3, and both in specification (4).

Table 6: **Estimation of Local and General Response, TravelDist**

	TravelDist			
	(1)	(2)	(3)	(4)
$LocalCases_{jt-1}$	456.873^{***}	426.596^{***}	-124.491^{***}	-61.433^{***}
	(24.014)	(23.718)	(14.374)	(14.429)
$RegionCases_{jt-1}$	-361.327^{***}	-532.699^{***}	-91.125^{***}	-28.045^{***}
	(10.660)	(13.691)	(5.713)	(6.776)
Individual FE	No	No	Yes	Yes
Date FE	No	Yes	No	Yes
N	9,482,376	9,482,376	9,482,376	9,482,376

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Notes: The table presents results of regression model (1). One observation corresponds to an individual on a specific date. Each model specification corresponds to the outcome variable of $TravelDist$, travel distance in meters. $LocalCases$ are the number of local cases in a subregion announced in the evening of $t-1$. $RegionCases$ are the cases of the region announced. Standard errors are reported in parentheses and clustered at the individual level.

Table 7: **Estimation of Local and General Response, HomeStay**

	HomeStay			
	(1)	(2)	(3)	(4)
$LocalCases_{jt-1}$	0.021^{***}	0.018^{***}	0.003^{***}	0.001^{***}
	(0.001)	(0.001)	(0.0003)	(0.0003)
$RegionCases_{jt-1}$	-0.004^{***}	-0.013^{***}	0.004^{***}	0.0001
	(0.0002)	(0.0003)	(0.0001)	(0.0002)
Individual FE	No	No	Yes	Yes
Date FE	No	Yes	No	Yes
N	9,482,376	9,482,376	9,482,376	9,482,376

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Notes: The table presents results of regression model (1). One observation corresponds to an individual on a specific date. Each model specification corresponds to the outcome variable $HomeStay$, a dummy variable that takes the value one if an individual remains at their home subzone for an entire day. $LocalCases$ are the number of local cases in a subregion announced in the evening of $t-1$. $RegionCases$ are the cases of the region announced. Standard errors are reported in parentheses and clustered at the individual level.

Table 8: **Estimation of Local and General Response, IndComRet**

	IndComRet			
	(1)	(2)	(3)	(4)
$LocalCases_{jt-1}$	0.709^{***}	0.614^{***}	-0.166^{***}	-0.117^{***}
	(0.053)	(0.052)	(0.035)	(0.035)
$RegionCases_{jt-1}$	-0.429^{***}	-0.609^{***}	-0.099^{***}	-0.083^{***}
	(0.022)	(0.029)	(0.013)	(0.016)
Individual FE	No	No	Yes	Yes
Date FE	No	Yes	No	Yes
N	9,482,376	9,482,376	9,482,376	9,482,376

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Notes: The table presents results of regression model (1). One observation corresponds to an individual on a specific date. Each model specification corresponds to the outcome variable $IndComRet$, a dummy variable that takes the value one if an individual enters at least one industrial, commercial or retail area. $LocalCases$ are the number of local cases in a subregion announced in the evening of $t-1$. $RegionCases$ are the cases of the region announced. Standard errors are reported in parentheses and clustered at the individual level.

Table 9: **Estimation of Local and General Response, Residential**

	Residential			
	(1)	(2)	(3)	(4)
$LocalCases_{jt-1}$	3.462^{***}	3.116^{***}	-0.097^{***}	-0.055^{**}
	(0.044)	(0.044)	(0.026)	(0.026)
$RegionCases_{jt-1}$	-1.208^{***}	-2.617^{***}	0.325^{***}	-0.029^{**}
	(0.022)	(0.028)	(0.011)	(0.013)
Individual FE	No	No	Yes	Yes
Date FE	No	Yes	No	Yes
N	9,482,376	9,482,376	9,482,376	9,482,376

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Notes: The table presents results of regression model (1). One observation corresponds to an individual on a specific date. Each model specification corresponds to the outcome variable *Residential*, a dummy variable that takes the value one if an individual enters a residential area except their own residence. $LocalCases$ are the number of local cases in a subregion announced in the evening of $t-1$. $RegionCases$ are the cases of the region announced. Standard errors are reported in parentheses and clustered on the individual level.

2.2 Excluding Dates

As we observe data anomalies on the 4th of February, this section shows results from our outflow regression analysis without observations from that date. Table 10 replicates the outward flow analysis summary from the main analysis. In comparison those results here we observe a slight change in the coefficients but not substantial enough to alter our interpretations of the results.

2.3 Definition of Local Cases

In our main analysis we study the travel behavior of individuals by considering COVID-19 cases close to an individual's home location, estimated by Lifesight. While we trust the residence estimates, the analysis necessarily drops individuals for whom home location estimates are not available. In addition, as we argue in the paper, individuals may not necessarily consider geographical distance from announced cases as much as the risk of potential contact with infected individuals. Thus, individuals may change their behavior not only for cases closes to home but also if they have visited areas outside home where a positive COVID-19

Table 10: **Estimation of Local and General Response without the 4th of February**

	TravelDist	StayHome	IndComRet	Residential
	(1)	(2)	(3)	(4)
$LocalCases_{jt-1}$	-59.918^{***}	0.132^{***}	-0.114^{***}	-0.047^{*}
	(14.445)	(0.034)	(0.035)	(0.026)
$AggregateCases_{jt-1}$	-32.771^{***}	0.019	-0.088^{***}	-0.037^{***}
	(6.792)	(0.015)	(0.016)	(0.013)
Individual FE	Yes	Yes	Yes	Yes
Date FE	Yes	Yes	Yes	Yes
Mean Outcome	13901	25.72	29.05	78.72
Mean Local Effect in Percent	-0.43	0.51	-0.39	-0.06
Mean Aggregate Effect in Percent	-0.24	0.07	-0.3	-0.05
N	9,306,704	9,306,704	9,306,704	9,306,704

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table presents results of regression model (1) excluding observations from the 4th of February. One observation corresponds to an individual on a specific date. Each model specification corresponds to a different outcome variable. $TravelDist$ is the travel distance in meters, $StayHome$ is a dummy variable that takes the value one if an individual remains at their home subzone for an entire day. $IndComRet$ is a dummy that takes the value one if an individual enters at least one industrial, commercial, or retail area. $Residential$ is a dummy that takes the value one if an individual enters a residential area excluding their own residence. Note, that we multiply outcome variables $StayHome$, $IndComRet$ and $Residential$ by 100 so that the coefficients are interpreted as percentage points. $LocalCases$ are the number of local cases in a subregion announced in the evening of $t-1$. $AggregateCases$ are the cases of the region announced. For all models we include individual and date fixed effects. We calculate the mean local effect and mean aggregate effect as the percentage difference from the average outcome. Standard errors are reported in parentheses and clustered at the individual level.

case has been announced.

Within this section we show the robustness of our results to specifications taking into account both of these considerations. We consider all observations, independent of their home estimate, and further build a new measure of local cases. We specifically consider cases announced in subzones where an individual has travelled during the last five days. We do not restrict that the subzone is within a specific region. As individual travel may span over different regions and as use date fixed effects to control for national trends in travel behavior, we are not able to identify local and aggregate responses separately. Nevertheless, this approach offers the possibility to show robustness of our local case estimates. The regression model is comparable to equation (1) in the main article:

$$\mathbf{a}_{ijt} = \beta_1 LocalCases_{jt-1} + \gamma_i + \rho_t + \varepsilon_{ijt}, \tag{1}$$

where a_{it} is a set of outcome variables: travel distance in meters ($TravelDist_{ijt}$); a dummy which takes the value one if i visits an area with an industrial-, commercial-, or retail-use classification ($IndRetCom_{ijt}$); a dummy which takes the value one if i visits an area with a residential-use classification ($Residential_{ijt}$). Note that we exclude the outcome $StayHome_{ijkt}$ since we consider the entire sample, including individuals without a home estimate. $LocalCases_{jt-1}$ are the sum of the number of cases of those subzones j announced in the evening of $t-1$ that an individual has visited during the last five days.

We present results of this alternative model in Table 11. We find results quantitatively similar to those in the main paper for the effect of case announcements on travel distance; the probability of entering at least one industrial, retail, or commercial area in a day; and on the probability of entering a residential area.

Table 11: **Estimation of Local Response, New Definition**

	TravelDist	IndComRet	Residential
	(1)	(2)	(3)
$LocalCases_{jt-1}$	-61.747^{***}	-0.090^{***}	-0.072^{***}
	(5.281)	(0.011)	(0.009)
Individual FE	Yes	Yes	Yes
Date FE	Yes	Yes	Yes
Mean Local Effect in Percent	-0.5	-0.33	-0.1
N	11,294,646	11,072,665	11,072,665

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Notes: The table presents results of regression model without considering the home location estimate as well as a new definition of local cases. One observation corresponds to an individual on a specific date. Each model specification corresponds to a different outcome variable. $TravelDist$ is the travel distance in meters, $IndComRet$ is a dummy that takes the value one if an individual enters at least one industrial, commercial or retail area. *Residential* is a dummy that takes the value one if an individual enters a residential area. Note, that we multiply outcome variables $IndComRet$ and *Residential* by 100 such that the coefficients are interpreted in percentage points. $LocalCases$ are the number of local cases in a subregion announced in the evening of $t-1$ which an individual has visited during the last five days. For all models we include individual and date fixed effects. We calculate the mean local effect as the percentage difference from the average outcome. Standard errors are reported in parentheses and clustered at the individual level.