UNIVERSITY of HAWAI'I®
MĀNOA

University of Hawai`i at Mānoa Department of Economics Working Paper Series

Saunders Hall 542, 2424 Maile Way,
Honolulu, HI 96822
Phone: (808) 956 -8496
www.economics.hawaii.edu

Working Paper No. 23-08

Estimating Intergenerational Health Transmission in Taiwan with Administrative Health Records

By
Harrison Chang
Timothy J. Halliday
Ming-Jen Lin
Bhashkar Mazumder

October 2023

Estimating Intergenerational Health Transmission in Taiwan with Administrative Health Records*

Harrison Chang
University of Toronto

Timothy J. Halliday
University of Hawai'i at Mānoa
IZA

Ming-Jen Lin
National Taiwan University

Bhashkar Mazumder†
Federal Reserve Bank of Chicago

October 21, 2023

Abstract

We use population-wide administrative health records from Taiwan to estimate intergenerational persistence in health, providing the first estimates for a middle income country. We measure latent health by applying principal components analysis to a set of indicators for 13 broad ICD categories and quintiles of visits to a general practitioner. We find that the rank-rank slope in health between adult children and their parents is 0.22 which is broadly in line with results from other countries. Maternal transmission is stronger than paternal transmission and sons have higher persistence than daughters. Persistence is also higher at the upper tail of the parent health distribution. Persistence is lower when using inpatient data or when using total medical expenses and may overstate mobility. Health transmission is almost entirely unrelated to household income levels in Taiwan. We also find that that there are small geographic differences in health persistence across townships and that these are modestly correlated with area level income and doctor availability. Finally, by looking at persistence within health conditions that vary in their genetic component, we find little evidence that health persistence is driven by genetic factors.

Key Words: intergenerational mobility, health, administrative data, genetics

*We thanks participants at the 2020 HCEO Conference on New Approaches to Intergenerational Mobility and seminar participants at the Tinbergen Institute, the University of Hohenheim, the Florida Applied Economic seminar, the 2023 IHEA conference and the Research Network on Intergenerational Mobility. The views expressed here are not those of the Federal Reserve Bank of Chicago or the Federal Reserve system.

†Corresponding Author: bhash.mazumder@gmail.com.

1 Introduction

Equality of opportunity has emerged as one of the most salient issues in many countries. With the rise of inequality in much of the world, policymakers are increasingly concerned about whether a person's chances of socioeconomic success are constrained by their family background. This has led to an enormous interest in studies of intergenerational mobility in socioeconomic status and the mechanisms that underpin it.

Over the last few decades there has also been increasing recognition that health is highly related to socioeconomic status (Case et al., 2005) and is a critical component of welfare (Sen, 1998; Jones and Klenow, 2016). Accordingly, a new line of research has begun to examine intergenerational persistence in health. This is challenging as health is a latent concept with potentially many dimensions that may not be easily captured in data. Several studies have used long-running panel data with survey-based measures such as the SF-12 or self-reported health status, to combine multiple measurements over as many years of the lifespan as possible, to proxy for health (Halliday et al., 2021, Bencsik et al., 2023, and Vera-Toscano and Brown, 2021). One study that we are aware of, Andersen (2022), has used administrative health records (for Denmark) to construct a broad-based measure of latent health, though it relied heavily on inpatient records.[1]

We make a number of contributions to the literature. First, we are the first to utilize population-wide administrative data containing outpatient health records on both parents and their adult children to estimate intergenerational health persistence. Second, we produce broad-based health mobility estimates for an Asian middle income country. Third, due to our extremely large samples we are able to identify statistically significant and meaningful differences by parent and child gender. Fourth, we highlight notable non-linearities in the intergenerational relationship. In particular, parents in the top decile of the health distribution have especially healthier adult children. Fifth, using

[1]Butikofer et al (2023) use administrative data on mental health in Norway but do not examine a measure of overall health status

data on inpatient visits or medical expenses leads to lower estimates of persistence and overstates health mobility. Sixth, we find remarkably little difference in health mobility across families in different deciles of the income distribution. We also find relatively small differences across cities and townships in Taiwan, though the availability of doctors can account for some of these small mobility gaps. Finally, using categories of more specific health conditions, we find no relationship between intergenerational transmission in these categories and their genetic relatedness (as proxied by twin correlations). This suggests that environmental factors play an important role in the intergenerational transmission of health.

Our data come from Taiwan's National Health Insurance (NHI) system. We cull records on all inpatient and outpatient visits from 2000 to 2019. We couple this with household registration data on family relationships, which allow us to link nearly all children with at least one parent. Our main sample consists of children born between 1979 and 1981 and their parents.

Our approach is very similar to Andersen (2022) who used principal component analysis (PCA) analysis to extract a measure of latent health. We use 18 different variables consisting of 13 broad categories of health conditions based on International Classification of Disease (ICD9 and ICD10) codes, and 5 indicator variables, one for each quintile in the distribution of the general practitioner visits. A distinguishing feature of our analysis compared to Andersen (2022), is that we are able to use data on *outpatient* as well as inpatient utilization. The latter tends to pick up only more severe illness. We show that using inpatient data yields much smaller estimates of intergenerational health persistence.

Our main measure of intergenerational persistence is the rank-rank slope in latent health. This provides a measure of *positional mobility* and shows the average association between a percentile change in the health rank of parents with that of the child in adulthood. This measure allows us to understand how much reshuffling occurs in the relative positions of families over a generation in terms of latent health. This provides a natural

metric for mobility that can be compared across different domains (e.g. education, income or wealth). We also present figures showing the rank-rank relationship at each centile of the parent distribution, providing descriptive estimates of upward and downward mobility throughout the distribution. In addition, we show estimates of the intergenerational health association (IHA) in levels which are generally quite similar in magnitude.

Our first finding is that the intergenerational rank-rank slope for health in Taiwan is 0.22 when we pool sons and daughters and combine the health of both parents. The analogous estimate of the IHA is 0.26. We also show similar results using a simpler measure of health persistence, the correlation in the number of broad health conditions experienced in each generation. We further document that persistence of maternal health is stronger than paternal health and that there is higher transmission to sons than to daughters. We also estimate a sharp increase in the rank-rank relationship in the upper tail of the parent health distribution that is robust to several methods to deal with potential measurement error. This suggests that very healthy parents in Taiwan are uniquely able to transmit their health to their children.

We examine heterogeneity in health persistence across several key dimensions. First, we examine families at different deciles of the parent income distribution and find surprisingly little difference. We also look across cities and townships in Taiwan and find small geographic differences. A portion of these geographic gaps can be accounted for by the availability of doctors, highlighting one potential policy lever. Finally, we look across our 13 broad categories of health conditions to see if there are important differences by their degree of genetic relatedness, proxied for by twin correlations, and find no relationship.

We also make several methodological contributions. First, we show that using more factors in the PCA analysis is critical to better estimating latent health. The estimates of persistence increase in magnitude as the number of factors increases. Thus, the availability of more health conditions in administrative data is akin to having more years of self-reported health data in studies using survey data such as Halliday et al. (2020) and

Halliday et al. (2021). Second, we show that including outpatient data on health conditions delivers much larger estimates than using only inpatient data. This is because outpatient data provides a more comprehensive measure of latent health that includes less severe conditions. Finally, we are the first study that we are aware of to estimate the intergenerational rank persistence in outpatient medical expenses, which provides a largely independent metric for proxying for health. We estimate correlations ranging from 0.10 to 0.15. However, medical expenses may be more accurately thought of as a measure of health care utilization, and prices, rather than as a measure of latent health.

Our estimates add to the growing body of cross-country estimates of intergenerational health persistence and mobility. Notably, we provide new evidence from population-wide administrative records, that shows that persistence in health is quite low and may be lower than that in income. Similar findings have previously been found for the US (Halliday et al., 2021), the UK (Bencsik et al., 2023), Germany (Graeber, 2020) and Denmark (Andersen, 2022). This is a provocative finding given how important health is to overall welfare and should serve as a complement to the vast majority of studies of intergenerational mobility which rely on outcomes such as income or education. When considering the extent to which there is equality of opportunity in any society, policy makers should extend their reach to consider health, given how powerfully it reflects human welfare.

2 Data and Measurement

Our data come from Taiwan's government-run single payer health insurer, the NHI, which was introduced in 1995. We use inpatient and outpatient records from three modules of NHI data spanning the years 2000 to 2019.

2.1 Family Relationships and Sample Construction

To construct family relationships, we use the NHI enrollment files supplemented with household registration data from Taiwan's Hukou system.[2] In most cases, the NHI data is sufficient. However, when there are multiple insurance plans within the household, we supplement the NHI data with the household registration data.[3]

Figure A.1, illustrates match rates by the birth year of the child. Since the NHI was only introduced in 1995, parents of earlier birth cohorts (born prior to 1980) are much less likely to have enrolled their children and therefore are much less likely to have been matched.[4] For cohorts born by the 1990's, match rates of children to at least one parent are extremely high. However, if we used these cohorts for our analysis, then we would only observe child health outcomes at quite young ages which would likely not be ideal.[5] The literature has not yet established the ideal age range to observe latent health, but a reasonable supposition is that it is valuable to observe health as far out into the lifecycle as possible, given that many chronic illnesses such as heart disease and diabetes may not manifest until individuals are in their fifties.[6] Accordingly, there is a potential trade-off between match rate and a possible "life-cycle bias".

Therefore, we use cohorts born between 1979 and 1981. These cohorts have a relatively high match rate of 86% but are also observed into their late thirties which should minimize any life-cycle bias. Moreover, because we have access to outpatient records, we are able to pick up some less severe conditions that could materialize earlier in the life-cycle.

In section 4.6, we show that our estimates are not sensitive to these cohort restrictions. In Figure A.2, we provide the distributions of the parents' birth years for mothers and

[2]We provide additional detail on how we constructed family relationships in Appendix C.

[3]Working parents pay progressive insurance fees based on their own wage. Children can then be insured by either parent. In extremely rare cases, children can also be insured by other relatives.

[4]For example, a child born in 1977 would have been 18 in 1995 and possibly living on their own. Therefore, they are much less likely to have been covered by a parent.

[5]For example, a child born in 1992 would only be 25 years old in 2019, the last year of our data.

[6]In the income mobility literature observing children at too young an age ("life-cycle bias") can lead to considerable attenuation bias (Haider and Solon, 2006).

fathers. Fathers were mostly born in the mid- and late-1950's and are slightly older than mothers who were mostly born in the late 1950's and early 1960's.

We report additional detailed information on the match rates for the 1979-1981 cohorts in Table B.2. Among these cohorts 31% are matched with one parent while the remaining 55% are matched with two parents. In total, we have 1,281,502 children including 637,962 daughters and 643,540 sons.

2.2 Latent Health Measurement

In order to study health mobility, it is crucial to create a reliable measure of latent health. Several studies that use survey data such as Halliday et al. (2020, 2021) use long time averages of a single self-reported health status (SRHS) variable to extract a latent health measure. Bencsik et al. (2023) use a similar approach but they compute time averages of the multidimensional SF-12. Our strategy is guided instead by Andersen (2022), who also uses administrative data, but largely relied on inpatient claims. Specifically, we use ICD9/10 codes from NHI claims data from 2000 to 2019 on all outpatient visits, along with indicator variables for occupying each of five quintiles in the number of general practitioner (GP) visits.[7] Rather than using time averaging to proxy for latent health, as with survey data, we extract the first principal component (PC1) from the variation in our data. We find that the number of conditions (factors) plays a role similar to the length of the time average in studies using survey data *e.g.* Halliday et al. (2021).

Importantly, our access to outpatient records ensures that we can identify conditions that are not severe enough to warrant a hospitalization but still proxy for latent health. This is because outpatient data provides more subtle insights into health earlier in the life-course, prior to the onset of many more serious conditions that warrant a hospitalization. We also conduct a supplementary analysis with inpatient data to show how our results

[7]The NHI used ICD9 codes from 2000 to 2016 and then transitioned to ICD10 codes after 2016.

compare.

We focus on a set of indicator variables for each of 13 ICD categories. These take on a value of one if an individual had outpatient contact with the medical system for that category, and zero, otherwise. We report descriptive information for outpatient utilization for both generations in Table 1. For comparison, we also report the same information for inpatient utilization in Table B.1. As expected, there is substantially more outpatient utilization. For example, 19% of sons in the data had an outpatient visit to treat cancer but only 2% were hospitalized due to cancer. This illustrates why we focus on outpatient utilization in this paper.

GP visits provide additional information on the severity of illnesses. We exclude GP visits for pregnancy, injury, poisoning, and preventive exams since we are interested in capturing sickness. Since there is universal health insurance in Taiwan, there is no financial cost for a GP visit. Therefore, our measure proxies for health status as opposed to just health *access*. Doorslaer et al. (2004) show that there is no income-based inequality in GP visits in 12 European countries with health care systems similar to Taiwan, suggesting that GP visits measure health status, and not economic status, in these countries.

In total, we use 18 variables for the PCA procedure (13 ICD categories and five indicators for GP visit quintiles). We compute PC1 separately for each generation. We report descriptive statistics in Tables 1 (based on outpatient utilization) and B.1 (based on inpatient utilization). In Figure A.3, we display the scree plots for daughters, sons, mothers, and fathers. The figure clearly demonstrates that PC1 accounts for a much larger proportion of the total variance of the 18 outcomes. Its variance ranges between 3.5 and four whereas the second component has a variance in the neighborhood of 1.5.

Finally, in Figure A.4, we display the factor loadings used to compute PC1. We find that the loadings range from roughly 0.1 to 0.3 among our ICD conditions. The loadings for the GP visit quintiles follow the expected pattern with the lowest quintile having a negative coefficient (indicating better health) and the highest quintile with a positive

coefficient that is on par with many of the loadings for the more serious conditions.

3 Empirical Strategy

To measure health mobility we estimate standard rank-rank mobility regressions based on ranks of PC1 (Chetty et al., 2014; Halliday et al., 2021; Andersen, 2022). To fix ideas, we let y_i^{MP} and y_i^{MC} denote the rankings of PC1 for the parents and children respectively where the principal component was computed from a total of M proxies. Rankings of PC1 are computed separately by generation and gender. Our preferred estimates use $M = 18$, but we also provide estimates using fewer proxies to investigate the role that the number of proxies plays in estimates of health mobility. This is similar to an exercise from (Halliday et al., 2021) except that it uses the number of proxies in lieu of the number of periods used in the time average. Our rank-rank estimates are based on equation

$$y_i^{MC} = \alpha + \beta y_i^{MP} + X_i'\theta + \upsilon_i$$

where β delivers the rank-rank correlation. The vector, X_i, only includes a parsimonious set of covariates including a quadratic in average ages of the parents and children and a gender dummy when appropriate. For our dependent variable, we use either all children pooled, only daughters or only sons. We run each of these against PC1 measured for: only the mother; only the father; or a combination of both parents. This yields a total of nine different estimates.[8] We will focus much of our attention on the rank-rank slope when pooling children and combining both parents' health. We will also compare and test for differences in coefficients by gender and parent type. Throughout, we compute robust standard errors.

[8]For our combined parents measure we average the ranks of mothers and fathers and then re-rank this new averaged measure so that it has the properties of a rank, i.e. a uniform distribution.

4 Results

4.1 Full Sample Estimates

In Figure 1, we show a bin scatter plot for our main estimate in which we pool children and combine both parents' health. The rank-rank slope is 0.218. This is shown as the linear slope in the figure. For comparison, estimates based on survey data and using a different methodology are 0.26 for the U.S. (Halliday et al., 2021), 0.17 for the UK (Bencsik et al., 2023) and 0.20 for Australia (Vera-Toscano and Brown, 2021).

Figure 1 also makes clear that there are notable non-linearities in the relationship, particularly at the top of the parent health distribution. This is apparent if we run a lowess smoother, but is also quite evident simply by looking at the dots representing each centile. For children whose parents health is at the 99th percentile, their expected health rank is the 70st percentile. However, a simple linear rank-rank slope would suggest that their expected rank would be the 62th percentile. Similarly, there is a notable gap of 3 percentiles in the expected rank of children starting at the very bottom centile of the parent health distribution relative to what the rank-rank slope would imply.

4.2 Differences by Gender

We present the full set of rank-rank slope estimates in Panel A of Table 2 for all parent-child combinations. The columns of the table correspond to the parent measure and the rows to the child measure. Each cell is an estimate from a separate regression. The estimates in the table vary from 0.134 for dads to daughters to 0.198 for moms to son.

An intriguing finding is that mothers have stronger associations with their children's health than fathers have. We see this in the fourth column that presents the difference between columns (1) and (2). The difference in the parent coefficients is especially large for daughters at 0.059. This suggests that the transmission from mothers to daughters is

44 percent higher than the transmission from fathers to daughters. All of these coefficients are highly statistically significant given our near population level samples for these cohorts.

In addition, sons have stronger associations with their parents' health than daughters. This is shown in the fourth row where we take the difference between sons and daughters. This is particularly pronounced for paternal transmission where the difference is 0.042. This implies that transmission from fathers to sons is about 31 percent higher than the transmission from fathers to daughters. We think further exploration of theses gender differences and their causes are ripe areas for future research.[9]

4.3 Inpatient vs Outpatient Data

While our estimates are consistent with other survey-based estimates, they are notably higher than estimates from Denmark in Andersen (2022). For example, the rank-rank slope for moms and sons in Denmark is 0.123 whereas it is 0.199 in Taiwan (see Table 2). At a first glance this suggests that there is more health mobility in Denmark than in Taiwan. However, a major difference between both of these studies is that Andersen (2022) uses inpatient data for health conditions while we use outpatient data. Both studies use quintiles of doctor visits which reflects outpatient care.

In Panel B of Table 3, we estimate the same models that we estimated in the main results from Table 2 but we use ICD codes for inpatient utilization in lieu of outpatient utilization. We find vastly lower persistence estimates when we use inpatient data. In fact, the estimates tend to be roughly half as large. For example, using inpatient data the estimate that pools sons and daughters and combines parent health is 0.114 compared to our primary estimate of 0.218. This suggests that using detailed data on outpatient

[9]We attempted to explore one explanation using our data. Specifically we examined whether the relationship between parent health and health investments in children, differed by gender. We proxied for investments by using health claims for an "infant check-up". We used more recent children born in 2004-2006 whose parents were between the ages of 30-39 but found no evidence that this relationship differed by gender. However, given that this was just one measure and was only available for more recent cohorts, we did not think this was dispositive evidence.

utilization is critical if we want to obtain a more comprehensive picture of latent health.

4.4 Changing the Number of Health Proxies

We also investigate how changing the number of health proxies in the PCA analysis impacts our findings. We find that this does have an important effect on our results. We show this in Figure 2. We find a roughly linear increase in our rank-rank slope estimates as we increase the number of proxies. For example, the estimate is 0.16 with seven proxies, 0.185 with 12 proxies, and 0.22 with 18 proxies. This indicates that using more proxies lowers the attenuation bias in the persistence estimates. This is akin to how the increasing the length of the time average when using survey panel data delivers larger persistence estimates (Halliday et al., 2021).

4.5 Other Measures of Health Persistence

In Panel B of Table 2, we present estimates of the IHA which comes from a regression of PC1 for the child onto PC1 for the parent. In principle, these estimates could be quite different than rank-based estimates. This may be easier to see in the context of income where there are vast differences in the levels of income in many countries when moving from the median to the 99th percentile. In contrast, with health, there is a clear intuitive biological ceiling on how healthy one can be. In practice, the rank-rank slope estimates have tended to be quite similar to IHA estimates in prior studies. Ours is no exception as we find generally quite similar results. In some cases, IHA estimates are lower. For example the father-daughter IHA estimate is 0.110 compared to a rank-rank slope of 0.134. On the other hand when we pool both parents health, the IHA estimates are quite a bit larger. For example, the IHA estimate for both parents to sons is now 0.293 compared to a rank-rank sloe estimate of 0.230.

In Panel A of Table 3, we use an alternative health measure. Specifically, we replace

PC1, which came from a principle component analysis, with the sum of the broad-based ICD categories in which an individual ever had a health claim. We do not include GP visits. The estimates are very similar to the estimates in Table 2.

Finally, we present rank-rank estimates using medical expenditures in Panel C of Table 3. These estimates are somewhat smaller than what we obtained using the 18 health proxies in Table 2. When using a sample of pooled children and both parents, we obtain an estimate of 0.17 in rank persistence in medical expenditures. The comparable estimate using PC1 is 0.22 in Table 2. One interpretation of this result is that using medical expenditures provides a less complete picture of health status than using a more comprehensive measure that includes information on a full set of specific health conditions.

4.6 Robustness Checks

We perform a number of robustness checks in Figure 3 for both the rank-rank and IHA estimates. We report the baseline estimate at the bottom of each panel in the figure and then show how the estimates change for each robustness check. We also report the same calculations in Figure A.5 for each parent-child gender combination. We begin by showing how our estimates change when only one parent is identified in our data. In the case of the rank-rank slope, using just one parent lowers the coefficient from roughly 0.22 to 0.18. In the case of the IHA, it reduces the estimate quite a bit more, from about 0.26 to 0.18. Both of these estimates are statistically significantly different from our baseline results. The explanation for these reductions is that there is more measurement error in the measure of parent latent health when using just one parent to proxy for *overall* parent latent health. This is also evident in Table 2, where estimates that uses both parents are larger than those that only use one parent. Nevertheless, the estimates for the rank-rank slope are not very different when using just one parent.

We find that our estimates are robust to restrictions on parent birth cohort (using

only those born between 1949 and 1958) or parent age (using only those between the ages of 52 and 62). Similarly, our results are highly robust to various restrictions on which children's birth cohorts we use. This suggests that although there is a trade off between the match rate and the age at which child's health is measured, as discussed earlier, this doesn't appear to impact our conclusions regarding health persistence.

We also show that excluding observations in which medical expenditures were in the bottom and top 5% also appears to lead to lower estimates relative to our baseline results. In the case of the rank-rank slope, the estimate declines from roughly 0.22 to just over 0.18. In the case of the IHA, the estimate falls from about 0.26 to roughly 0.23. An explanation for the declines can be found in Figure 1 which highlights the strong non-linearities in the relationship between child and parent health rank at both tails of the parent health distribution. In particular, we find stronger associations at both the very bottom and very top of the parent health distribution. Thus, excluding the tails of the distribution will tend to lower the estimates. Therefore, keeping the tails of the distribution of medical expenditures is likely important to obtain accurate estimates of health persistence.

5 Heterogeneity

We now explore heterogeneity across a number of domains including income, geography, health conditions, and genetics.

5.1 Household Income

First, we explore heterogeneity in the relationship between child and parent health rank by *parent household* income. A growing literature has highlighted how health disparities may be rooted in differences in various aspects of socioeconomic status including income (e.g., Case, Paxson and Lubotsky, 2002). Our data allows us to approximate income

based on the schedule of health insurance rates that are based in part on income.[10] In order to examine heterogeneity, we divide our sample into ten deciles of parent income and estimate separate rank-rank slopes for each decile. The ranks of PC1 are still based on the full sample. Figure 4 plots the predicted rank-rank relationship for each decile. As is evident in the figure, the ten lines representing each decile, are virtually indistinguishable from each other suggesting that there is very little heterogeneity in rank-based measures of health mobility in Taiwan. Notably, not only are the slopes similar, but the entire profile of conditional expected ranks are similar. This suggests that there is little heterogeneity in both relative mobility and absolute mobility.

5.2 Geographic Differences

We next turn to examining geographic differences in health mobility. This is not only useful for descriptive purposes, but can potentially also provide insight into possible mechanisms that explain any differences. There are now studies in many countries highlighting within country geographic differences in income mobility (e.g. Chetty et al. (2014); Corak (2019); Deutscher and Mazumder (2020); Eriksen and Munk (2020); Acciari et al. (2022), but only one study that we are aware of (Fletcher and Jajtner, 2021) that has looked at this issue for health mobility in the US. To address this, we begin by looking across cities.

Specifically, we show the rank-rank relationship in PC1 for three selected cities in Figure 5. In this exercise, we first compute P25, or the predicted national rank of the children conditional their parents' national rank equaling 25, for each of 22 cities in Taiwan. We then rank the cities by their P25 estimates. Finally, we plot the rank-rank slopes for the city with the highest P25 (Taipei), the median P25 (Keelung), and the smallest P25 (Penghu). This provides a general sense of the range in health mobility across the country.

[10]Health insurance rates are also based on occupation. We take the median value of household income over all years in which income is observed.

Looking across these cities, we observe a modest degree of heterogeneity. The difference in the expected health rank of children between Taipei and Penghu ranges between 5 and 10 percentiles over the entire range of the parent health distribution. So, although parent-level household income is not associated with differences in health mobility, there appear to be some notable geographic differences.

The next step in our analysis is to consider what geographic factors could play a role in accounting for geographic mobility differences. To analyze this further, we turn to a more granular level of geography, the township level, that allows us to leverage greater variation. There are a total of 368 townships in Taiwan and we are able to assemble a few candidate variables for each township, that a priori, might be hypothesized to affect health mobility.

We start by considering the role that access to healthcare may play in health mobility. Specifically, we use the total number of physicians as a proxy for healthcare access and rank townships by this measure. We then divide the sample into deciles by the number of physicians in one's township. We then estimate rank-rank regressions for each decile group once again, using national ranks. We plot this relationship in Figure 6. The figure suggests that there is clearly a higher expected rank throughout the parent health distribution, for those children living in the top decile of townships. However, for those living in townships below the top decile, there appears to be little difference in mobility. This provides suggestive evidence that health mobility in Taiwan is only loosely related to access to healthcare.

We further delve into the sources of health mobility differences at the township level by examining a host of other factors. This analysis is modeled on Chetty et al. (2014), who examine a wide variety of correlates of upward income mobility across geographic areas in the US. Specifically, we use the indigenous share of the population, the share of workers that are in the private sector, the share of workers in the public sector, mean educational attainment, median income levels, the number of doctors per capita, the total

number of doctors, and the total population in the township. We calculate the expected rank at p25 for each of the townships and then show the correlation coefficient between this measure of upward mobility and each of the factors in Figure 7. In the figure, we plot the estimate of the correlation and its 90% confidence interval.

We find that healthcare access in a township, as proxied by either measure of physician availability, and median income, are the only factors with statistically significant associations. However, these factors are only modestly associated with upwards health mobility with correlation coefficients of around 0.18 to 0.26. In contrast, Chetty et al. (2014), show much stronger correlations between a wide variety of factors and income mobility. Mean education levels and township population are also positively associated with upward mobility with correlation coefficients of 0.12 and 0.13, respectively, but these are not statistically significant.

Overall, Taiwan appears to exhibit a great deal of equality when it comes to health mobility. There are virtually no differences by parent income and only small geographic differences. The small geographic differences that do exist can only be modestly accounted for by income differences and access to healthcare. This analysis reveals that policies the widen access to physicians in under served areas may be a promising avenue to promote upward health mobility.

5.3 Differences by Health Conditions

In this section, we focus on heterogeneity by health conditions. Here we ask if there are particular types of health problems that have higher rates of intergenerational persistence than others. If so these might be potential drivers of the broader intergenerational association in latent health.

To examine this, we estimate the degree of intergenerational health persistence in each of our 13 broad ICD categories. Since our measures are binary indicators for ever having

a health condition that falls into each category, a simple correlation coefficient is not appropriate. So, we supplement the correlation with two additional measures, the odds ratio (OR) and Yule's Y. The odds ratio is:

$$OR = \frac{\frac{P(\text{Child Disease=1}|\text{Parent Disease=1})}{P(\text{Child Disease=1}|\text{Parent Disease=0})}}{\frac{P(\text{Child Disease=0}|\text{Parent Disease=1})}{P(\text{Child Disease=0}|\text{Parent Disease=0})}}.$$

Yule's Y, or the coefficient of colligation, is a transformation of the odds ratio. Its formula is:

$$Y = \frac{\sqrt{OR} - 1}{\sqrt{OR} + 1}.$$

Yule's Y is computed to associate two binary random variables. Like Pearson's correlation, this statistic has an absolute value less than one aiding in its interpretation.

Figure 8 presents the estimates for each of our three measures for each condition. We display analogous estimates in Figure A.7 for each specific type of parent-child pair. Here we focus on Yule's Y (middle panel) as our preferred measure, and note that it exhibits much greater variation than the simple correlation coefficient. Yule's Y is less than 0.1 for genitourinary conditions (related to urinary or genital organs), blood conditions, circulatory conditions, mental conditions and neoplasms. The coefficient is between 0.1 and 0.2 for ill-defined conditions, musculoskeletal conditions, skin conditions, digestive issues, nervous conditions and infectious diseases. The one outlier with a much higher Yule's Y of close to 0.4, is respiratory conditions. Respiratory conditions also have a much higher odds ratio than all of the other conditions. Thus, this analysis highlights the potentially important role of respiratory issues in driving the intergenerational persistence of health. It is also important to note that respiratory conditions are almost universally experienced in the Taiwanese population with incidence rates of above 90 percent in both generations.

5.4 Is Persistence Related to Genetics?

An important question is how much of the intergenerational persistence in health is related to genetic versus environmental factors. A growing number of studies have considered this with respect to intergenerational mobility in income, wealth and education, but we are unaware of studies that have considered this with health mobility. To be clear, this is purely a descriptive exercise meant to illuminate our understanding of mechanisms, rather than to necessarily drive policy discussion. Health conditions that are highly genetically related can still be treated by medical interventions and/or influenced by public policies. Moreover, conditions can be driven by an interaction of both genetic and environmental factors.

We take advantage of the heterogeneity in intergenerational persistence across health problems described above to see if those conditions with higher associations also have a higher degree of genetic relatedness.

In order to proxy for the degree of genetic relatedness of each condition, we use our health insurance records to compute associations across same sex twins in whether they ever had a health condition in each category.[11] We use same sex twins because we do not observe monozygosity in our data, and this increases the share of twins that share the same genes.[12] Monozygotic twin correlations are often used as markers of genetic relatedness. Since these are binary random variables, we again use the same three measures of associations (correlation coefficient, Yule's Y and the Odds Ratio), but as before, focus on the Yule's Y measure. We then correlate our Yule's Y intergenerational persistence estimate with our Yule's Y twin correlation estimate for each of the 13 health conditions.

In Figure 9, we display a scatter plot between the intergenerational Yule's Y and the twins' Yule's Y. The figure depicts a *negative* relationship. This is inconsistent with

[11] Twins are identified as those individuals with the same parents who also share the same date of birth

[12] Calculations based on Heuser (1967) and James (1975) suggest that the share of monozygotic twins among same-sex twins is roughly around 0.45

the hypothesis that health conditions with higher intergenerational persistence are more genetically related.[13] If anything, our results imply the opposite conclusion, that more genetically related health conditions have lower intergenerational persistence.

At first glance, this result might seem counter-intuitive. However, this finding may be reconciled by the fact many human characteristics are highly *polygenic*, meaning that they are determined by possessing certain combinations of a fairly large number of genes. While this means that this combination will be shared by identical twins, it is precisely for this reason that it is also highly unlikely that this exact permutation of genes will be passed down to the next generation.[14] This explains why many complex human traits can possess large genetic components but still exhibit low intergenerational persistence.

6 Conclusion

We estimate the degree of intergenerational health mobility in Taiwan using administrative claims data spanning 20 years. The nascent literature on health mobility thus far has focused on advanced economies mainly in the West including: the US, the UK, Australia, Denmark, and Germany. This is the first study of its kind set in Asia, and the first to examine a middle income country. In addition, we are among the first studies to leverage administrative data to gauge a society's degree of intergenerational mobility in latent health.

Importantly, Taiwan provides universal national health insurance with limited out-of-pocket costs. This allows us to obtain comprehensive measures of health that do not simply reflect an individual's ability to *access health care*. We use principal components

[13]The underlying data are also shown in Figure A.6

[14]See Mukherjee (2016) who writes: "If you possess a particular combination of genes, the chance of developing the illness (schizophrenia) is extremely high: hence the striking concordance among identical twins. On the other hand, the inheritance of the disorder across generations is quite complex. Since genes are mixed and matched in every generation, the chance that you will inherit that exact permutation of variants from your father or mother is dramatically lower"(p. 446)

analysis on a set of indicator variables for ever experiencing outpatient care for each of a broad set of health conditions, along with quintiles of visits to a general practitioner, to create a proxy measure of latent health.

Our preferred rank-rank estimate is 0.22 which we obtain by pooling sons and daughters and combining both parents' health. The intergenerational association (IHA) in levels is about 0.26. We further show that maternal transmission is stronger than paternal transmission and that persistence to sons is higher than persistence to daughters. What distinguishes our analysis from prior work is that with our population-wide samples we can show that these differences are highly statistically significant.

We further document important non-linearities. Most notably, we show that parents in the top 10% of the latent health distribution are better able to transmit their health to their children than the remaining 90% of parents and that this is most evident for mothers. Thus, there is a health "premium" associated with being born to a very healthy mother above what the linear rank-rank model predicts. Future research is needed to better understand the source of this nonlinearity.

We also make some methodological contributions. We show that using a larger number of proxies increases the degree of estimated health persistence across generations. We demonstrate that using data on outpatient visits in specific health categories roughly doubles the estimate of intergenerational persistence relative to using solely inpatient data. This is likely due to the greater "signal" of latent health contained in outpatient data. Notably, we find vastly higher rates of utilization of outpatient care than inpatient care. We also obtain somewhat smaller estimates when we use medical expenditures as a proxy for health.

Studies on intergenerational mobility have focused heavily on earnings and education, but until recently have paid scant attention to health status. This is despite the fact that health is an important predictor of human welfare (Sen, 1998; Jones and Klenow, 2016). This work adds to a body of evidence demonstrating that intergenerational transmission

of health is lower than that of earnings or education. When considering the degree of equality of opportunity in any society, policy makers ought to consider paying closer attention to health given its strong implications on social welfare.

References

Acciari, P., Polo, A., and Violante, G. L. (2022). And yet it moves: Intergenerational mobility in italy. *American Economic Journal: Applied Economics*, 14(3):118–63.

Andersen, C. (2022). Intergenerational Health Mobility: Evidence From Danish Registers. *Health Economics*, 30(12):3186–3202.

Bencsik, P., Halliday, T. J., and Mazumder, B. (2023). The Intergenerational Transmission Of Mental And Physical Health In The United Kingdom. *Journal of Health Economics*, page 102805.

Case, A., Fertig, A., and Paxson, C. (2005). The Lasting Impact Of Childhood Health And Circumstance. *Journal of health economics*, 24(2):365–389.

Chetty, R., Hendren, N., Kline, P., and Saez, E. (2014). Where Is The Land Of Opportunity? The Geography Of Intergenerational Mobility In The United States. *The Quarterly Journal of Economics*, 129(4):1553–1623.

Corak, M. (2019). The Canadian Geography of Intergenerational Income Mobility. *The Economic Journal*, 130(631):2134–2174.

Deutscher, N. and Mazumder, B. (2020). Intergenerational mobility across australia and the stability of regional estimates. *Labour Economics*, 66:101861.

Doorslaer, E. v., Koolman, X., and Jones, A. M. (2004). Explaining Income-related Inequalities In Doctor Utilisation In Europe. *Health Economics*, 13(7):629–647.

Eriksen, J. and Munk, M. D. (2020). The geography of intergenerational mobility — danish evidence. *Economics Letters*, 189:109024.

Fletcher, J. and Jajtner, K. M. (2021). Intergenerational health mobility: Magnitudes and importance of schools and place. *Health economics*, 30(7):1648–1667.

Graeber, D. (2020). Intergenerational Health Mobility in Germany. *Working Paper*.

Haider, S. and Solon, G. (2006). Life-cycle Variation In The Association Between Current And Lifetime Earnings. *American economic review*, 96(4):1308–1320.

Halliday, T., Mazumder, B., and Wong, A. (2021). Intergenerational Mobility In Self-reported Health Status In The US. *Journal of Public Economics*, 193:104307.

Halliday, T. J., Mazumder, B., and Wong, A. (2020). The Intergenerational Transmission Of Health In The United States: A Latent Variables Analysis. *Health Economics*, 29(3):367–381.

Heuser, R. L. (1967). Multiple Births, U.S. - 1964. *National Center for Health Statistics*, 21(14).

James, W. H. (1975). Sex Ratio In Twin Births. *Annals of Human Biology*, 2(4):365–378.

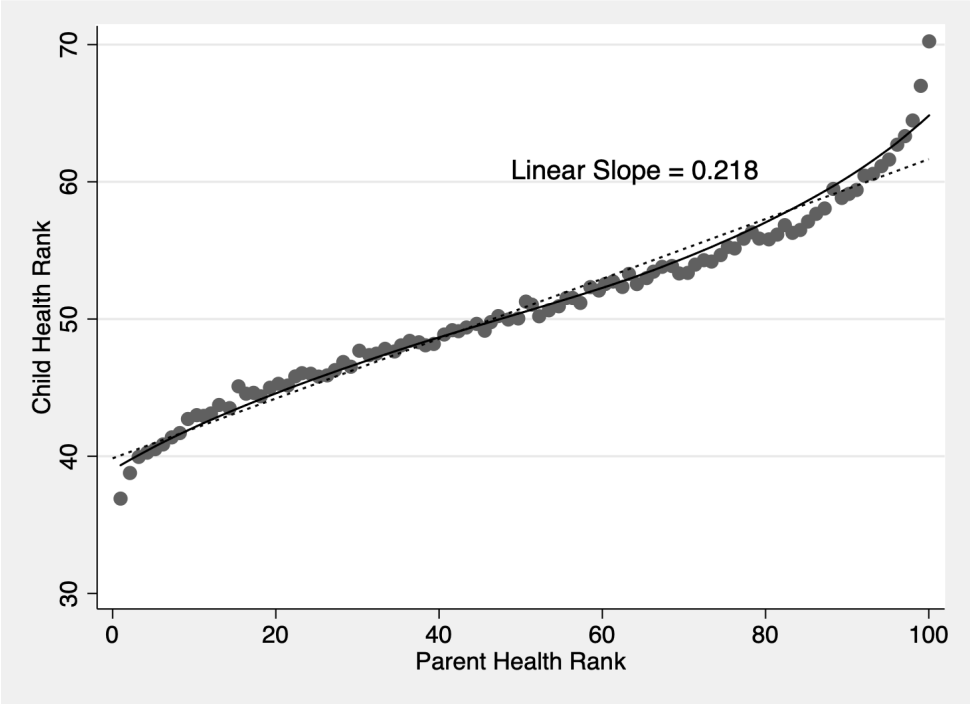
Jones, C. I. and Klenow, P. J. (2016). Beyond GDP? Welfare Across Countries And Time. *American Economic Review*, 106(9):2426–57.

Mukherjee, S. (2016). *The Gene*. Scribner, New York.

Sen, A. (1998). Mortality As An Indicator Of Economic Success And Failure. *The economic journal*, 108(446):1–25.

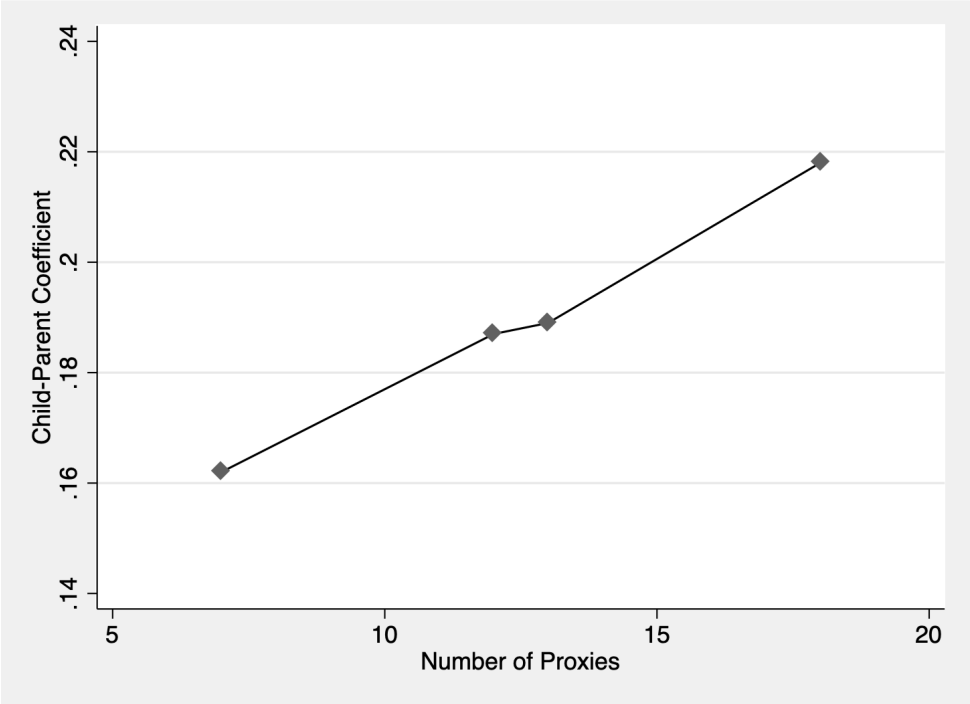
Vera-Toscano, E. and Brown, H. (2021). The Intergenerational Transmission of Mental and Physical Health in Australia: Evidence Using Data From the Household Income and Labor Dynamics of Australia Survey. *Frontiers in Public Health*, 9.

Figure 1: Intergenerational Rank-rank Slope between Parents and Children



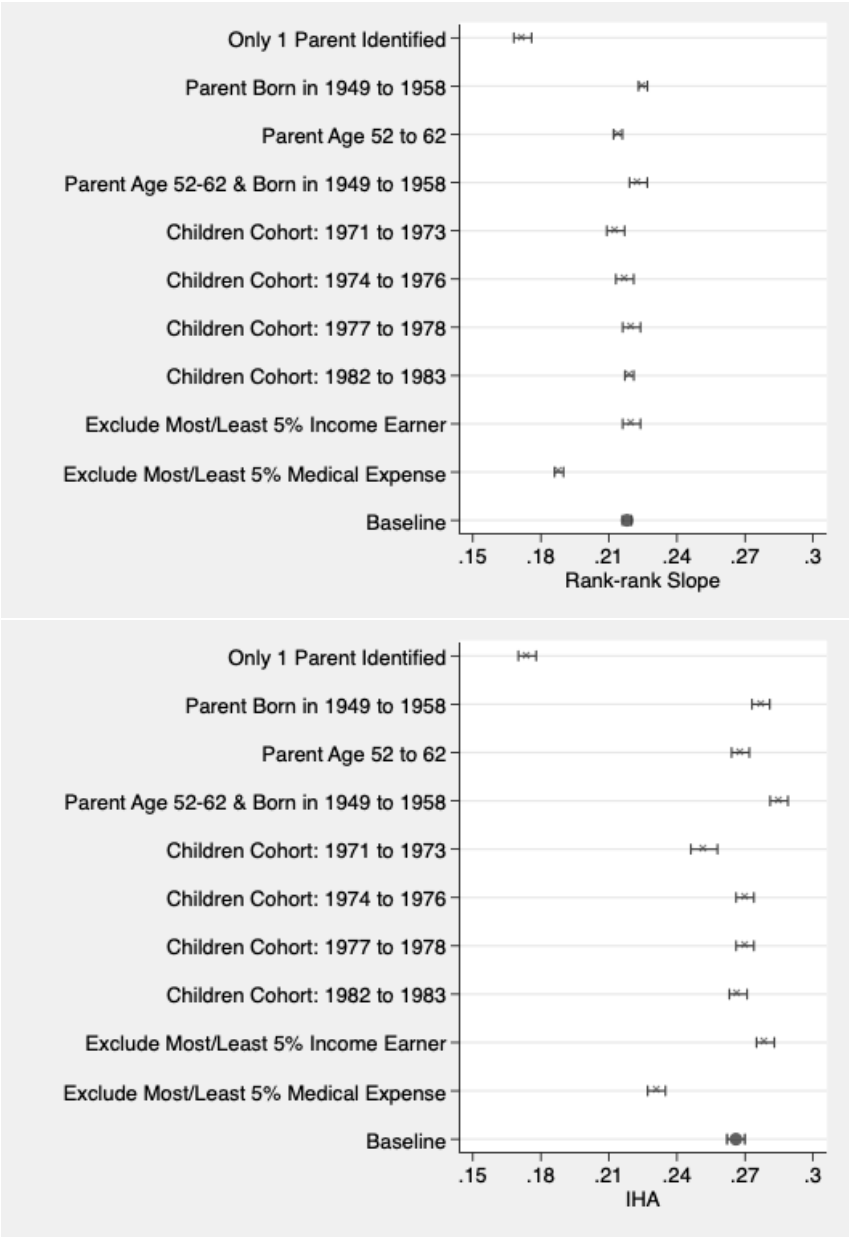
Notes: This plot utilizes 18 proxies from the outpatient data and shows the intergenerational rank-rank slope between parents and children. A solid line lowess smoother is estimated and plotted through the dots. The dashed line corresponds to a linear approximation and has a slope of 0.218.

Figure 2: Intergenerational Rank-Rank Slope by the Number of Proxies



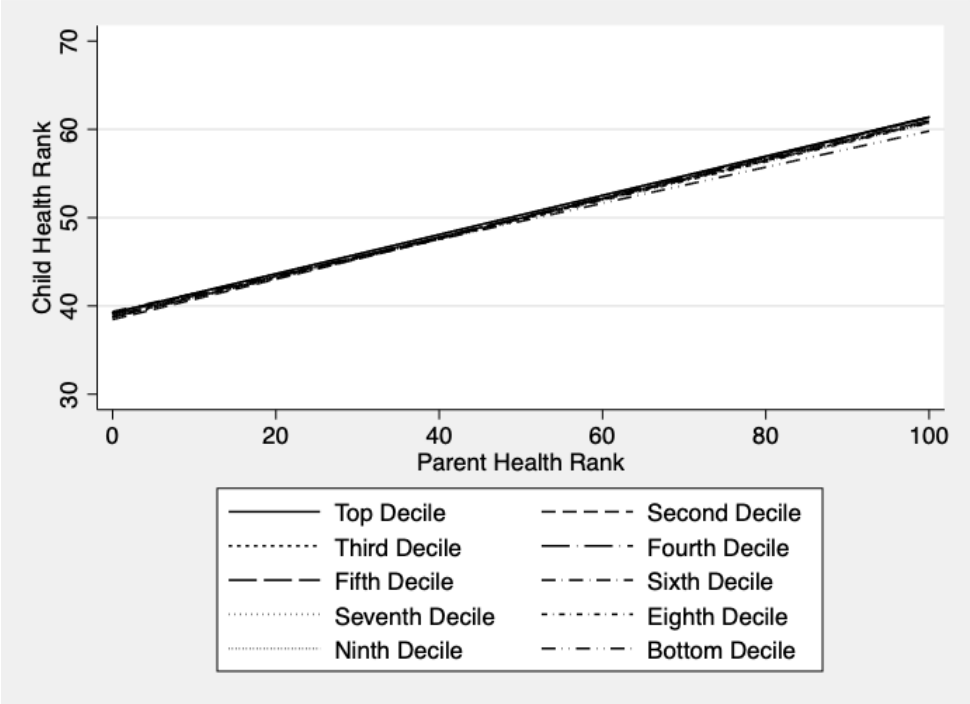
Notes: This figure leverages outpatient data and shows the evolution of estimates among child-parent rank-rank slopes as we progressively add conditions. The first seven ICD categories included as proxies are: *Infectious, Endocrine, Mental, Nervous, Circulatory, Skin, and Musculoskeletal.* When the number of proxies is 12, we added another five ICD categories: *Neoplasms, Blood, Digestive, Genitourinary, and Ill-defined Conditions.* Next, we added *Respiratory* increasing the number of proxies to 13. Finally, we added five quintile dummies approximating visit intensity bringing the number of proxies to 18.

Figure 3: Robustness Checks in Rank-Rank Slopes and IHA



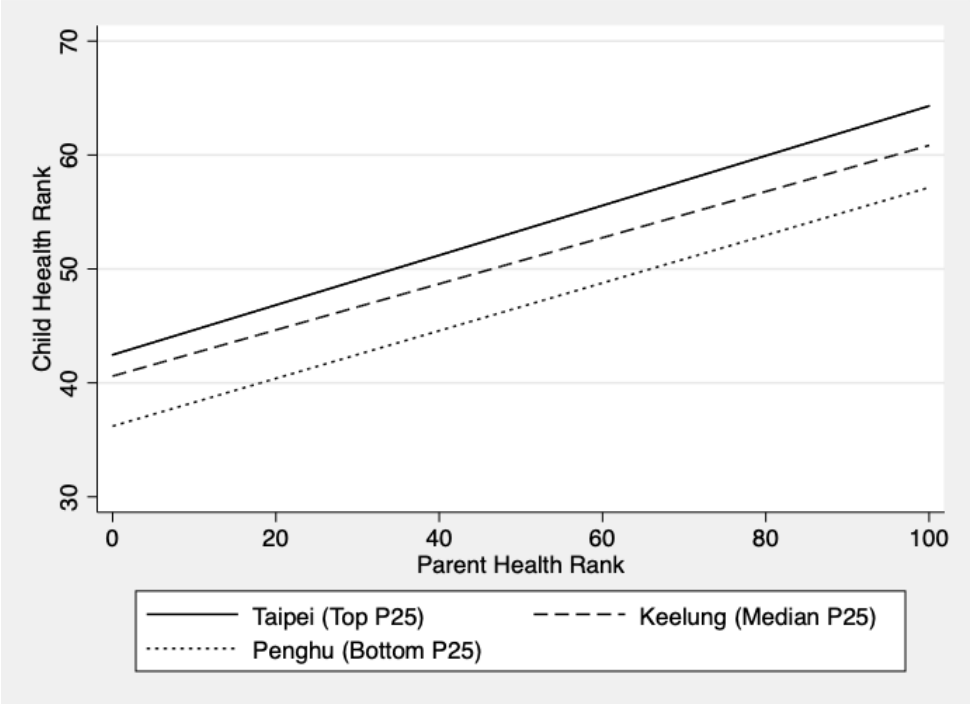
Notes: This figure estimates the rank-rank slope and the IHA for a number of different sample selection restrictions.

Figure 4: Differences in Absolute Intergenerational Health Mobility by Parent Household Income



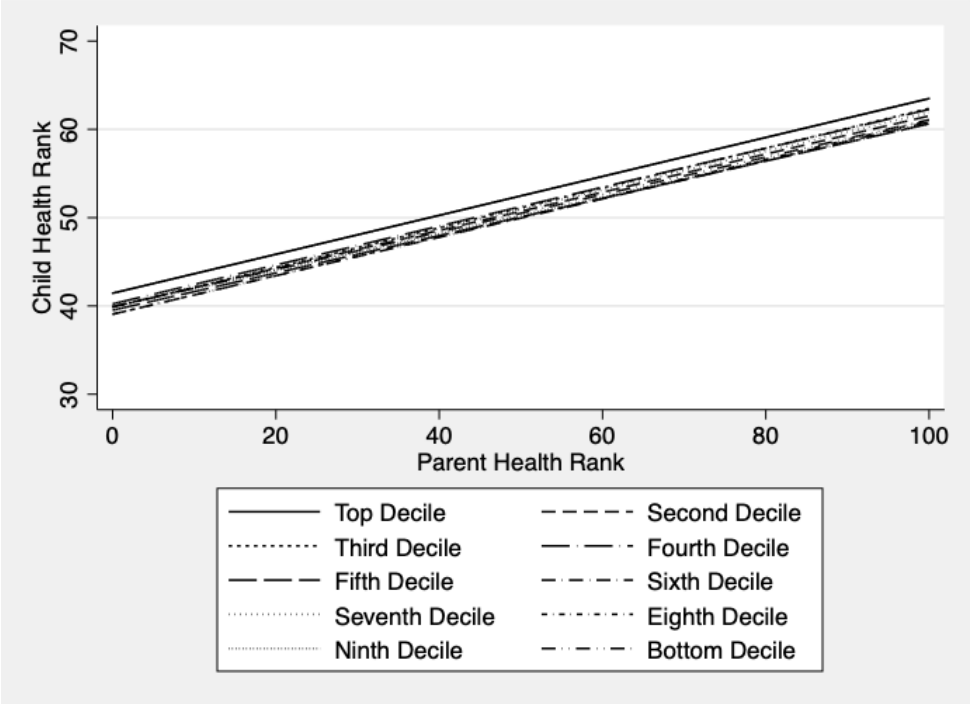
Notes: All parent-child pairs are divided into 10 subgroups based on parent household income levels. We estimated rank-rank slopes for each of these groups.

Figure 5: Geographic Differences in Intergenerational Health Mobility by City



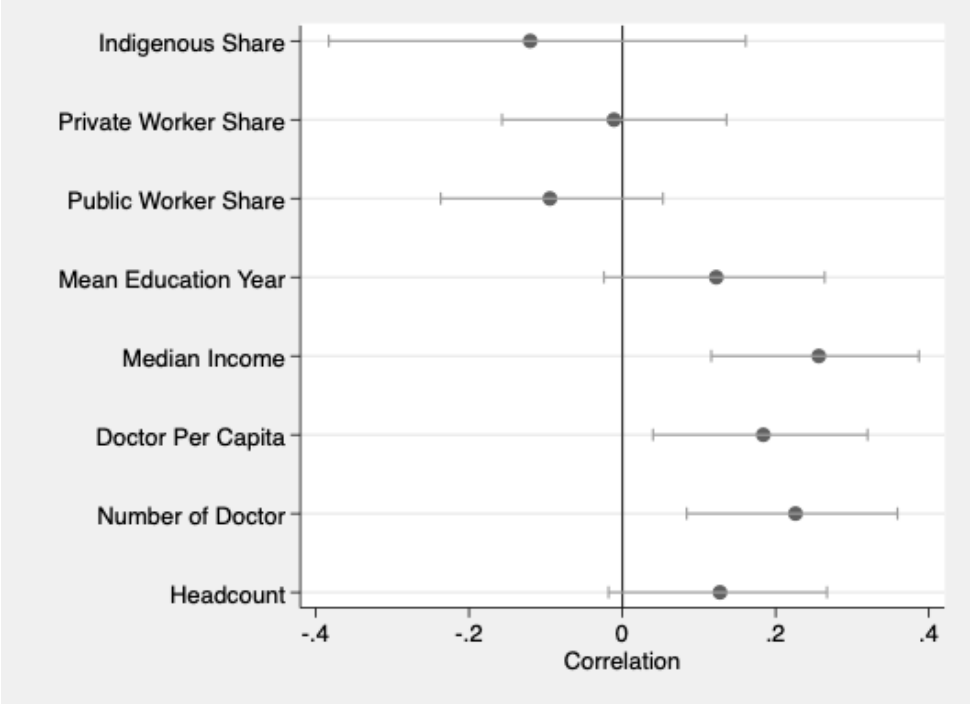
Notes: Among all 22 cities in Taiwan, this figure shows rank-rank slopes and intercepts of three cities representing the *top, median* and *bottom* rank of absolute mobility in the 25th percentile (P25).

Figure 6: Differences in Absolute Intergenerational Health Mobility by Decile of Number of Doctors across Townships



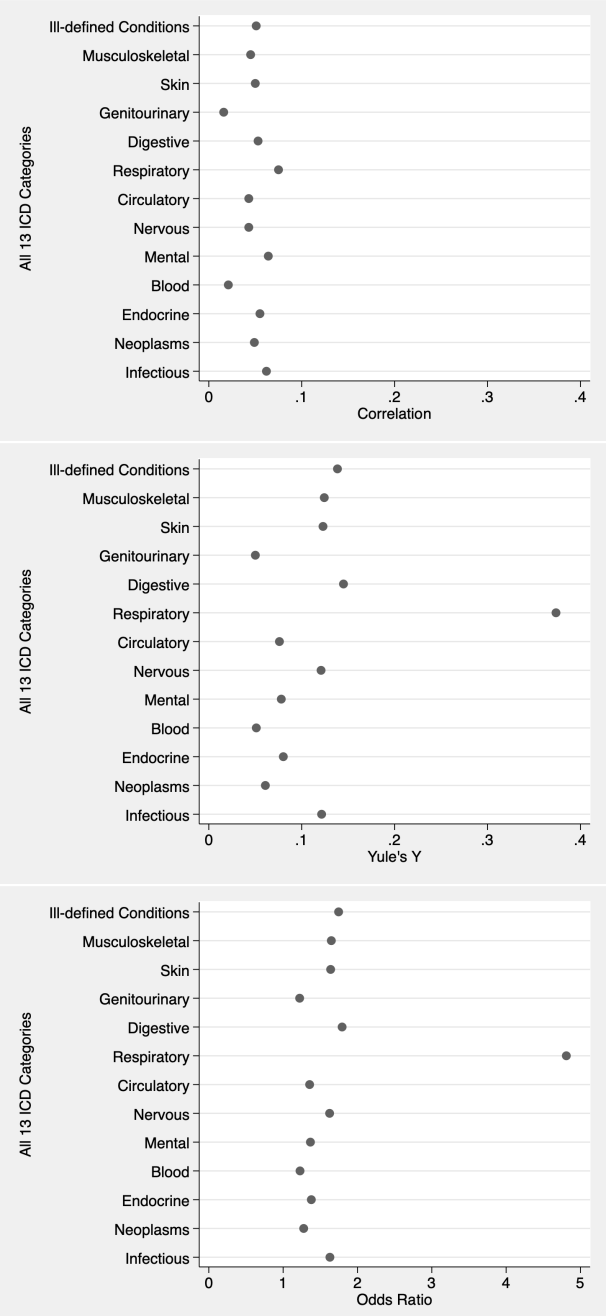
Notes: All 368 townships are divided into 10 subgroups by the number of doctors. We estimated rank-rank slopes for each of these groups.

Figure 7: Correlates of Intergenerational Absolute Mobility (P25) across Townships



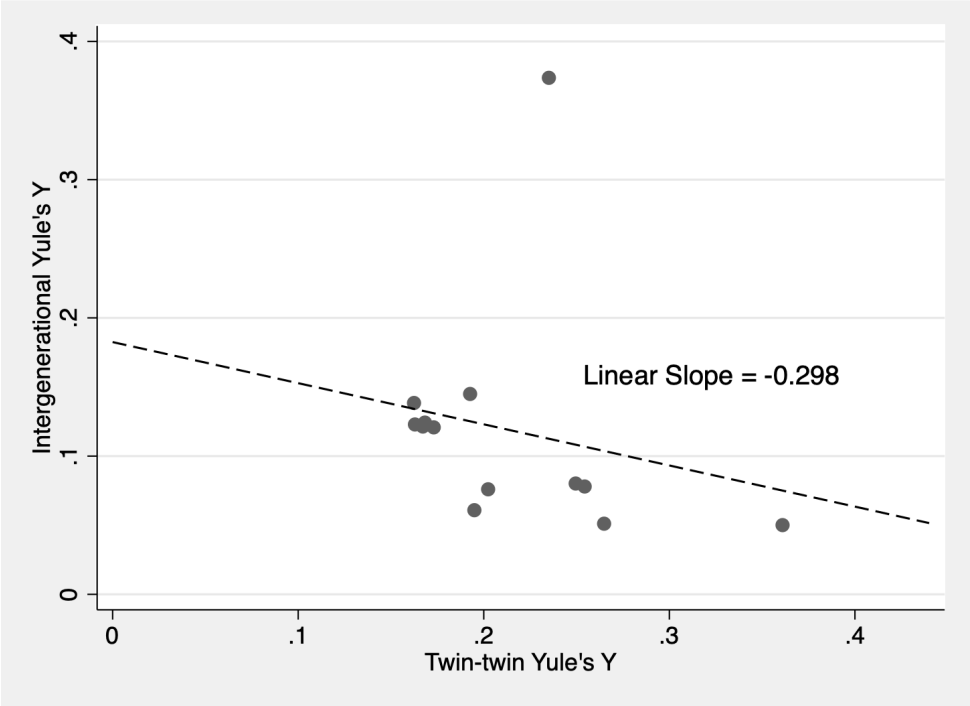
Notes: This figure shows the correlations between absolute mobility (P25) and various township-level variables. The horizontal bars represent a 90% confidence interval of our estimates.

Figure 8: Heterogeneity in Intergenerational Health Persistence by Health Conditions: Intergenerational Correlations, Yule's Y, and Odds Ratios by ICD Categories



Notes: This set of figures estimates intergenerational health persistence across 13 ICD categories. All parent-child pairs include children born in 1979-1981. Categorical dummy variables are utilized when calculating correlation, Yule's Y and odds ratio.

Figure 9: Relationship between the Intergenerational Yule's Y and the Twin's Yule's Y by ICD Categories



Notes: This figure shows the linear relationship between twin-twin Yule's Y and intergenerational Yule's Y; the slope of our linear approximation is -0.298 with a standard error of 0.202 which is not statistically different from zero. Each of the 13 dots represent the ICD categories depicted in Figure 8.

Table 1: Summary Statistics of Outpatient Visits

	Children		Parent	
	Son	Daughter	Father	Mother
Selected Cohort	1979-1981	1979-1981	1920-1968	1920-1968
Observed Age	30-39	30-39	40-75	40-75
Number of Observed Years	10	10	20	20
Parent Match Rate	0.91	0.80	-	-
Panel A: Outpatient Ailment Share (All 13 Categories)				
Infectious	0.72	0.81	0.80	0.88
Neoplasms	0.19	0.39	0.48	0.63
Endocrine	0.25	0.30	0.64	0.67
Blood	0.03	0.11	0.08	0.15
Mental	0.24	0.25	0.41	0.51
Nervous	0.69	0.83	0.89	0.94
Circulatory	0.25	0.24	0.77	0.73
Respiratory	0.93	0.97	0.96	0.98
Digestive	0.74	0.83	0.90	0.93
Genitourinary	0.35	0.88	0.66	0.92
Skin	0.73	0.88	0.85	0.92
Musculoskeletal	0.64	0.63	0.89	0.95
Ill-defined Conditions	0.69	0.84	0.88	0.94
Panel B: Health PC1				
Mean	0.00	0.00	0.00	0.00
S.E.	1.89	1.89	1.99	1.93
Min	-5.19	-7.31	-8.09	-11.00
Max	3.70	2.91	2.40	1.94
Panel C: Health Rank				
Mean	50.42	50.30	50.09	49.80
S.E.	28.86	28.76	28.42	28.17
Min	1	1	1	1
Max	100	100	100	98
Observation	571,079	509,596	706,040	778,943

Notes: Panel A displays the share of individuals who ever recorded outpatient visits for the listed ICD categories. Panels B and C display summary statistics for PC1 and its rank. Both were calculated from the 13 ICD indicators and the five GP quintile dummies.

Table 2: Intergenerational Rank-rank Slope Estimates by Parent-child Combinations with Outpatient Data, Number of proxies = 18

	(i) Mother	(ii) Father	(iii) Both Parents	(i)-(ii) Mother - Father
Panel A: Rank-rank Slope				
Sons	0.199***	0.176***	0.230***	0.022***
	(0.001)	(0.002)	(0.002)	(0.002)
	499,755	454,473	383,149	954,228
Daughters	0.193***	0.134***	0.203***	0.059***
	(0.002)	(0.002)	(0.002)	(0.002)
	428,822	389,041	308,327	817,923
Pooled Children	0.196***	0.157***	0.218***	0.039***
	(0.001)	(0.001)	(0.001)	(0.001)
	928,637	843,514	691,476	1,772,151
Sons - Daughters	0.006***	0.042***	0.027***	
	(0.002)	(0.002)	(0.002)	
	928,637	843,514	691,476	
Panel B: IHA Correlation				
Sons	0.201***	0.169***	0.293***	0.034***
	(0.002)	(0.002)	(0.002)	(0.002)
	499,755	454,473	383,149	954,228
Daughters	0.174***	0.110***	0.231***	0.066***
	(0.002)	(0.002)	(0.002)	(0.002)
	428,822	389,041	308,327	817,923
Pooled Children	0.190***	0.144***	0.266***	0.049***
	(0.001)	(0.001)	(0.002)	(0.001)
	928,637	843,514	691,476	1,772,151
Sons - Daughters	0.028***	0.060***	0.065***	
	(0.002)	(0.002)	(0.003)	
	928,637	843,514	691,476	

Notes: Both panels of estimates utilize 18 proxies from outpatient data that include *Infectious, Neoplasms, Endocrine, Blood, Mental, Nervous, Circulatory, Respiratory, Digestive, Genitourinary, Skin, Musculoskeletal, and Ill-defined Conditions* as well as the GP quintile dummies. *** Represents a 1% level of statistical significance. Each set of estimates includes three numbers: the coefficient estimate, the standard deviation, and the number of observations.

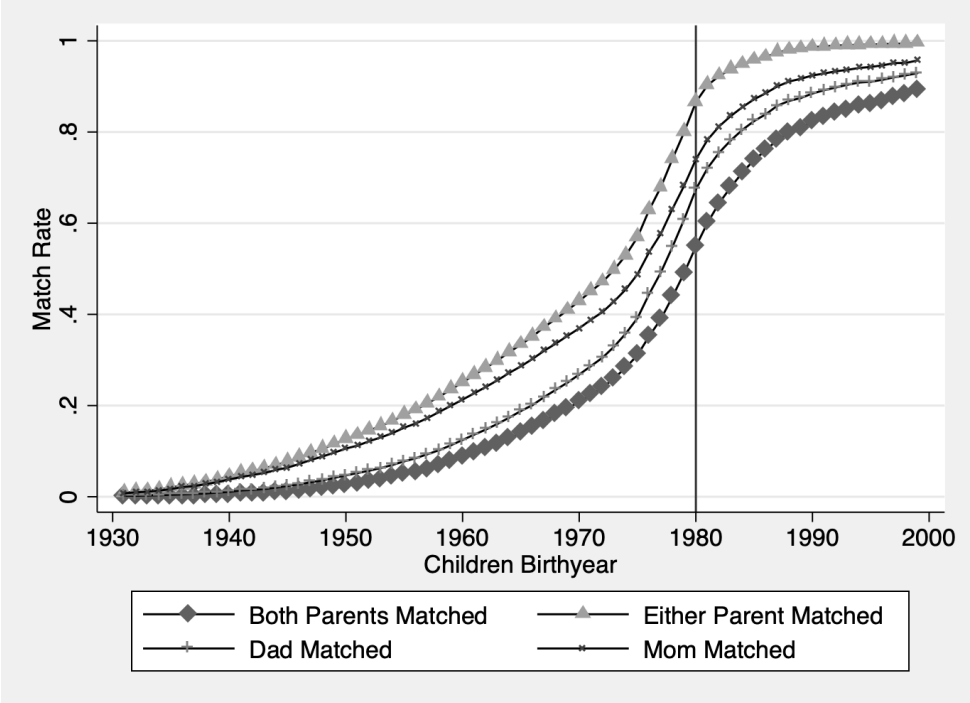
Table 3: Other Intergenerational Health Transmission Estimates by Parent-child Combinations

	(i) Mother	(ii) Father	(iii) Both Parents	(i)-(ii) Mother - Father
Panel A: Outpatient IHA Correlation Without PCA				
Sons	0.211***	0.160***	0.295***	0.053***
	(0.002)	(0.002)	(0.002)	(0.003)
	499,755	454,473	383,149	954,228
Daughters	0.194***	0.114***	0.247***	0.083***
	(0.002)	(0.002)	(0.003)	(0.002)
	428,822	389,041	308,327	817,923
Pooled Children	0.206***	0.141***	0.275***	0.067***
	(0.001)	(0.001)	(0.002)	(0.002)
	928,637	843,514	691,476	1,772,153
Sons - Daughters	0.018***	0.047***	0.050***	
	(0.003)	(0.002)	(0.003)	
	928,637	843,514	691,476	
Panel B: Inpatient Rank-rank Slope				
Sons	0.078***	0.080***	0.120***	−0.001
	(0.002)	(0.002)	(0.002)	(0.002)
	453,352	412,312	346,874	865,664
Daughters	0.075***	0.066***	0.106***	0.01***
	(0.002)	(0.002)	(0.002)	(0.002)
	392,373	355,906	282,995	748,279
Pooled Children	0.077***	0.073***	0.114***	0.003
	(0.001)	(0.001)	(0.002)	(0.002)
	845,725	768,218	629,829	1,613,943
Sons - Daughters	0.004***	0.014***	0.015***	
	(0.002)	(0.002)	(0.003)	
	845,725	768,218	629,829	
Panel C: Outpatient Expense Rank-rank Slope				
Sons	0.149***	0.126***	0.182***	0.023***
	(0.001)	(0.001)	(0.002)	(0.002)
	507,753	461,806	388,369	969,556
Daughters	0.147***	0.095***	0.156***	0.053***
	(0.002)	(0.002)	(0.002)	(0.002)
	432,393	392,117	310,595	824,510
Pooled Children	0.148***	0.111***	0.170***	0.037***
	(0.001)	(0.001)	(0.001)	(0.001)
	940,116	853,923	698,964	1,794,039
Sons - Daughters	0.002	0.031***	0.026***	
	(0.002)	(0.002)	(0.003)	
	940,116	853,923	698,964	

Notes: Panel A displays estimates of the IHA correlation using 13 ICD-based proxies from the outpatient data without applying PCA. The proxies include *Infectious, Neoplasms, Endocrine, Blood, Mental, Nervous, Circulatory, Respiratory, Digestive, Genitourinary, Skin, Musculoskeletal, and Ill-defined Conditions.* Panel B displays estimates of the rank-rank slopes using the 18 proxies from the *inpatient* data using PCA. These estimates include the additional quintile dummies indicating clinical severity. Panel C displays estimates of the IHA correlation using all the outpatient expenses recorded. *** Represents a 1% level of statistical significance. Each set of estimates includes three numbers: the coefficient estimate, the standard deviation, and the number of observations.

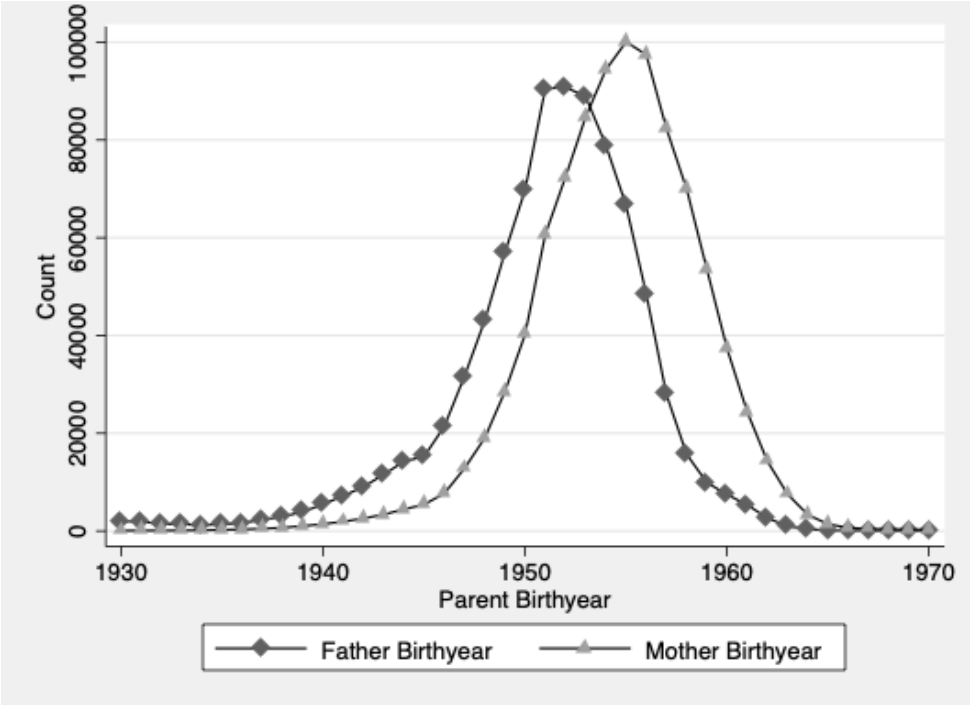
A Appendix Figures

Figure A.1: Parent Match Rate by Children's Cohort



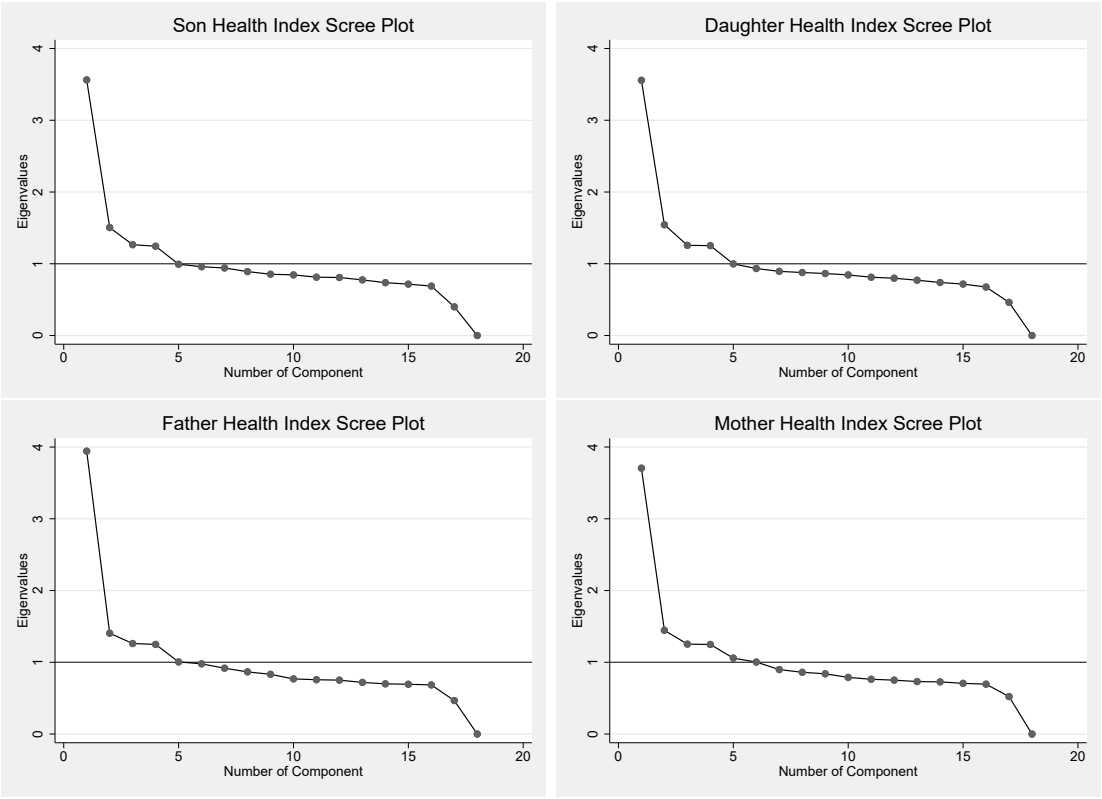
Notes: Plots parent match rates for each child cohort in the NHI claims data and Household registration data.

Figure A.2: Parental Birthyear Distribution



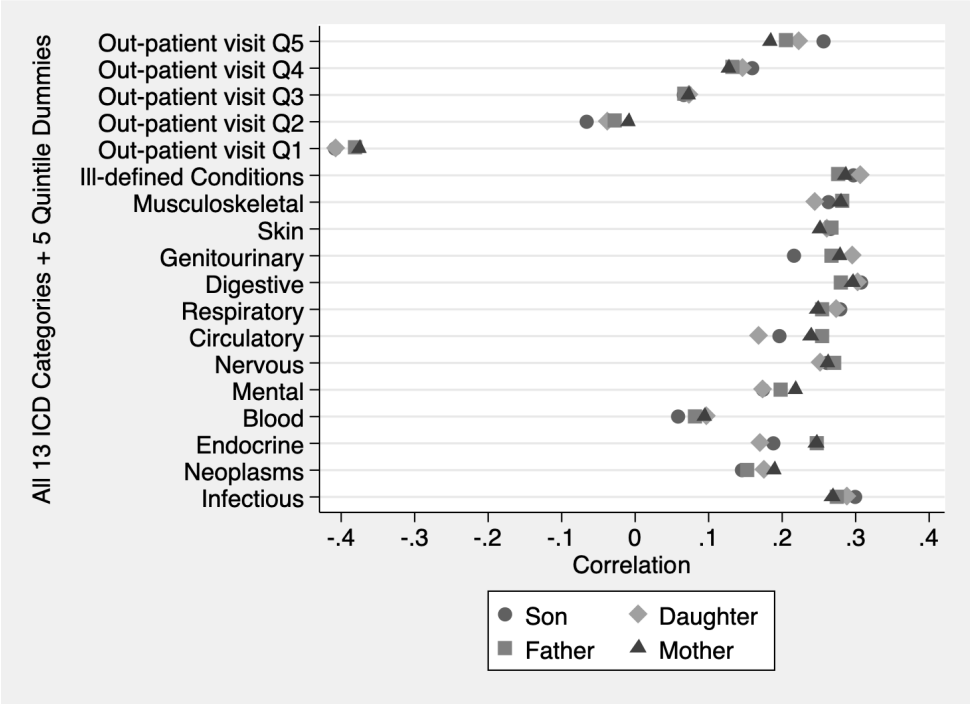
Notes: Plots the parent birth year distribution for the children's cohort of 1979-1981.

Figure A.3: Scree Plots



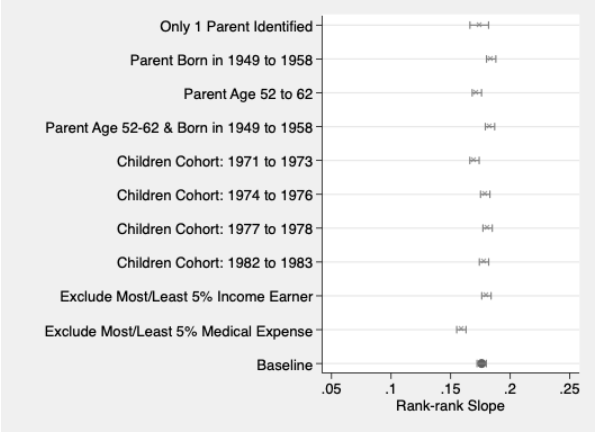
Notes: Displays scree plots for different child-parent relationships.

Figure A.4: Component Loadings

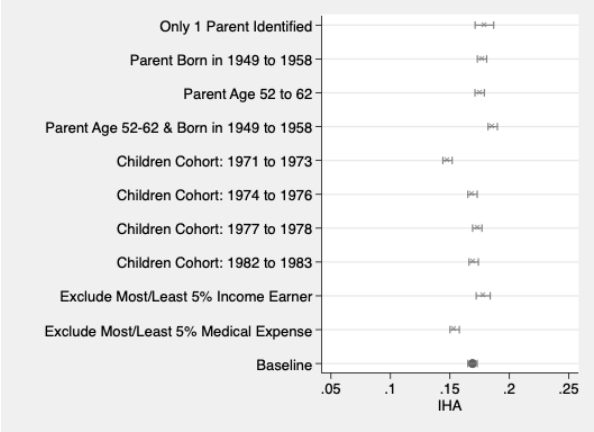


Notes: Plots loadings from the first principal component of the 18 clinical outcomes.

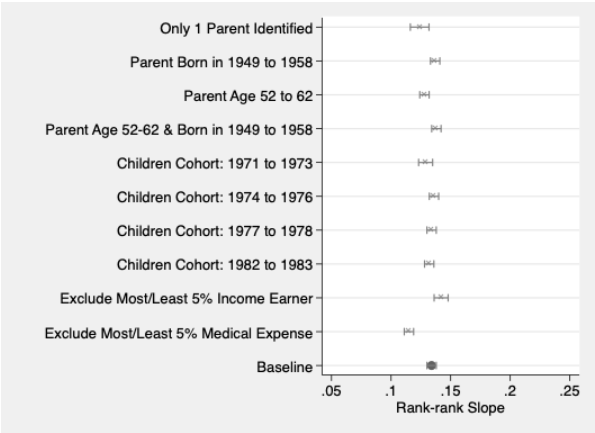
Figure A.5: Robustness Checks in Rank-rank Slopes and IHA Across Different Parent-Child Relationships



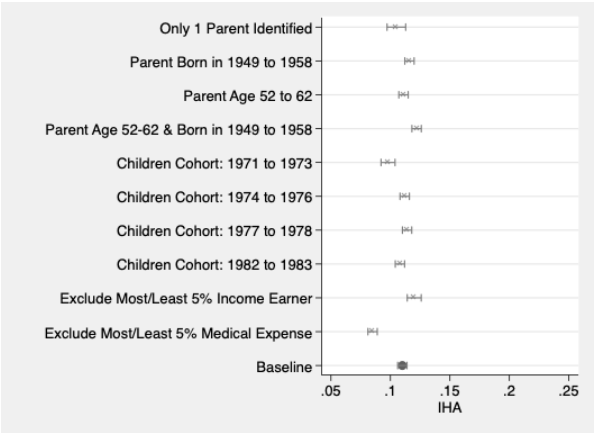
(a) Dad-Son Rank



(b) Dad-Son IHA

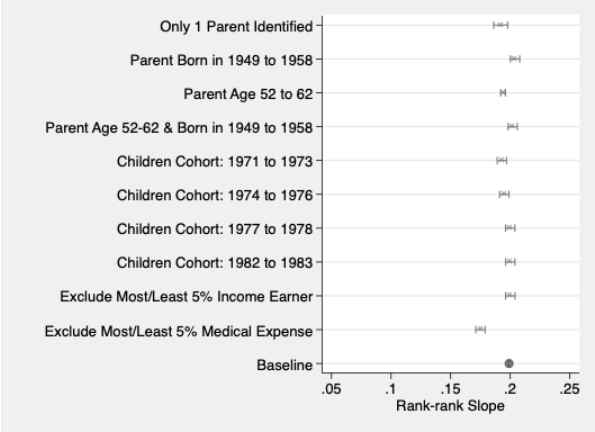


(c) Dad-Daughter Rank

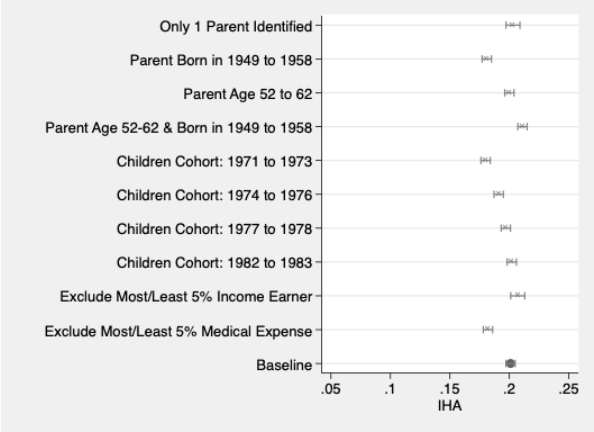


(d) Dad-Daughter IHA

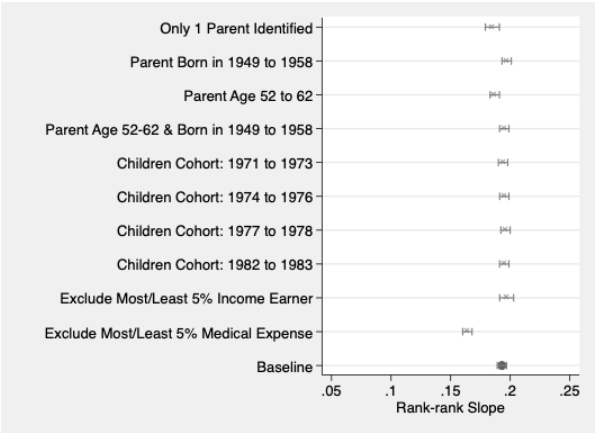
Figure A.5: Robustness Checks in Rank-rank Slopes and IHA Across Different Parent-Child Relationships (Continued)



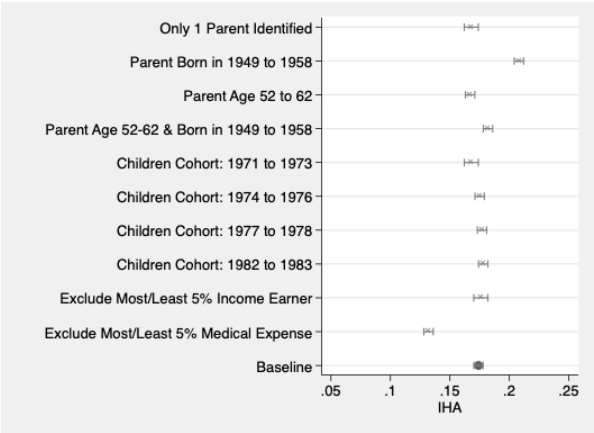
(e) Mom-Son Rank



(f) Mom-Son IHA



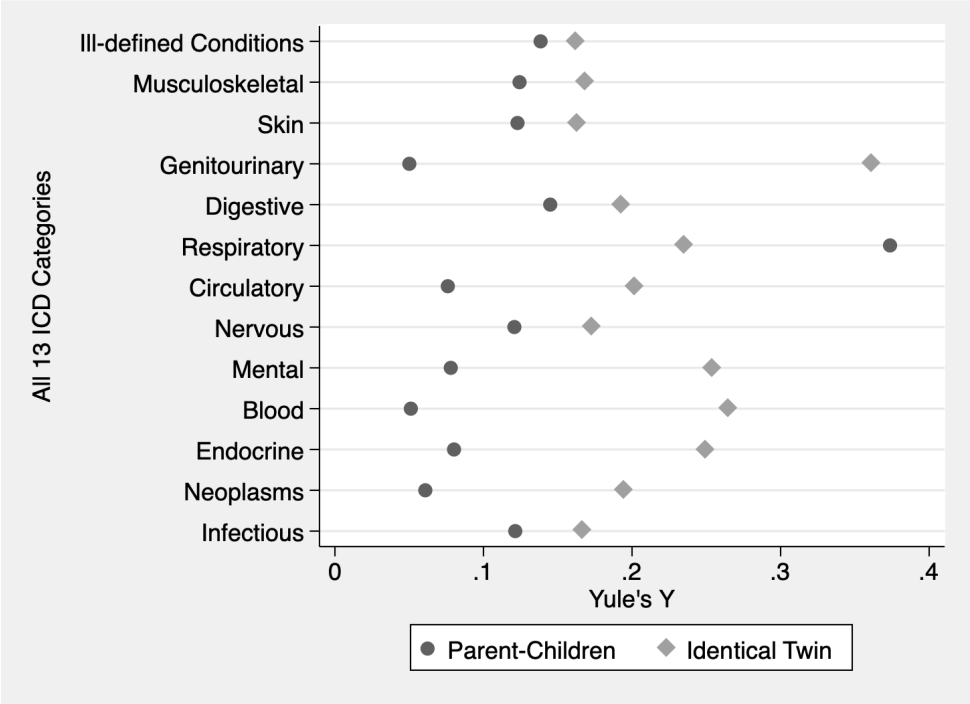
(g) Mom-Daughter Rank



(h) Mom-Daughter IHA

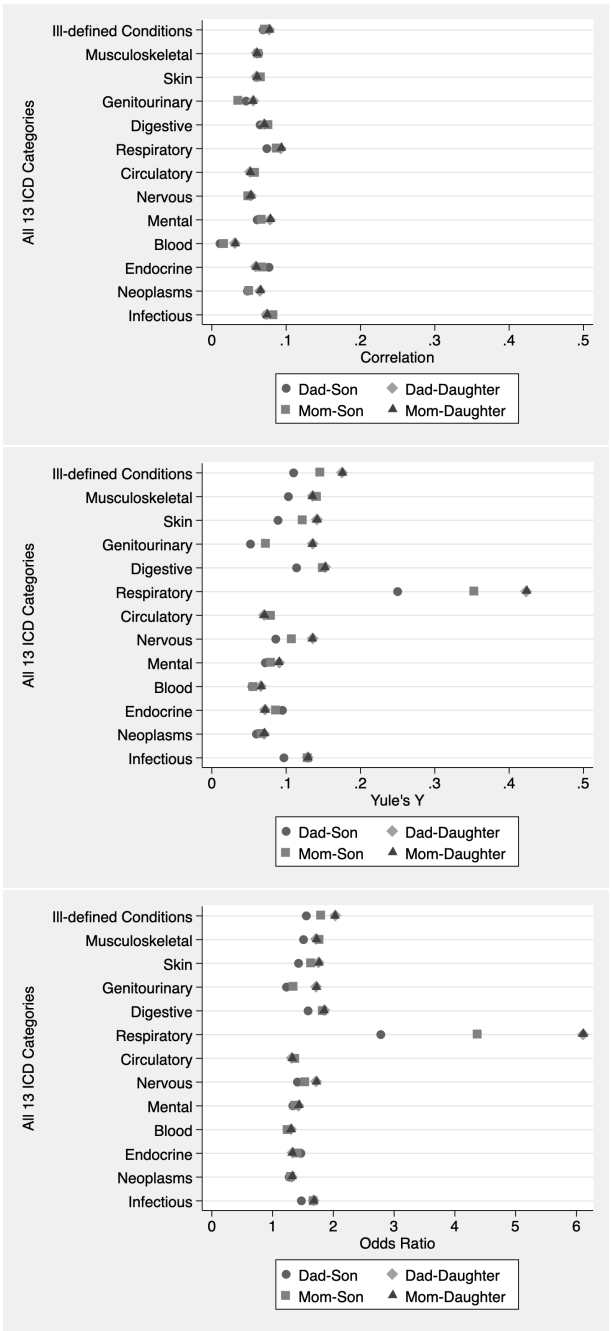
Notes: This figure compares rank-rank slope and IHA among a variety of robustness checks that include different sample selections and restrictions across Father-Son, Father-Daughter, Mother-Son, and Mother-Daughter pairs.

Figure A.6: Intergeneraional and Twin-twin Yule's Y by ICD Categories



Notes: This figure estimates Yule's Y across 13 ICD categories using outpatient data for all parent and all children pairs as well as identical twins in the 1979-1981 cohort.

Figure A.7: Heterogeneity in Health Persistence by Health Conditions by Parent-Child Pairs: Correlation, Odds Ratio and Yule's Y by ICD categories



Notes: Per Figure 8.

B Appendix Tables

Table B.1: Summary Statistics of Inpatient Visits

	Children		Parents	
	Son	Daughter	Father	Mother
Selected Cohort	1979-1981	1979-1981	1920-1968	1920-1968
Observed Age	30-39	30-39	40-75	40-75
Number of Observed Years	10	10	20	20
Parent Match Rate	0.91	0.80	-	-
Panel A: Inpatient Ailment Share (All 13 Categories)				
Infectious	0.02	0.01	0.08	0.05
Neoplasms	0.02	0.06	0.17	0.19
Endocrine	0.01	0.01	0.06	0.05
Blood	0.00	0.00	0.01	0.01
Mental	0.01	0.01	0.01	0.01
Nervous	0.01	0.01	0.07	0.06
Circulatory	0.03	0.02	0.25	0.13
Respiratory	0.04	0.03	0.16	0.09
Digestive	0.06	0.04	0.25	0.13
Genitourinary	0.03	0.06	0.15	0.15
Skin	0.02	0.01	0.06	0.03
Musculoskeletal	0.04	0.02	0.14	0.16
Ill-defined Conditions	0.01	0.01	0.05	0.04
Panel B: Health PCA				
Mean	0.00	0.00	0.00	0.00
S.E.	1.76	1.75	1.76	1.76
Min	-0.94	-0.94	-1.82	-1.64
Max	14.61	14.62	10.44	10.49
Panel C: Health Rank				
Mean	21.31	21.80	45.03	41.80
S.E.	37.01	37.17	34.58	36.67
Min	1	1	1	1
Max	100	100	100	100
Observation	518,790	465,284	706,040	778,943

Notes: Per Table B.1 but uses inpatient information in lieu of outpatient information.

Table B.2: Parent Match Rate Breakdown of Children Cohort 1979-1981

Children Birth Cohort	Match Rate			Observation
	At Least One Parent	1 Parent	2 Parents	
Panel A: Pooled Sons and Daughters				
Cohort 1979	0.80	0.31	0.49	431,971
Cohort 1980	0.86	0.31	0.55	422,782
Cohort 1981	0.90	0.30	0.60	426,749
Cohort 1979-1981	0.86	0.31	0.55	1,281,502
Panel B: Sons				
Cohort 1979	0.87	0.31	0.56	215,436
Cohort 1980	0.92	0.31	0.61	209,848
Cohort 1981	0.94	0.29	0.65	212,678
Cohort 1979-1981	0.91	0.30	0.61	637,962
Panel C: Daughters				
Cohort 1979	0.73	0.31	0.42	216,535
Cohort 1980	0.81	0.33	0.48	212,934
Cohort 1981	0.86	0.32	0.54	214,071
Cohort 1979-1981	0.80	0.32	0.48	643,540

Table B.3: Family type I

Obs	ID	INS-ID	RELATIONSHIP
1	(non-working) Son	Dad	Kid
2	(non-working) Daughter	Mom	Kid
3	(non-working) Mom	Dad	Spouse
4	(non-working) Grandpa	Dad	Parent
5	(working) Dad	Dad (Self-insured)	NA

Table B.4: Family type II

Obs	ID	INS-ID	RELATIONSHIP
1	(non-working) Son	Dad	Kid
2	(non-working) Daughter	Mom	Kid
3	(working) Dad	Dad (Self-insured)	NA
4	(working) Mom	Mom (Self-insured)	NA

Table B.5: Family type III

Obs	ID	INS-ID	RELATIONSHIP
1	(non-working) Son	Dad	Kid
2	(non-working) Daughter	Dad	Kid
3	(working) Dad	Dad (Self-insured)	NA
4	(working) Mom	Mom (Self-insured)	NA

C Family Tree Construction

In this section, we document how familial relationships are identified in our administrative data. We begin by constructing Family Tree 1.0 using only the NHI data. Next, we improve upon this by constructing Family Tree 2.0 in which we supplement the NHI data with the household registration data. Ultimately, due to its completeness, we adopt Family Tree 2.0. The differences between Family Tree 1.0 and 2.0 are elaborated next.

C.0.1 Family Tree 1.0: Only the NHI data

The NHI dataset includes a personal identification number (ID), the identification number of the person who insures them (INS-ID), and details about the person's relationship with the insured individuals (RELATIONSHIP). This data allows us to construct a family tree for a significant number of households, although not all. To illustrate this, we present a hypothetical household in Table B.3 where we can generate a complete family tree solely based on the NHI data. In this particular case, all household members are insured under the father. However, relying solely on the NHI data leads to lower match rates, averaging around 80% across interested cohorts.

C.0.2 Family Tree 2.0: Complement the NHI Data with Household Registration Data

Tables B.4 and B.5 showcase hypothetical scenarios where the NHI data alone fails to identify a complete family tree. In these instances, both the mother and father insure themselves, leading to a need for the family registry to establish their presence in the same household.[15] After recovering the relationships that cannot be identified in the NHI data, the matching rate increases by 10 percent across interested cohorts. Throughout the paper, we only adopt Family Tree 2.0 when identifying familial relationships.

[15]The household registration data comes from the Hukou system in Taiwan which partially resembles the one in China.