Substantial transmission of SARS-CoV-2 through casual contact in retail stores: Evidence from matched administrative micro-data on card payments and testing

Niels Johannesen (Oxford University and University of Copenhagen)
Alessandro Tang-Andersen Martinello (Danmarks Nationalbank)
Bjørn Bjørnsson Meyer (Danmarks Nationalbank)
Emil Toft Vestergaard (Danmarks Nationalbank)
Asger Lau Andersen (University of Copenhagen and CEBI)
Thais Lærkholm Jensen (Danmarks Nationalbank)

5 April 2024

SUPPLEMENTARY MATERIAL

1. BACKGROUND

The first case of Covid-19 in Denmark was confirmed in February 2020, roughly coinciding with the first cases in most other European countries. Figure S1 illustrates the epidemiological dynamics over the following two years by plotting infection rates (Panel A), hospitalization rates (Panel B) and mortality rates (Panel C) over time. It is easy to discern three distinct waves: In Spring 2020 following the initial outbreak, in Winter 2020-2021, and in Winter 2021-2022 coinciding with the arrival of the Omicron variant.

The government affected the epidemiological dynamics with different sets of policies, notably *restrictions* on activities involving physical proximity, free and widely available *testing* for Covid-19 combined with self-quarantine of those testing positive, and a population-wide *vaccination* program.

The restrictions varied significantly in scope over the pandemic. At different points in time, non-essential parts of the public sector were shut down; private sector employees were urged to work from home; non-essential retail and personal services were shut down; borders were closed for foreign nationals; congregations of more than 10 individuals were banned; schools and universities were limited to online teaching; bars, nightclubs and concert venues were shut down; restaurants were limited to take-away service; and face masks were required in public buildings, shops and public transport. The restrictions were relatively mild compared to most European countries, but significantly stricter than in neighboring Sweden [1]. Supermarkets and grocery stores were open throughout the pandemic.

Except for the very earliest stage of the pandemic with severe shortages of test equipment, Covid-19 tests were generally free and easily accessible for individuals with symptoms or exposure to infected persons. Further, the government quickly adopted an aggressive test strategy by which individuals were encouraged to test frequently, even in the absence of symptoms and known exposure [2]. Figure S1 illustrates how the aggregate number of tests evolved over the pandemic (Panel D). By the end of our sample period in January 2022, public and private test centers had performed around 56 million molecular tests, around 10 tests per inhabitant, and roughly as many antigen tests. By comparison, the United Kingdom had performed around 6 tests per inhabitant; France around 3; the United States around 2; and Germany around one [3]. Two Danish seroprevalence studies estimate that the share of infections diagnosed in tests was around 50% in 2020 and around 80% in 2021 [4-5].

Starting in December 2020, the government rolled out a comprehensive vaccination program. While the scale of the program was initially limited by scarce supply of vaccines, Denmark gradually achieved a high vaccination rate by international standards. By the end of our sample period in January 2022, the health authorities had administered more than 12 million doses, i.e. around 2 per inhabitant. Around 80% of the population had completed a vaccination program, which compares to 74% in Germany, 72% in the United Kingdom and 65% in the United States [6].

2. DATA CREATION

2.1 Data sources

We combine micro-data from three sources: *Danske Bank* that is the largest retail bank in Denmark [7]; *Statens Serum Institut* in Denmark that collect and process Covid-19 test data from public as well as private test providers [8]; and *Statistics Denmark* that compiles administrative micro-data from a range of government registers [9-13]. All the data sources use the same unique personal identifiers, which makes it possible to combine them at the

level of individuals. Personal identifiers are encrypted and the datasets contain no other information that directly identifies individuals (e.g. names and addresses).

From *Danske Bank*, we obtain comprehensive transaction data for each of the bank's customers for the period between 1 January 2018 and 15 January 2022. We focus on two types of transactions: payments by card and money transfers through a mobile application.

First, information about card transactions allows us to determine the time and place of consumers' in-store purchases. We distinguish between transactions made in a physical store and transactions made online by exploiting that the two transaction types differ with respect to the authorization process in the payment system. To determine the place, we use a unique identifier for the store where the card was used. When multiple stores constitute a retail chain, each physical store has its own store identifier. When a store has multiple payment terminals, i.e. typically one terminal per cash register, they share the same store identifier. To determine the time, we use a time stamp in the transaction records. For technical reasons, the precision of the time stamps varies across the two major cards used by the bank customers. For the most commonly used card, *Visa/Dankort*, transaction information arrives in batches. This implies that one can identify a time interval in which a transaction is made, typically an hour or longer, but not the exact time. For the other major card, *MasterCard*, the time stamps indicate the date, hour, minute and second that the payment was made in the store. As our empirical design requires the timing of transactions to be determined with a high degree of precision, our analysis only uses transactions made with the latter type of cards.

Second, we exploit that money transfers carry information about social networks, a potential confounder of the analysis. For the sample of Danske Bank customers, we observe all ingoing and outgoing transactions through the mobile application MobilePay. This is the dominant mobile tool for person-to-person money transfers in Denmark, used by around 95% of Danish individuals aged 16-69 [14]. An existing study documents that these transactions map a network with structural properties very similar to large-scale social networks such as Facebook [15]. To integrate this data into our analysis, we identify pairs of individuals where, at some point during the sample period, one received the exact same amount through the application as the other sent in the exact same second. We refer to such pairs of individuals as members of the same payment network. In the infrequent cases where more than one individual sent or received the exact same amount in the same second, we do not consider them members of the same payment network.

From *Statens Serum Institut*, we obtain administrative information about the Covid-19 tests performed by all public test providers as well as the private providers offering free tests under a government contract. We do not have information about the relatively small number of tests performed by private providers for a fee nor about the tests performed at home. We observe the unique identifier of the individual taking the test, the date at which the test was performed, the type of test (i.e. antigen or molecular) and the test result.

From *Statistics Denmark*, we obtain administrative micro-data from a range of government registers. The administrative registers contribute to the identification of social networks and further provide detailed information about background characteristics.

Starting with the social networks, the population register contains information about the parents of each individual, which we use to identify extended families. We define extended family members as siblings, parents,

children, grandparents and grandchildren. The population register also contains an encrypted identifier for the address of each individual's main residence, allowing us to identify co-habiting individuals. We define household members as individuals who share the same registered address in the beginning of the quarter. Further, the employment register contains information about the physical workplaces of each individual. We define work colleagues as individuals who were at the same workplace in 2020, the most recent year with available data, provided no more than 100 individuals worked at the workplace. Finally, the education register contains detailed information about enrolment in educational institutions at all levels. For primary and secondary schools, we define school friends as individuals who attend the same institution in the same school year and who belong to the same birth cohort. For tertiary education, we define school friends as individuals who attend the same institution and are enrolled in the same degree program, except for degree programs with more than 100 individuals.

For background characteristics, we obtain information about each individual's age, gender and place of residence from the population register; about the industry of employment from the employment register; and about total taxable income from the income register. The income measure is highly reliable, as it is mostly based on compulsory reporting by employers and financial institutions, which makes underreporting negligible [16].

In summary, our analysis uses payment data, test data and government register data for a gross sample of around 630,000 individuals who are customers at Danske Bank and have a MasterCard. As shown in Table S1, these individuals are not entirely representative of the general population in terms of socio-demographics. While they mirror the population almost perfectly in terms of gender and is highly similar in terms of household composition and size, they are somewhat younger, with an average age of around 38 years compared to around 41 years in the general population, and they have a lower average income. As Danske Bank is a national bank with a customer base that is generally highly representative of the overall population [17], the high share of young and low-income people reflects that we only use payment data from MasterCard, a product that is particularly popular in the young customer segment.

2.2 Sample selection

To create the estimation sample, we first identify all the instances where an in-store transaction on day d was made by an individual with a positive Covid-19 test in the 7-day period between day d-4 and day d+2. Assuming that individuals infected with Covid-19 are contagious from around two days before the onset of symptoms until five days after, these individuals are potential infectors who were likely contagious when making the purchase on day d. We limit the search to transactions with MasterCard, implying that we observe the precise time of the purchase, and to transactions in supermarkets and grocery stores, corresponding to the merchant category code 5411.

Next, we identify individuals who made a transaction with MasterCard in the same store on the same day as a transaction of a potential infector and measure the time distance between the individual's own transaction and the potential infector's transaction. In the rare cases where we find two infector transactions in the same store on the same day, we define the time distance relative to the closest one.

Drawing on this measure, we define a group of exposed individuals who made a transaction within 5 minutes of the potential infector's transaction. To be precise, if the infector's transaction is recorded within the first minute after 10am, e.g. at 10:00:30, an individual is counted among the exposed if they make a transaction in the 11-minute interval from 9:55:00 to 10:06:00.

Analogously, we define two alternative groups of exposed individuals comprised by individuals who make transactions within 1 minute and within 10 minutes of the potential infector respectively. We also define a reference group of non-exposed individuals who made a transaction between 16 and 30 minutes before the potential infector's transaction. In all cases, we exclude from these groups individuals who are themselves potential infectors, i.e. individuals with a positive test between day d-4 and day d+2.

To create the estimation sample, we combine exposed and non-exposed individuals, according to one of the definitions above, and exclude individuals with social connections to the potential infector: members of the same extended family, members of the same household, employees at the same workplace, students at the same educational institution and members of the same payment network. We note that the same individual can appear in the estimation sample multiple times, by making transactions in supermarkets or grocery stores close to a potential infector more than once over the sample period.

Table S2 offers an overview of the sample selection. We identify around 126,000 transactions made by a potential infector, representing around 53,000 unique individuals. We find 1,517,000 transactions made within 30 minutes of a potential infector, on the same day and in the same store, by individuals who are not themselves potential infectors. From this gross sample, we exclude around 11,000 transactions where the card owner has social ties to the potential infector: around 1,300 family members, 800 household members, 1,300 school colleagues, 800 work colleagues and 7,000 with transfers through the mobile payment application. From the remaining transactions, we define the baseline estimation sample, which comprises around 328,000 exposed individuals who made a transaction within 5 minutes of the potential infector and around 340,000 individuals who made a transaction between 16 and 30 minutes before the potential infector. The alternative exposed groups comprise 96,000 and 598,000 individuals who made a transaction within 1 minute and 10 minutes of the potential infector, respectively.

We note that the number of transactions is not completely proportional to the length of the time interval. In the baseline sample, the intervals defining exposure and non-exposure are 11 minutes and 15 minutes respectively, but the difference in the number of transactions is relatively smaller. This reflects a general pattern whereby the number of transactions per minute decreases slightly the longer the distance to the potential infector, which is at least partly due to stores' opening hours. If a potential infector makes a purchase 5 minutes after a store opens or 10 minutes before it closes, there will be no transactions more than 5 minutes before in the former case and more than 10 minutes after in the latter case.

2.3 Descriptive statistics

Individuals socially connected to the potential infector represent a possible confounder of the analysis. By definition, they are likely to be exposed to the potential infector outside the store, e.g. in the household, at school or at work. Thus, if they are more likely to be exposed than non-exposed inside the store, the key assumption underlying the research design, i.e. no correlation between exposures inside and outside the store, breaks down.

Figure S2 documents that individuals socially connected to the potential infector are indeed more likely to be exposed than non-exposed in the store. For each type of social connection, the figure shows the fraction of transactions made by socially connected individuals by the number of minutes between their own transaction and the transaction of the potential infector. In all four cases, there is sharp bunching around 1 minute. For instance, work colleagues of social infectors account for around 0.04% of the

transactions at the same minute as the infector's transaction and less than 0.01% of the transactions at all times more than 1 minute away.

After excluding family members, household members, school colleagues, work colleagues and individuals with money transfers from the estimation sample, some individuals with other social links to the potential infector may remain in the sample. In a robustness test, we further exclude around 142,000 observations from the gross sample where the individual in the sample had a transaction within 1 minute of the potential infector at some other time during the sample period. While the vast majority of these individuals are likely not social connections, the patterns in Figure S2 suggest that there is a much higher fraction of social connections in this group than in the full sample.

Next, we describe the timing of the potential infector transactions in our sample and document that it closely tracks the dynamics in overall Covid-19 infections in Denmark. Figure S3 shows the number of potential infector transactions in our sample by month. There is a spike with just below 10,000 infector transactions in December 2020 at the peak of the second wave and another spike with more than 35,000 infector transactions in December 2021 at the peak of the third wave. Figure S4 documents a tight correlation between the weekly number of confirmed cases in the population and the weekly number of potential infector transactions (blue markers) and transactions within 30 minutes of potential infector transactions (red markers) in our sample.

Finally, we compute for each potential infector transaction the number of exposed and non-exposed individuals in the baseline estimation sample and illustrate the distribution in Figure S5. Most commonly, there are one or two individuals in each group.

3. ESTIMATION FRAMEWORK

The goal of the empirical framework is to compare the outcomes of individuals who made a transaction within 5 minutes of a potential infector ("exposed") to individuals who made a transaction on the same day and in the same store between 16 and 30 minutes before the same potential infector ("non-exposed").

Letting i denote an individual in our estimation sample and letting q denote the transaction of a potential infector that assigns individual i to the group of exposed or non-exposed, we estimate the following model:

$$infected_{i,q} = \alpha_q + \beta exposed_{i,q} + \epsilon_{i,q}$$

On the left-hand side, $infected_{i,q}$ is an indicator that individual i tests positive for Covid-19 between day 3 and day 7 after transaction q. On the right-hand side, α_q represents a separate intercept for each infector transaction q. It captures, separately for each infector transaction q, the average infection rate among the non-exposed between day 3 and day 7 after the transaction and thus absorbs the background infection risk. The variable of interest, $exposed_{i,q}$, is an indicator for individual i being exposed at transaction q. Thus, the parameter β captures the *differential infection rate* for the exposed, measured relative to the non-exposed associated with the same infector transaction q.

We interpret β as the probability of transmission from the potential infector to exposed individuals. This interpretation requires two assumptions. First, we assume that other infection risks are uncorrelated with exposure across individuals associated with the same infector transaction (Assumption #1). This requires that individuals transacting within 5 minutes of a potential infector are not systematically different from individuals transacting

between 16 and 30 minutes before the same potential infector, at least not in ways that correlate with exposures outside the store. Second, we assume that there is no transmission from the potential infector to the non-exposed individuals (Assumption #2).

We also estimate the model using a number of alternative dependent variables. First, we use ex ante characteristics such as age, occupation and testing frequency as outcomes to investigate whether individuals associated with the same infector transaction are similar in these dimensions across exposed and non-exposed individuals. Second, we use indicators for infection in other periods to estimate how differential infection rates evolve dynamically over a longer time window. Both types of analysis serve to probe the assumption that individuals who are exposed and non-exposed in the store are not differentially exposed to other infection risks in a systematic way (i.e. Assumption #1). To the extent that the exposed and non-exposed groups are highly similar in terms of observable characteristics correlating with such external exposures (i.e. age, occupation, testing frequency) and in terms of infection rates in other periods than the one following the differential in-store exposure, it is suggestive that this assumption holds.

Further, we estimate variants of the model that absorb differences in observable characteristics between the exposed and non-exposed groups. We take a non-parametric approach by defining a set of indicators for each dimension of heterogeneity and augmenting the model with interactions between each of these indicators and a set of calendar day dummies. For instance, we control for age differences by including interactions between birth-year indicators and calendar day dummies, which implies that the model allows for a separate and fully flexible trend in infection rates for each birth-year cohort.

Finally, we employ alternative definitions of the exposure indicator based on the notion that individuals who make transactions nearer to the potential infector are more likely to have close contact in the store and therefore more likely to be infected. First, we vary the interval around the potential infector's transaction that delineates exposure. This enables us to corroborate that the transmission rate decreases with the time distance to the potential infector. Second, we estimate the model with a placebo measure of exposure that covers transactions between 11 and 15 minutes before the potential infector. This serves to probe the assumption that individuals categorized as non-exposed indeed have no exposure to the potential infector (i.e. Assumption #2). To the extent that the estimated effect of this placebo treatment is zero, i.e. that the individuals with transactions between 11 and 15 minutes before the infector have the same infection risk as those with transactions between 16 and 30 minutes before the infector, it is suggestive that this assumption holds.

4. ANALYSIS

4.1 Comparing exposed and non-exposed individuals

The empirical design critically assumes that individuals that were exposed and non-exposed to the potential infector in the store were not differentially exposed to other infection risks. While these infection risks are not directly observable, one can probe the assumption by comparing characteristics that correlate with infection risks across the two groups.

Table S3 compares the socio-demographic characteristics (Panel A) and behavior relevant for infection risks and the ability to detect infections (Panel B) across exposed and non-exposed. For each characteristic, the table displays the raw means in the two groups (Columns 1-2), the difference in the raw means (Column 3) and the difference conditional on a separate

intercept for each infector transaction (Column 4). The latter is the most relevant diagnostic. It compares exposed and non-exposed who made a transaction in the same store on the same day with only a slight difference in the timing, which is the exact same comparison used to estimate the in-store transmission rate. In practice, we obtain this diagnostic by estimating the empirical model using the observable characteristics of interest as dependent variable in separate regressions [SM section 3].

Starting with the socio-demographic characteristics, we find a small age difference, 33.57 years in the exposed group vs. 33.76 years in the non-exposed group, which may reflect a tendency for younger individuals to shop later in the day. This is potentially important, as age was a strong correlate of infection risk throughout the pandemic, and it motivates a robustness test with exhaustive controls for age differences. The two groups are almost perfectly balanced on other socio-demographic variables, such as the share of females, 50.3% vs 50.2%; household size, 2.91 vs 2.89 individuals; and income, DKK 140,262 vs 139,667.

Turning to behavior, we first note that exposed and non-exposed individuals exhibited exactly the same infection rates over the 30 days prior to exposure, 3.3% vs 3.3%. Hence, this summary measure of baseline infection risk shows no indication of behavioral differences across the two groups. However, infections are only detected if a test is performed, which makes it important to compare the test behavior of the two groups. Reassuringly, they exhibited almost identical test rates over the 30 days prior to exposure, 50.6% vs 50.3%, and almost the same average number of tests over the same period, 1.528 vs 1.507. Moreover, individuals differ in virus exposure due to their occupation. We find that the share employed in high-exposure sectors was almost the same share for the two groups, 6.9% vs 7.0% in the health sector and 6.7% vs 6.5% in education, suggesting that their average work-related exposure was similar.

Consumer activities may be associated with infection risk and it is therefore comforting that the two groups exhibit almost identical shopping behavior. In 2019, they had the same number of in-store card transactions, 495.6 vs 495.6, aggregating across all types of stores. Focusing only on supermarkets and grocery stores, the number of in-store card transactions remains highly similar for the two groups, 204.0 vs 205.0, and this continues to be the case when zooming in on the peak hour between 4pm and 5pm where supermarkets have the highest number of transactions, 19.64 vs 19.71.

In summary, exposed and non-exposed individuals are strikingly similar in terms of observable characteristics even without restricting the comparisons to individuals associated with the same potential infector transaction. This is not surprising, despite the presence of selection into stores and transaction times, given that the two groups are sampled from precisely the same set of stores and almost the same set of transaction times.

When we restrict comparisons to individuals associated with the same potential infector transaction, the two groups often become even more similar. For instance, the small difference in annual income of DKK 594 shrinks to DKK -126 and the small difference in the number of Covid-19 tests over the past 30 days of 0.021 narrows to 0.005. In some other cases, the differences become marginally larger.

4.2 Main results

In the main analysis, we estimate the differential infection rate for exposed individuals relative to non-exposed individuals associated with the same potential infector transaction (see section 2.1). The infection rates are computed between day d+3 and day d+7 after the transaction, which corresponds

to the typical period where symptoms would emerge in the case of transmission from the potential infector on day d. The estimates are illustrated in Figure 2 in the main text and reiterated with precise coefficients and sample sizes in Table S4.

Our main result is a differential infection rate in the exposed group of around 0.12%-points ($p<0.000$). This estimate of the in-store transmission rate compares to a baseline infection rate of 1.3% in the non-exposed group over the same period. When we vary the time interval that defines exposure, we continue to find statistically significant differential infection rates in the exposed group. Specifically, our estimate increases to around 0.18%-points ($p=0.002$) for a narrower definition covering transactions within 1 minute of the infector and decreases to around 0.08%-points ($p=0.002$) for a broader definition covering transactions within 10 minutes of the infector. The gradient in these estimates confirms the intuitive notion that an individual's effective exposure in the store is decreasing in the time difference between the individual's own transaction and the potential infector's transaction. Finally, the estimate drops to 0.01%-point ($p=0.77$) when we employ a placebo measure of exposure that covers transactions between 11 and 15 minutes before the potential infector. This result suggests that individuals transacting more than 10 minutes before the infector were virtually non-exposed and, by implication, that the reference group with transactions between 16 and 30 minutes before the potential infector are not contaminated by exposure to the infector.

The results are robust to including a range of control variables that address the potentially confounding effect of differences in observable characteristics, as shown in Figure S6. The main concern is age: the exposed are slightly younger than the non-exposed and as overall infection rates correlated significantly with age throughout the pandemic, with higher infection rates for younger cohorts in most periods, this slight imbalance introduces a risk that age-related differences in out-of-store exposures add to the estimated effect of differential in-store exposure. However, controlling non-parametrically for age with interactions between birth-cohort indicators and calendar-day indicators, barely changes the estimated effects (red bars). Another potential concern is geography, as infection rates varied strongly across different parts of Denmark, with more densely populated areas generally experiencing more infections. While the empirical design removes much of the geographical variation by identifying from within-store comparisons, it is conceivable that some systematic differences in residential patterns remain. However, the results are similar to the baseline when adding interactions between 99 indicators of the individual's municipality of residence and calendar-day indicators (blue bars). Further, income is a potential confounder, as it correlates with many other factors associated with infection risks such as housing conditions, occupation and awareness about health risks and disease prevention. Again, to the extent that individuals at different income levels sort into different stores, the empirical design addresses this issue by restricting the identifying variation to within-store comparisons. Indeed, the results are very similar to the baseline when augmenting the model with interactions between 100 income indicators, based on total taxable income in 2019, and calendar-day indicators (green bars). As a final robustness test, we include all three sets of controls – age, geography and income – at the same time. While the precision of the estimates decreases slightly in this augmented model, the estimated effect sizes remain highly similar to the baseline.

The results are also robust to addressing the potentially confounding effect of social networks with further sample restrictions. The baseline sample already excludes individuals connected to the infector through family, household, work and education as well as individuals who sent money to or received money from the potential infector through a mobile app at any point

during the sample period. However, one may be concerned that, as these measures of network connections are inherently incomplete, our estimates could still pick up out-of-store transmissions in social networks. We address this concern with two additional tests and illustrate the results in Figure S7.

First, we re-estimate the model while excluding all individuals who, at least once during the sample period, made a transaction within 1 minute of the potential infector (not counting the transaction that defines the individual as exposed or non-exposed). We observed in Figure S2 above that connected individuals are highly overrepresented among those who make transactions within 1 minute and we should therefore expect this restriction to reduce the prevalence of social connections in the estimation sample. As this approach defines possible social connections based on closeness in stores (at other times), it may reduce the prevalence of precisely the social connections that potential infectors are more likely to go to stores with, which are also the connections that could bias the results. The estimates remain highly similar to the baseline when we impose this additional sample restriction (blue bars). When we follow the same procedure, but exclude a wider range of individuals, i.e. those who ever made a transaction within 5 minutes of the potential infector, the point estimates generally become somewhat smaller, but remain statistically indistinguishable from the baseline estimates (green bars).

Second, we re-estimate the model while excluding all individuals whose age is within 5 years of the potential infector. Social networks generally exhibit strong homophily in age [18], which implies that individuals with a similar age as the potential infector are much more likely to be social connections than others. This additional sample restriction does not materially change the estimated effects (brown bars).

While our baseline specification restricts the non-exposed group to individuals who transacted 16-30 minutes *before* the potential infector, we check whether the results are robust to using a symmetrically defined non-exposed group, which also includes individuals who transacted 16-30 minutes *after* the potential infector. The estimated effect is somewhat smaller with this specification (purple bars), which is consistent with the notion that individuals who transacted after the potential infector may have experienced some transmission through exposure to contaminated air or surfaces in the store and therefore do not constitute a valid reference group. Restricting the non-exposed group to individuals who transacted 20-30 minutes before or after the potential infector yields highly similar but, if anything, slightly larger estimates (gray bars).

4.3 Infection dynamics

In addition to the main results, which concern infection rates shortly after exposure, we also conduct a dynamic analysis of infection rates over a longer time horizon. Letting *d* denote the day of exposure, our main outcome is an indicator for testing positive in the 5-day period [d+3, d+7]. None of the individuals in our estimation sample tested positive in the period [d-4, d+2] by construction - if they did they would be potential infectors and therefore excluded from the estimation sample. To study dynamics, we therefore construct indicators for testing positive in other 5-day periods, both before exposure, i.e. [d-9, d-5], [d-14, d-10], [d-19, d-15] etc., and after exposure, i.e. [d+8, d+12], [d+13, d+17], [d+18, d+22] etc. We use these infection indicators as outcomes in a series of separate regressions and illustrate the results in Figure 3 in the main text.

The results indicate that exposed and non-exposed individuals generally followed similar infection trajectories both before and after exposure. The

period [d+3, d+7] stands out as the only period where the exposed experienced materially higher infection rates than the non-exposed. This is consistent with our interpretation that the differential infection rate of the exposed in the period [d+3, d+7] reflects in-store transmission on day d and not general differences in infection risk.

Figure S8 illustrates the dynamics at a higher frequency, with separate estimates for each day rather than 5-day periods. The precision is generally low, which is the main reason why we use 5-day periods in the main analysis. The largest estimate is for day d+3 where the differential infection rate of the exposed is close to 0.004 percentage points (p=0.012). The estimates for days d+4, d+5, d+6 and d+7 are all positive and larger than 0.001 percentage points. In the pre-exposure period [d-30, d-5], the daily estimates are generally smaller and carry no indication of a systematically higher infection risk in the exposed group. While a few of the daily estimates for the period after day d+7 are numerically large, this does not appear to reflect systematic differences, as the estimates for the adjacent days are generally much smaller and often even have the opposite sign.

Figure S9 reports dynamic estimates for the two alternative definitions of exposure, i.e. transactions within 1 minute (Panel A) and 10 minutes (Panel B) of the potential infector. We observe a similar pattern as in the main dynamic analysis. Exposed and non-exposed individuals do not differ materially with respect to their infection dynamics, except for the differential infection rate of the exposed group in period [d+3, d+7]. Not surprisingly, the pattern is somewhat noisier for the narrowest definition of exposure.

Finally, we probe the robustness of the dynamic estimates by including non-parametric controls for age, region and income. Figure S10 illustrates the results. Overall, the estimates are highly similar to the analogous results without controls illustrated in Figure 3.

4.4 Differences across Covid-19 variants

We investigate to what extent the transmission rate in stores varied with the dominant variant of Covid-19. Specifically, we split the sample period into four subperiods, each corresponding to a Covid-19 variant. To delimit the subperiods, we use the first days on which a new variant accounted for more than half of the genome-sequenced tests, i.e. 19 January 2021 for Alpha; 28 June 2021 for Delta; and 17 December 2021 for Omicron.

Figure 4 illustrates the estimates for the baseline definitions of exposure as well as the two alternative measures. The estimated transmission rates for Omicron are generally much larger than for the other three variants and much larger than the headline estimates based on the full sample. There is much less variation across the other three variants.

One may be concerned that the striking differences in the estimated variant-specific transmission rates do not reflect differences between the variants themselves, but confounding differences in the environments, in which they operated. The main concern is seasonal factors. We only observe Omicron in the cold season where in-store transmission rates may generally be more elevated, although conditions inside supermarkets and grocery stores are relatively stable throughout the year. Another concern is changing policy interventions: Face masks were required through the entire period where we observe Omicron, but not in many other phases of the pandemic. Finally, there are concerns about confounding changes in the composition of those infected.

We address these concerns with a model that identifies the excess transmission rate of Omicron relative to other variants while also allowing

the transmission rate to depend on other relevant factors. We implement this idea by interacting the exposure variable with an indicator for Omicron being the dominant variant on the day of the exposure as well as other variables that capture confounding factors. First, we interact with indicators for calendar months. This implies that the model effectively compares the transmission of Omicron in December 2021-January 2022 to the transmission of the Index and Alpha variants in December 2020-January 2021. Second, we interact exposure with an indicator for a facemask requirement in retail stores. This implies that the model compares the transmission of Omicron to transmission in other periods with a facemask requirement. Third, we interact exposure with an indicator for the age of the exposed individual. This implies that the model compares the transmission of Omicron to transmission of other variants for individuals at a similar age.

We find that the excess transmission rate of Omicron is highly robust, as illustrated in Figure S11. For the main definition of exposure, the estimated excess transmission rate is around 0.3% without controlling for confounding factors. The estimate does not fall below 0.25% when allowing the transmission rate to vary by calendar month, face masks requirements and age.

4.5 Heterogeneity by individual characteristics

We investigate how transmission rates in stores varied with individual characteristics. We implement this idea by splitting the sample according to the characteristic of interest and estimate the baseline model separately for each subsample. Figure S12 illustrates the results for the baseline definition of exposure (transactions within 5 minutes).

The estimated transmission rate is strongly decreasing in the age of the exposed individual (red bars), increasing in the age of the potential infector (blue bars), and almost the same for the two genders (green bars).

This result relates to a large literature that investigates how the susceptibility to Covid-19 infection as well as the onward transmissibility of the infection vary with age [19–23]. While many of these studies suggest that both susceptibility and transmissibility are increasing in age, differences in contact patterns and the share of asymptomatic cases across age groups are important potential confounders. Moreover, recent studies find that both age gradients might differ substantially across variants, resulting in much higher infection rates of children and adolescents under Delta and Omicron.

Taken at face value, our results suggest that susceptibility is decreasing in age whereas transmissibility is increasing in age. An important advantage of our research design is that transmission rates are compared across age groups while holding the physical environment approximately constant: All exposures happen in a supermarket or a grocery store. However, there are also notable caveats, which make this causal interpretation less straightforward. First, the comparison may be influenced by behavioral differences, e.g. elderly people may observe distancing requirements more rigorously and wear a protective facemask more frequently because of the higher risk of severe illness in case of infection. Second, vaccinations were offered to the elderly first and take-up rates were generally increasing in age. Third, systematic differences in the physical environment may remain, as younger people may tend to visit less spacious stores at times where they are more congested.

4.6 Testing

As detection of Covid-19 is imperfect, differential testing across exposed and non-exposed individuals could potentially confound our estimates. While we showed above that the two groups are very similar in terms of their baseline testing behavior, one may be concerned that exposed individuals tested more in the days following exposure. In principle, the contact-tracing app operated by the health authorities may have alerted exposed individuals about the potential infector in the store and induced some of them to get tested. To the extent that such additional tests detected infections contracted elsewhere, we would over-estimate the in-store transmission rate.

We gauge the importance of this potential confounder by estimating our model using as the dependent variable an indicator for having taken a test with a negative result between day 3 and day 7 after exposure. To the extent that our main result were driven by alerts from the contact-tracing app, we should expect a large differential effect of exposure on testing. Specifically, according to data released by the Danish Ministry of the Interior and Health, around 99% of the tests triggered by alerts from the app were negative [24], so to explain the entire estimated effect on positive tests of around 0.12 %-points, the estimated effect on negative tests should be around two orders of magnitude larger, i.e. around 12 %-points. By contrast, if the contact-tracing app plays no role in explaining our results, we should see no effect on negative tests.

Figure S13 illustrates the results for negative tests (green bars) and compares them to the analogous results for positive tests (red bars). Regardless of the definition of exposure, the effect on negative tests is essentially zero. This result is hard to reconcile with any material confounding effect of the contact-tracing app. A plausible explanation is that casual contact in stores is typically associated with much less than the 15 minutes' close contact required for the contact tracing app to send an alert.

4.7 Multiple exposures

To the extent that individuals are exposed to multiple potential infectors in the same store, our estimates could in principle overstate the transmission risk associated with a single exposure.

For the sample of individuals with a transaction within 10 minutes of a potential infector, Figure S14 shows the number of potential infectors who made a transaction within these 10 minutes.

The vast majority of cases involve only one potential infector (around 90%) and while a non-negligible number of cases involve two potential infectors (around 9%), there are very rarely three or more (around 1%).

Figure S15 illustrates that our estimates of in-store transmission rates are not sensitive to considering only exposed individuals with a single potential infector making a transaction within 10 minutes of their own transaction.

4.8 Socially connected individuals

Our main analysis excludes individuals who are socially connected to the potential infector based on the notion that they are likely to be exposed to the potential infector not just inside the store, but also outside of the store [See Section S2.2]. This suggests that leaving such individuals in the estimation sample would cause a severe upward bias in the estimated probability of transmission inside the store.

We substantiate this argument by estimating the model for precisely the individuals who are socially connected to the potential infector. Concretely, the estimation sample then consists of non-exposed individuals (same as in the baseline specification) as well as exposed individuals who are family members, household members, school colleagues, or work colleagues of the potential infector or are linked to the potential infector through money transfers.

Figure S16 illustrates the results. The 5-day infection rate of individuals who made a transaction within 5 minutes of a potential infector to whom they are socially connected is 8 percentage points above the infection rate of the non-exposed (blue bar). This estimate is more than 60 times higher than the corresponding estimate from the baseline specification where socially connected individuals are excluded (red bar). There is considerable variation across social networks with the highest estimates for household members (around 17 percentage points) and family (around 12 percentage points), mid-range estimates for individuals linked by money transfers (around 8 percentage points) and the lowest estimates for school colleagues and work colleagues (both around 4 percentage points).

We emphasize that these results *should not* be interpreted as estimates of the transmission probability associated with the interaction in the store, because socially connected individuals are also likely to interact outside the store.

We conduct the same exercise for individuals who transacted within 1 minute or 5 minutes of the potential infector on another occasion, indicating that they may be socially connected (green bars). The estimates are close to the results from the baseline specification suggesting that at most a small fraction of these individuals is in fact socially connected to the potential infector.

5. IMPLICATIONS

We use the estimated transmission rates from the regression analysis to gauge the individual and aggregate risks associated with casual contact in supermarkets and grocery stores. Specifically, we provide estimates of the probability of getting infected for the average non-infected individual making a purchase in a store (Section 5.2) and the number of transmissions in stores for the average infected individual, i.e. reproduction (Section 5.3).

The key challenge is that we do not observe all card payments for the full population. We only observe the payments made by customers at Danske Bank and we only have detailed information about the timing of the payments when they use a MasterCard. This has implications for both risk metrics. When we estimate the probability of getting infected in a store for an individual in our sample, we need to account for the exposures we do not observe because the potential infectors are not customers at Danske Bank or do not pay with a MasterCard. Similarly, when we estimate the average number of in-store transmissions caused by a potential infector in our sample, we need to account for the exposures we do not observe because the exposed individuals are not customers at Danske Bank or do not pay with a MasterCard.

We address this challenge by scaling up the exposures we observe in the sample with an estimate of the inverse sampling probability, i.e. the number of similar transactions in the population corresponding to a given transaction in the sample. When estimating the inverse sampling probabilities, we account for the non-representativeness of the sample in

terms of age and for the fact that individuals can have multiple banks and multiple payment cards.

5.1 Inverse sampling probabilities

To implement this approach, we first delineate a set of *primary customers*, i.e. individuals whose primary bank is Danske Bank. We exploit that all adults in Denmark must designate a bank account for transactions with the public sector, that is an account for receiving child benefits, tax refunds, pensions and so on. For adults (18 years and older), we require that this so-called NemKonto is at Danske Bank. For children (below 18 years) who presumably only rarely have multiple banks, we simply require that they have an account at Danske Bank. Next, for each primary customer, we measure the share of the payments in supermarkets and grocery stores that they make with their MasterCard.

Drawing on these steps, we compute the number of individuals for whom we observe transactions in supermarkets and grocery stores measured in full-customer equivalents. This simply amounts to counting the primary customers while weighting with their MasterCard share. For instance, primary customers who pay for all their transactions with their MasterCard contribute one full-customer equivalent while primary customers paying with their MasterCard every third time contribute one third full-customer equivalent. We make this computation for each quarter separately, thus allowing for changes in bank relations and payment behavior over time, and for each birth cohort.

Finally, we estimate the inverse sampling probabilities, by quarter and by birth cohort, by dividing the number of individuals in the population by the number of primary customers measured in full-customer equivalents. Figure S17 illustrates the resulting estimates for the full sample period. The estimate is around 13 for the average person, but varies significantly by age group, from less than 5 at age 18 to more than 50 at age 70.

We use the estimated inverse sampling probabilities to account for the incompleteness of our card payment data. First, when we observe an individual in our sample exposed to a potential infector with inverse sampling probability X, we assume that the individuals in our sample were in fact exposed to X-1 potential infectors that we do not observe, because these potential infectors were not customers at Danske Bank or because they did not pay with a MasterCard. Second, when we observe a potential infector in our sample exposing an individual with inverse sampling probability Y, we assume that the potential infectors in our sample exposed Y-1 other individuals that we do not observe, because they were not customers at Danske Bank or because they did not pay with a MasterCard.

The key assumption is that selection into being a Danske Bank customer and into paying with MasterCard is random conditional on age. In other words, two same-aged individuals are equally likely to be exposed in a store when they are uninfected and equally likely to expose others in a store when they are infected, regardless of whether they are customers at Danske Bank or another bank, and regardless of whether they pay with MasterCard or another card.

5.2 Individual risk

We estimate the infection risk associated with a purchase in a supermarket and grocery store for the primary customers in our sample in the following steps. First, we identify the instances where we observe that a primary customer in our sample was exposed. Second, we scale up each instance with the inverse sampling probability of the potential infector and aggregate to obtain the expected number of times the primary customers in our sample were

exposed to any potential infector in the population. Third, we multiply by the estimated in-store transmission rate to obtain the expected number of times the primary customers were infected in a store. Finally, we divide by the number of payments made by the primary customers in supermarkets and grocery stores to estimate the infection risk per store visit.

In these computations, we use the broadest definition of exposure covering transactions within 10 minutes of a potential infector. We employ the methodology described above to obtain estimates of in-store transmission rates while splitting the sample to allow the estimates to vary by the age of the potential infector, the age of the exposed individual and the dominant Covid-19 variant. Concretely, we distinguish between three age groups (i.e. <25 years, 25-45 years and >45 years) as well as between two types of variants (i.e. Omicron and other), yielding 18 distinct transmission rates.

This methodology yields an estimated infection risk of around 0.00005 for the average store visit in the estimation period, i.e. around one infection per 20,000 store visits. The estimate varies significantly across months, from less than one infection per 1 million store visits in June 2020 to around one infection per 2,000 store visits in December 2021. The striking variation in the infection risk over time is primarily due to changes in the probability of in-store exposure, reflecting largely the number of infected individuals in the population (see Figure S4), and to a lesser extent changes in the estimated transmission rate (see Figure 4).

We also provide an estimate of the infection risk associated with a purchase in a supermarket and grocery store for the average individual in the population. This estimate differs from what we found for the sample of primary customers presented above because the sampled individuals, being non-representative of the overall population, face different probabilities of being exposed conditional on making a purchase as well as different probabilities of infection conditional on being exposed.

The approach differs from the one applied above in two respects. First, we scale up each instance where a primary customer is exposed not just with the inverse sampling probability of the potential infector, but also with the inverse sampling probability of the exposed individual. This yields the expected number of exposures for the population as opposed to for the sample of primary customers. Moreover, it yields the expected number of in-store infections in the population when multiplied by the relevant transmission rates. Second, we use the inverse sampling probabilities to scale the number of payments observed in the sample up to the expected number in the population.

We obtain an estimated infection risk of around 0.000025 for the average store visit, i.e. around one infection per 40,000 store visits, which is lower than in the sample of primary customers. This is intuitive as the sample is younger and therefore more likely to be exposed conditional on visiting a store and more likely to be infected conditional on being exposed (see Figure S12).

5.3 Reproduction

We take a similar approach to estimating the average number of transmissions in stores per infection in the sample of primary customers. First, we identify the instances where we observe that a primary customer is a potential infector and exposes another individual. Second, we scale up each of these instances with the inverse sampling probability of the exposed individual and aggregate to obtain the expected number of exposures due to the potential infectors in our sample. Third, we multiply by the estimated transmission rates to obtain the expected number of in-store transmissions due to the potential infectors

in our sample. Finally, we divide by the number of infections in our sample to estimate the number of in-store transmissions per infection.

We note that also infected individuals who do not make card payments between day d-4 and day d+2, and who are thus not potential infectors, contribute to the number of infections in the denominator of the reproduction number. We also note that individuals may contribute more than one infection to the denominator if they suffer multiple distinct infections during the sample period.

The estimated reproduction number for the full sample period is 0.038, suggesting that every 100 infections gave rise to around 4 transmissions through casual contact in supermarkets and grocery stores. As shown in Figure 5, the estimated reproduction number was roughly constant around 0.02 for most of the sample period, but climbed to around 0.06 at the arrival of the Omicron variant in December 2021. This striking increase in in-store reproduction is mostly due to an increase in the estimated transmission rate population (see Figure 5).

We also provide estimates of in-store reproduction numbers for the average individual in the population. Conceptually, this estimate differs from what we found for the sample of primary customers because the sampled individuals, being non-representative of the overall population, may exhibit different propensities to visit stores around infection, different probabilities of exposing others conditional on visiting a store, and different probabilities of infecting others conditional on exposing them.

To produce population estimates, we alter our methodology in two respects. First, we scale up each instance where a potential infector in our sample exposes another individual not just with the inverse sampling probability of the exposed individual, but also with the inverse sampling probability of the potential infector. This yields the expected number of exposures for the population as opposed to the sample of primary customers. Moreover, it yields the expected number of in-store infections in the population when multiplied by the relevant transmission rates. Second, we use the number of confirmed cases in the population in the denominator as opposed to the infections observed in the sample.

The estimated reproduction number for the population is 0.039 for the full sample period and exhibits a similar dynamics over the course of the pandemic (see Figure 5).

A key assumption underlying this methodology is that our gross sample of Danske Bank customers is representative in terms of the risk of being exposed in a store conditional on age. One may be concerned about potential selection mechanisms that invalidate this assumption. For instance, if MasterCard is used primarily by males, and males are more likely to be exposed than females due to different shopping patterns, our baseline approach would overestimate the reproduction number by assigning a too small sampling probability to males who have higher-than-average likelihood of exposure and a too large sampling probability to females who have a lower-than-average likelihood of exposure.

We address this concern by taking an alternative approach to obtaining sampling probabilities. For each quarter, we estimate a probit model where the sample is the full population of Denmark, the outcome is an indicator for being in the gross sample, and the explanatory variables are indicators capturing a range of observable characteristics, i.e. age cohort, gender, income decile, occupation, and region. This regression yields sampling probabilities for each individual in the sample and each quarter, which vary not just with age, but also with gender, income, occupation and region. For

individuals with multiple payment cards, we scale down the sampling probability with the non-MasterCard transaction share to reflect that we do not observe all their purchases in our transaction data. For instance, we assign a sampling probability of 7.5% to an individual who has an estimated 10% probability of being in the customer sample and uses MasterCard for 75% of their transactions in supermarkets and grocery stores. Figure S17 illustrates the distribution of the resulting sampling probabilities by age cohort.

This approach yields an estimated reproduction number for the full sample period of 0.037, only slightly lower than the baseline estimate.

REFERENCES:

[1] Sheridan, A., Andersen, A. L., Hansen, E. T., & Johannesen, N. (2020). Social distancing laws cause only small losses of economic activity during the COVID-19 pandemic in Scandinavia. *Proceedings of the National Academy of Sciences*, 117(34), 20468-20473.

[2] Gram, M. A., Steenhard, N., Cohen, A. S., Vangsted, A. M., Mølbak, K., Gorm Jensen, T., ... & Ethelberg, S. (2023). Patterns of testing in the extensive Danish national SARS-CoV-2 test set-up. medRxiv, 2023-02.

[3] Our World in Data available at https://ourworldindata.org/covid-cases

[4] Krogsgaard, L. W., Espenhain, L., Tribler, S., Sværke Jørgensen, C., Hansen, C. H., Møller, F. T., ... & Ethelberg, S. (2023). Seroprevalence of SARS-CoV-2 Antibodies in Denmark: Results of Two Nationwide Population-Based Surveys, February and May 2021. *Infection and Drug Resistance*, 301-312.

[5] Espenhain, L., Tribler, S., Sværke Jørgensen, C., Holm Hansen, C., Wolff Sönksen, U., & Ethelberg, S. (2021). Prevalence of SARS-CoV-2 antibodies in Denmark: nationwide, population-based seroepidemiological study. *European journal of epidemiology*, 36, 715-725.

[6] Our World in Data available at https://ourworldindata.org/covid-cases

[7] Danske Bank (2022). Dataset on customer transactions. Data provided by Danske Bank, Accessed on 5 April, 2022.

[8] Statens Serum Institut (2022). Dataset on COVID-19 tests. Data provided by Statens Serum Institut, Accessed on 9 March, 2022.

[9] Statistics Denmark (2022a). Befolkningen (BEF register, Population). Data provided by Statistics Denmark, Research Service. Accessed on 4 February 2022.

[10] Statistics Denmark (2022b). Ansættelser (IDAN register, Employment). Data provided by Statistics Denmark, Research Service. Accessed on 30 September 2022.

[11] Statistics Denmark (2022c). Education (UDDA register, Education). Data provided by Statistics Denmark, Research Service. Accessed on 4 February 2022.

[12] Statistics Denmark (2022d). Indkomster (IND register, Incomes). Data provided by Statistics Denmark, Research Service. Accessed on 4 February 2022.

[13] Statistics Denmark (2022e). Arbejdsklassifikationsmodulet (AKM register, Job Classifications). Data provided by Statistics Denmark, Research Service. Accessed on 4 February 2022.

[14] MobilePay, 2022. Statistics available at https://www.mobilepay.dk/

[15] Sheridan, A., 2019. Learning About Social Networks from Mobile Money Transfers. Available at: https://sites.google.com/view/adamsheridan

[16] Alstadsæter, A., Johannesen, N., & Zucman, G. (2019). Tax evasion and inequality. *American Economic Review*, 109(6), p. 2073-2103.

[17] Andersen, A. L., Johannesen, N., & Sheridan, A. (2020). Bailing out the kids: new evidence on informal insurance from one billion bank transfers. CEPR Working Paper 14867.

[18] McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology* 27(1), p. 415-444.

[19] Davies, N. G., Klepac, P., Liu, Y., Prem, K., Jit, M., & Eggo, R. M. (2020). Age-dependent effects in the transmission and control of COVID-19 epidemics. *Nature medicine*, 26(8), 1205-1211.

[20] Ludvigsson, J. F. (2020). Children are unlikely to be the main drivers of the COVID-19 pandemic–a systematic review. *Acta Paediatrica*, 109(8), 1525-1530.

[21] Gaythorpe, K. A., Bhatia, S., Mangal, T., Unwin, H. J. T., Imai, N., Cuomo-Dannenburg, G., ... & Ferguson, N. M. (2021). Children's role in the COVID-19 pandemic: a systematic review of early surveillance data on susceptibility, severity, and transmissibility. *Scientific reports*, 11(1), 13903.

[22] Chun, J. Y., Jeong, H., & Kim, Y. (2022). Identifying susceptibility of children and adolescents to the Omicron variant (B. 1.1. 529). *BMC medicine*, 20(1), 1-9.

[23] Chun, J. Y., Jeong, H., & Kim, Y. (2022). Age-varying susceptibility to the Delta variant (B. 1.617. 2) of SARS-CoV-2. *JAMA network open*, 5(3), e223064-e223064.

[24] Danish Ministry of the Interior and Health. Twitter on January 25, 2021 at 5:01pm from @DKsundhed

Figure S1: The Covid-19 pandemic in Denmark. The figure the daily number of new infections (Panel A), Covid-19 related hospitalizations (Panel B), Covid-19 related deaths (Panel C) and performed Covid-19 tests (Panel D) all measured per 100.000 inhabitants. The data is available at the website of *Statens Serum Institut*, https://covid19.ssi.dk/overvagningsdata.

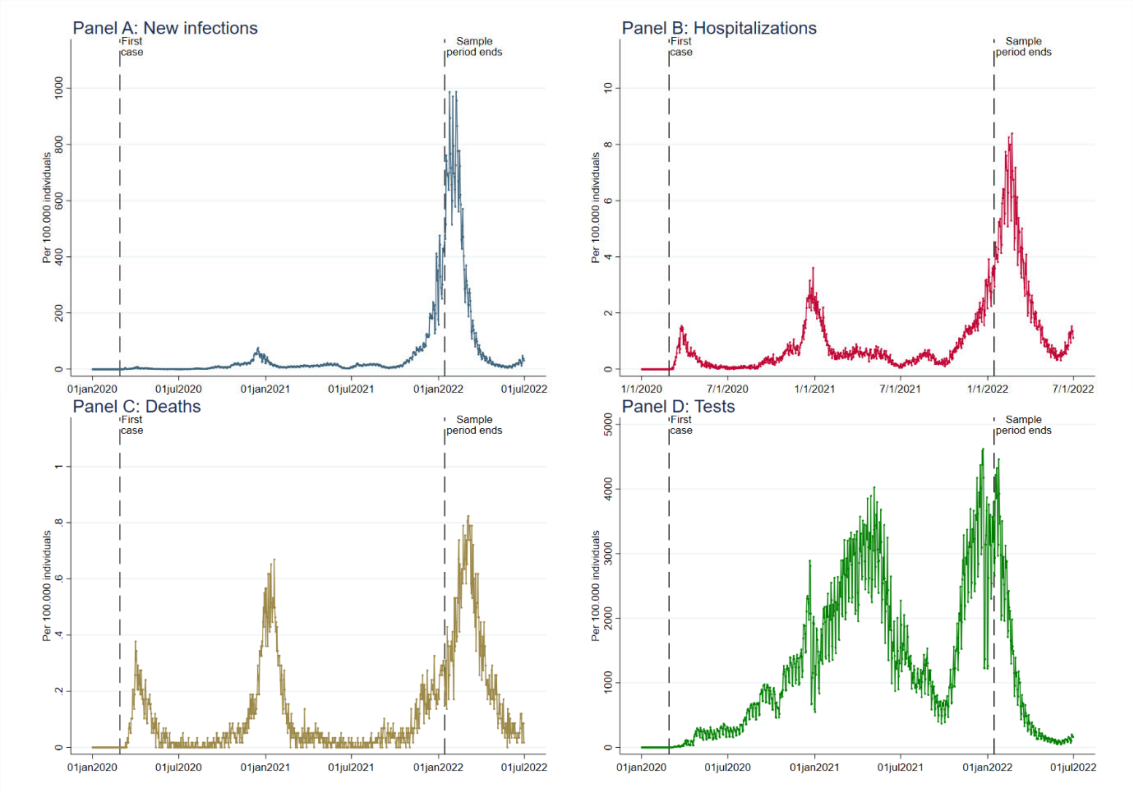

Figure S2: Social network indicators. The figures shows the prevalence of socially connected individuals in transactions by the time relative to the transactions of the potential infector for family and household members (Panel A), individuals working at the same workplace (Panel B), individuals attending the same educational institution (Panel C) and individuals who send or receive money on the money transfer app *MobilePay*.

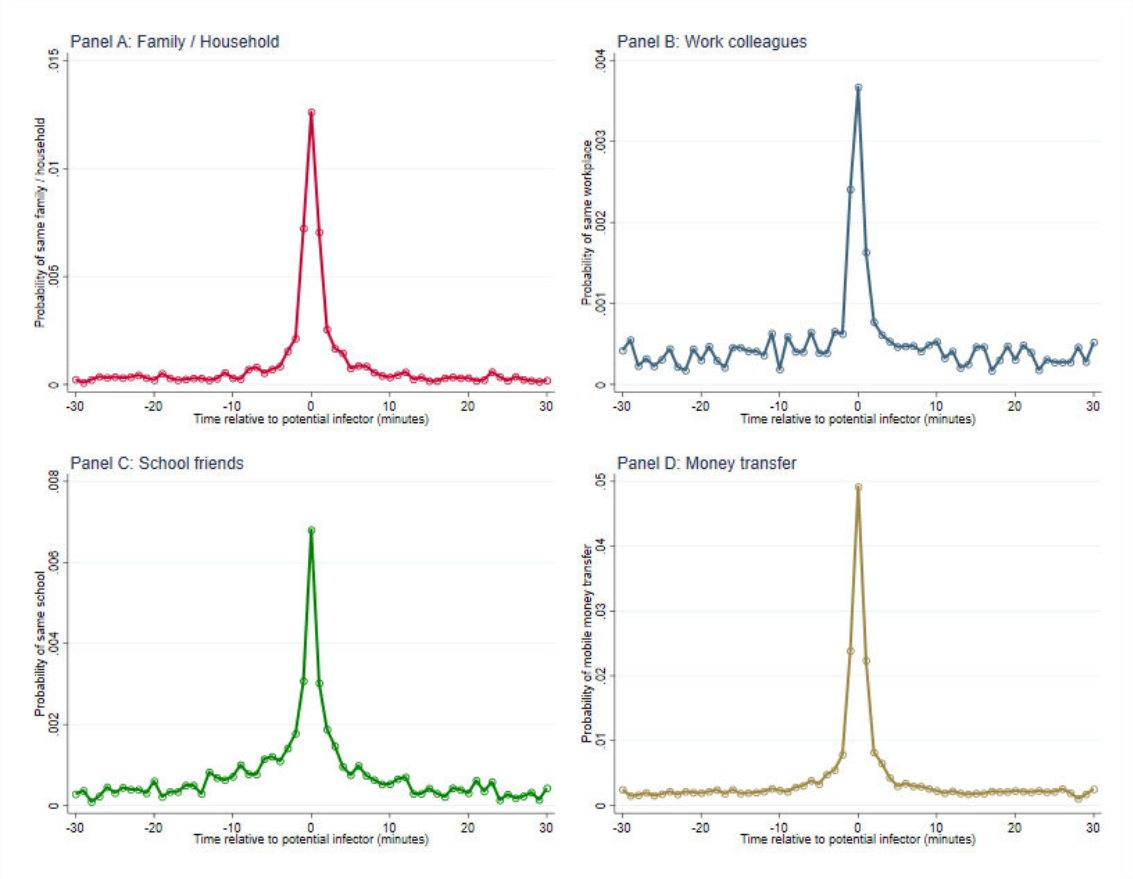

Figure S3: Potential infectors over time. The figure shows the number of potential infectors' payments in supermarkets and grocery by month.

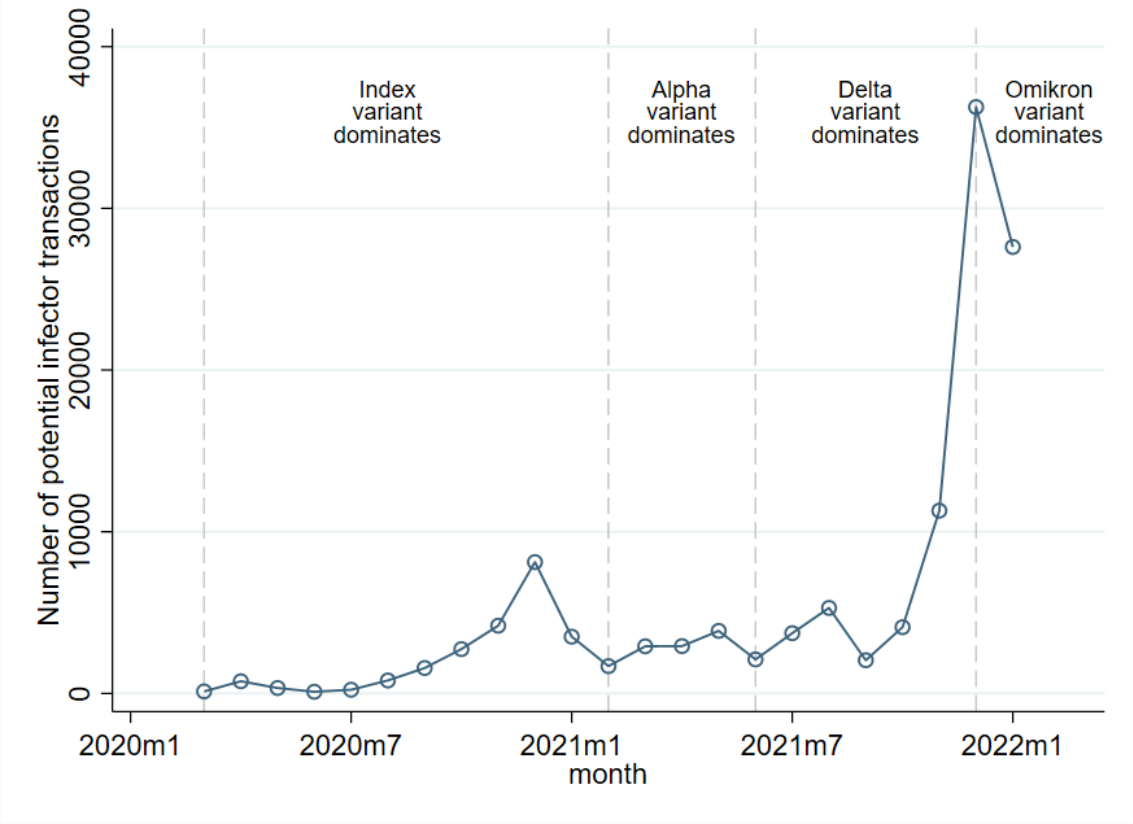

Figure S4: Covid-19 cases vs potential infectors. The figure plots the weekly number of Covid-19 cases in the population against the weekly number of potential infector transactions in our sample (blue dots) and the weekly number of transactions within 30 minutes of a potential infector transaction.

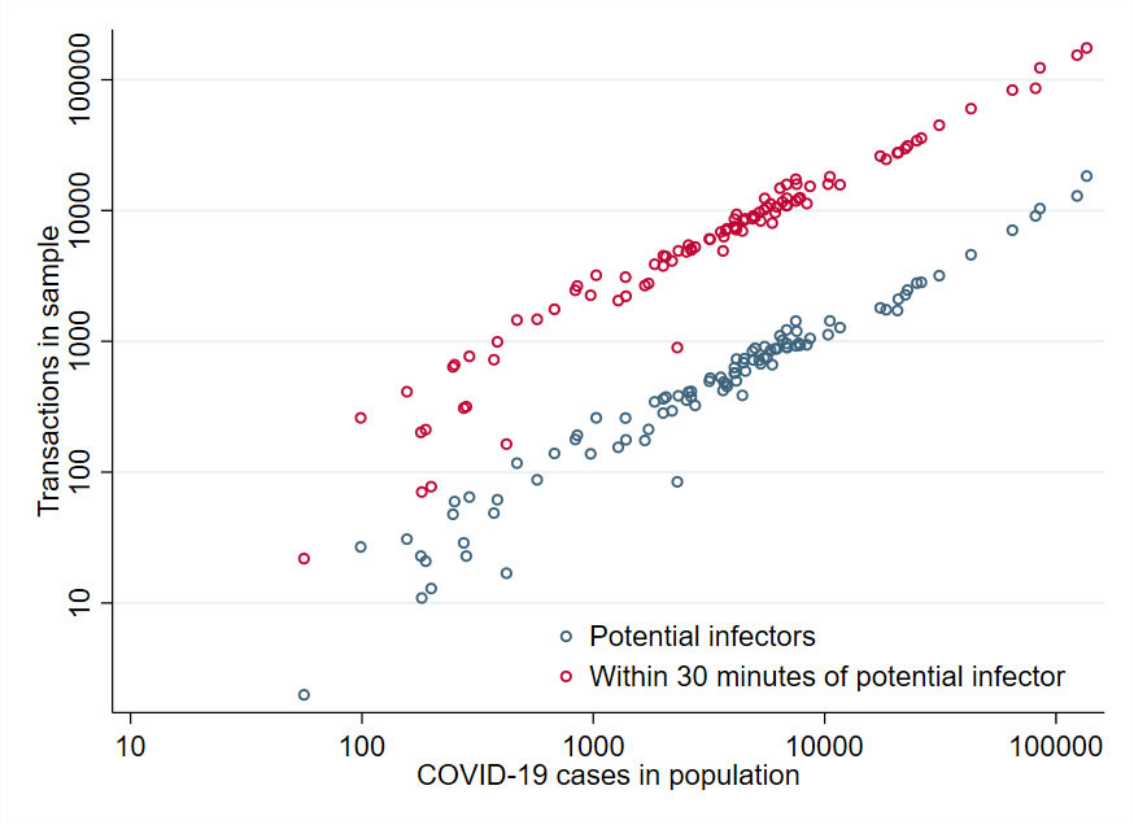

Figure S5: Exposed and non-exposed individuals per potential infector. The figure illustrates how the number of exposed and non-exposed individuals per potential infector is distributed.

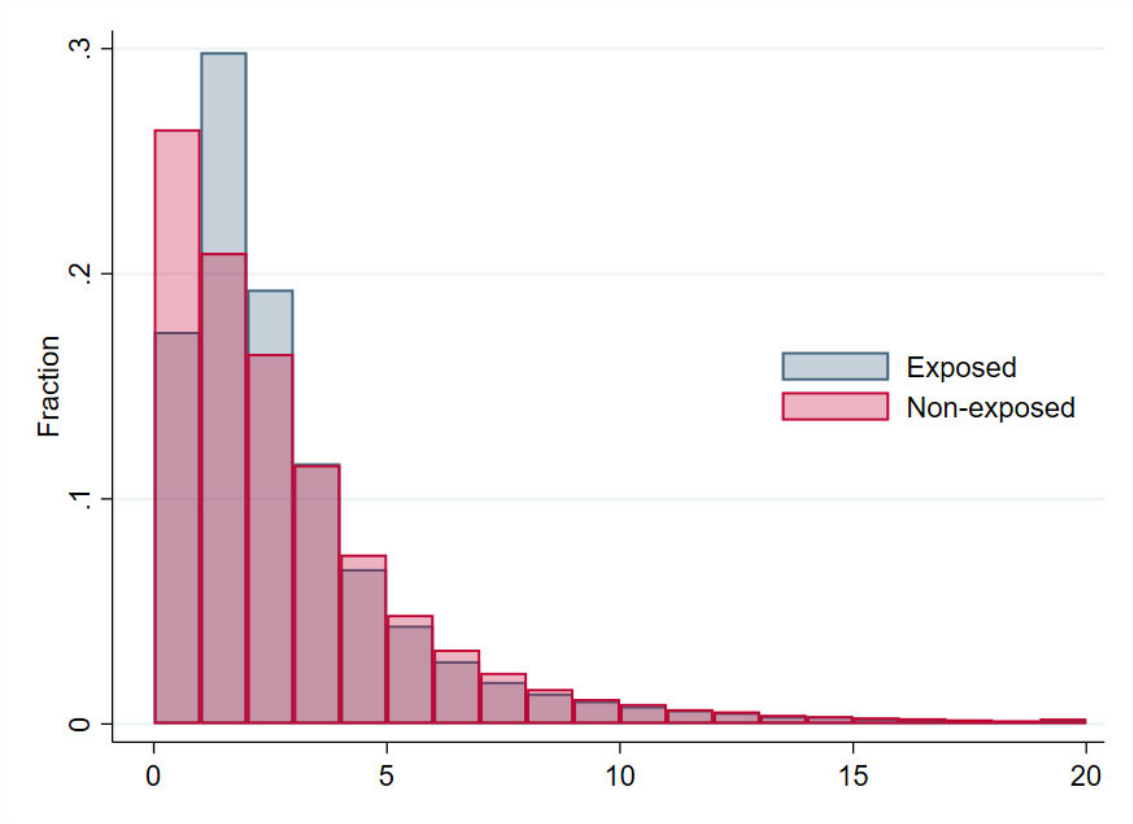

Figure S6: Robustness to controls. The figure illustrates how the main estimates change when our model is augmented with controls. For each of the three exposure measures, the figure shows the excess probability of testing positive between day d+3 and day d+7 for exposed individuals relative to non-exposed individuals estimated with five different sets of controls: no controls (red columns), indicators for birth year interacted with calendar days (blue bars), indicators for income percentile interacted with calendar days (green bars), indicators for municipality of residence interacted with calendar days (brown bars), all three sets of controls jointly (gray bars). The estimated coefficients and standard errors are reported in Table S4.

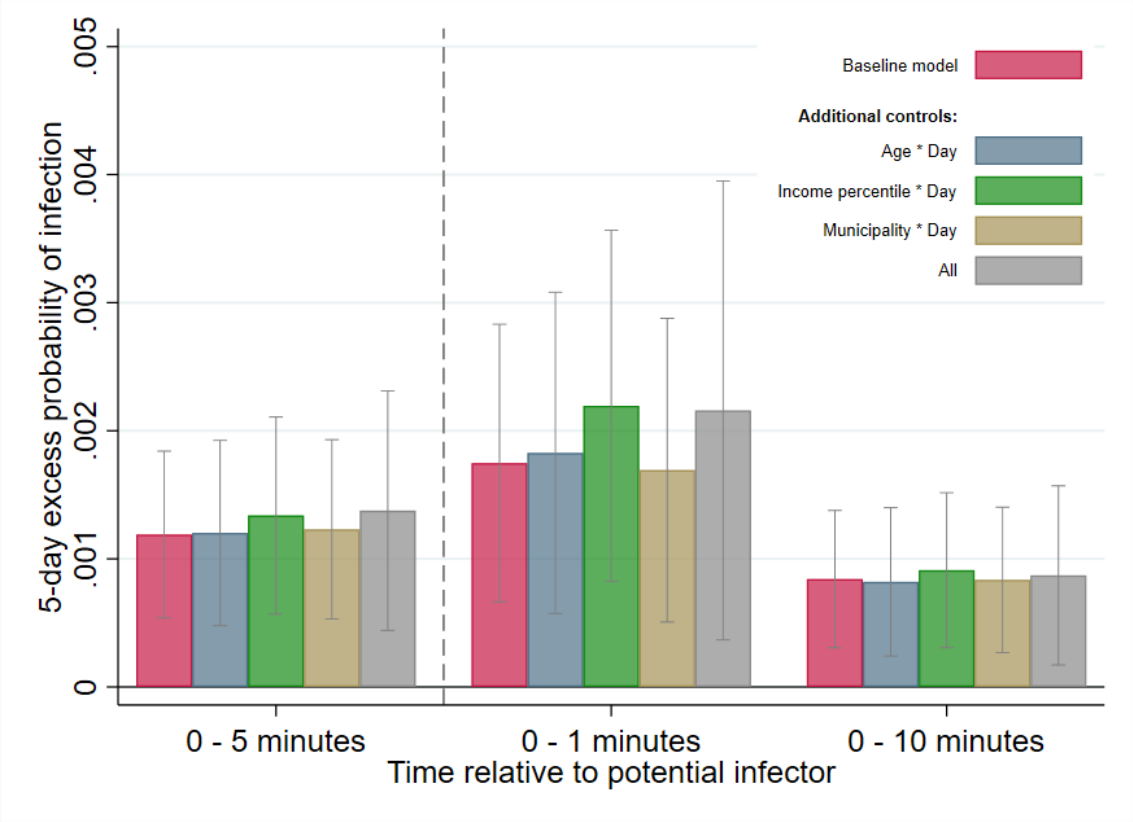

Figure S7: Robustness to sample restrictions. The figure illustrates how the main estimates change with the sample restrictions. For each of the three exposure measures, the figure shows the excess probability of testing positive between day d+3 and day d+7 for exposed individuals relative to non-exposed individuals estimated with five different sample restrictions: baseline (red columns), exclude individuals with another transaction within 1 minute of the potential infector (blue bars), exclude individuals within 5 years of age to the potential infector (green bars), include transactions between 16 and 30 minutes before and after the potential infector (brown bars), include transactions between 21 and 30 minutes before and after the potential infector (gray bars). The estimated coefficients and standard errors are reported in Table S4.

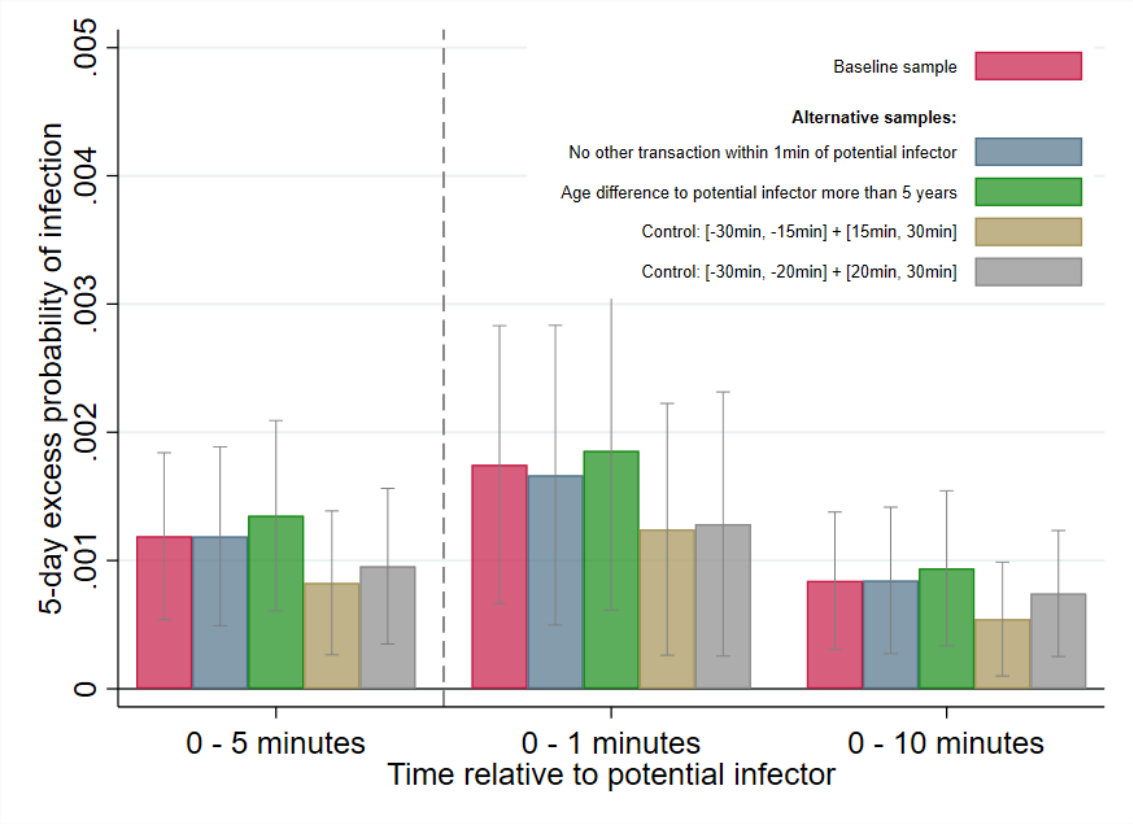

Figure S8: One-day infection dynamics. The figure shows the daily excess probability of testing positive for exposed relative to non-exposed individuals.

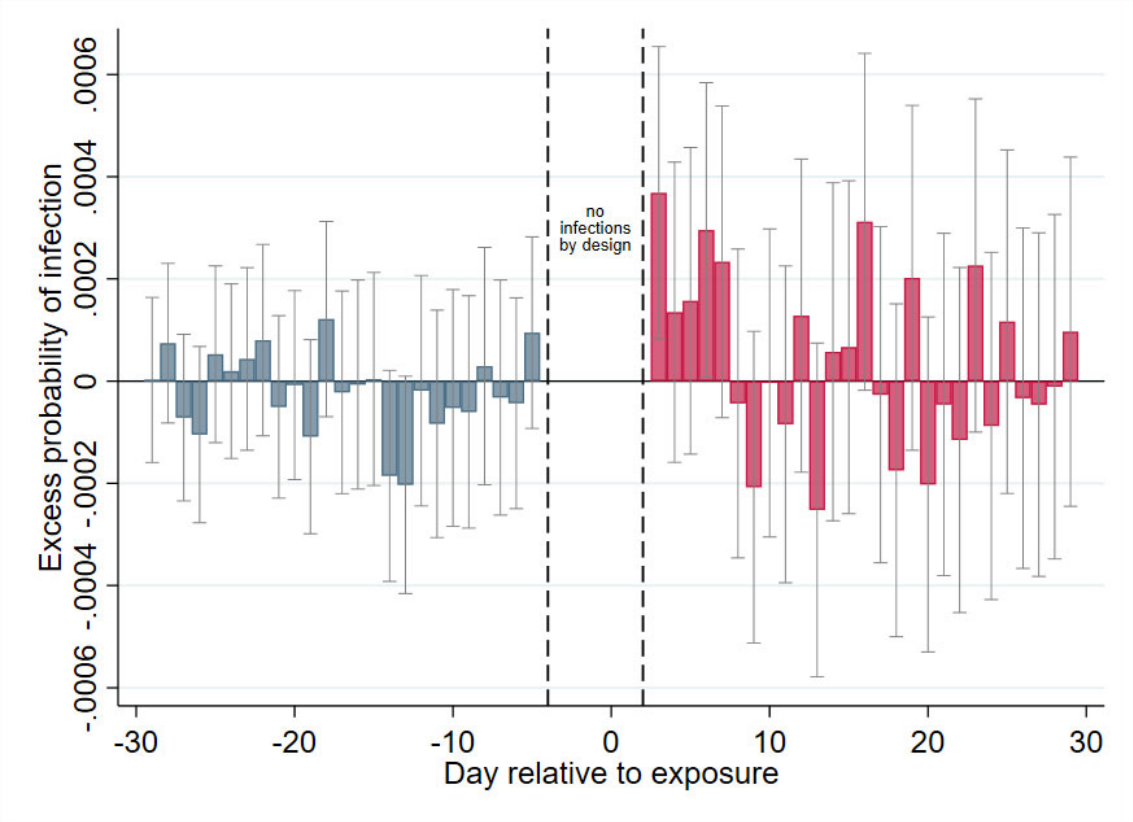

Figure S9: Dynamic results with alternative exposure measures. The bars indicate the excess probability of testing positive in different 5-day periods for exposed relative to non-exposed individuals using two alternative exposure measures: transactions with 1 minute (Panel A) and 10 minutes (Panel B) of the potential infector.

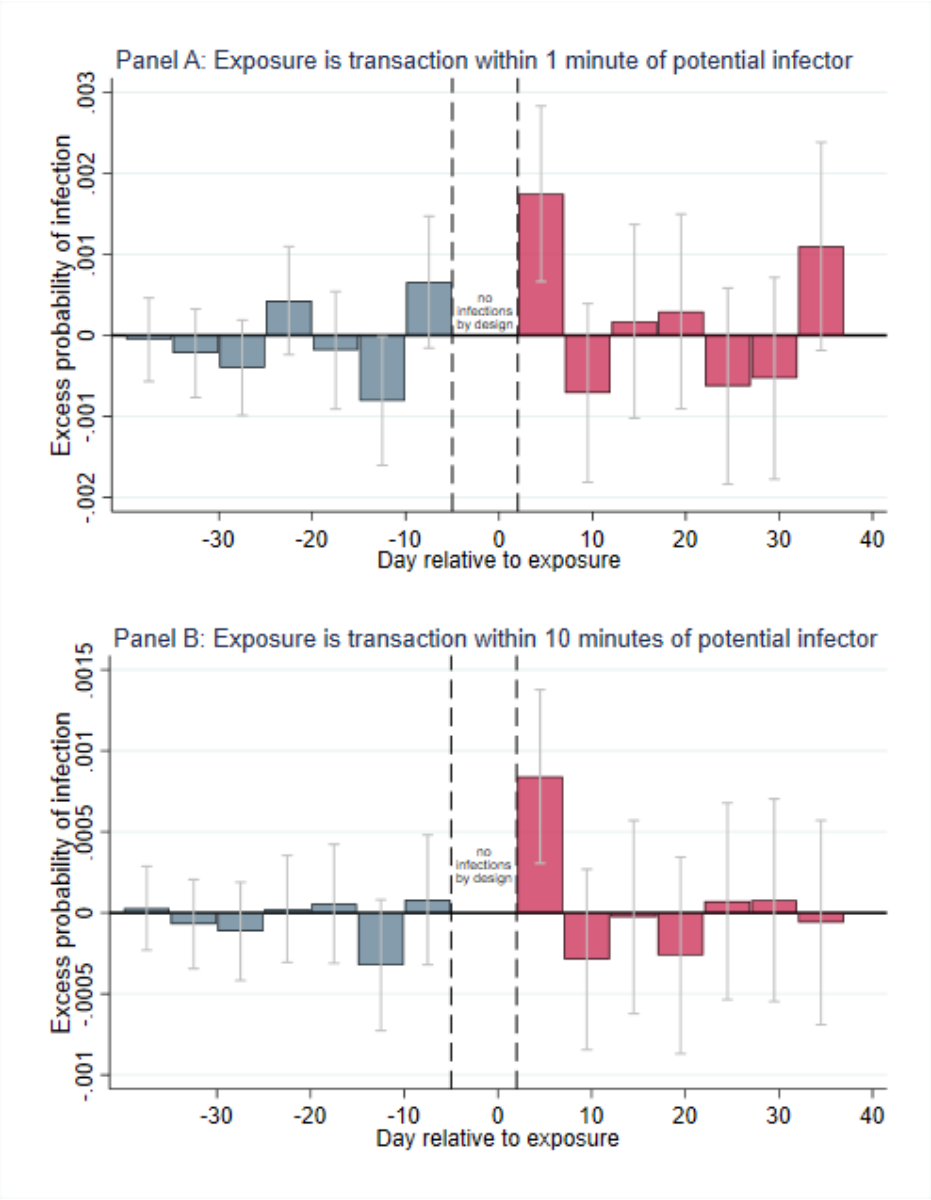

Figure S10: Dynamic results with controls. The bars indicate the excess probability of testing positive in different 5-day periods for exposed relative to non-exposed individuals estimating in a model with additional controls: indicators for birth year interacted with calendar days, indicators for income percentile interacted with calendar days, indicators for municipality of residence interacted with calendar days.

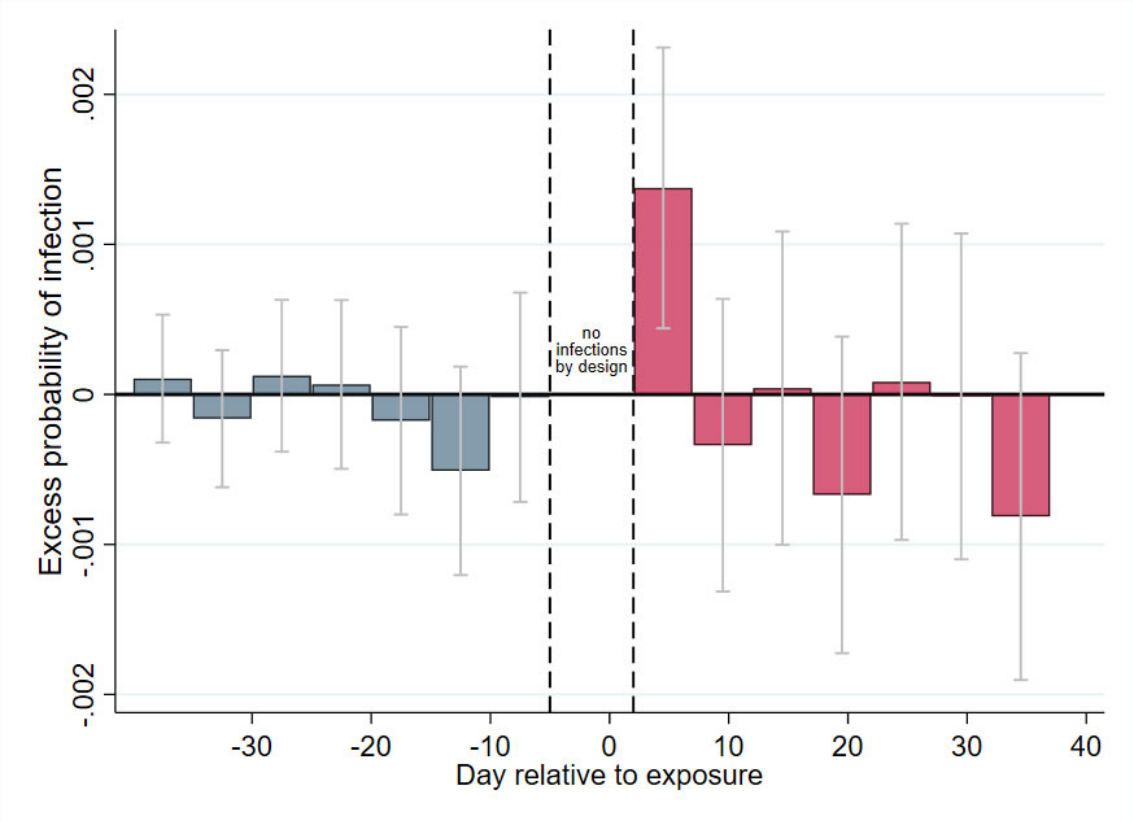

Figure S11: Differential transmission rate of Omicron. The figure shows the differential transmission rate of Omicron relative to other variants estimated in five different models: No controls (red columns), controls for calendar month (blue bars), controls for mask requirement (green bars), controls for the age of the exposed individual (brown bars), all three controls jointly (gray bars).

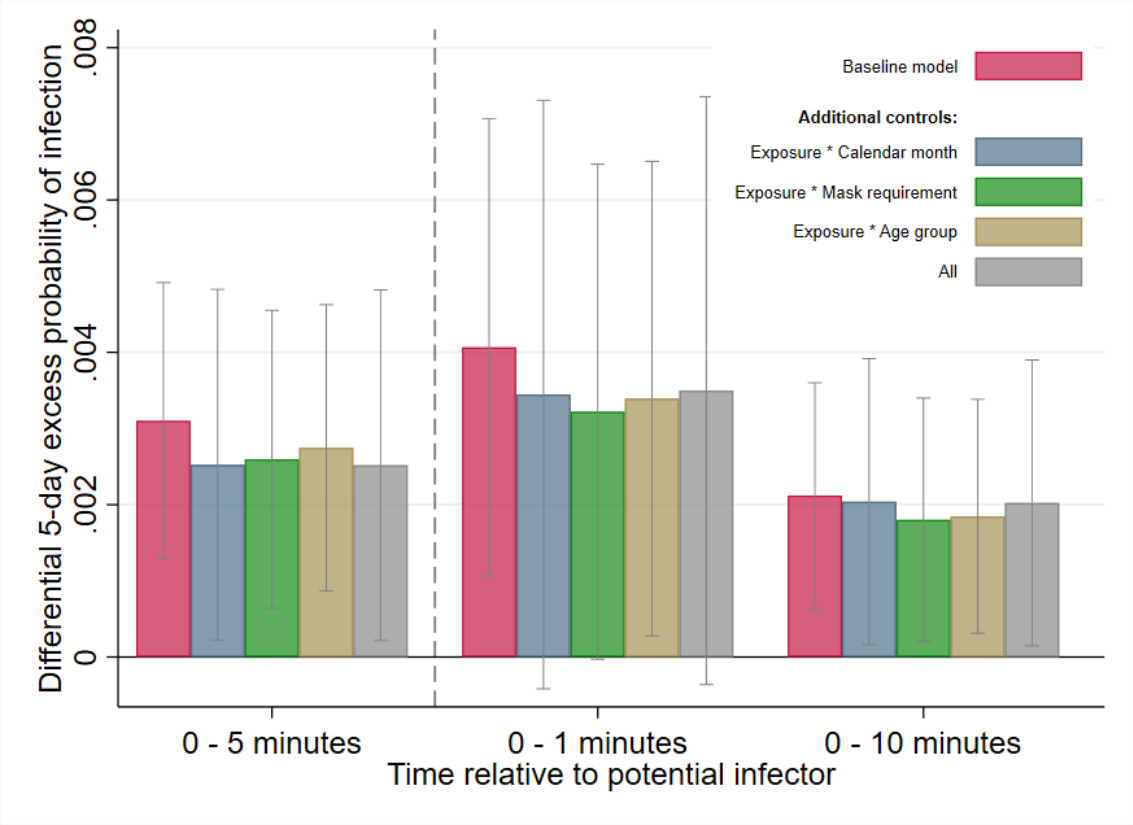

Figure S12: Heterogeneity by individual characteristics. The figure shows the excess probability of testing positive between day d+3 and day d+7 for exposed individuals relative to non-exposed individuals by the age of the exposed individual (red bars), the age of the potential infector (blue bars) and the gender of the exposed individual (green bars). The estimates are obtained by estimating the baseline model while restricting the sample to individuals with a given characteristic.

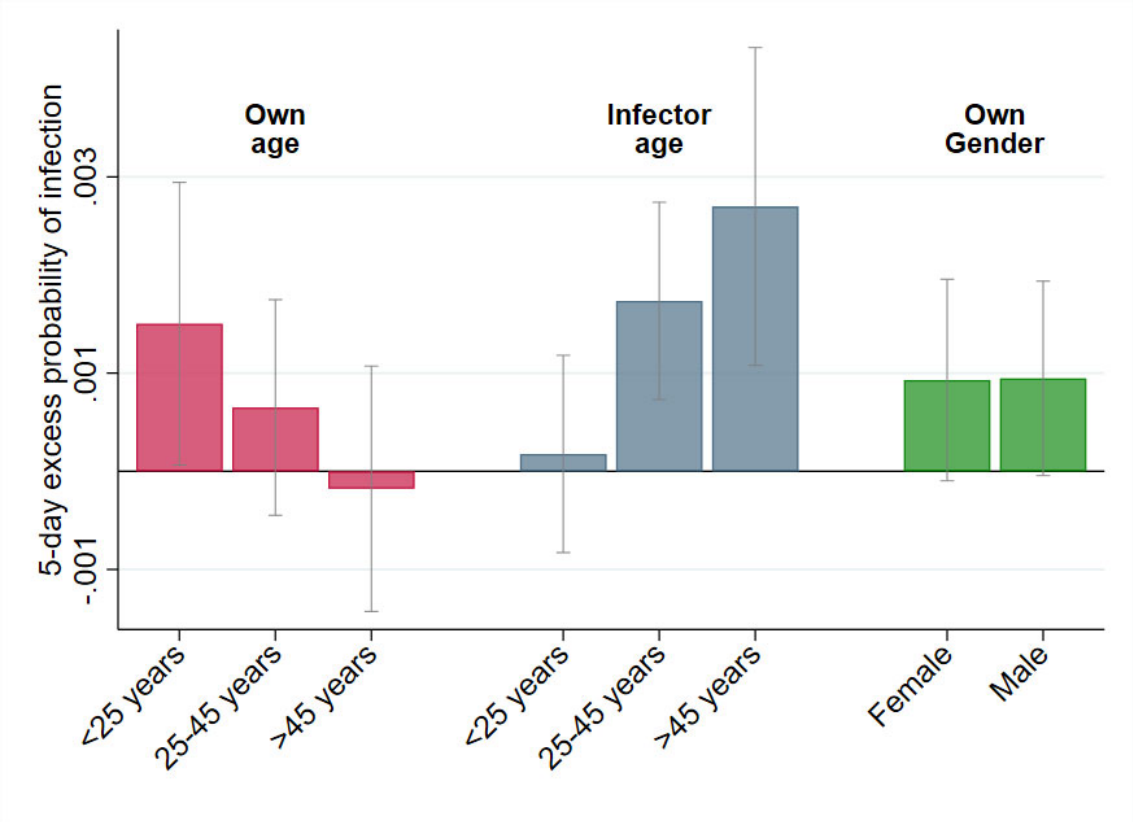

Figure S13: Endogenous testing. The figure shows the excess probability of testing positive (red bars) and testing negative (green bars) between day d+3 and day d+7 for exposed individuals relative to non-exposed individuals.

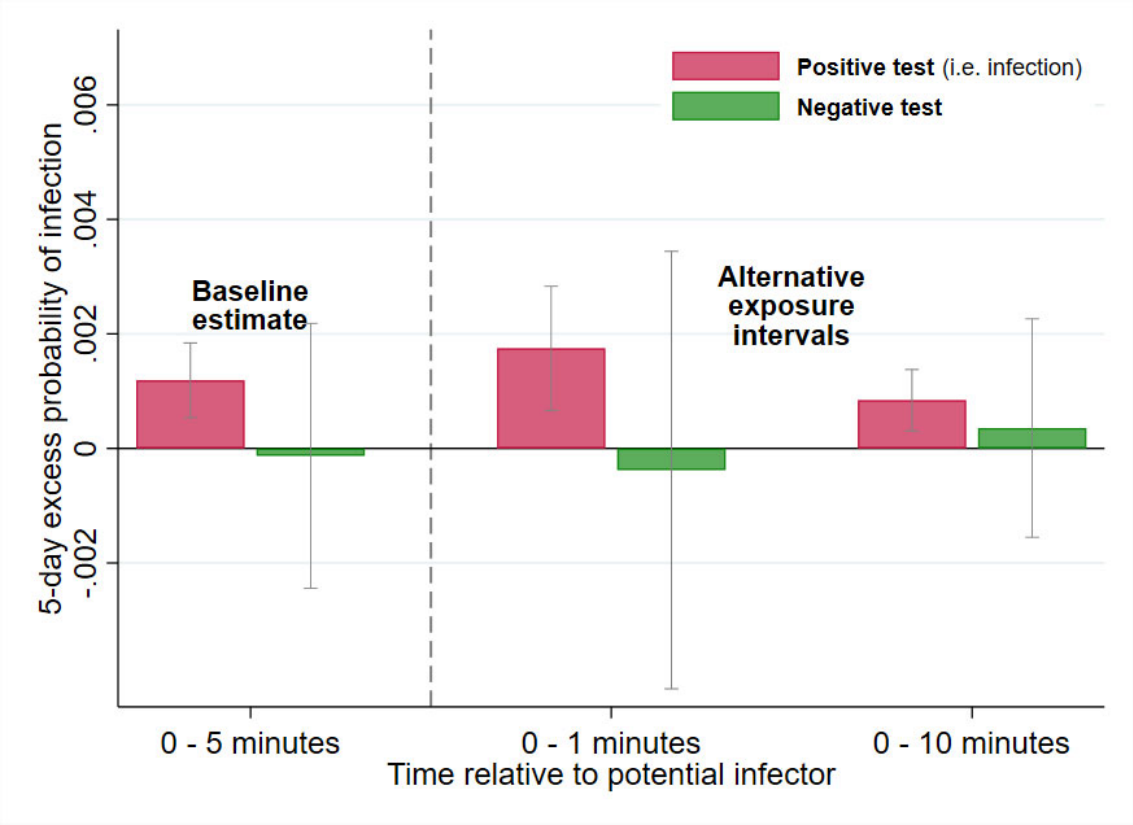

Figure S14: Number of potential infectors. The figure illustrates the distribution of the number of potential infectors with a transaction within 10 minutes of the exposed individual.

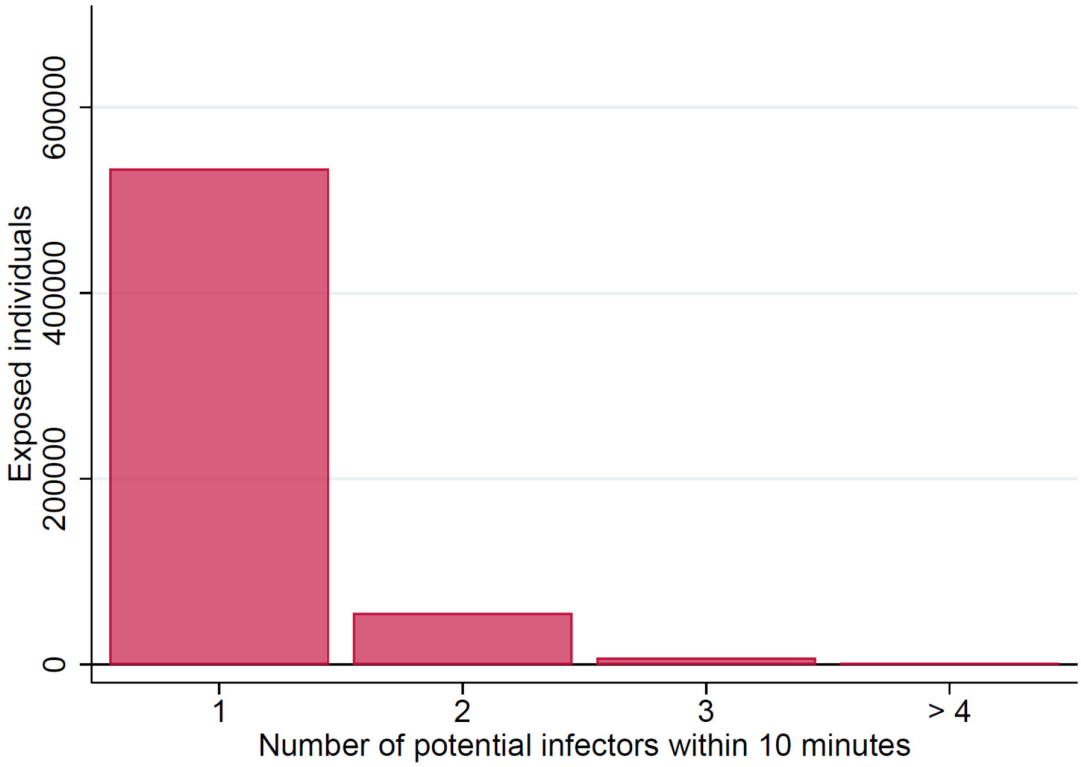

Figure S15: Robustness, subsample with only one potential infector. The figure shows the excess probability of testing positive between day d+3 and day d+7 for the baseline sample of exposed individuals (red bars) and for the subsample of exposed individuals with only one potential infector making a transaction within 10 minutes (green bars).

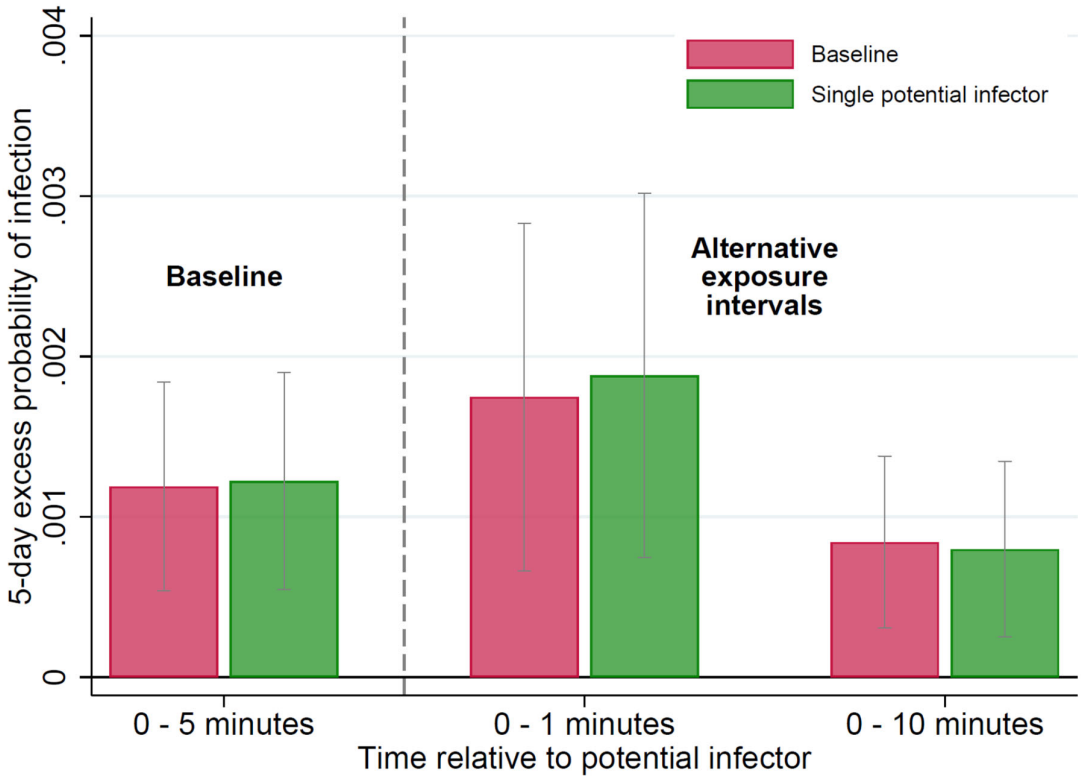

Figure S16: Estimation results for socially connected individuals. The figure shows the excess probability of testing positive between day d+3 and day d+7 for the baseline sample of exposed individuals (red bars); for the sample of socially connected individuals (blue and gray bars); and for the sample of individuals who made a transaction within 1 minute and 5 minutes of the potential infector on some other occasion (green bars). We emphasize that these are not estimates of in-store transmission probabilities as socially connected individuals are likely to be exposed to the potential infector outside of the store.

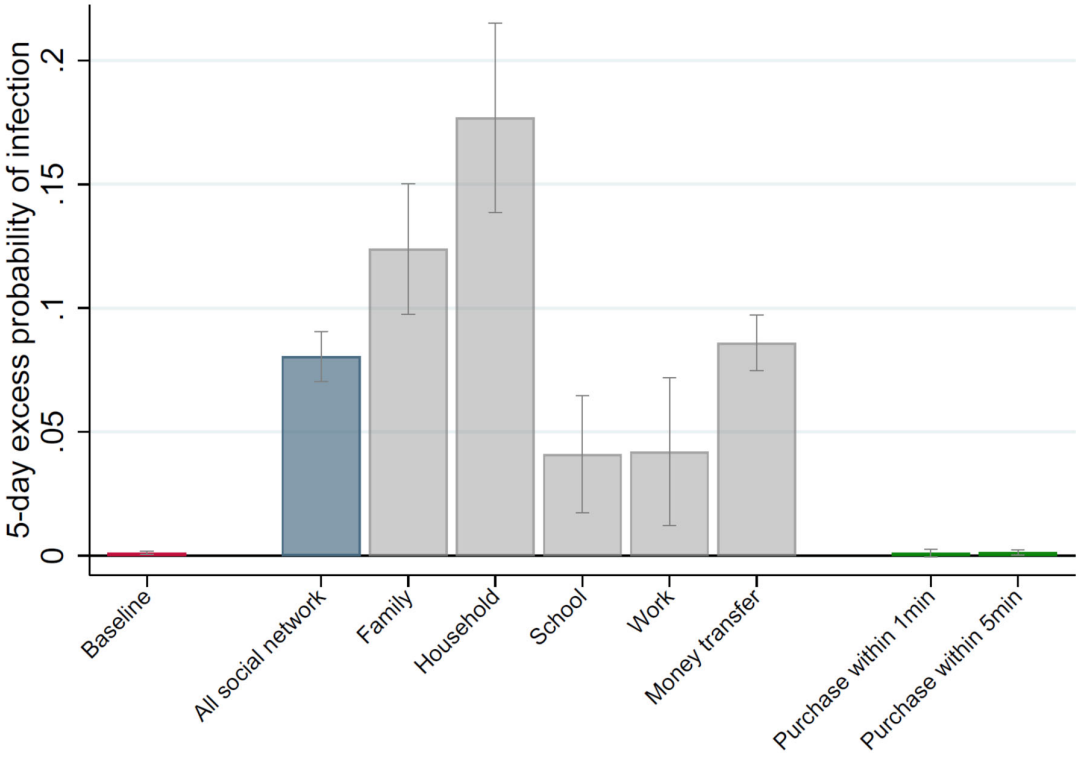

Figure S17: Inverse sampling probabilities. The figure illustrates two sets of estimates of the inverse sampling probabilities. The baseline estimates of the sampling probabilities are the ratio between the number of individuals at a given age in the population relative to the number of individuals whose main bank is Danske Bank in our sample, weighted by the share of MasterCard payments in their supermarket and grocery store transactions. The alternative approach obtains sampling probabilities as the predicted outcomes of a probit regression of an indicator for being in the sample of MasterCard holders on a range of observable characteristics: age, gender, income deciles, occupation dummies, and regional dummies. The resulting probabilities are scaled by the share of MasterCard payments in their overall supermarket and grocery store transactions. The gray area illustrates the range between the 90th and 10th percentile at each age with the alternative approach.

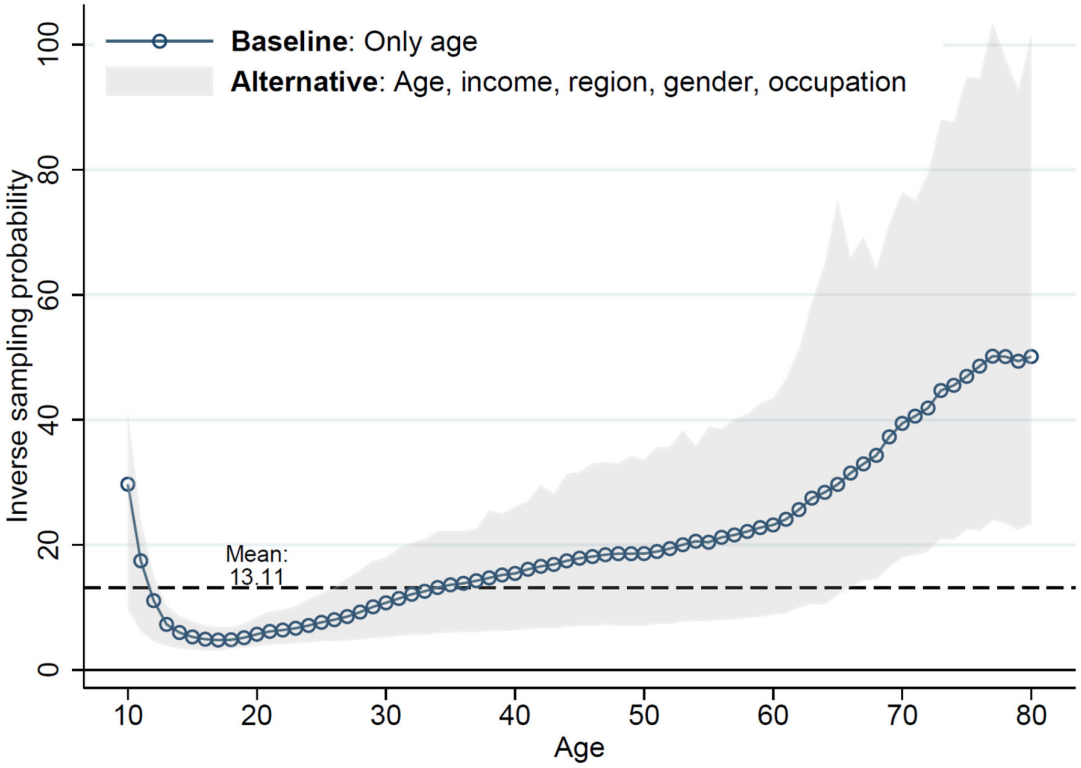

Table S1: Descriptive statistics. The table compares the sample of Danske Bank customers with MasterCard to the general population in Denmark in terms of sample size (Panel A) and individual characteristics (Panel B). Annual income is total pre-tax income including government transfers in 2019.

	Danske Bank sample	Population in Denmark
Panel A: Sample size		
Number of individuals	630,042	5,822,742
Panel B: Individual characteristics		
Age	37.7	41.3
- under 25	0.34	0.29
- between 25 and 45 years	0.30	0.26
- above 45	0.36	0.45
Female	0.48	0.50
Any children in household	0.41	0.43
Household size	2.86	2.58
Annual income (DKK)	181,690	266,072

Table S2: Sample selection. The table describes the selection of our estimation sample. We identify transactions by potential infectors: around 126,000 transactions and 54,000 unique individuals (Panel A). We define a gross sample of around 1,517,000 transactions that occur within 30 minutes of a potential infector transaction (Panel B). We exclude the roughly 11,000 of these transactions made by individuals socially connected to the potential infector (Panel C). In the remaining sample, we identify around 328.000 transactions occurring within 5 minutes of a potential infector and classify the card owner as exposed (Panel D) and around 340,000 transactions occurring between 16 and 30 minutes before the potential infector where we classify the card owner as non-exposed (Panel E).

A: Potential infectors	
- transactions	126,418
- unique individuals	53,506
B: Gross sample	
- within 30 minutes	1,517,105
C: Social connections	
- same family	1,326
- same household	827
- same school	1,291
- same workplace	829
- same payment network	6,958
D: Exposure	
- within 5 minutes	327,850
- within 1 minute	96,937
- within 10 minutes	598,006
E: Non-exposure	
- 16-30 minutes before	340,199

Table S3: Balancing tests. The table compares the characteristics of exposed and non-exposed individuals. Specifically, for a range of socio-demographic variables (Panel A) and behavioral variables (Panel B), the shows the sample mean for exposed and non-exposed individuals (Columns 1-2), the raw difference in means (Column 3) and the difference in means conditional on a separate intercept for each potential infector transaction (Column 4).

	Exposed (mean)	Non-exposed (mean)	Raw difference	Conditional difference
Panel A: Socio-demographics				
Age	33.57	33.76	-0.18	-0.30
	(0.03)	(0.03)	(0.04)	(0.04)
Female	0.503	0.502	0.001	-0.004
	(0.001)	(0.001)	(0.001)	(0.001)
Single	0.206	0.210	-0.004	-0.004
	(0.001)	(0.001)	(0.001)	(0.001)
Children	0.627	0.628	-0.001	-0.001
	(0.002)	(0.002)	(0.003)	(0.003)
Household size	2.91	2.89	0.01	0.02
	(0.003)	(0.003)	(0.004)	(0.004)
Annual income (DKK)	140,261	139,668	593	-126
	(516)	(405)	(656)	(756)
Panel B: Behavior				
Positive test (30 days)	0.0328	0.0325	0.0003	-0.0007
	(0.0004)	(0.0003)	(0.0005)	(0.0005)
Number tests (30 days)	1.5284	1.5074	0.0211	0.0050
	(0.0049)	(0.0042)	(0.0064)	(0.0064)
Any tests (30 days)	0.5061	0.5032	0.0029	-0.0002
	(0.0010)	(0.0009)	(0.0013)	(0.0014)
Health worker	0.0695	0.0699	-0.0004	-0.0013
	(0.0005)	(0.0004)	(0.0007)	(0.0008)
Education worker	0.0667	0.0654	0.0013	0.0011
	(0.0005)	(0.0004)	(0.0007)	(0.0007)
All in-store transactions (2019)	495.61	495.62	-0.01	0.69
	(0.65)	(0.56)	(0.85)	(0.96)
- in supermarkets	204.00	205.01	-1.01	-1.12
	(0.35)	(0.30)	(0.47)	(0.52)
- in supermarkets, peak-hour	19.65	19.71	-0.06	-0.08
	(0.04)	(0.04)	(0.06)	(0.06)

Table S4: Main results and robustness. The table reports the estimated coefficients and standard errors illustrated in Figures 2, S6 and S7.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Baseline Figure 2		**Robustness to controls** Figure S6					**Robustness to samples** Figure S7		
		Age controls	Area controls	Income controls	All controls	No purchase within 1min	No purchase within 5min	Age difference > 5 years	Non-exp: [-30,-15]+[15,30]	Non-exp: [-30,-20]+[20,30]
Baseline (<5min)	0.00119***	0.00120**	0.00125**	0.00123***	0.00132**	0.00119***	0.000871*	0.00135***	0.000826**	0.000956**
	(0.000)	(0.001)	(0.001)	(0.001)	(0.006)	(0.001)	(0.029)	(0.000)	(0.004)	(0.002)
Observations	547,435	539,935	534,356	539,328	515,702	493,575	415,335	410,935	873,868	649,077
Alt. exposure (<1min)	0.00175**	0.00183**	0.00215**	0.00169**	0.00218*	0.00167**	0.00153*	0.00186**	0.00124*	0.00128*
	(0.002)	(0.004)	(0.002)	(0.005)	(0.017)	(0.005)	(0.024)	(0.003)	(0.013)	(0.015)
Observations	369,976	361,491	354,922	361,857	332,718	334,378	283,219	274,672	698,050	471,893
Alt. Exposure (<10min)	0.000842**	0.000842**	0.000820**	0.000891**	0.000836**	0.000845**	0.000644*	0.000939**	0.000544*	0.000743**
	(0.002)	(0.002)	(0.006)	(0.004)	(0.004)	(0.004)	(0.045)	(0.002)	(0.016)	(0.003)
Observations	782,496	775,766	769,717	774,698	753,903	708,514	598,195	592,051	1,106,532	883,512

Table S5: Covid-19 variants. The table reports the estimated coefficients and standard errors illustrated in Figure 4.

	(1)	(2)	(3)	(4)
		Figure 4		
	Index	**Alpha**	**Delta**	**Omicron**
Baseline (<5min)	0.000548	-0.000103	0.000351	0.00310***
	(0.082)	(0.750)	(0.415)	(0.001)
Observations	121,063	61,384	190,672	174,314
Alt. exposure (<1min)	0.00110*	-0.000431	0.000838	0.00407**
	(0.048)	(0.406)	(0.248)	(0.009)
Observations	82,658	42,025	129,314	115,977
Alt. Exposure (<10min)	0.000634*	-0.000177	0.000175	0.00212**
	(0.013)	(0.514)	(0.624)	(0.005)
Observations	171,218	86,828	270,970	253,478