A Theory of Disease and Development: Productivity, Exchange, and Social Segregation

Javier A. Birchenall*

University of California at Santa Barbara

May 5, 2021

Abstract

This paper views disease as more than a biological process and argues that it plays a central role in long-run economic development. I examine a model where disease spillovers limit sociality and the extent of the market. I characterize the aggregate efficiency of production and exchange arrangements with a differential scope for disease spillovers and market centralization. The model provides a framework to understand why and how disease-prone societies counter contagion through social segregation. I present many historical examples of social and legal enforcement of distance-based rules whose objective, as in the model, appears to be social segregation by disease.

Keywords: Disease, Economic Development, Total Factor Productivity


Communications to: Javier A. Birchenall
Department of Economics, 2127 North Hall
University of California, Santa Barbara CA 93106
Phone/fax: (805) 893-5275
Javier.Birchenall@ucsb.edu

*I would like to thank the seminar participants at the University of Chicago, University of California Merced, Yonsei University, American University of Beirut, UC Santa Barbara Symposium on Inequality, University of Pittsburgh, Gurven Laboratory for Evolutionary Anthropology and Biodemography at UC Santa Barbara, Universidad Javeriana, and Federal Reserve Bank of Atlanta, for their many suggestions. I have, over the years, also benefited greatly from comments and remarks by Aashish Mehta, Hoyt Bleakley, Robert Fogel, Pete Klenow, Karen Kopely, Shelly Lundberg, Marla Ripoll, Raaj Sah, and Alwyn Young.
1 Introduction

This paper studies the influence of disease on long-run economic development. Its focus is not on how individuals respond to disease, but on the social effects of disease on aggregate efficiency and output. From an individual perspective, disease is a biological process. On physiological grounds, infectious and parasitic diseases reduce the amount of energy and time available for work, study, and physical and mental growth; see, e.g., Fogel (1994), Miguel and Kremer (2004), Bleakley (2007, 2010), and Weil (2007). Communicable diseases, however, rarely affect individuals in isolation. Throughout history, endemic and epidemic diseases have typically acted on large groups of individuals.

I explore the influence of disease spillovers on the social organization of production and exchange in a highly stylized assignment model. The model provides a framework to understand why many disease-prone societies in the past and at present have sought to counter spillovers (i.e., contagion) through social and economic arrangements that foster social segregation. In the model, individuals are healthy or diseased, production and exchange are social activities, and meetings take place through random matchings. The main economic problem is that of sorting individuals across a continuum of communities. I show that when disease spillovers are sufficiently negative, it is (socially) efficient to segregate healthy and diseased individuals across communities (Proposition 1).

The social organization of production and exchange matters because aggregate efficiency is a function of the technology for production and the sorting of workers across communities (Proposition 2). The model is therefore able to explore how disease influences economic development beyond a purely biological mechanism. On a social front, disease matters because it inhibits the integration of different communities (i.e., social groups or geographic regions) within an economy. On an economic front, disease matters because it limits the extent of the market and the degree centralization of economic activity. Both social integration and economic centralization are crucial for a country’s economic development.

The model views disease spillovers as a non-technological influence in aggregate efficiency. I call this a productivity effect of disease and systematically characterize it across production and exchange arrangements that differ in their degree of centralization and dis-
ease exposure (Propositions 3 to 5). I also show that capital and labor markets tackle disease spillovers through a “social barrier.” This “social barrier” is an endogenous response to disease spillovers. Therefore, the absence of factor price equalization across individuals and communities is not indicative of misallocation, as it is in Hayashi and Prescott (2008) and Hsieh et al. (2019), say. Calibrating the disease spillovers to be consistent with the randomized treatment interventions for deworming in Miguel and Kremer (2004), and the quasi-experimental findings in Bleakley (2010), I find that the productivity effect of disease is at least as large as the proximate effect measured by Fogel (1994) and Weil (2007). The aggregate efficiency of centralized markets without disease spillovers is higher than with spillovers by roughly a factor of three.

There are many historical examples of social and legal enforcement of distance-based rules whose objective, in part, appears to be to counter contagion through social segregation. I briefly catalogue some of these examples here. The objective of the paper is not to rule out competing views, or to justify or validate social segregation. Rather, the examples offer historical perspectives of the general tension between aggregate efficiency and disease spillovers sketched by the model in seemingly disconnected social and economic arrangements. These historical perspectives are consistent with the productivity effect of disease, as their scope is not that of an individual (i.e., biological) response. They are also still relevant in today’s world. The social influence of disease on production and exchange is often path-dependent and difficult to change.

A prominent example of social segregation by disease is the Indian caste system in which direct and indirect physical contact has been socially regulated since ancient times. As a disease-avoidance mechanism, the ‘practice of untouchability’ seems to focus on social barriers for workers engaged in ‘polluting’ and unhealthy sectors such as animal skinning, leather-work, scavenging, and removing sewage and corpses; see, e.g., Hutton (1969). On epidemiological grounds, McNeill (1976) views the emergence of caste as a social response to an epidemiological imbalance between invaders unfamiliar with the diseases and parasitic infections of the Indus valley and natives with acquired tolerance to local infections.¹

¹Briefly stated, the central argument in McNeill (1976) is that human populations and their disease environments reach a stable equilibrium which, when disrupted (i.e., because of trade or travel), motivates social, economic, demographic, and political changes. Crosby (1972) and Zinsser (1935) are equally relevant.
Other familiar forms of segregation by disease rely on legal barriers. Quarantines and isolation in medieval Europe, where social interactions were temporarily restricted to protect coastal cities from bubonic plague, were based on state enforcement. Restrictions on factor mobility and exchange represented a “better safe than sorry” strategy against a very dangerous and invisible disease; see, e.g., Birchenall (2021). Additional state-enforced restrictions included the permanent physical separation into dedicated spaces for individuals afflicted with tuberculosis and Hansen’s disease, and the permanent spatial segregation of land-use in cities and towns. Rapid growth in industrializing Western cities exacerbated an “urban mortality penalty” present since ancient times. Cities, in response, often limited or eliminated physical contact between individuals and unhealthy land-uses in production through zoning regulations; see, e.g., Pinter-Wollman et al. (2018).

Segregation by disease also operated through the use of physical barriers. In pre-colonial Africa, peripheral markets carried out trade in such a highly decentralized way that exchange featured no direct physical interactions between partners, i.e., trade was ‘silent’ and used alternating offers without face-to-face contacts. Hartwig and Patterson (1978) argue that decentralized exchange with no interactions between traders reduced the risk of disease transmission in Central Africa. Woodburn (2016) also stressed the disease-avoidance nature of a silent trade protocol. Colonial policies that promoted integration and long-distance trade in sub-Saharan Africa inadvertently lead to outbreaks of indigenous epidemic diseases such as sleeping sickness (trypanosomiasis) and malaria; see, e.g., Azevado (1978) and Brown (1978).

Finally, I consider the emergence of high altitude cities in pre-Columbian empires in the New World. High altitude is protective against disease vectors and bacteria prevalent in disease-prone coastal areas; see, e.g., Coatsworth (2008). During pre-colonial times, the large tributary empires of the New World established their capital cities at high elevations. Even the military campaigns by the Incas recognized the unhealthiness of the low lands and limited the exposure of soldiers to tropical locations; see, e.g., Monge (1948). During colonial times, the British army in Jamaica and India relied on stations at high elevations as a way to counter endemic malaria and periodic bouts of yellow fever; see, e.g., Curtin (1989). As discussions on the ‘disruptive’ role of disease.
with other forms of segregation, high elevation has protected large urban areas from tropical diseases at the expense of distancing from ports and coastal areas. The distance between production and exchange centers has limited the ability of tropical countries, even those not landlocked, to engage in global trade.

**Some related literature.** Economists have been typically concerned with the biological and demographic effects of disease.\(^2\) In a development accounting framework, Weil (2007) measured the significance of the *proximate effect* of health in explaining income differences between rich and poor countries. I extend Weil’s (2007) measurement strategy. Weil (2007) treats health human capital in efficiency units of labor. Under efficiency units, the sorting of workers is irrelevant since only mean (or total) health human matters for aggregate output.

Indirect links between disease and long-run economic and political outcomes include the *colonial origins hypothesis* advanced by Acemoglu et al. (2001) and Alsan’s (2015) detailed case study of the TseTse fly on the political centralization and demography of sub-Saharan Africa. Work in this area is highly empirical and naturally focuses on sub-Saharan Africa. It has been established that historical disease environments contributed over the long-run to Africa’s underdevelopment likely through political economy distortions, greater ethnic diversity, and the suboptimal provision of public goods; see, e.g., Cervellati, Esposito and Sunde (2017); Cervellati, Sunde, and Valmori (2017); Depetris-Chauvin and Weil (2018); Fenske (2014); Gennaioli and Rainer (2007); and Weil (2018) among others. This paper develops a framework to interpret evidence already presented and offers parallels with other historical instances where disease spillovers and “social barriers” also appear to have limited sociality and the extent of the market.\(^3\)

---

\(^2\)Demographic effects are linked to the burden of higher population growth. Disease loads that reduce population growth are bound to increase temporarily output per capita due to diminishing returns to labor; see, e.g., Acemoglu and Johnson (2007), Birchenall (2007), and Young (2005). The influence of disease on education and physical capital has been amply studied; see, e.g., Bleakley (2007, 2010), Bleakley and Lange (2005), Evans and Miguel (2005), and Soares (2005). Disease influences capital markets by influencing, among other aspects, savings for retirement and education; see, e.g., De Nardi et al. (2009). Bloom et al. (2019) seeks to reconcile the empirical effects of health on economic growth at disaggregate and aggregate levels.

\(^3\)Physical barriers figure prominently in studies of ethnic and linguistic fractionalization because they facilitate the identification of causal effects; see, e.g., Michalopoulos (2012); Nunn and Puga (2018). Ethnolinguistic fractionalization is often viewed as non-hierarchical social segmentation, whereas some of the examples I consider have a clear social stratification (i.e., castes). Since the “social barriers” discussed here
There is a large literature in economics that examines socioeconomic segregation but without touching on the role of disease; see, e.g., Schelling (1971) for a foundational work on sorting and segregation. Segregation is often examined in the realm of education and location theory; see, e.g., Benabou (1996), Kremer and Maskin (1996), Fernández and Rogerson (2001), and Becker and Murphy (2003). I bring insights from this literature to study disease spillovers with a focus on the effects of worker sorting on aggregate production efficiency. The paper is also related to Jones (2014) who used a general aggregator to study imperfect substitutability and the role of human capital in development accounting for educational outcomes; see also Caselli and Ciccone (2019). Negative disease spillovers are at the core of the strong consensus supporting public health programs worldwide, but they have received less attention in the long-run growth literature relative to, say, the study of externalities for human capital in the form of education.

In a complementary paper, Fogli and Veldkamp (2018) proposed a network approach to study the effect of different social networks on aggregate output and technology adoption. Their model is dynamic and rich in terms of the social connections between individuals. In their model, the diffusion of ideas and germs through the population serves to trace the structure of the social network, and the structure of the social network serves to further diffuse ideas and germs in the population. The rich dynamics and social structure make the characterization and implementation of a network setting difficult. Overall, Fogli and Veldkamp’s (2018, p. 24) “bottom line is that the way in which networks affect economic growth depends on the disease environment.” This paper makes the social structure much simpler, so the effects of the disease environment on production and exchange protocols are very tractable. I have also taken the level of technology as exogenous. This makes aggregate efficiency comparisons direct and more revealing.

---

4Fogli and Veldkamp (2018) characterize the average path length of the network, i.e., the distance between nodes, and limiting properties of the network, i.e., the existence of a stationary disease-free. The mapping between social networks across countries is also challenging. Cross-country measures of individualism-collectivism, individual mobility, and friendship links are drawn from attitude surveys made by a multinational US company (IBM) in host countries, the cross-state mobility of first-generation US immigrants from different countries, and the number of close friends US residents with ancestors from different countries report. Their framework involves many moving parts. For example, since high-diffusion networks help technology adoption but foster disease, Fogli and Veldkamp (2018, p.22) find quantitatively that the effect of high-diffusion networks on technology and output “can be positive, negative, or zero.”
2 Model Economy

This section exposes channels, many not tacitly recognized in economics, through which disease influences aggregate efficiency. The starting point of the model is the social nature of economic transactions. In the model, disease spillovers influence aggregate output and efficiency depending on the social organization of production and exchange. The organizational arrangements, and their aggregate efficiency, are the focus of this section.

Table 1. Production and exchange arrangements to be considered.

<table>
<thead>
<tr>
<th>Social arrangement</th>
<th>Brief description</th>
<th>Aggregate efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Segregated</td>
<td>Social segregation by disease across communities with disease spillovers and no capita exchange</td>
<td>$A_s$</td>
</tr>
<tr>
<td>Centralized</td>
<td>Centralized production without individual sorting, with disease spillovers and no capita exchange</td>
<td>$A_c$</td>
</tr>
<tr>
<td>No spillovers</td>
<td>Segregation by disease across communities with centralized capital exchange and no disease spillovers</td>
<td>$A_n$</td>
</tr>
<tr>
<td>Decentralized</td>
<td>Segregation by disease across communities with decentralized capital exchange and no disease spillovers</td>
<td>$A_d$</td>
</tr>
<tr>
<td>Integrated</td>
<td>Social integration across communities with decentralized capital exchange and disease spillovers</td>
<td>$A_i$</td>
</tr>
</tbody>
</table>

Note: The details behind each protocol are presented in the text.

To aid with the presentation of the results, Table 1 lists the production and exchange arrangements I consider with their defining characteristics in terms of individual sorting, disease spillovers, and capital exchange. The details behind each arrangement are discussed below. The table presents the arrangements in the order they are presented, and not in terms of their aggregate efficiencies, which are explicitly derived and ranked in this section.

**Individuals and communities.** There are $N$ individuals and a continuum of communities (e.g., intermediate sectors, geographic regions, or social divisions) of measure 1. Aggregate capital is $K > 0$. Each individual is endowed with the same amount of physical capital, $k = K/N$. Capital is tradable, but I delay the study of capital trade until later in this section. Individuals are of two types, healthy and diseased. Their respective health human capitals are $h$ and $d$. 
**Assumption 1.** Healthy individuals have higher labor productivity than diseased individuals, i.e., \( h > d \geq 0 \).

The fraction of healthy and diseased individuals in the economy are \( \gamma \) and \( 1 - \gamma \) respectively. The economic problem is the sorting of individuals across communities. The fraction of healthy individuals to be assigned to community \( e \in [0, 1] \) is \( n(e) \). The assignment’s feasibility condition is

\[
\int_0^1 n(e) de = \gamma. \tag{1}
\]

Individuals cannot produce in isolation. Within each community, production takes place in pairs of *randomly matched* individuals. (I discuss the desirability of pairwise meetings in Proposition 3.) When two individuals meet, the pair’s pooled physical capital is the (arithmetic) mean of the two individual’s capital. Health human capital is also pooled, but not necessarily linearly. The way health human capital is averaged depends on disease spillovers. I assume that when individual of type \( h \) and \( d \) meet, the pair’s health human capital is

\[
\mu^\rho(h, d) = \left[ (h^\rho + d^\rho) / 2 \right]^{1/\rho}, \tag{2}
\]

where the CES aggregator is representative, i.e., \( \mu^\rho(h, h) = h \) and \( \mu^\rho(d, d) = d \), linearly homogeneous, i.e., \( \mu^\rho(\theta h, \theta d) = \theta \mu^\rho(h, d) \), and symmetric, i.e., \( \mu^\rho(h, d) = \mu^\rho(d, h) \).

As will be evident below, the parameter \( \rho \) is central to determine individual sorting and the social organization of production (Proposition 1). The reason is that \( \rho \) captures the disease spillovers in a matched pair. In particular, the pairwise health human capital \( \mu^\rho(h, d) \) is increasing in \( \rho \). For example, (2) satisfies: \( \mu_{-\infty}(h, d) = \min\{h, d\} = d \), which represents strongly *negative* spillovers, and \( \mu_{+\infty}(h, d) = \max\{h, d\} = h \), which represents strongly *positive* disease spillovers.\(^5\)

The level of technology is given, denoted by \( z \), and the pairwise production function is Cobb-Douglas. When alike types meet they produce \( f(h, h) = zh^{1-\alpha}k^\alpha \) or \( f(d, d) = zd^{1-\alpha}k^\alpha \) with \( \alpha \in (0, 1) \). When healthy and diseased individuals meet they produce \( f(h, d) = z[(h^\rho + d^\rho) / 2]^{(1-\alpha)/\rho}k^\alpha \).

\(^5\)Other values of \( \rho \) yield traditional means, i.e., \( \mu_{-1}(h, d) = 2(hd)/(h + d) \) is the harmonic mean of \( h \) and \( d \), \( \mu_0(h, d) = \sqrt{hd} \) is the geometric mean, and \( \mu_1(h, d) = (h + d)/2 \) is the arithmetic mean.
A healthy individual meets another healthy individual in proportion to the fraction of healthy individuals in the community. The *meeting probability* between healthy individuals is \(n(e)^2\). Likewise, the meeting probability between diseased individuals is \((1 - n(e))^2\). The two (symmetric) ways healthy and diseased individuals meet yields \(2n(e)(1 - n(e))\).

Community \(e\)'s expected output per capita, denoted by \(q(n(e))\), is

\[
q(n(e)) \equiv \{n(e)^2f(h, h) + 2n(e)(1 - n(e))f(h, d) + (1 - n(e))^2f(d, d)\},
\]

(3)

where, obviously, \(n(e)^2 + 2n(e)(1 - n(e)) + (1 - n(e))^2 = 1\).

**Social segregation.** Let \(n^*(e)\) denote the optimal value of \(n(e)\). Since communities are homogeneous, there are only two ways to sort individuals: healthy and diseased individuals should be either integrated or segregated; see Figure 1. Segregation is more likely to maximize aggregate output when disease exerts negative spillovers. The following standing assumption captures these spillovers.

**Assumption 2.** Disease spillovers are sufficiently negative, i.e., \(\rho\) and \(\alpha\) satisfy \(\rho < 1 - \alpha\).

The sorting of individuals across communities is characterized by the next proposition.

**Proposition 1** *It is (socially) optimal to segregate healthy and diseased individuals, with \(n^*(e) = 1\) in a measure \(\gamma\) of communities and \(n^*(e) = 0\) in a measure \(1 - \gamma\) of communities.*

**Proof.** The Lagrangean of the assignment problem is

\[
\mathcal{L} \equiv \int_0^1 q(n(e))de + \lambda_\gamma \left[\gamma - \int_0^1 n(e)de\right] + \int_0^1 \{\lambda_0(e)n(e) + \lambda_1(e)[1 - n(e)]\}de,
\]

(4)

where \(\lambda_\gamma\) is the Lagrange multiplier on (1), and \(\lambda_0(e)\) and \(\lambda_1(e)\) are the Lagrange multipliers on \(n(e) \geq 0\) and \(1 - n(e) \geq 0\), respectively. The first-order conditions include

\[
\frac{\partial q(n^*(e))}{\partial n(e)} - \lambda_\gamma + \lambda_0(e) - \lambda_1(e) = 0,
\]

(5)

\(n^*(e) \geq 0, 1 - n^*(e) \geq 0, \lambda_0(e)n^*(e) = 0, \lambda_1(e)[1 - n^*(e)] = 0, \lambda_0(e) \geq 0, \text{ and } \lambda_1(e) \geq 0.\)
Since the constraints are linear, the assignment depends on concavity-convexity of $q(n(e))$ in (4). If $q(n(e))$ is strictly convex in $n(e)$, the only solution to (5) is $n^*(e) = \{0, 1\}$, where the measure of all communities with $n^*(e) = 1$ equals $\gamma$ to satisfy (1); see Figure 1. Convexity requires $\partial^2 q(n^*(e)) / \partial n(e)^2 = f(h, h) - 2f(h, d) + f(d, d) > 0$. Using (2), this condition is
\[
[(h^\rho + d^\rho) / 2]^{1/\rho} < [(h^{1-\alpha} + d^{1-\alpha}) / 2]^{1/(1-\alpha)}.
\] (6)

By Jensen’s inequality, (6) requires $\rho < 1 - \alpha$, which is Assumption 2. ■

Proposition 1 is the foundation for the rest of the paper. It shows that disease spillovers determine the sorting of healthy and diseased individuals. Assumption 2 implies that output per capita is maximized by a corner solution in which each community has only healthy or diseased individuals with no room for mixing. The distribution of these segregated communities along $[0, 1]$ is indeterminate. An assignment with $n^*(e) = 1$ for $e \in [0, \gamma]$ and $n^*(e) = 0$ for $e \in (\gamma, 1]$ where all healthy pairs are ‘clustered’ (right panel of Figure 1) is equivalent to a ‘polarized’ assignment with $n^*(e) = 1$ for $e \in [0, \gamma/2] \cup [1 - \gamma/2, 1]$ and $n^*(e) = 0$ for $e \in (\gamma/2, 1 - \gamma/2)$ where the (equally-sized) segregated communities are in the extremes of the $[0, 1]$ interval.

Segregation by disease is optimal for all values of $(h, d)$ and $(\rho, \alpha)$ that satisfy Assumptions 1 and 2. The literature on segregation by skill typically finds that segregation depends on particular values of the skills. For instance, in Kremer and Maskin (1996) and McCann and Trokhimtchouk (2010, Example 1), $h$ and $d$ would have to be sufficiently different for segregation to occur.6 In here, even small differences in $h$ and $d$ lead to social segregation.

The pairwise health human capital aggregate is similar to the general human capital aggregator in Jones (2014).7 The pairwise health human capital aggregate (2) is the key margin

---

6Consider $f(h, d) = d^\beta h^\alpha$. Worker segregation is socially efficient iff differences between workers’ skills are sufficiently large. The second-order condition (6) yields $h^{\alpha+\beta} - 2d^\beta h^\alpha + d^{2\beta}$. For $\delta \equiv h/d \geq 1$, the relevant terms needed to determine if segregation is optimal can be written as $d^{\beta+\delta} (\delta^{\alpha+\beta} - 2\delta^{\beta} + 1)$. Under $\beta > \eta > 0$, the term in brackets is zero if skills are equal or $\delta = 1$. This term decreases monotonically as $\delta$ increases until it reaches a trough at $(2\beta/(\eta + \beta))^{1/\eta}$. Then it increases monotonically until it eventually becomes positive. A positive value is necessary for segregation.

7Jones (2014) used a generalized human capital measure in a development accounting framework. The empirical implementation of a general aggregator depends on the sources of variation underlying the different inputs; see, e.g., the comment on Jones (2014) by Caselli and Ciccone (2019). One of these sources of variation is the relative efficiency of skilled workers. In the case of health human capital, relative efficiency seems to be driven largely by physiological considerations; see, e.g., Fogel (1994).
Figure 1: Segregated and integrated assignments. In the left panel, $n^*(e) = \gamma$ for all $e \in [0, 1]$ in an integrated assignment. In the right, $n^*(e) = 1$ in a fraction $\gamma$ of communities in a segregated assignment. The shaded areas represent the feasibility condition (1).

where disease spillovers act.

The CES aggregator (2) showcases the tension between disease spillovers and diminishing returns. For instance, only $\alpha$ and $\rho$ determine if the assignment leads to segregation or integration (Assumption 2). Even a small difference between $1 - \alpha$ and $\rho$ is sufficient to induce segregation. To illustrate the tension between $\rho$ and $\alpha$, consider two values of $\rho$. First, under $\rho = 1$, (2) becomes the arithmetic mean of $h$ and $d$: $\mu_1(h, d) = (h + d)/2$. Assumption 2 (and hence Proposition 1) does not hold in this case. The function $q(n(e))$ is strictly concave in $n(e)$ so it is optimal to integrate healthy and diseased individuals by setting $n^*(e) = \gamma < 1$ for all $e \in [0, 1]$. In this case, the disease spillovers are small so, to counter the diminishing returns that arise when healthy workers are paired together, the efficient assignment requires social integration.

Consider next $\rho = -\infty$, which represents Kremer’s (1993) O-ring technology. Under $\rho = -\infty$, meeting a diseased individual renders healthy individuals unproductive as (2) becomes $\mu_{-\infty}(h, d) = \min\{h, d\}$, and (3) becomes $q(n(e)) = n(e)^2 f(h, h) + (1 - n(e)^2)f(d, d)$. Since $f(h, h) > f(d, d)$, it is optimal to ‘protect’ the labor productivity of healthy individuals by setting $n^*(e) = 1$ in the largest possible number of communities. As these examples show, the tension between $1 - \alpha$ and $\rho$ in Assumption 2 is a tension between diminishing returns to scale and (negative) disease spillovers.
Figure 2: Pairwise productivity and disease spillovers. Under $\rho > 1 - \alpha$, integration is socially efficient as the gains exceed the losses while under $\rho < 1 - \alpha$, social segregation is efficient.

Figure 2 illustrates this tension. Under efficiency units, $\rho = 1$, the pairwise health human capital is $\mu_1(h, d)$ and pairwise output is $f(\mu_1(h, d))$. Integration is optimal since the losses from ‘breaking’ a healthy pair $f(h, h) - f(\mu_1(h, d))$ are smaller than the gains from ‘forming’ an integrated pair $f(\mu_1(h, d)) - f(d, d)$. Under $\rho = 1 - \alpha$, pairwise health human capital is $\mu_{1-\alpha}(h, d)$ and the gains and losses from integration are identical and equal to $[f(h, h) - f(d, d)]/2$. For $\rho < 1 - \alpha$, production efficiency dictates social segregation.

**Output and aggregate efficiency.** Social segregation matters for aggregate output and efficiency. Let $H \equiv \gamma h + (1 - \gamma)d$ denote mean health human capital, and let the subscript $(s)$ denote the allocation under social segregation. Aggregate output per capita is

$$y_s = \int_0^1 q(n^*(e))de.$$ (7)

Aggregate output is $Y_s = y_sN$, and $A_s$ is aggregate efficiency, i.e., measured TFP.

**Proposition 2** Under social segregation, aggregate output satisfies $Y_s = A_s K^\alpha (HN)^{1-\alpha}$, for

$$A_s \equiv z \left[ \gamma \left( \frac{h}{H} \right)^{1-\alpha} + (1 - \gamma) \left( \frac{d}{H} \right)^{1-\alpha} \right],$$ (8)

where aggregate efficiency is lower than technologically possible, i.e., $A_s < z$. 

12
Proof. Using $n^*(e)$ in (3) yields $q(n^*(e)) = \gamma f(h, h) + (1 - \gamma)f(d, d)$. This and (7) yield $y_s = z[\gamma h^{1-\alpha} + (1 - \gamma)d^{1-\alpha}]k^\alpha$. Simple rearrangements and (8) yield $y_s = A_s H^{1-\alpha} k^\alpha$, from which $Y_s$ follows. The inequality $A_s < z$ in (8) follows from Jensen’s inequality: $\gamma h^{1-\alpha} + (1 - \gamma)d^{1-\alpha} < [\gamma h + (1 - \gamma)d]^{1-\alpha} \equiv H^{1-\alpha}$ for $\alpha \in (0, 1)$.

The aggregate production function is Cobb-Douglas with the aggregate labor input as the product of the quantity and quality of the labor force, $N$ and $H$ respectively. Labor quality $H$ captures the proximate effect of disease on aggregate output. The proximate effect depends on mean health human capital, but not on how individuals are sorted across communities. In Proposition 2, disease is a non-technological influence in TFP. The effect of health human capital on $A_s$ acts as a productivity effect of disease. As (8) shows, aggregating across segregated communities lowers aggregate efficiency relative to the level of technology.

The productivity effect of disease is independent of mean health human capital. If, for example, $h$ and $d$ increase by a factor of $\theta > 0$, $H$ increases by $\theta$ and $Y_s$ increases by $\theta^{(1-\alpha)}$, and $A_s$ remains unchanged. Instead, $A_s$ depends on how health human capital could spillover across production units. Consider, for example, a mean preserving improvement of health $\Delta H = \gamma \Delta h + (1 - \gamma)\Delta d = 0$ with $\Delta d > 0 > \Delta h$. While $H$ remains constant, aggregate output and aggregate efficiency (i.e., measured TFP) increase by

$$\Delta Y_s = \Delta A_s = (1 - \alpha)(1 - \gamma)\{d^{-\alpha} - h^{-\alpha}\}\Delta d > 0.$$  \hspace{1cm} (9)

As health dispersion vanishes $d \to h$, sorting becomes indeterminate in the sense that all allocations become equally productive. In this case, $A_s$ and $z$ coincide.

The productivity effect of disease arises because the aggregate labor input is represented by total health human capital $NH$. Representing the aggregate labor input using an aggregate that takes into account the distribution of health human capital (not just its mean) would ‘eliminate’ the role of disease in aggregate efficiency. Changes in health dispersion would then be assigned to changes in the labor input, as in Jones (2014). Proposition 2, however, maps the aggregate inputs $K$ and $HN$ into an aggregate output $Y_s$. This representation is consistent with the actual measurement of aggregate TFP and the accounting framework in Weil (2007).
A centralized marketplace. To further study the productivity effects of disease, abandon temporarily the pairwise matching assumption. Assume that all individuals meet in a centralized marketplace subject to disease spillovers. Let subscript \(c\) denote the variables under this alternative production arrangement. There is no sorting as all individuals are in a single community. In a centralized marketplace, \(n(c)\) is not a choice variable as it is under segregation or integration. I will return to integration later on.

In per capita terms, physical capital in a centralized marketplace is \(k_c = \gamma k + (1 - \gamma)k = k\), health human capital is \(\mu_{c,\rho}(h, d) = [\gamma h^\rho + (1 - \gamma)d^\rho]^{1/\rho}\), and output per capita is \(y_c = z[\mu_{c,\rho}(h, d)]^{(1 - \alpha)k^\alpha}\). Aggregate output is also Cobb-Douglas, \(Y_c = A_c K^\alpha H N^{1 - \alpha}\), with

\[
A_c \equiv z \left[ \gamma \left( \frac{h}{H} \right)^\rho + (1 - \gamma) \left( \frac{d}{H} \right)^\rho \right]^{(1 - \alpha)/\rho}. \tag{10}
\]

This Cobb-Douglas production function differs from that in Proposition 2 only due to differences in measured TFP. In particular,

**Proposition 3** Aggregate efficiency in a centralized marketplace with disease spillovers is lower than under social segregation, i.e., \(A_c < A_s\).

**Proof.** From (8) and (10), \(A_c < A_s\) iff \([\gamma h^\rho + (1 - \gamma)d^\rho]^{(1 - \alpha)/\rho} < \gamma h^{1 - \alpha} + (1 - \gamma)d^{1 - \alpha}\). This inequality, which is a general version of (6) that holds for any value of \(\gamma \in (0, 1)\), can be written as

\[
[\gamma h^\rho + (1 - \gamma)d^\rho]^{1/\rho} < [\gamma h^{1 - \alpha} + (1 - \gamma)d^{1 - \alpha}]^{1/(1 - \alpha)}. \tag{11}
\]

By Jensen's inequality, (11) is satisfied when \(\rho < 1 - \alpha\), which is Assumption 2.

Since segregation by disease is socially efficient, pairwise meetings and individual sorting are a source of aggregate efficiency gains. The reason is that less efficiency is lost in the aggregation of segregated pairs than in the pooling of all individuals in a centralized marketplace with disease spillovers. For instance, Propositions 2 and 3 imply that \(A_c < A_s < z\). On the contrary, when integration is optimal, inequality (11) is reversed and centralizing...
production in a single marketplace is a source of efficiency gains. In the extreme case of \( \rho \to -\infty, A_c \to A_\infty \equiv z (d/H)^{(1-\alpha)} \), which is a lower bound for aggregate efficiency.

**Capital exchange.** So far, there is no exchange in physical capital. Since physical capital is more productive in the hands of healthy individuals, aggregate output and aggregate efficiency would differ if capital could be exchanged across communities. The efficiency gains, however, depend on how capital markets are introduced. To study the social organization of exchange, I next consider several trading arrangements that vary in the degree of centralization and exposure to disease spillovers. The starting point is a segregated assignment. Capital is therefore exchanged between segregated communities. (Segregated worker-pairs are also individual types.)

Consider first a benchmark in which capital is exchanged in a centralized market with no disease spillovers. The subscript \((n)\) denotes allocations with no spillovers:

**Proposition 4** Suppose that segregated communities exchange capital in a centralized marketplace with no disease spillovers. Aggregate output, \( Y_n = A_n K^\alpha (HN)^{1-\alpha} \), is as efficient as technologically possible, i.e., \( A_n = z \).

**Proof.** A competitive capital market between segregated communities attains the same allocation as a social planner that solves: 
\[
z \max \{ \gamma h^{1-\alpha} k(h)^\alpha + (1 - \gamma) d^{1-\alpha} k(d)^\alpha \}, \text{s.t., } \gamma k(h) + (1 - \gamma) k(d) = k.
\]
Capital-labor ratios are equalized, i.e., \( k_n(h)/h = k_n(d)/d \), and substitutions yield \( Y_n = z K^\alpha (HN)^{1-\alpha} \) hence \( A_n = z \).

Capital exchange increases aggregate efficiency because capital endowments are inefficiently assigned, i.e., individuals endowments are independent of health human capital. In Proposition 4, capital exchange undoes the negative effects of disease. There are, however, no disease spillovers when individuals interact in a centralized marketplace.

To introduce disease spillovers, consider again pairwise trading. Assume that segregated communities can exchange capital but that trading opportunities arrive stochastically as segregated worker-pairs search for trading partners. The probability that an individual from a healthy community meets a healthy individual is \( \gamma^2 \). The meeting probability between individuals from diseased communities is \( (1 - \gamma)^2 \), and the meeting probability between individuals from healthy and diseased communities is \( 2\gamma(1 - \gamma) \).
There are no gains from trade when alike individuals meet. Gains from trade occur when capital is reallocated between diverse communities. Consider further two possible outcomes in these decentralized meetings depending on the absence or presence of disease spillovers. Assume first that pairwise trading opportunities arise stochastically but suppose that capital is exchanged in decentralized markets with no disease spillovers, similar to Proposition 4 but at a pairwise level. Let subscript \( (d) \) denote this decentralized trading arrangement. Aggregate output is Cobb-Douglas with \( A_d \) described below.

**Proposition 5** Suppose that segregated communities exchange capital in decentralized marketplaces with no disease spillovers. Aggregate efficiency is higher than under segregation and no capital exchange (Proposition 2) but lower than in a centralized marketplace with no disease spillovers (Proposition 4), i.e., \( A_s < A_d < A_n \).

**Proof.** The expected output of alike pairs is \( \gamma^2 z h^{1-\alpha} k^\alpha + (1 - \gamma)^2 z d^{1-\alpha} k^\alpha \). The expected output of a meeting between individuals from diverse communities is given by \( 2\gamma(1 - \gamma)z \max \{[h^{1-\alpha} k(h) + d^{1-\alpha} k(d)]/2, \text{s.t.}, k(h) + k(d) = 2k \} \). The efficient allocation of capital satisfies \( k_d(d) = [2d/(h + d)]k \) and \( k_d(h) = [2h/(h + d)]k \), so the value of the maximization problem is \( k^\alpha [(h + d)/2]^{1-\alpha} \). Aggregate efficiency satisfies

\[
A_d \equiv z \left[ \gamma^2 \left( \frac{h}{H} \right)^{1-\alpha} + (1 - \gamma)^2 \left( \frac{d}{H} \right)^{1-\alpha} + 2\gamma(1 - \gamma) \left\{ \frac{1}{2} \left( \frac{h + d}{H} \right) \right\} \right]^{(1-\alpha)}. \tag{12}
\]

From (8) and (12), \( A_s < A_d \) iff \( \gamma h^{1-\alpha} + (1 - \gamma)d^{1-\alpha} < \gamma^2 h^{1-\alpha} + (1 - \gamma)^2 d^{1-\alpha} + 2\gamma(1 - \gamma)((h + d)/2)^{1-\alpha} \). This inequality can be written simply as: \( [(h^{1-\alpha} + d^{1-\alpha})/2]^{1/(1-\alpha)} < (h + d)/2 \), which holds by Jensen’s inequality. Recall that \( A_n = z \). Hence for \( A_d < A_n \), one simply needs \( \gamma^2 h^{1-\alpha} + (1 - \gamma)^2 d^{1-\alpha} + 2\gamma(1 - \gamma)((h + d)/2)^{1-\alpha} < H^{1-\alpha} \) in (12). Since \( H^{1-\alpha} = (\gamma^2 h + (1 - \gamma)^2 d + 2\gamma(1 - \gamma)((h + d)/2))^{1-\alpha} \), the inequality holds by concavity.

Without disease spillovers, capital exchange is a source of efficiency gains relative to social segregation by disease, i.e., \( A_s < A_d \). Aggregate efficiency improves because capital is directed toward higher productivity communities. For instance, heathier communities receive more capital relative to their endowment, while diseased communities end up with less capital, i.e., \( k_d(h)/k_d(d) = k_n(h)/k_n(d) = h/d > 1 \). Since these gains only apply to a
fraction $2\gamma(1-\gamma)$ of meetings, aggregate efficiency is still lower than in a central marketplace with no spillovers, i.e., $A_d < A_n$.

For the second capital exchange arrangement, assume that capital is exchanged in decentralized marketplaces with disease spillovers. I denote this trading arrangement by the subscript (i). I use (i) because decentralized capital exchange under disease spillovers coincides with an integrated assignment. For instance, when alike types meet, expected output is $\gamma^2 z h^{1-\alpha} k^\alpha + (1-\gamma)^2 z d^{1-\alpha} k^\alpha$, and when diverse types meet, expected output is $2\gamma(1-\gamma) z [\{(h^\rho + d^\rho)/2\}^{(1-\alpha)/\rho} k^\alpha]$. These expressions coincide with (3) when $n^*(e) = \gamma$ for all $e \in [0,1]$. Aggregate output is also Cobb-Douglas with aggregate efficiency $A_i$.

**Proposition 6** Suppose that segregated communities exchange capital in decentralized marketplaces with disease spillovers. Aggregate efficiency is higher than in a centralized marketplace with disease spillovers (Proposition 3) but lower than under segregation and no capital exchange (Proposition 2), i.e., $A_c < A_i < A_s$.

**Proof.** Under integration, aggregate output per capita is $\gamma^2 z h^{1-\alpha} k^\alpha + (1-\gamma)^2 z d^{1-\alpha} k^\alpha + 2\gamma(1-\gamma) z [\{(h^\rho + d^\rho)/2\}^{(1-\alpha)/\rho} k^\alpha]$; see (3). Aggregate efficiency is

$$A_i \equiv z \left[ \gamma^2 \left( \frac{h}{H} \right)^{1-\alpha} + (1-\gamma)^2 \left( \frac{d}{H} \right)^{1-\alpha} + 2\gamma(1-\gamma) \left\{ \frac{1}{2} \left( \frac{h}{H} \right)^\rho + \frac{1}{2} \left( \frac{d}{H} \right)^\rho \right\}^{(1-\alpha)/\rho} \right].$$

To show that $A_i < A_s$, one needs $\gamma^2 h^{1-\alpha} + (1-\gamma)^2 d^{1-\alpha} + 2\gamma(1-\gamma) [(h^\rho + d^\rho)/2]^{(1-\alpha)/\rho} < \gamma h^{1-\alpha} + (1-\gamma) d^{1-\alpha}$. Rearrangements show that this inequality implies $[(h^\rho + d^\rho)/2]^{1/\rho} < \{(h^{1-\alpha} + d^{1-\alpha})/2\}^{1/(1-\alpha)}$, which holds due to Assumption 2; see (6).

Next, rewrite $[\gamma h^\rho + (1-\gamma) d^\rho]^{(1-\alpha)/\rho}$ in $A_c$ as $\{\gamma^2 h^\rho + (1-\gamma)^2 d^\rho + 2\gamma(1-\gamma) [(h^\rho + d^\rho)/2]\}^{(1-\alpha)/\rho}$. To show that $A_c < A_i$, notice that the relevant comparison becomes $\{\gamma^2 h^\rho + (1-\gamma)^2 d^\rho + 2\gamma(1-\gamma) [(h^\rho + d^\rho)/2]^{\rho/\rho}\}^{1/\rho} < \{\gamma^2 h^{1-\alpha} + (1-\gamma)^2 d^{1-\alpha} + 2\gamma(1-\gamma) [(h^\rho + d^\rho)/2]^{(1-\alpha)/\rho}\}^{1/(1-\alpha)}$, which again holds by Assumption 2. □

Proposition 1 has shown that it is inefficient to integrate healthy and diseased individuals for production purposes. Proposition 6 says that it socially inefficient to integrate segregated communities for exchange purposes. Exchanging capital in a decentralized marketplace exposes traders to disease spillovers and undoes the gains from segregation by disease.
As noted before, Proposition 6 coincides with an integrated assignment. An integrated assignment is more efficient than a centralized marketplace with disease spillovers, i.e., $A_c < A_i$. The reason is that the disease spillovers in a centralized marketplace (Proposition 3) apply to all individuals. In decentralized marketplaces (Proposition 6), only a fraction $2\gamma(1 - \gamma)$ of worker-pairs experience negative spillovers. This fraction is at most $1/2$.

The previous propositions complete the characterization of the production and exchange arrangements listed in Table 1. Overall, these propositions say that aggregate efficiency and output can be ranked according to the exposure of disease spillovers and the degree of centralization, as in:

$$A_c < A_i < A_s < A_d < A_n.$$  \hspace{1cm} (13)

The extremes cases in (13) involve centralized marketplaces. The case $A_c$ generates the largest scope for disease spillovers and the lowest aggregate efficiency. In $A_c$, all individuals meet in a centralized marketplace with disease spillovers. In the other extreme, $A_n$, communities exchange capital in a centralized marketplace with no spillovers at all. In the intermediate cases, some aggregate efficiency is always lost in the process of production and exchange. As a consequence of disease spillovers, all intermediate production and exchange protocols are less efficient than technologically possible.

“Social barriers” and misallocation. The assignment problem takes a social planner perspective. Competitive markets tackle disease spillovers through a “social barrier.” Social segregation is inconsistent with factor price equalization across communities or individuals.

Consider a market decentralization of the segregated assignment. Suppose, to obtain a contradiction, that diseased and healthy workers coexist in a location and benefit from forming a production pair. If wages for healthy and diseased workers are $w_s(h)$ and $w_s(d)$, a healthy worker would prefer a diseased worker, over a healthy one, if the net gains of forming such pair are large enough, i.e., $f(d, h) - w_s(d) \geq f(h, h) - w_s(h)$. Likewise, a diseased worker would prefer a healthy worker, over a diseased one, if $f(h, d) - w_s(h) \geq f(d, d) - w_s(d)$. These inequalities require $2f(d, h) \geq f(d, d) + f(h, h)$ which contradicts Assumption 2; see (6). Diseased individuals benefit from pairing with healthy workers, but are unable to ‘bribe’ them. Healthy individuals prefer pairing only with other healthy individuals.
When physical capital is rewarded its marginal product and wages exhaust income, i.e.,
\[ w_s(h) = (1 - \alpha)zk^\alpha h^{1-\alpha} \] and \[ w_s(d) = (1 - \alpha)zk^\alpha d^{1-\alpha}, \] the log-wage gap between healthy and diseased workers is
\[
\ln \left[ \frac{w_s(h)}{w_s(d)} \right] = (1 - \alpha) \ln \left[ \frac{h}{d} \right].
\] (14)

The inequality in marginal products in (14) is not evidence of misallocation. In standard concave economies, misallocation is associated with differences in marginal products for the same input. Hsieh et al. (2019), for example, studied “social barriers” associated with racial and gender discrimination in the US, and Hayashi and Prescott (2008) considered a “social barrier” that limited regional and occupational mobility in prewar Japan.

Even though the aggregate production function is a standard Cobb-Douglas, the economy in this paper is not a standard concave economy. Differences in marginal products are therefore not evidence of misallocation. The main issue here is that the “social barrier” is endogenous and not imposed, as in Hayashi and Prescott (2008) and Hsieh et al. (2019). Thus, marginal products are different, but there are no gains from factor reallocations. Segregation by disease is socially efficient and policies that promote social integration in production or exchange reduce aggregate efficiency and output.

The log-wage gap (14) is similar to the log-wage gap under efficiency units (e.g., Weil (2007)), but efficiency units assume that healthy and diseased workers are perfect substitutes. If \( \rho = 1 \), Assumption 2 is not satisfied and the assignment problem becomes a standard concave economy. Integration would be optimal and the baseline price for human capital would be \( w_i = (1 - \alpha)z[k/\mu_1(h, d)]^\alpha \).

**Some remarks.** I ignored dynamic spillovers, but in the Appendix I introduce disease transmission mechanics as in Anderson and May (1992). Disease transmission generates a dynamic spillover effect and an endogenous prevalence of disease. The dynamic spillovers essentially serve to weaken Assumption 2. The Appendix also examines a multi-class society where social hierarchies depend, as in reality, on multidimensional attributes beyond just health human capital. In a multi-class society, it is necessary to modify Assumption 2 to account for the number of potential classes but a similar set of results as the ones presented here follows.
A key assumption in this paper is that health types are observable. Partially observable types introduce a signal-extraction problem that makes complete segregation impossible. I consider partially observable health types elsewhere; Birchenall (2021). I show there that the danger and visibility of disease are central for understanding the social response to past epidemics, including exclusionary errors in the form of “false alarms.”

3 Quantitative Findings

This section quantifies the productivity effects of disease. In particular, I now examine how much larger aggregate efficiency would be in the arrangements ranked in (13).

Parameterization. I rely on parameter values consistent with observations on long-run economic growth and causal estimates of disease spillovers. I use a growth accounting framework and $z = 1$. Along a constant capital-output ratio, the productivity effect of disease contributes to aggregate output by $A^{1/(1-\alpha)}$.

I consider proximate changes in health from Fogel (1994) and Weil (2007). Fogel (1994) estimated that health human capital in Great Britain “increased by a factor of 1.95” over the period 1780-1980. Weil (2007) measured disease’s proximate effect on aggregate output across countries using differences on adult mortality and latent measures of health (i.e., age of menarche and height). Using cross-country differences in the age of menarche of 3.7 years, the range in his sample, Weil (2007, p. 1289) found differences in the “labor input per worker of a factor of 2.73.” Using cross-country differences in adult mortality between Iceland and Botswana, Weil (2007, p. 1292) found differences in the “labor input per worker by a factor of 1.59.” These numbers suggest that $H = 1$ and $H' = 2.5$ are a reasonable benchmark.

I take $\alpha = 0.40$ based on factor payments. I assume that $\gamma = 0.8$. As Fogel (1994, pp. 373-374) notes, toward the end of the 18th century, about 20 percent of the population in France and England was excluded from the labor force due to the lack of energy for work.

---

9I ignore complementarities between health and human or physical capital, and technology. I also ignore intertemporal and intergenerational complementarities associated with health gains that originate as early as during fetal development; Fogel (1994). The main offsetting effect not considered here is the demographic response to disease; see, e.g., Acemoglu and Johnson (2007). Increases in population counter the productivity and proximate effects of disease on aggregate income. Population growth, however, is likely transitory as during the demographic transition when fertility adjusted too slowly to mortality reductions.
i.e., “beggars constituted as much as a fifth of the populations of ancien régimes.” This value is conservative as it considers extremely malnourished individuals.

To calibrate the CES parameter $\rho$, I consider the following thought experiment. A population of diseased individuals randomly receives a health intervention with gains of $h = \Delta_T d$ where $\Delta_T > 1$ is the treatment effect. Untreated individuals meet with the directly treated and, as in (2), their health human capital improves to $\mu_\rho(d, \Delta_T d) = \Delta_U d$, with

$$\Delta_U = \mu_\rho(1, \Delta_T) = \left\{ \frac{1 + \Delta^\rho_T}{2} \right\}^{1/\rho},$$

(15)
capturing the disease spillovers, i.e., the treatment on the untreated.

I use the disease spillovers for deworming of intestinal parasites in school-aged children by Miguel and Kremer (2004, [2014]). They measured the direct and indirect benefits of treating moderate-to-heavy helminth infections with spatial spillovers at a radius of 3km from treated schools. Their direct and indirect treatment effects are $-0.333$ (s.e. 0.052) and $-0.102$ (s.e. 0.043); see Miguel and Kremer ([2014], Table B1).\(^{10}\) Assuming a one-for-one increase in health human capital yields $\Delta_T = 1.333$, $\Delta_U = 1.102$, and $\rho = -4.8$ in (15).\(^{11}\)

It is not possible to measure health human capital directly. I first consider $(h, d) = (1.2, 0.2)$ and $(h, d) = (1.15, 0.4)$ consistent with $H = 1$. The ratio $h/d = 6$ in the first case is based on a six-fold “urban mortality penalty” at young ages between industrializing cities and rural areas in eighteenth century England; see Birchenall (2007). The second case matches the ratio between the top and bottom decile in daily kcal consumption in France c.1785 and England c.1790 from Fogel (1994, Table 2).

**Findings.** Table 2 lists the production and exchange arrangements in (13). I focus on the efficiency of centralized trading with and without spillovers, i.e., $A_c$ and $A_n$, and the efficiency gains from social segregation, i.e., $A_s$ and $A_c$.

\(^{10}\) The intestinal infections were very common and the treatment very successful. The original estimates in Miguel and Kremer (2004) contained a coding error so I rely on their 2014 guide for replication. Bleakley (2010) considered spillovers in eradication campaigns against parasitic disease in the Americas. The reduced-form spillovers of direct health interventions on the income of untreated individuals are around “one third to one half of the direct effect on treated cohorts,” which is of the same order of magnitude as the estimates from Miguel and Kremer ([2014], Table B1).

\(^{11}\) The value of $\rho$ depends on the size of treated population. If the population receiving the direct intervention is $\gamma$, (15) yields $\Delta^\rho_U = 1 - \gamma + \gamma \Delta^\rho_T$. I use $\gamma = 1/2$ since the groups in Miguel and Kremer (2004, [2014]) are balanced and their estimates are weighted by population size.
Table 2. Proximate and productivity effects of disease.

<table>
<thead>
<tr>
<th></th>
<th>$H = 1$</th>
<th>$H' = 2.5$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(i)</td>
<td>(ii)</td>
</tr>
<tr>
<td>$h$</td>
<td>1.20</td>
<td>1.15</td>
</tr>
<tr>
<td>$d$</td>
<td>0.20</td>
<td>0.40</td>
</tr>
<tr>
<td>$A_{c}^{1/(1-\alpha)}$</td>
<td>0.28</td>
<td>0.56</td>
</tr>
<tr>
<td>$A_{c}^{1/(1-\alpha)}$</td>
<td>0.78</td>
<td>0.87</td>
</tr>
<tr>
<td>$A_{c}^{1/(1-\alpha)}$</td>
<td>0.95</td>
<td>0.98</td>
</tr>
<tr>
<td>$A_{d}^{1/(1-\alpha)}$</td>
<td>0.98</td>
<td>0.99</td>
</tr>
<tr>
<td>$A_{d}^{1/(1-\alpha)}$</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Note: The values are based on $\rho = -4.8$, $z = 1$, $\gamma = 0.8$, $\alpha = 0.4$.

Disease spillovers limit the efficiency of centralized production and exchange. Consider (i), which yield the same values for $A$ but different values for $H$ and $H'$. Under centralized production with disease spillovers, $A_c$ is 0.28 of the aggregate efficiency of centralized exchange without spillovers, i.e., $A_n = 1$. If disease spillovers are removed, aggregate efficiency and output per capita would triple. The difference between $A_c$ and the lower bound $A_- = (d/H)^{1-\alpha}$ implies that $\rho = -4.8$ induces strong spillover effects. (Recall that $A_c \rightarrow A_-$ as $\rho \rightarrow -\infty$.) Under $\rho \rightarrow -\infty$, eliminating disease spillovers leads to a five-fold increase in aggregate efficiency in (i). Under a less unequal distribution of health human capital in (ii), removing disease spillovers doubles and triples aggregate efficiency from $A_c$ and $A_-$. In Table 2, disease spillovers are more important in production arrangements than in exchange. Consider social segregation. The difference between $A_s$ and $A_c$ is ‘large,’ i.e., $A_{s}^{1/(1-\alpha)} = 0.95$ in (i), so segregation by disease for production purposes is effective at dealing with disease spillovers. In contrast, the marginal gains from decentralized capital exchange without disease spillovers are ‘small’ relative to the segregated assignment, as $A_{d}^{1/(1-\alpha)} = 0.98$ in (i). The marginal gains from no spillovers in centralized exchange $A_n$ are even smaller. A more equal distribution of health human capital in (ii) also yields ‘large’ gains from social segregation.

The productivity effect of disease depends on the distribution of health human capital. Instead of eliminating spillovers, columns (iii) and (iv) consider an even less unequal distribution of health human capital than (i) or (ii). The changes between (ii)-(iv) are mean-
preserving improvements relative to (i). These cases progressively increase $d'$ by $\Delta d = 0.5$ and lower $h'$ so as to maintain $H' = 2.5$. Reducing health dispersion increases aggregate efficiency by large orders of magnitude along any row. Efficiency in centralized markets with disease spillovers reaches $A_c = 0.96$ in (iv).

**Overview.** Mean health human capital has increased over time and health disparities have narrowed across individuals and countries. The levels of chronic malnutrition documented by Fogel (1994) are no longer widespread in the world and life expectancy in poor countries has converged to that of rich countries; see, e.g., Birchenall (2007). A change from $A_c$ under $(h, d) = (1.2, 0.2)$ to $A_n$ under $(h', d') = (2.6, 2.0)$ is perhaps the most representative change during modern times. In such comparison, output per capita increases by a factor of 9 as $H' \rightarrow 2.5 \times H$ and $A_n^{1/(1-\alpha)} \rightarrow 3.6 \times A_c^{1/(1-\alpha)}$. This increase goes a long way in accounting for economic growth in developed countries. Even in moderate cases such as (ii), the productivity gains are similar to the proximate gains, i.e., $A_n^{1/(1-\alpha)} \rightarrow 1.78 \times A_c^{1/(1-\alpha)}$.

The social organization of production and exchange has undergone a drastic change with the rise of modern cities. It is easy to rationalize large efficiency gains by considering urbanization. The degree of centralization in production and exchange, as measure by population density, has increased in modern times. In the past, cities played a central role in disease transmission. Until the twentieth century, disease spillovers and the close proximity of individuals induced a considerable “urban mortality penalty” that made urban growth unsustainable by natural increase; see, e.g., Birchenall (2007). It is perhaps obvious but the current scale and concentration of workers in cities would not be possible with the disease spillovers of the past. Overall, this section suggests large efficiency gains associated with the reduction of disease spillovers and health human capital investments.

### 4 Historical Perspectives

This section offers some brief historical perspectives to establish the plausibility of the model. The goal of this section is to catalog social responses to disease aligned with the

---

12 It is also important for the current differences between poor and rich countries. For example, the 85th/15th percentile income ratio across countries today is 16.9 and the 75th/25th ratio is 6.3; see, e.g., Jones (2014, Table 1).
paper’s thesis. A detailed study of the role of segregation by disease in the production and exchange accounts noted here requires at a minimum a series of dedicated papers.

Caste. The canonical example of social segregation is the caste system. Caste organizes society into groups according to principles of separation in terms of direct and indirect contact, specialization and division of labor, and social hierarchy; see, e.g., Dumont (1970, p. 21).

The best-known example of caste is the Indian caste system. The Indian caste system segregates individuals using the opposing concepts of pollution and purity derived from Hindu principles; see, e.g., Bayly (1999), Dumont (1970), Hutton (1969), and Singh (2005) for some general and diverse perspectives of the caste system in India. Very broadly, caste in India separates individuals into mutually exclusive hierarchical groups: the ‘forward caste’ that includes Brahmins (traditionally priest and teachers), Kshatriyas (traditionally warriors and rulers), and Vaisyas (traditionally money lenders and traders); and the Shudras (traditionally servants) which constitute the ‘other backward castes.’ These groups are above the ‘pollution line.’ The fifth group, the dalits, is regarded as being outside the caste system (i.e., outcasted) and below the ‘pollution line’ as they are “untouchable” because their physical proximity is polluting.

Pollution and purity are prevalent ideas in historical narratives; see, e.g., Douglas (1966). Across societies in the past, as Dumont (1970, p. 49) notes, “dangerous contact acts directly

---

13 Many social institutions are based on some of these principles, but strict separation is the defining feature of the caste system. For example, specialization and division of labor guided the Medieval guild system in Europe and the occupational code of the Western Roman Empire, i.e., the Theodosian Code which mandated children to follow their father’s occupation. Many other examples of hereditary functionalism that do not constitute caste are discussed by Hutton (1969, Chapter IX). This category also includes the Japanese ie discussed by Prescott and Hayashi (2008). Racial relationships between whites and blacks feature a social hierarchy but they are not part of a caste system. As discussed by Hutton (1969, P. 136), “there is strong prejudice on the part of the whites […] but the question of taboo and pollution by touch hardly arises. A negro servant to a white man is no strange anomaly, but a Brahman with a Chandal [who deals with disposal of corpses] cook is unthinkable.”

14 There are many examples of outcaste groups outside of India; see Hutton (1969, Chapter IX) and the discussion of unranked systems in Horowitz (2000, Chapter 1). In Burma, under the old rule, outcaste classes included beggars and vagabonds, lepers and others suffering from disease, as well as the sandala in charge of grave-digging and the disposal of corpses. (As Hutton (1969, p. 145) notes, sandala “is clearly the same word as the Hindi chandal.”) Outcastes are also prevalent in sub-Saharan Africa for individuals that perform occupations that are regarded as unclean (i.e., leather-working). Outcastes in Japan (i.e., the Burakumin) were in charge of animal slaughter and removing sewage. The Philippines and Sri Lanka also featured unclean outcastes.
on the person involved, affecting his health.” The view of pollution in the caste system, however, is more extreme than in traditional societies and tribes. Impurity is often temporary as in the case of seclusions after births or burials, or during menstruation; see Dumont (1970, p. 50). The ‘practice of untouchability’ in the Indian caste system is permanent to the point of even being hereditary. Moreover, the separation underlying the ‘practice of untouchability’ restricts direct physical contact but also indirect contact including sharing and preparing meals as well as entering the same physical space (i.e., kitchen or temple). Ambedkar’s (1936) famous *Annihilation of Caste* goes further to describe the untouchables as those “whose presence, whose touch, whose very shadow is considered to be polluting by privileged-caste Hindus.”

**Discussion.**— Although caste is fluid and serves many social, economic, and political purposes, the model interprets the regulation of direct and indirect physical contact as a disease-avoidance mechanism to enforce social segregation by disease. The social barriers involved in the ‘practice of untouchability,’ for example, center on the physical separation for workers engaged in “dirty tasks” or unhealthy production sectors associated with “taboo infections” such as animal skinning, leather-work, scavenging, and removing sewage and corpses; see, e.g., Hutton (1969, p. 89).

McNeill (1976) argues that the emergence of the caste system, and the hierarchical relationship between castes, originated as a social response to disease spillovers in populations with differential prior disease exposure. Accounts of the Indian caste system date back to the Vedic period somewhere around 1500-1000 BC. According to McNeill (1976, pp. 81-84), the caste system developed when “intrusive Aryans” from central Asia encountered “small, self-contained communities of forest-dwelling peoples.” He views this encounter as an “epidemiological standoff” between a population unfamiliar with the disease environment of the Indus valley and ‘forest folk’ who experienced a variety of tropical diseases and parasitic infections. Such population, familiar with the disease environment, “had acquired tolerances for formidable local infections.” McNeill (1976, pp. 66-67) argues that “the taboos on

---

15 The Indian caste system has adapted to historical circumstances during the Mughal era, British colonial rule, and independence; see, e.g., Bayly (1999). There are many views on the origin and persistence of the Indian caste system. For example, Freitas (2006) argues that the caste system was a social mechanism to share information, enforce contracts, and facilitate trade. Freitas (2006) also discusses alternative hypotheses from anthropology and sociology.
personal contact across caste lines, and the elaborate rules for bodily purification in case of inadvertent infringement of such taboos, suggests the importance fear of disease probably had in defining a safe distance between the various social groups that became the castes of historic Indian society.” The end result was that “the homogenizing process fell short of the ‘digestive’ pattern characteristic of the other Old World civilizations.” The hierarchies in the caste system also align with an increased distance between higher castes and outcaste populations with higher exposure to disease.

Disease and the avoidance of polluting production tasks as contributing factors in the caste system were also noted by Singh (2005, p. 36) since “[m]ost of the communities that were in India before the arrival of the Aryans were integrated in the Sudra varna or were made outcastes depending on the professions of these communities. Communities who professed non-polluting jobs were integrated into Sudra varna. And communities who professed polluting professions were made outcastes. The Brahmans are very strict about cleanliness. In the past, people believed that diseases can also spread through the air and not only through physical touch. Perhaps because of this reason the untouchables were not only disallowed to touch the high caste communities but they also had to stand at a certain distance from the high castes.”

**Quarantines.** Caste regulates social and physical distance for ‘polluting’ production tasks based on social norms. As in the lifetime isolation of lepers, discussed in the *Old Testament*, the enforcement of the physical separation between diseased and healthy individuals in the Indian caste system is based on social consensus and not legal sanctions. (Legal sanctions nowadays operate against the Indian caste system.) Quarantines are the earliest example of segregation by disease enforced by the authority of the state, i.e., by legal barriers.

Quarantines began as a strategy by Mediterranean city-state governments to minimize the risk of infection during the Black Death. The bubonic plague’s arrival to Europe in the 14th century is among the best documented (and deadliest) instances of “what can happen when an unfamiliar infection attacks a population for the first time;” see, e.g., McNeill (1976, p. 7). Plague was introduced to Europe by merchants from central Asia and it spread along trade routes (Biraben, 1975). Coastal cities and towns in medieval
Europe institutionalized quarantine regulations for travelers and shipping. Quarantines prevented individuals presumed to be diseased (and their cargo) from entering cities and ports. According to McNeill (1976, p. 151), the idea of quarantines “stemmed from biblical passages prescribing the ostracism of lepers; and by treating plague sufferers as though they were temporary lepers.”

**Discussion.**— The model interprets quarantines as strong restrictions on any form of social interactions between communities. Quarantines were preventive measures to restrict factor mobility and exchange between individuals inside and outside ports and cities. Quarantines separated individuals who had the potential for transmitting plague including asymptomatic individuals who may or may not have been exposed to disease. (Semantically, isolation represents the physical separation between symptomatic individuals or diseased individuals.) Trade and factor mobility restrictions were often supplemented with public health practices (i.e., bills of health that certified that the last port visited by a traveler was disease-free), penalties against infringement, and by the ‘persecution’ and isolation of plague victims in pesthouses. Quarantines and isolation measures were eventually embedded in a legal structure and enforced by centralized authorities with enough coercive power.

Segregation by disease was based on the mistaken hypothesis of the existence of a ‘miasma’ or poison cloud emerging from rotten corpses or rotting matter in the earth. Physical distancing, however, did minimize direct contact between healthy and diseased individuals over a sufficiently long period. Forty days eventually became standard in Italian cities, although some ports considered shorter durations (i.e., Dubrovnik in Croatia first implemented a thirty day isolation or ‘trentina’ in 1377). Although there was no understanding of disease transmission, “[i]n many cases such precautions must have checked the spread of plague. [...] Quarantine rules were therefore well founded” according to McNeill (1976, p. 151). Quarantines were not universally adopted in medieval Europe, but they became “general in Christian ports of the Mediterranean.” Quarantines and the permanent *cordon sanitaire* between the Habsburg and Ottoman empires established by Austria in the

---

16 The plague also served to stigmatize and persecute “minority groups” that were “easily identifiable, already unpopular, widely scattered and lacking any powerful protector.” Examples include pilgrims and lepers all over Europe, and Jews and Arabs in Germany and Spain; see Ziegler (1969, p. 97). Quite often, as these examples illustrate, quarantine principles took a class, ethnicity, or race character.
mid-eighteenth century are often credited for the disappearance of the plague from Western Europe; see, e.g., Biraben (1975). Plague remained a recurrent event in the Ottoman Empire centuries after its last visitation to European ports and cities.

Prior to the advent of the germ theory, segregation by disease applied to many additional diseases such as tuberculosis and Hansen’s disease. The segregation of infected individuals into dedicated hospitals (i.e., sanatoria and preventoria for tuberculosis, and leprosaria for Hansen’s disease) followed the principles of physical separation of quarantines but, due to the chronic nature of these diseases, the separation was permanent. In Birchenall (2021), I present a more detailed discussion of the “better safe than sorry” nature of quarantines, and their use to counter Yellow Fever and plague pandemics in the West.

Zoning. Quarantines are not the only form of segregation by disease enforced by legal barriers. During the age of miasmas, public health officials considered ways to counter disease spillovers within cities. As in the response to the Black Death, early public health approaches relied on a notion of contagion that “misrepresented the root cause of disease.” Eventually, with the advent of the germ theory of disease, “filth was recognized as the medium for transmitting disease instead of the primary source of contagion.” The regulation of urban spaces collectively known as zoning was one of the central public health strategies to counter the harmful effects of disease within cities; see, e.g., Melosi (2008, pp. 41-43) and Pinter-Wollman et al. (2018).

Discussion.— Zoning is defined as the spatial segregation of residential, commercial, and industrial land-uses. The model does not consider the optimal distribution and use of land within a city, but as in the segregation of individuals by disease, zoning permanently separates ‘dirty’ and ‘clean’ activities by restricting the location of economic activity.

Cities feature a number of positive and negative agglomeration externalities due to the close proximity between individuals, and between individuals and firms. Since antiquity, and until the early twentieth century, cities had much higher infectious disease loads than rural areas. Mortality rates before the age of 15 in industrializing European cities, for example, exceeded mortality in rural areas by at least a factor of 6 to 1; see Birchenall (2007). As noted by Pinter-Wollman et al. (2018, p. 3) “London, Paris, New York City and Chicago were densely populated and characterized by residences in proximity to factories,
animal yards, slaughterhouses and crowded tenement houses with little airflow or light.” Rapid and unchecked urbanization during the nineteenth and early twentieth century produced greater potential for disease spillovers and facilitated the spread of communicable diseases (Birchenall, 2007). The “urban mortality penalty” was reversed only after major urban infrastructure projects, such as sewers and water sanitation, and widespread land-use regulation.

Modern zoning principles have a clear public health basis with precedents on public nuisance laws that prohibit one’s property use to harm neighbors or the neighborhood. As with other forms of segregation by disease, zoning limited or eliminated physical contact between individuals and unhealthy land-uses in production. The most common form of separation is between residential, commercial, and industrial uses. In New York City, for example, regulation included the prohibition of industries such as tanneries, tallow makers, distilleries, and slaughterhouses in heavily populated areas as early as 1664. The regulation of skyscrapers also considered the need for sunlight and ventilation in offices to prevent the spread of disease. Overall, zoning served to create distance between residential areas and polluting firms, noxious fumes, and animal waste as a prophylactic measure; see, e.g., Schilling and Linton (2005, p. 98).

‘Silent trade.’ Zoning focuses on the legal physical separation between residential and industrial units. Disease-avoidance often involved the physical separation between healthy and diseased individuals. An early form of decentralized exchange, often labeled silent trade, takes place with no direct physical contact between trading partners and with a total absence of social interactions.

The earliest account of silent trade protocols comes from Herodotus of Halicarnassus, around 500 BC. As reported by Price (1980, p. 75), he described Carthaginian traders in West Africa:17 “The Carthaginians also tell us that they trade with a race of men who live in a part of Libya beyond the Pillars of Heracles. On reaching this country, they unload...”

---

17Grierson (1903), Sundström (1974), Kurimoto (1980), and Price (1967, 1980) discuss additional examples of similar trade protocols in monetary and barter economies. This includes accounts by Ibn Battuta in Northern Siberia in the 1300s, by Cosmas of Alexandria in present-day Ethiopia in the 500s, and many other narratives in the Old World. Batra and Birchenall (2019) provide a more detailed survey of existing narratives and a discussion of the different views on silent trade. Several remarks listed here are taken from that source.
their goods, arrange them tidily along the beach, and then, returning to their boats, raise a smoke. Seeing the smoke, the natives come down to the beach, place on the ground a certain quantity of gold in exchange for the goods, and go off again to a distance. The Carthaginians then come shore and take a look at the gold; and if they think it represents a fair price for their wares, they collect it and go away; if, on the other hand, it seems too little, they go back on board and wait, and the natives come and add to the gold until they are satisfied. There is perfect honesty on both sides; the Carthaginians never touch the gold until it equals in value what they have offered for sale, and the natives never touch the goods until the gold has been taken away.”

Discussion.– According to the model’s predictions, exchange based on iterative transactions that feature no direct physical contact between traders is likely the result of severe disease spillovers. (Language barriers and general concerns for safety are also seen as reasons for the silent trade.) Woodburn (2016, p. 473), in a recent overview and reassessment of the silent trade, stressed “the entirely rational desire of isolated hunter-gatherers and others to keep their groupings healthy and free from deadly diseases by systematically avoiding direct contact with outsiders. Fear of illness could have been a significant factor favoring such avoidance when trading, exchanging, or sharing goods with outsiders.”

Some of the best documented instances of decentralized exchange took place in sub-Saharan Africa. It is known that tropical parasitic and infectious diseases have been heavily concentrated in sub-Saharan Africa. In pre-colonial Africa, as Hartwig and Patterson (1978, pp. 6-7) note, exchange tended to be conducted at the border of territories rather than at centralized markets. In sub-Saharan Africa, “some long-distance trade involved relatively little human movement because specific communities prohibited transit traffic. Instead, these communities acquired goods on one frontier and carried what they did not want to keep across their own territory to exchange with yet another neighboring community. Such practices reduced the possibility that contagious diseases would spread over large areas.” Moreover, “[b]efore the European arrival in the interior of Central Africa, trade

---

18 Diamond and Panosian’s (2006, pp. 32-33), in their discussion of the geographic origins of tropical disease, note that “even the most significant diseases which originated in the New World tropics, Chagas’ disease and leishmaniasis (the latter also arose in the Old World tropics), have much less human impact than any of the six leading diseases of the Old World tropics (yellow fever, falciparum malaria, vivax malaria, cholera, dengue fever, and East African sleeping sickness).”
was localized and organized in such a way that one ethnic group transported its goods to the limits of its district, and the next group did the same;” see Azevado (1978, p. 127). Decentralized exchange was also common in the trans-Saharan trade networks as “aliens encountered severe health problems in West Africa, and merchants tended to turn back at the desert’s edge;” see Curtin et al. (1995, pp. 93-94). Alsan (2015) contains a detailed study of the role of disease spillovers emerging from the TseTse fly in political centralization and urbanization in sub-Saharan Africa.

During colonial times, integration and long-distance trade (i.e., the use of rivers) was particularly costly in Central Africa. As Azevado (1978, p. 119) notes, indigenous epidemic diseases including sleeping sickness (trypanosomiasis) and malaria “accompanied the breakdown of barriers between previously isolated peoples.” As Hartwig and Patterson (1978, p. 11) note, “[e]fficient movement of people and goods was essential for administrative control and economic exploitation of the newly acquired territories. Internal trade barriers were eliminated and banditry suppressed. Road, railways, and harbors were built and greater use was made of navigable waterways. [...] Clearly, greater intercommunication meant a more rapid spread of disease.” Many additional accounts of the relationship between increased intercommunication and epidemic diseases in early colonial Africa are discussed in Brown (1978). It has been long been understood, for example, that railways and transport routes accelerated and facilitated the spread of disease in Africa.

**Altitude.** The ‘silent trade’ protocol counters contagion spillovers with geographic distancing between trading partners. Establishing settlements at high altitudes is another form of geographic distancing to counter tropical diseases. In Latin America, high altitude cities have been continuously used as administrative, religious, and cultural centers (e.g., as capital cities) since pre-colonial times. Tenochtitlán (currently Mexico city) and Cuzco, the capital cities of the Aztecs and Incas, and the most densely populated cities in pre-

---

19 Azevado (1978) studied the Sara of Southern Chad. Hodder (1977) provides additional accounts of how sub-Saharan African markets worked “before the colonial period.” For example, Hodder (1977, pp. 256-257) notes that “there is in Yorubaland a remarkable lack of correspondence between the location of traditional periodic markets on the one hand and the location and hierarchy of settlements in the other.” Hodder (1977, p. 258) writes “a few isolated traditional markets may often be found around the periphery of the tribal lands where inter-tribal trade could take place: a few markets, for instance, seem for long to have existed along the Ubangi river where it forms the boundary between the Ngbandi and Banda peoples. Similar peripheral markets are to be found along the borders of the Ruandi and Urundi groups.”
Columbian America, were established at high altitudes: 2,420 and 3,399 meters above sea level, respectively. Many other pre-Columbian cities were also located at high altitudes and remained in use during colonization and independence. As Hardoy (1993, p. 100) pointed out, “[m]any Latin American capitals were founded in what we could define as areas of comparatively dense indigenous settlements. These include Guatemala city, San Salvador, Bogotá, Caracas, Quito, Lima, Asunción, and Santiago de Chile.” The vast majority of these pre-modern settlements were established in the interior of each territory and at relatively high altitudes; e.g., Guatemala city (1,500 m.), San Salvador (1,893 m.), Bogotá (2,625 m.), Quito (2,800 m.), La Paz (3,650 m.), and Tegucigalpa (980 m.).

The administrative centers of African empires in pre-colonial times were of a smaller scale compared to the empires of the New World. “[S]ome African monarchies deliberately avoided urbanization. [...] In early Kanem and Mali, capitals were even more mobile and existed wherever the king made his residence at a particular time;” see Hull (1976, pp. 2-7). Nonetheless, the majority of the Ethiopian population still resides in the northern highlands in the interior of the country. Addis Ababa (2,355 m.), Nairobi (1,795 m.), Maseru (1,673 m.), Kigali (1,567 m.), Gitega (1,504 m.) and Harare (1,483 m.) are among Africa’s high elevation capital cities.

Discussion.—According to the theory, the logic behind high altitude settlements is simple. At higher altitudes, disease environments are less hostile since several tropical disease vectors and bacteria are sensitive to altitude and UV radiation. As Coatsworth (2008, p. 556) notes, “[t]he concentration of pre-Columbian populations in the highlands occurred because the disease environment was less deadly and the soils richer in minerals, both of which helped cities grow larger.” A detailed account of how the expansion of the Inca empire to the Peruvian lowlands was regulated to counter the incidence of disease is available in Monge (1948). The Incas were well aware that Indians “do very badly away from the region to which they are native [...]. Those of hot regions die when they go to cold regions, and if they are out in the cold country they die on going to the hot.” Strict mobility restric-

---

20 Only two countries are landlocked in Latin America (Bolivia and Paraguay) but there were “very few ports at the time of the Spanish conquest; these were mostly located in the Yucatan Peninsula [...] and on the Caribbean and Pacific coasts of Guatemala”; see Hardoy (1993, p. 100). Ports in the Caribbean, such as the ports of Habana and Veracruz, became central during colonial times.
tions prevented “upland Indians” from going to “coastal plains” and vice versa. In their military campaigns at lower altitudes, tropical exposure was limited since Inca soldiers were replaced “every two months since the coastal lands are unhealthy for the uplanders;” see Monge (1948, pp. 7-8).

During colonial times, European settlers in the Caribbean established ‘hill stations’ for military personnel as a way to avoid tropical diseases. The understanding of the association between elevation and disease was imperfect, as the cool weather of higher altitudes was seen as the reason for a protective effect against disease and putrefaction. European ports were particularly exposed to endemic malaria and periodic episodes of yellow fever. As noted by Curtin (1989, p. 62) “[t]he normal military response once the disease reached a Jamaica port, was flight.” Relocating soldiers to the mountains succeed in the Caribbean; “[t]he death rate for European troops on Jamaica dropped from 128 per thousand in 1817-1836 to 60.8 per thousand in 1837-46, a change credited to the move [to a hill station], no doubt correctly;” see Curtin (1989, p. 49). Hill stations were not successful in Algeria and India due to the misunderstood nature of malaria; see Curtin (1989, pp. 47-50).

Tropical diseases were also central, according to McNeill (1974, p. 76), in explaining the “slow pace of Chinese settlement in more southerly parts of what is today China.” Climate and disease environments are sharply different between Northern and Southern China. McNeill (1974, p. 77-78) claims that malaria and dengue fever “may have constituted the principal obstacle to early Chinese expansions southward.” Schafer (1985) contains a rich account based on administrative records and travel guides of the attitudes and interactions between the “Man,” aborigines of Nam-Viet, and the “Hua,” Chinese intruders during the T’ang Dynasty in the seventh century. As noted by Schafer (1985, Chapter 7), malaria and dengue fever played a prominent role as a deterrent in military intervention and settlements in tropical areas. A similar line of thought was developed by Acemoglu et al. (2001).

5 Conclusions

This paper examined the social organization of production and exchange in economies subject to disease spillovers. I sketched out a simple analytical framework in which aggregate
efficiency ranks according to basic concavity principles (i.e., Jensen’s inequality). I charac-
terized aggregate efficiency across centralized and decentralized production and exchange
protocols with differential exposure to disease. Quantitatively, disease spillovers have a
large effect on aggregate efficiency and output. In an accounting framework, a three-fold
difference in aggregate efficiency is significant. To validate the theory’s predictions, I also
catalogued how past societies, from very different parts of the world, tried to avoid contagion
through social segregation by disease in production and exchange.

Having discussed some theoretical and practical implications of social segregation by
disease, it may be useful to summarize some general principles in the theory and the per-
spectives offered by the historical examples. Following these general principles, I lay out
some caveats to the analysis presented here.

The previous instances are not individual but social responses to disease. The baseline
model and the historical perspectives highlight that production and exchange are social
activities. Contagion arises because individuals are not self-sufficient in dealing with ‘dirty’
production or exchange tasks. The social responses noted here are not about the demogra-
phy of the economy, i.e., the size of the different types of agents in the economy or mean
health human capital. Instead, social responses to disease target the distribution of the
different types in the population across space, broadly defined. Understanding and measur-
ing the aggregate economic effects of social segregation by disease requires a distributional
framework.

The social responses to disease are organized around the notion of distance. In the
past, as in the simple theory, physical distance provided the only protective barrier against
communicable diseases. Even with a very imperfect understanding of disease, distance took
many forms: permanent social quarantines in the caste system, temporary legal quarantine
orders and isolation in medieval Europe, and permanent land-use separation in industrial
cities. Disease and distancing also posed limits to the extent of the market and the location of
economic activity. In pre-colonial times, tropical diseases limited the geographic proximity
between economic centers of production, i.e., capital cities at high altitudes and trading
ports in coastal areas. Disease exposure even limited the proximity between economic
agents themselves, i.e., buyers and sellers in a ‘silent trade’ protocol.
The historical perspectives noted here are selected, but they matter for long-run economic development. As the theory makes clear, all production and exchange protocols are less efficient than technologically possible. Efficiency costs take many forms, but they are likely to introduce path dependencies. The social enforcement of caste-based rules, for instance, remains much alive in India today despite the many legal sanctions against ‘the practice of untouchability’ and discrimination. High altitude cities in Latin America similarly served to consolidate large tributary empires in pre-colonial times. The increased distance from coastal areas and the rugged terrain, however, have led to much higher transportation and integration costs in today’s globalized world. This is perhaps why, since colonial times, Latin American exports have concentrated in commodities with high value per-weight (i.e., gold, silver, tin, oil, sugar, flowers, and illegal drugs). Finally, difficulties in centralized state-building prevalent in pre-colonial times appear to still be important in sub-Saharan Africa today where severe, and often fatal, infectious diseases (i.e., Ebola) are recurrent phenomena.

The paper narrowly considered the efficiency of social arrangements in which individuals are sorted into their most advantageous use in production and exchange. I focused on production efficiency since, on a positivist take, efficient arrangements are more likely to survive than inefficient ones. The segregation of land-use functions, for instance, is the most prevalent land-use planning tool in modern cities and towns. Likewise, in the absence of vaccines, quarantines and isolation remain one of the most important and effective preventive public health measures against communicable disease outbreaks. The paper, however, is not a defense of social segregation or a justification for past social arrangements such as caste. In practice, the social organization of production often takes class, ethnicity, religion, gender, nationality, and race dimensions. There are countless examples where ‘pollution,’ not just in a physical or moral sense, has been used to justify a “social order” and subordination. Economics in general, and this paper in particular, provide no normative guidance as to how society should be ‘ordered.’

Even within the proper realm of social segregation by disease, many past segregationists arrangements are no longer desirable (if ever). Modern changes in technology are evident in the production process, but also in the knowledge available to counter disease. The way
we now define, explain, organize, and treat disease rests in more solid epistemic grounds than in the past. Still, outdated and outlawed social arrangements such as caste and the ‘practice of untouchability’ remain difficult to dismantle. The paper does not address these difficulties, but its lessons may be helpful for understanding the emergence of social impediments to long-run economic development.

References


6 Appendix: Some Extensions

Dynamic spillovers. The baseline model is not explicit about the progression and transmission of disease. One can even reformulate the baseline theory to consider attributes that trigger social exclusion and marginalization for reasons unrelated to disease contagion.
I next consider disease progression and transmission as in mathematical epidemiology, e.g., Anderson and May (1992). Let $\gamma$ and $1 - \gamma$ denote the current fraction of healthy but susceptible individuals and the fraction of infected individuals, respectively. There is a stationary population of size $m > 0$ of infected disease vectors (or nonhuman hosts such as mosquitoes). These vectors are uniformly distributed across all communities. Susceptible (i.e., healthy) individuals become infected through contact with a disease vector or infected individuals.

Contagion takes place randomly. The probability that a healthy individual meets a disease vector is $mn(e)$. The probability that such meeting leads to a disease transmission is represented by the vector transmission rate $\pi > 0$ (i.e., the biting rate of the vector). As before, the meeting probability between a healthy and a diseased individual is $2n(e)(1 - n(e))$, and the human transmission rate is $\tau > 0$, i.e., the probability that the susceptible individual becomes infected. Infected individuals recover at a rate $\sigma$, i.e., a fraction of $\sigma(1 - \gamma)$ individuals recover from the infection during the current period. The fraction of susceptible (i.e., healthy) individuals evolves over time as follows:

$$\gamma' - \gamma = \sigma (1 - \gamma) - \pi \int_0^1 mn(e)de - \tau \int_0^1 2n(e)(1 - n(e))de,$$

where the first term in the right-hand-side is the fraction of infected individuals that recover, and the second and third are the vector and human transmissions per capita. (Births and deaths are implicitly cancelled in (16).)

Let $0 < \beta < 1$ denote the discount factor. The social planner allocates healthy workers across communities to maximize the discounted present value of aggregate output,

$$v(\gamma) = \max \left\{ \int_0^1 q(n(e))de + \beta v(\gamma') \right\}, \text{ s.t., (1) and (16)}.\]

The following proposition characterizes the dynamic assignment:

**Proposition 7** Segregation is more socially desirable in a dynamic environment than in the benchmark static model. The desirability of segregation is independent of the vector transmission rate $\pi$, but it increases with the human transmission rate $\tau$. The (stable) stationary fraction of healthy individuals in the population is $\gamma_s = \sigma/(\sigma + \pi m)$.

**Proof.** As in the static problem, the second-order condition determines the assignment of individuals across communities. In addition to the static terms in the baseline model (4), the assignment problem now contains a Lagrange multiplier $\lambda'$ on the constraint (16). The second-order condition for $n(e)$ depends on

$$[f(h, h) - 2f(h, d) + f(d, d)] + \lambda' 2\tau.$$  

---

21 Canonical examples of diseases transmitted by physical contact include respiratory infections such as tuberculosis and pneumonia. Sexually transmitted diseases also feature direct transmission. Many parasitic diseases, notably malaria, are not contagious (spread by contact). These diseases, however, exhibit cross infection: a vector infects the human host and then the human host infects another vector.

22 The population of vectors is not modeled explicitly. There is a large literature that examines how disease evolves. See Anderson and May (1992) for an introduction to this literature.
The first term in (17) captures the static spillovers of disease (6) and the second the dynamic spillovers. This second term is positive since \( \lambda' > 0 \). The added constraint (16) thus increases the social planner’s incentives for segregation. The transmission rate \( \tau \) does not influence (17) because the vector transmission is uniform across locations. The human transmission rate \( \tau \), however, induce segregation in (17) even if \( q(n(e)) \) is linear, i.e., \( 1 - \alpha = \rho \). Under segregation, (16) becomes \( \gamma' - \gamma = \sigma(1 - \gamma) - \pi m \gamma \), which converges monotonically to \( \gamma_s \) from any \( \gamma_0 > 0 \).

Segregation is more socially desirable in a dynamic environment. The social planner is now concerned with maximizing aggregate production (Proposition 1) but also with increasing mean health over time. These two objectives are complementary in \( v(\gamma) \). The dynamic spillovers imply that segregation would be beneficial even if there are no static disease spillovers. In this sense, dynamic considerations weaken Assumption 2. The incentive to discourage meetings between healthy and diseased individuals is larger for easily transmittable diseases than for diseases with low transmission rates. For instance, if \( \tau = 0 \) in (16) or if there is segregation, disease evolves exogenously from the perspective of a social planner.

In Proposition 7, the fraction of healthy individuals is endogenous in a dynamic environment. Under segregation, (16) becomes \( \gamma_0 - \gamma = \sigma(1 - \gamma) - \pi \mu \gamma \), which converges monotonically to \( \gamma_\sigma = \sigma/(\sigma + \pi \mu) \). This value is increasing in the recovery rate \( \sigma \) and decreasing in the product of the vector transmission rate \( \pi \) and the prevalence of disease vectors \( \mu \). The disease vector here sustains an endemic steady state, with \( \gamma_s < 1 \).

**A Multi-Class Society.** The baseline model considers two types of health human capital. Social stratification typically takes place in hierarchies, so it is important to know the conditions that might lead to segregation in more general settings. This Appendix considers \( J \geq 2 \) types of individuals. The (positive) health human capital endowments of these different types are represented by a (column) vector \( h \equiv (h_1, ..., h_J) \). Health human capitals are ordered so that \( h_1 < h_2 < ... < h_J \). The given number of workers of type \( j \) is \( N_j \) and their fraction in the population is \( \gamma_j = N_j/N \), with

\[
\sum_{j=1}^{J} \gamma_j = 1.
\]

Individuals are endowed with physical capital as in the benchmark model, i.e., \( k_j = k = K/N \).

To keep the analysis comparable to the baseline model, production takes place in pairs. Pairwise production can be summarized by a \( J \times J \) matrix \( F \) defined such that its \((i, j)\) element is given by \( F(i, j) \equiv f(h_i, h_j) \). That is,

\[
F \equiv \begin{bmatrix}
f(h_1, h_1) & f(h_1, h_2) & \cdots & f(h_1, h_J) \\
f(h_2, h_1) & f(h_2, h_2) & \cdots & f(h_2, h_J) \\
\vdots & \vdots & \ddots & \vdots \\
f(h_J, h_1) & f(h_J, h_2) & \cdots & f(h_J, h_J)
\end{bmatrix}
\]

The assignment problem is the allocation of workers across communities. Let \( n_j(e) \) denote the fraction of individuals of type \( j \in \{1 : J\} \) to be assigned to community \( e \in E \).
The choice vector is \( n(e) \equiv (n_1(e), ..., n_J(e)) \), which is consistent with the baseline model. Since workers meet randomly, expected output in community \( e \) is given by a quadratic form \( 2q(n(e)) \equiv n(e)F_n(e) \). A social planner maximizing aggregate output per capita solves
\[
\max_{n(e)} \int_0^1 q(n(e))de, \text{ s.t., } \int_0^1 n_j(e)de = \gamma_j,
\]
and \( 0 \leq n_j(e) \leq 1, \) for \( j \in \{1: J\} \).

A generalization of Assumption 2 is sufficient to induce a multi-class society:

**Proposition 8** For a given \( J \geq 2 \) types of workers, assume that the parameters \( J, \rho \) and \( \alpha \) satisfy
\[
\left( \frac{h_i^\rho + h_j^\rho}{2} \right)^{1/\rho} < \left( \frac{1}{J-1} \right)^{1/(1-\alpha)} \left( \frac{h_i^{1-\alpha + h_j^{1-\alpha}}}{2} \right)^{1/(1-\alpha)},
\]
for all \( h_i \) and \( h_j \) pairs with \( i, j \in \{1: J\} \). Then it is (socially) optimal to segregate individuals into \( J \) classes with \( n_j(e) = 1 \) in \( \gamma_j \) communities, and \( n_j(e) = 0 \) otherwise.

**Proof.** The matrix \( F \) is symmetric due to the symmetry of the mean function \( \mu_\rho(h, d) \), as in the benchmark model. Since the entries \( F_{(i,j)} \) are real, the matrix \( F \) is also Hermitian. Therefore, if pairwise comparisons satisfy
\[
F_{(i,j)} < \frac{1}{J-1} \sqrt{F_{(i,j)}F_{(j,i)}},
\]
for all \( (i, j) \in \{1: J\} \), then \( F \) is positive definite; see Bernstein (2005, Fact 8.7.34). The standard inequality of arithmetic and geometric means, i.e., \( \sqrt{F_{(i,j)}F_{(j,i)}} \leq (F_{(i,i)} + F_{(j,j)})/2 \), imply that
\[
F_{(i,j)} < \frac{1}{J-1} \frac{F_{(i,i)} + F_{(j,j)}}{2},
\]
is a sufficient condition for \( F \) to be positive definite. Since \( F_{(i,i)} = zh_i^{1-\alpha}k^\alpha \), \( F_{(j,j)} = zh_j^{1-\alpha}k^\alpha \), and \( F_{(i,j)} = z|(h_i^\rho + h_j^\rho)/2|^{(1-\alpha)/\rho}k^\alpha \), some simple algebra shows that (19) can be written as (18). Since the feasibility constraints are linear, the second-order conditions associated with output maximization can be related solely to the properties of \( F \). If \( F \) is positive definite, the optimal allocation is a corner solution. \( \blacksquare \)

When \( J = 2 \), expression (18) holds as long as \( \rho < (1-\alpha) \), which is precisely Assumption 2. But Assumption 2 might not be sufficient for the general case, especially when the number of types \( J \) is large. There is no straightforward intuition to describe the general condition (18) other than to reiterate that the spillovers must be sufficiently negative to outweigh the benefits of integrating different types of individuals. Special cases such as \( \rho = -\infty \) do not help much with intuition.

Proposition 8 leads to generalizations of the baseline model. Under segregation, the aggregate production function is still Cobb-Douglas \( Y_s = A_sK^\alpha(HN)^{1-\alpha} \), and aggregate
efficiency is given by
\[ A_s \equiv z \left[ \sum_{j=1}^{J} \gamma_j \left( \frac{h_j}{H} \right)^{1-\alpha} \right], \quad \text{with} \quad H \equiv \frac{1}{N} \sum_{j=1}^{J} \gamma_j h_j. \] (20)

Expression (20) extends (8) to the case of \( J \) classes and it carries the same interpretation. Likewise, aggregate efficiency in a centralized market place with disease spillovers can be defined as in the benchmark model by
\[ A_c \equiv z \left[ \sum_{j=1}^{J} \gamma_j \left( \frac{h_j}{H} \right)^{\rho \frac{(1-\alpha)}{\rho}} \right], \]
where \( A_c < A_s < z \) due to Jensen’s inequality. If, once segregated, capital is exchanged in a central marketplace with no disease spillovers, as if a planner is solving
\[ z \max_{k_j} \left( \sum_{j=1}^{J} \gamma_j h_j^{1-\alpha} k_j^\alpha \right), \text{ s.t. }, \sum_{j=1}^{J} \gamma_j k_j = K, \]
one recovers an aggregate production function \( Y_n = A_n K^\alpha (HN)^{1-\alpha} \), with \( A_n = z \). These expressions generalize the two-class model of Section 2 and so they carry the same economic implications.

Proposition 8 retains the pairwise matching assumption but allows for different \( J \) types of workers. Society is segregated in \( J \) classes and there is a clear hierarchy as some types are more productive than others. A multi-class society is not especially interesting from the perspective of different degrees of contagion because there is no obvious disease hierarchy. Proposition 8, however, is interesting because there are typically multiple traits associated with a stratified society (i.e., occupation, race, gender, religion, and so on). Human capital is often viewed as a composite of multiple types, as in Weil (2007) where individuals differ in terms of education and health. If these multiple types are multiplicative, one can establish a hierarchy based on education, holding health constant, and another based on health, holding education constant. Hierarchies can be established more generally using the intersection of many relevant social attributes.

7 Appendix [Not for Publication]

The paper focused on aggregate efficiency and the social organization of production and exchange. Normative comments on social segregation, even on the realm of disease, require controversial welfare considerations possibly involving the use of preferences that should not be validated. In this Appendix I consider a very narrow welfare evaluation of social segregation by disease. Essentially, I show that segregation by disease might improve welfare if disease spillovers are sufficiently negative relative to the degree of inequality that is socially tolerated.

Consider the static assignment but suppose that production is equally consumed within each pair of workers. Let the \textit{ex post} utility of a meeting between healthy individuals in the community be \( \phi(h, h) = (1 - \epsilon)^{-1} [f(h, h)]^{1-\epsilon} \). Likewise, define \( \phi(h, d) \) and \( \phi(d, d) \) as the \textit{ex
post utility of \((h,d)\) and \((d,d)\) meetings, respectively. Let \(u(n(e))\) denote community \(e\)'s utility, i.e.,

\[
u(n(e)) = \{n(e)^2 \phi(h, h) + 2n(e)(1 - n(e))\phi(h, d) + (1 - n(e))^2 \phi(d, d)\},
\]

where the economy’s social welfare function is the CES aggregate of expected utilities across communities, i.e.,

\[
U(u) = \frac{1}{1 - \varphi} \int_0^1 u(n(e))^{1 - \varphi} de.
\]  

(21)

The welfare maximization problem now involves maximizing \(U(u)\) s.t. (1) and \(0 \leq n(e) \leq 1\).

The assignment of workers across communities is more involved than in Section 2 because \(\epsilon\) and \(\varphi\) introduce concavity at the community level (i.e., in \(\phi\) relative to \(f\)) and at the societal level (i.e., in \(U\) relative to \(y\)). Inequality matters in two separate ways. First, \(\epsilon \in [0, 1]\) captures social concerns for within-community inequality. As \(\epsilon\) increases, society gains from observing healthy and diseased workers paired together in any community. Second, \(\varphi \in [0, 1]\) captures social concerns for between-community inequality. As \(\varphi\) increases, society gains from observing more homogeneous communities.

The values \(\epsilon = \varphi = 0\) correspond to the assignment problem studied in the text. The presence of \(\epsilon > 0\) and \(\varphi > 0\) reduce the incentives of social segregation, so Assumption 2 must be strengthened for social segregation by disease to be welfare enhancing. The next proposition provides two illustrations of the conditions in production and the social welfare functions that make segregation desirable from a welfare perspective.

**Proposition 9** Suppose that \(\varphi = 0\) and amend Assumption 2 as follows \(\rho < (1 - \alpha)(1 - \epsilon)\). Alternatively, assume that \(\rho \to -\infty\) and that \(0 < \varphi \leq 1/2\). In either case, segregation by disease maximizes social welfare (21).

**Proof.** The proof of the first case is standard because the objective function (21) has the same form as (7). Thus, as in (6) and Proposition 1, Proposition 9 can be proven using Jensen’s inequality, i.e., by showing that the previous joint restriction on \((\rho, \alpha, \epsilon)\) implies that community \(e\)'s expected utility is convex in \(n(e)\), i.e., \(\phi(h, h) + \phi(d, d) > 2\phi(h, d)\).

Consider next the second case. When \(\rho \to -\infty\), \(u(n(e)) = n(e)^2 \phi(h, h) + (1 - n(e))^2 \phi(d, d)\). It is straightforward to show that \(\varphi \leq 1/2\) is sufficient for \(U(u)\) to be a strictly convex function of \(n(e)\). The first derivative of \(U(u)\) with respect to \(n(e)\) is \([n(e)^2 \phi(h, h) + (1 - n(e)^2)\phi(d, d)]^{-\varphi}\{2n(e)\phi(h, h) - \phi(d, d)\}\). After some manipulations, the second derivative can be written as

\[
\frac{\partial^2 U(u(n(e)))}{\partial n(e)^2} = \frac{2\phi(h, h) - \phi(d, d)}{[n(e)^2 \phi(h, h) + (1 - n(e)^2)\phi(d, d)]^{1 + \varphi}}.
\]

The only source of ambiguity lies in \((1 - 2\varphi)\). Under \(0 < \varphi \leq 1/2\), \(\partial^2 U(u(n(e)))/\partial n(e)^2 > 0\) leading to a convex objective function and a corner solution as the one in the text. 

The previous propositions shows that a strengthening of Assumption 2 is sufficient for social segregation to maximize social welfare. In the first case, when \(\varphi = 0\), society is indifferent to the inequality between communities. If the social concerns for within-community
inequality are small, i.e., \( \epsilon < 1 - \rho/(1 - \alpha) \), then segregation becomes desirable. In the second case, when \( \varphi > 0 \), the social welfare function is a strictly concave function of a community’s expected utility \( u(n(e)) \). Concavity implies a preference for social integration. Integration when \( \rho \to -\infty \) and \( \varphi \leq 1/2 \), however, has such adverse effects in a community’s output that society is better off trading off output and equality.