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

Javier A. Birchenall∗
University of California at Santa Barbara

January 20, 2021

Abstract

The thesis of this paper is that disease plays a fundamental role in long-run economic development through the social organization of production and exchange. I study an assignment problem in which disease spillovers induce social segregation and limit the extent of the market. I characterize the productivity effects of disease across production and exchange arrangements with a differential scope for disease spillovers. Eliminating disease spillovers increases aggregate productive efficiency by roughly a factor of 3. Historical accounts describe many instances in which past societies dealt with disease through social segregation in production and exchange. Some of these instances are catalogued here.

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, and UC Santa Barbara Symposium on Inequality for their many comments. I have, over the years, also benefited from comments and remarks by Hoyt Bleakley, Robert Fogel, Pete Klenow, Shelly Lundberg, 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 influence of disease on aggregate productivity and output. From an individual perspective, disease is a biological process. Economists know that 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, [2014]), Bleakley (2007, 2010), and Weil (2007). Disease, however, rarely affects individuals in isolation. Throughout history, endemic and epidemic diseases have typically acted on large groups of individuals. The goal of this paper is to characterize and quantify the influence of disease spillovers on the social organization of production and exchange.

I examine a highly stylized assignment model. There are two types of individuals, healthy and diseased. Production and exchange are social activities subject to disease spillovers (i.e., contagion). Individuals meet randomly and production and exchange take place in pairwise meetings. The main economic problem is that of sorting individuals across a continuum of communities. Communities are then consolidated into an aggregate economy summarized by an aggregate production function. When disease spillovers are sufficiently negative, relative to the diminishing returns to labor, it is (socially) efficient to segregate healthy and diseased individuals across communities (Proposition 1). Social segregation matters for a country’s output because aggregate efficiency is a function of the technology for production and the sorting of workers across communities. In the model, for example, aggregating across production units that face disease spillovers reduces measured total factor productivity relative to the technological possibilities in the economy (Proposition 2). Since disease acts as a non-technological influence in total factor productivity, I call this a productivity effect of disease and characterize it across several production and exchange arrangements that differ in their degree of centralization and disease exposure (Propositions 3 to 6).

The social organization of production and exchange is not usually on the radar of economists. Economists have been typically concerned with the biological and demographic effects
of disease. Since healthier individuals are better workers, the biological response to disease represents a *proximate effect* of health on a country’s output; see, e.g., Fogel (1994) and Weil (2007). In a development accounting framework, Weil (2007) measured the significance of health human capital in explaining income differences between rich and poor countries. Weil’s (2007) measurement strategy treats health human capital in efficiency units of labor. Efficiency units, however, are inconsistent with disease spillovers. Under efficiency units, the sorting of workers is irrelevant since only mean (or total) health human matters for aggregate output. I build on Fogel (1994) and Weil’s (2007) measurement strategy and calibrate the disease spillovers to be consistent with the estimated treatment externalities in deworming interventions in Miguel and Kremer (2004, [2014]) and Bleakley (2010). The effect of disease spillovers on a country’s aggregate output are potentially large. Keeping mean health human capital constant, measured aggregate TFP in economies with centralized markets without disease spillovers are higher than in centralized markets with spillovers by roughly a factor of 3. These productivity gains are at least as large as the proximate health benefits measured by Fogel (1994) and Weil (2007).

In contrast to economists, historians have long recognized the importance of disease environments in shaping social and economic institutions; see, e.g., McNeill (1976). To illustrate the indirect effects of disease environments, I catalogue how some past societies sought to avoid contagion through social segregation and decentralized exchange. These historical accounts include prominent examples such as the caste system in India in which direct and indirect physical contact was socially regulated as a disease-avoidance mechanism; legal quarantines and isolation in medieval Europe where trade was temporarily restricted to protect coastal cities from plague epidemics; peripheral markets in pre-colonial Africa in

---

1Demographic 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., Acemoğlu 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), Miguel and Kremer (2004), 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.

2Briefly 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 discussions on the ‘disruptive’ role of disease.
which trade was highly decentralized to the point of featuring no social interactions between traders (i.e., trade was ‘silent’); and the emergence of high altitude cities in pre-Columbian empires in the New World which faced less hostile environments than the disease-prone coastal areas. These historical perspectives are consistent with the productivity effect of disease sketched by the model as their scope is beyond that of an individual (i.e., biological) response. The social organization of production and exchange in the past is also still relevant in today’s world. As I remark below, the social influence of disease on production and exchange is often path-dependent and difficult to change.

Some related literature. 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 is known that the most significant tropical diseases in history have arisen overwhelmingly in sub-Saharan Africa with characteristics that largely differ from tropical diseases elsewhere; see Diamond and Panosian (2006). It has also 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. The purpose of this paper is not to bring new evidence in favor or against the role of disease in African development, but to develop a plausible theoretical framework to interpret some evidence already presented and to offer parallels with other historical instances where disease spillovers and “social barriers” also appear to have limited sociality and the extent of the market.3

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

---

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 are endogenous, causal estimates are more difficult to obtain.
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 human capital externalities in education.

In a related and 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 very 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. In the present model, disease acts as non-technological influence in productivity. In here, capital and labor markets tackle disease spillovers through a “social barrier.” This “social barrier” is endogenous thus the absence of factor price equalization across individuals is not indicative of misallocation, as it is in Hsieh et al. (2019) and Hayashi and Prescott (2008), for example.

---

4 Fogli 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 The Model Economy

This section studies the emergence of segregation and stratification in a two-class society. The goal of the model is to expose channels, many not tacitly recognized in economics, through which disease determines the social organization of production and exchange. The starting point of the model is that production and exchange are social activities subject to contagion spillovers. Disease spillovers differ according to the organization of production and exchange. The different organizational arrangements, and their effects on aggregate production efficiency, are the focus of this section.

**Individuals and communities.** There are \( N \) individuals and a continuum of communities (e.g., ethnicities, religions, linguistic groups, or other social divisions) of measure 1. 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. Aggregate capital is \( K > 0 \). 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., \( h > d \geq 0 \).

There is a given number of healthy individuals in the economy, \( \Gamma \equiv \gamma N \). The number of diseased individuals is \( N - \Gamma = (1 - \gamma)N \) with \( \gamma \in (0, 1) \). The disease prevalence rate is therefore \( 1 - \gamma \). The economic problem is the assignment of individuals across communities. The number of healthy individuals to be assigned to community \( e \) is \( n(e)N \), where \( n(e) \in [0, 1] \) is the fraction of healthy individuals in the community. The assignment’s feasibility condition is

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

Within each community, production takes place in pairs of *randomly matched* individuals. Individuals cannot produce in isolation. The desirability of pairwise meetings is discussed below (Proposition 3). When two individuals meet, they pool their physical capital. By construction, the pair’s pooled capital is the (arithmetic) mean of the two individual’s capital, i.e., \( k = (k + k)/2 \). A pair of randomly matched individuals may have different
health human capitals. The health human capital of a matched pair is also pooled, but not necessarily linearly. The way health human capital is averaged depends on disease spillovers. I assume that when individuals of type $\eta$ and $\delta$ meet, the pair’s health human capital is

$$
\mu_\rho(h, d) = [(h^\rho + d^\rho)/2]^{1/\rho},
$$

(2)

where the aggregator $\mu_\rho(h, d)$ is: linearly homogeneous, $\mu_\rho(\theta h, \theta d) = \theta \mu_\rho(h, d)$ for all $\theta > 0$; symmetric, $\mu_\rho(h, d) = \mu_\rho(d, h)$; and representative, $\mu_\rho(h, h) = h$ and $\mu_\rho(d, d) = d$ for all $\rho$.

The pairwise human capital aggregator $\mu_\rho(h, d)$ is increasing in $\rho$ and yields a variety of means as $\rho$ varies. For example, (2) satisfies: $\mu_{-\infty}(h, d) = \min\{h, d\} = d$; $\mu_{-1}(h, d) = 2(hd)/(h + d)$, e.g., the harmonic mean of $h$ and $d$; $\mu_0(h, d) = \sqrt{hd}$, e.g., the geometric mean of $h$ and $d$; $\mu_1(h, d) = (h + d)/2$, e.g., the arithmetic mean of $h$ and $d$; and $\mu_{\infty}(h, d) = \max\{h, d\} = h$. As will be evident below (Proposition 1), the parameter $\rho$ is central to determine the social organization of production.

Production takes place with a Cobb-Douglas production function. The level of technology is given and denoted by $z > 0$. When healthy individuals meet they produce $f(h, h) = zh^{1-\alpha}k^\alpha$ with $\alpha \in (0, 1)$; when diseased individuals meet they produce $f(d, d) = zd^{1-\alpha}k^\alpha$; and when a healthy and a diseased individual meet they produce $f(h, d) = z[\mu_\rho(h, d)]^{1-\alpha}k^\alpha$ or

$$
 f(h, d) = z[(h^\rho + d^\rho)/2]^{(1-\alpha)/\rho}k^\alpha.
$$

(3)

The probability that a healthy individual meets another healthy individual is proportional to the fraction of healthy individuals in the community, $n(e)$. Thus, the meeting probability between two healthy individuals is $n(e)^2$. Likewise, the meeting probability between two diseased individuals is $(1 - n(e))^2$. The meeting probability between healthy and diseased individuals is $2n(e)(1 - n(e))$, since there are two (symmetric) ways such meetings could occur. Obviously, $n(e)^2 + 2n(e)(1 - n(e)) + (1 - n(e))^2 = 1$.

Community $e$’s expected output per capita, denoted by $q(n(e))$, is

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

(4)
Sorting and social segregation. Let \( n^*(e) \) denote the optimal value of \( n(e) \). There are only two ways to assign individuals across communities: healthy and diseased individuals should be either integrated or segregated. Segregation makes sense when a healthy individual’s productivity increases more when paired with another healthy individual than when paired with a diseased one, e.g., when disease exert 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 social organization of production 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

\[
L \equiv \int_0^1 q(n(e))de + \lambda_\gamma \left[ \gamma - \int_0^1 n(e)de \right] + \int_0^1 \left\{ \lambda_0(e)n(e) + \lambda_1(e)[1 - n(e)] \right\}de, \quad (5)
\]

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

\[
\frac{\partial q(n^*(e))}{\partial n(e)} - \lambda_\gamma + \lambda_0(e) - \lambda_1(e) = 0, \quad (6)
\]

\( n^*(e) \geq 0, 1 - n^*(e) \geq 0, \lambda_0(e)n^*(e) = 0, \lambda_1(e)[1 - n^*(e)] = 0, \lambda_0(e) \geq 0, \) and \( \lambda_1(e) \geq 0. \)

Since the constraints are linear, the nature of the assignment depends on concavity-convexity of \( q(n(e)) \) in (5). If \( q(n(e)) \) is strictly convex in \( n(e) \), the only solution to (6) is \( n^*(e) = \{0, 1\} \), where the measure of all communities with \( n^*(e) = 1 \) equals \( \gamma \) to satisfy (1). Convexity requires \( \partial^2 q(n^*(e))/\partial n(e)^2 = f(h, h) - 2f(h, d) + f(d, d) > 0. \) Using (3), this condition is

\[
[(h^\rho + d^\rho) / 2]^{1/\rho} < [(h^{1-\alpha} + d^{1-\alpha}) / 2]^{1/(1-\alpha)}. \quad (7)
\]

By Jensen’s inequality, (7) 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 social organization of production. The type of sorting of healthy and diseased
individuals into communities depends on the concavity-convexity of output per capita with respect to \( n(e) \). Assumption 2 implies that \( q(n(e)) \) is strictly convex in \( n(e) \). Thus 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. For instance, 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’ 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 is optimal for all values of \((\rho, \alpha)\) that satisfy the standing Assumptions 1 and 2. The literature on segregation by skill has typically found that worker segregation depends on particular values of the skills. For instance, in Kremer and Maskin (1996) and McCann and Trokhimtchouk (2010, Example 1), health human capitals \( h \) and \( d \) would have to be sufficiently different for segregation to occur.\(^5\) In the worker assignment here, even small differences in \( h \) and \( d \) lead to social segregation. The central difference is that I treat health human capital in a pairwise meeting as a single input of production. The pairwise health human capital approach is similar to the general human capital \textit{aggregator} in Jones (2014).\(^6\) An advantage of the CES aggregator is that it isolates the labor input as the key margin for disease spillovers, and showcases the relevant conflict between spillovers and diminishing returns in production.

For instance, only \( \rho \) and \( \alpha \) 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 \( \rho = 1 \),

\(^5\)Assume that \( z = k = 1 \) and consider \( f(d, h) = d^\rho h^\beta \). Worker segregation is socially efficient if and only if differences between workers’ skills are sufficiently large. The second-order condition (7) yields \( h^{\nu + \beta} - 2d^\beta h^\beta + d^{\nu + \beta} \). For \( \delta \equiv h/d \geq 1 \), the relevant terms needed to determine if segregation is optimal can be written as \( d^\beta \{ \delta^{\nu + \beta} - 2\delta^\beta + 1 \} \). As a function of \( \delta \), under \( \beta > \eta > 0 \), the difference in the term inside the brackets is zero if skills are equal or \( \delta = 1 \). It decreases monotonically as \( \delta \) increases until it reaches a trough at \( (2\beta/(\eta + \beta))^{1/\eta} \). Then this term increases monotonically until it eventually becomes positive. A positive difference in the term in brackets is necessary for segregation.

\(^6\)Jones (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).
(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. To minimize 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. This is an extreme case of Assumption 2. Under $\rho = -\infty$, meeting a diseased individual renders healthy individuals unproductive as (2) becomes $\mu_{-\infty}(h, d) = \min\{h, d\}$, and (4) 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’ 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 two negative forces: diminishing returns to scale in production and (negative) disease spillovers in pairwise matching.

**Output and aggregate efficiency.** Social segregation in Proposition 1 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 segregation. For example, $y_s$ is aggregate output per capita, i.e., the aggregate of $q(n^*(e))$ across segregated communities,

$$y_s \equiv \int_0^1 q(n^*(e))de,$$

aggregate output is $Y_s \equiv y_s N$, and $A_s$ is measured aggregate total factor productivity.

**Proposition 2** Under social segregation by disease, the aggregate production function is Cobb-Douglas $Y_s = A_s K^\alpha (HN)^{1-\alpha}$, where measured aggregate production efficiency $A_s$ is given by

$$A_s \equiv \exp \left[ \gamma \left( \frac{h}{H} \right)^{1-\alpha} + (1 - \gamma) \left( \frac{d}{H} \right)^{1-\alpha} \right], \text{ with } A_s < z. \quad (9)$$

**Proof.** Using $n^*(e)$ in (4) yields $q(n^*(e)) = \gamma f(h, h) + (1 - \gamma) f(d, d)$. This and (8) yield $y_s = z \gamma h^{1-\alpha} + (1 - \gamma) d^{1-\alpha} k^\alpha$. Simple rearrangements and (9) yield $y_s = A_s H^{1-\alpha} k^\alpha$, from which $Y^s$ follows. The inequality $A_s < z$ in (9) 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 health human capital is distributed across individuals or production units. Proposition 2 also shows that disease is a non-technological influence in total factor productivity. The effect of health human capital on \( A_s \) acts as a *productivity effect* of disease. This effect captures the fact that measured aggregate TFP is lower than the pairwise level of technology. As (9) shows, aggregating across pairwise meetings lowers aggregate efficiency, i.e., \( A_s < z \).

The productivity effect of disease is independent of the mean of 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 \( \theta^{(1-\alpha)} \), but \( A_s \) remains unchanged. Instead, \( A_s \) depends on how health human capital is distributed across individuals and 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 measured aggregate TFP increase by

\[
\Delta Y_s = \Delta A_s = (1 - \alpha)(1 - \gamma)\{d^{-\alpha} - h^{-\alpha}\}\Delta d > 0.
\] (10)

As health dispersion vanishes \( d \rightarrow h \), the social organization of production becomes irrelevant: sorting becomes indeterminate in the sense that all allocations become equally productive. In this case, \( A_s \) and \( z \) exactly 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 a more general 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, uses a standard representation for the aggregate production function, i.e., aggregate inputs \( K \) and \( HN \) are mapped into aggregate output \( Y_s \). This representation is consistent with the actual measurement practice for aggregate TFP. The measurement
strategy in Proposition 2 is also consistent with the accounting framework in Weil (2007), although he used an efficiency units framework.

**A central marketplace.** To further discuss the productivity effects of disease, abandon temporarily the pairwise matching assumption. Assume that capital and labor are pooled in a *central marketplace* subject to disease spillovers. Let subscript \(_{(c)}\) denote the variables under this alternative production arrangement. A central marketplace is not the same as an integrated assignment. Under integration, production takes place in pairs across different communities but with \(n(e) = \gamma\) across communities. In a central marketplace, \(n(e)\) is not a choice variable because there is no need to sort people when there is only one community. In per capita terms, the economy’s physical capital is \(k_c = \gamma k + (1 - \gamma)k = k\), the economy’s health human capital is \(\mu_{c,\rho}(h, d) = (\gamma h^\rho + (1 - \gamma) d^\rho)^{1/\rho}\), and the economy’s per capita production function is \(y_c = z[\mu_{c,\rho}(h, d)]^{(1-\alpha)}k^\alpha\). In a central marketplace, the values \(\mu_{c,\rho}(h, d)\) and \(y_c\) follow by definition.\(^7\)

In a central marketplace, aggregate output is also Cobb-Douglas, \(Y_c = A_c K^\alpha(HN)^{1-\alpha}\), with

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

This Cobb-Douglas production function differs from that in Proposition 2 only due to differences in measured TFP. This means that the productivity effect of disease can be stated in terms relative to the production efficiency of a central marketplace.

**Proposition 3** Aggregate production efficiency is lower in a central marketplace with disease spillovers than under segregated pairwise meetings, i.e., \(A_c < A_s\).

**Proof.** From (9) and (11), \(A_c < A_s\) 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 (7) that holds for any value of \(\gamma \in (0, 1)\), can be written as

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

By Jensen’s inequality, (12) is satisfied when \(\rho < 1 - \alpha\), which is Assumption 2. \(\blacksquare\)

---

\(^7\)The index \(\mu_{c,\rho}(h, d)\) is the natural representation of the economy’s health human capital when all factors of production are pooled. For instance, expression (2) in pairwise meetings is \(\mu_{c,\rho}(h, d)\) with \(\gamma = 1/2\).
Since segregation is socially efficient, pairwise production is 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 factors in a central marketplace subject to disease spillovers. For instance, Propositions 2 and 3 imply that $A_c < A_s < z$. On the contrary, when integration is optimal, inequality (12) 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 measured aggregate productivity across trading arrangements.

**Disease spillovers and market 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 production pairs. The efficiency gains from capital trade, however, depend on how exchange in capital markets is 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. In the trade protocols in this sub-section, capital exchange takes place between worker-pairs. When worker-pairs are segregated, worker-pairs are also individual types.

Consider first a benchmark in which capital can be exchanged in a centralized competitive market with no disease spillovers. Let the subscript $(n)$ denote allocations in a central marketplace with no spillovers.

**Proposition 4** Suppose that segregated workers exchange capital in a central marketplace with no disease spillovers. Aggregate output is Cobb-Douglas and production is as efficient as pairwise production, i.e., $A_n = z$.

**Proof.** A competitive capital market exchange between segregated pairs of workers attains the same allocation as a social planner that chooses $k(h)$ and $k(d)$ to solve $\max \{\gamma L_{h^{\alpha}}k(h)^{\alpha} + (1-\gamma) d^{1-\alpha}k(d)^{\alpha}\}$, 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 production efficiency because capital endowments are inefficiently assigned, i.e., all individuals have the same capital regardless of their health human capital. Capital exchange, in a sense, undoes the negative effect of disease and
segregation. There are, however, no disease spillovers when segregated workers interact in a centralized marketplace in Proposition 4.

To introduce disease spillovers, consider again pairwise trading arrangements. Assume that matched pairs of segregated workers can exchange capital but that trading opportunities arrive stochastically as segregated worker-pairs search for trading partners across communities. The probability that a matched pair of healthy individuals meets another pair of healthy individuals is $\gamma$, so their meeting probability is $\gamma^2$. The meeting probability between pairs of diseased individuals is $(1 - \gamma)^2$, and the meeting probability between healthy pairs and disease pairs is $2\gamma(1 - \gamma)$. Obviously, $\gamma^2 + (1 - \gamma)^2 + 2\gamma(1 - \gamma) = 1$.

There are no gains from trade when alike worker-pairs meet. Gains from trade occur when capital is reallocated between diverse pairs of segregated workers. Consider further two possible outcomes from these meetings.

Assume first that capital is exchanged in a decentralized marketplace with disease spillovers, as in Proposition 3. Expected output when healthy worker-pairs meet other healthy worker-pairs is $\gamma^2 z h^{1-\alpha} k^\alpha$. Diseased types meetings yield $(1 - \gamma)^2 z d^{1-\alpha} k^\alpha$. Expected output when healthy worker-pairs meet diseased worker-pairs is $2\gamma(1 - \gamma) z [(h^\rho + d^\rho)/2]^{(1-\alpha)/\rho} k^\alpha$, composed by the meeting probability $2\gamma(1 - \gamma)$ and the output when their labor and capital is pooled: $z [(h^\rho + d^\rho)/2]^{(1-\alpha)/\rho} k^\alpha$.

Let the subscript $(i)$ denote this trading arrangement. I use $(i)$ because pairwise capital trading under disease spillovers coincides with an integrated assignment in which $n^*(e) = \gamma$ for all $e \in [0, 1]$. Aggregate output under $(i)$ is also Cobb-Douglas $Y_i = A_i K^\alpha (HN)^{1-\alpha}$, and aggregate efficiency is characterized below.

**Proposition 5** Suppose that segregated workers exchange capital in decentralized marketplaces with disease spillovers. Aggregate production efficiency is higher than in a central marketplace for capital with disease spillovers (Proposition 3) but lower than under segregation and no capital trade (Proposition 2), i.e., $A_c < A_i < A_s$.

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

14
Aggregate production 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 \( A_i < A_4 \), one needs \( \gamma^2 h^{1-\alpha} + (1-\gamma)^2 d^{1-\alpha} + 2\gamma (1-\gamma) \left[ (h^\rho + d^\rho)/2 \right]^{(1-\alpha)/\rho} < \gamma h^{1-\alpha} + (1-\gamma) d^{1-\alpha} \). Rearrangements show that this inequality implies \( \left[ (h^\rho + d^\rho)/2 \right]^{1/\rho} < \left[ (h^{1-\alpha} + d^{1-\alpha})/2 \right]^{1/(1-\alpha)} \), which holds due to Assumption 2; see (7).

Next, rewrite \( [\gamma h^\rho + (1-\gamma) d^\rho]^{(1-\alpha)/\rho} \) in \( A_c \) as \( \{\gamma^2 h^\rho + (1-\gamma)^2 d^\rho + 2\gamma (1-\gamma) [(h^\rho + d^\rho)/2]^{\rho/\rho} \}^{1/\rho} < \{\gamma^2 h^{1-\alpha} + (1-\gamma)^2 d^{1-\alpha} + 2\gamma (1-\gamma) [(h^\rho + d^\rho)/2]^{(1-\alpha)/\rho} \}^{1/(1-\alpha)} \), which again holds by Assumption 2.

Proposition 1 has shown that it is inefficient to integrate healthy and diseased individuals for production purposes. Proposition 5 says that it also socially inefficient to integrate segregated individuals for exchange purposes. Exchanging capital in a decentralized marketplace essentially exposes traders to disease spillovers thereby undoing the efficiency gains from social segregation. Proposition 5 actually coincides with an integrated assignment. An integrated assignment, however, is more efficient than a central marketplace with disease spillovers, i.e., \( A_c < A_4 \). The reason is that the negative disease spillovers in a central marketplace apply to all individuals, whereas in the decentralized marketplaces of Proposition 5, only a fraction \( 2\gamma (1-\gamma) \) of worker-pairs experience the negative disease spillovers. This fraction is at most 1/2 when \( \gamma = 1/2 \).

For the second capital exchange arrangement, consider an intermediate point between the previous protocols. Assume, as in Proposition 5, that pairwise trading opportunities arise stochastically but suppose now that the capital of diverse matched pairs is exchanged in a decentralized competitive capital market with no disease spillovers, similar to Proposition 4 but at a pairwise level. Let subscript \( (d) \) denote this trading arrangement. The output of healthy and diseased worker-pairs is \( \gamma^2 z h^{1-\alpha} k^\alpha + (1-\gamma)^2 zd^{1-\alpha} k^\alpha \). The expected output of a meeting between different worker-pairs is

\[ 2\gamma (1-\gamma) z \max \{2^{-1} [h^{1-\alpha} k(h)^\alpha + d^{1-\alpha} k(d)^\alpha] \}, \text{ s.t., } k(h) + k(d) = 2k \].

\[(13)\]
Aggregate output can also be written as a Cobb-Douglas function $Y_d = A_d K^\alpha (HN)^{1-\alpha}$, with aggregate efficiency $A_d$ described below.

**Proposition 6** Suppose that segregated workers exchange capital in decentralized marketplaces with no disease spillovers. Aggregate production efficiency is higher than under segregation and no capital trade (Proposition 2) but lower than in a central marketplace for capital with no disease spillovers (Proposition 4), i.e., $A_s < A_d < A_n$.

**Proof.** The efficient allocation of capital satisfies $k_d(d) = [2d/(h + d)]k$ and $k_d(h) = [2h/(h + d)]k$. The value of the internal maximization problem in (13) is $k^\alpha[(h + d)/2]^{1-\alpha}$. Therefore, aggregate efficiency satisfies

$$A_d \equiv z \left[ \gamma^2 \left( \frac{h}{H} \right)^{1-\alpha} + (1 - \gamma)^2 \left( \frac{d}{H} \right)^{1-\alpha} + 2\gamma(1 - \gamma) \left\{ \frac{1}{2} \left( \frac{h + d}{H} \right) \right\}^{1-\alpha} \right]. \quad (14)$$

From (9) and (14), $A_s < A_d$ 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 (14). 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. \(\blacksquare\)

The absence of disease spillovers in worker-pair meetings makes capital exchange a source of efficiency gains relative to a segregated assignment, i.e., $A_s < A_d$. Efficiency gains arise as capital is directed to higher productivity worker-pairs. For instance, heathier worker-pairs receive more capital relative to their endowment, while diseased worker-pairs end up with less relative capital. As in the case of a central marketplace, capital-labor ratios are equalized, as in

$$\frac{k_d(h)}{k_d(d)} = \frac{k_n(h)}{k_n(d)} = \frac{h}{d} > 1.$$ 

Since these gains only apply to a fraction $2\gamma(1 - \gamma)$ of meetings, aggregate productivity is still lower than in a central marketplace in which all worker-pair exchanges are unaffected by disease, i.e., $A_d < A_n$.

**“Social barriers.”** The assignment problem takes a social planner perspective. Competitive markets tackle disease spillovers through a “social barrier.” For instance, segregation
is inconsistent with factor price equalization across communities or individuals of different types. The reason is that diseased individuals cannot ‘bribe’ healthy workers into forming a pairwise match. Diseased workers benefit from pairing with healthy workers, but healthy individuals are better off teaming only with other healthy individuals.

Consider a market decentralization of the segregated assignment. Physical capital is rewarded its marginal product, denoted by \( r_s(h) \) and \( r_s(d) \), i.e., \( r_s(h) = \alpha z^\alpha h^{1-\alpha} \). Wages for worker types, \( w_s(h) \) and \( w_s(d) \), exhaust income, i.e., \( f(h, h) = w_s(h) + r_s(h)k \) so that \( w_s(h) = (1-\alpha)f(h, h) = (1-\alpha)z^\alpha h^{1-\alpha} \) and \( w_s(d) = (1-\alpha)z^\alpha d^{1-\alpha} \). Wages across workers differ due to differences in health human capital. The log-wage gap between healthy and diseased workers is

\[
\ln \left[ \frac{w_s(h)}{w_s(d)} \right] = (1-\alpha)\ln[h/d].
\] (15)

Inequality in marginal products (15) 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, and showed that social integration, measured by convergence in the occupational and wage distribution, increased aggregate efficiency in the US since 1960. 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 Hsieh et al. (2019) and Hayashi and Prescott (2008). Thus, marginal products are different, but there are no gain from factor reallocations. Segregation by disease is socially efficient and policies that promote social integration in production or exchange reduce aggregate efficiency and income.

Assumption 2 is inconsistent with the efficiency units view. The log-wage gap (15) 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 workers are perfect substitutes, i.e., \( \rho = 1 \), Assumption 2 is not satisfied and the assignment problem becomes
a standard concave economy. When integration is optimal, there is a single baseline price for human capital \( w_i = (1 - \alpha)z[k/\mu_1(h, d)]^\alpha \), and the log-wage gap between healthy and diseased workers is of the same form as (15). Wage income, however, is more unequal under efficiency units as \( \ln \left[ \frac{w_i h}{w_i d} \right] = \ln \left[ \frac{w_s(h)}{w_s(d)} \right] (1 - \alpha) \).

Some remarks. Overall, the previous propositions say that disease spillovers reduce social interactions and aggregate efficiency. The efficiency losses in production and exchange vary with the scope of the disease spillovers. Aggregate efficiency, and consequently aggregate output, can be ranked according to the extent of disease spillovers and the degree of centralization:

\[
A_c < A_i < A_s < A_d < A_n. \tag{16}
\]

In (16), all production and exchange protocols are less efficient than technologically possible. The extremes involve central marketplaces. The case \( A_c \) generates the largest scope for disease spillovers and the lowest aggregate production efficiency. The other extreme, \( A_n \), assumes no disease spillovers. In the intermediate cases, some aggregate efficiency is lost in the process of production and exchange.

The model is highly stylized so it can be easily extended in multiple directions. For example, I focused on the social organization of production and exchange, but segregation might improve welfare if disease spillovers are sufficiently negative relative to the degree of inequality that can be socially tolerated. I also focused on a static assignment and ignored dynamic spillover associated with disease transmission. These issues do not contradict the main findings presented here, so they are discussed in more detail in the Appendix. In the Appendix, for example, I consider worker sorting under a concave utility function and a concave social welfare function. I show that a strengthening of Assumption 2 is sufficient for social segregation to maximize social welfare under disease spillovers. I also integrate the static assignment problem with standard epidemiological models with explicit disease transmission mechanics, e.g., Anderson and May (1992). Including the dynamics behind disease transmission induces a dynamic spillover effect as contacts between healthy and diseased workers increase the prevalence of disease in the population (i.e., \( \gamma \)). Dynamic spillovers reinforce the static spillovers considered here so they serve to weaken Assumption 2. The
Appendix also examines the emergence of a multi-class society where social hierarchies, as in reality, can depend on multidimensional attributes rather than just on 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.

An implicit assumption in this paper is that health types are fully observable. Partially observable types introduce a signal-extraction problem that makes complete segregation impossible. An analysis of worker sorting under imperfect information cannot be done in an Appendix, but I consider an exchange economy with partially observable health types in a separate paper; Birchenall (2021). I show there that the stochastic properties of disease categorized along danger and visibility lines are central for understanding the social response to past epidemics, including errors in the form of false alarms.

3 Some Quantitative Findings

I next examine how much larger aggregate efficiency and output would be under the different production and exchange arrangements in (16). I consider a very standard accounting framework and compare the long-run contribution of the productivity effect of disease to the proximate effect. Since the aggregate production function is Cobb-Douglas, output per capita can be written as $Y/N = A^{1/(1-\alpha)} (K/Y)^{\alpha/(1-\alpha)} H$. I assume that the capital-output ratio is independent of health human capital. With $z = 1$, the comparison between the productivity and the proximate effects of disease in output per capita is based on differences between $A^{1/(1-\alpha)}$ and $H$.

Parameterization. I rely on parameter values consistent with causal estimates of disease spillovers and a number of empirical observations on long-run growth. I consider changes in the aggregate labor input 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 observed 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.” Using $H = 1$, these numbers suggest that a difference in mean health human capital of $H' = 2.5$ is a reasonable benchmark.

I take $\alpha = 0.40$ based on aggregate factor payments to capital. 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, i.e., “beggars constituted as much as a fifth of the populations of ancien régimes.” A value of $\gamma = 0.8$ is conservative as it only considers extremely malnourished individuals even when endemic and epidemic diseases were highly prevalent.

I calibrate $\rho$ using the estimated disease spillovers for deworming of intestinal parasites in school-aged children in Kenya by Miguel and Kremer (2004, [2014]). Consider the following thought experiment. A population of diseased individuals receive an intervention that leads to health gains of $\Delta_T d$, with $\Delta_T > 1$ as the direct treatment effect. Untreated diseased individuals meet with the directly treated. As in (2), their health human capital improves to $\Delta_U d$, where

$$\Delta_U \equiv \left\{ \frac{1}{2} + \frac{1}{2} \left( \frac{\Delta_T d}{d} \right)^{\rho} \right\}^{1/\rho}, \tag{17}$$

represents the health gains due to the disease spillovers (i.e., the treatment on the untreated).

Miguel and Kremer (2004, [2014]) measured the direct and indirect benefits of treating moderate-to-heavy helminth infections. Their indirect effects are significant for interactions at a radius of 3km from treated schools. I interpret their estimates assuming that the decline in the probability of infection yields a one-for-one increase in health human capital. For instance, their estimate for the direct treatment effect is $\beta_T = -0.333$ (s.e. 0.052) and their indirect treatment effect is $\beta_U = -0.102$ (s.e. 0.043); see Miguel and Kremer ([2014], Table B1).\footnote{This interpretation is possible because the intestinal infections were very common and the treatment very successful. For instance, their linear probability estimates are very similar to their probit estimates. In their baseline sample, 52 percent in the untreated group (students in grades 3-8) had a moderate-to-heavy helminth infection after the intervention, compared to 27 percent of the students in the treatment group; see their Appendix Table A5. The original estimates in Miguel and Kremer (2004) contained a coding error. I rely on the updated estimates provided by their guide for replication in 2014.} Using $\Delta_T = 1.333$ and $\Delta_U = 1.102$ yields $\rho = -4.8$ in (17). Bleakley (2010)
considered spillovers in eradication campaigns against parasitic disease in the Americas. It is reassuring that 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).

It is possible to consider other values for \( \rho \) depending on the size of treated population. For example, if the population receiving the direct intervention is \( \gamma \), (17) yields \( \Delta_U = (1 - \gamma + \gamma \Delta_T)^{1/\rho} \). Under \( \gamma = 4/5 \), the implied value is \( \rho = -16.2 \), and under \( \gamma = 1/5 \), \( \rho = 4.7 \). If more people are treated (i.e., \( \gamma = 4/5 \)) but the indirect gains remain unchanged, one should see lower values for \( \rho \). Likewise, if a few individuals are treated but the indirect gains are unchanged, then one should have a high value for \( \rho \). I use \( \gamma = 1/2 \) since the groups in Miguel and Kremer (2004, [2014]) are balanced and their estimates are weighted by population size.

It is not possible to measure health human capital directly so I consider multiple values of \((h, d)\) and \((h', d')\) consistent with \(H = 1\) and \(H' = 2.5\). I consider \(H = 1\) in two forms: \((h, d) = (1.2, 0.2)\) and \((h, d) = (1.15, 0.4)\). The ratio \(h/d\) in 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). The ratio \(h/d\) in the first case is larger than in Fogel (1994), but it is consistent with a six-fold “urban mortality premium” at young ages between industrializing cities and rural areas in eighteenth century England; see Birchenall (2007). I consider \(h'\) and \(d'\) with the previous ratios and more equal values.

**Main findings.** Table 1 lists the production and trading arrangements in (16). I first focus is on the extremes that consider centralized trading with and without spillovers, i.e., \(A_c\) and \(A_n\). In (a), under centralized trading with disease spillovers, \(A_c\) represents 0.28 of aggregate TFP without spillovers, i.e., \(A_n = z = 1\). The difference between \(A_c\) and the lower bound \(A_-\) is relatively small, so \(\rho = -4.8\) already induces a strong spillover effect. (Recall that \(A_- = A_c\) when \(\rho \to -\infty\), so under \((1.2, 0.2)\), \(A_-^{1/(1-\alpha)} = d/H = 0.2\).) Productivity differences between segregated and integrated assignments are ‘small,’ i.e., \(A_i^{1/(1-\alpha)} = 0.78\) and \(A_n^{1/(1-\alpha)} = 0.95\), so segregation, even under unequal health human capital, is close to the upper bound \(A_n^{1/(1-\alpha)} = 1\). In general, segregation is quite effective at dealing with
disease spillovers. For instance, the gains from capital exchange without disease spillovers, i.e., $A_d^{1/(1-\alpha)} = 0.98$, are ‘small’ relative to the segregated assignment.

The productivity effect of disease depends on the distribution of health human capital. The second column in panel (a) considers a more equal distribution of $h$ and $d$. This is a mean-preserving improvement of health relative to the first column since both yield $H = 1$. As the second column shows, the productivity gains from the mean-preserving move to $(h, d) = (1.15, 0.4)$ are also large as $A_c^{1/(1-\alpha)}$ doubles, from 0.28 to 0.56. Overall, eliminating disease spillovers or reducing health human capital inequality leads to large productivity gains in centralized markets.

Table 1(b) considers different ways to reach $H' = 2.5$. The case $(h', d') = (0.5, 3.0)$ is exactly the same as $(h, d) = (0.2, 1.2)$ since measured productivities only depend on relative values. The following three cases progressively increase $d'$ by $\Delta d = 0.5$ and lower $h'$ so as to maintain $H' = 2.5$. As expected, reducing health dispersion increases measured productivity along any row as the scope of the disease spillovers is reduced in more homogeneous populations. The quantitative gains from more equal health human capital are of the same order of magnitude as the gains from eliminating disease spillovers. For example, when $(h', d') = (2.0, 2.6)$, the lower bound satisfies $A_c^{1/(1-\alpha)} = 0.8$ but centralized exchange with spillovers $A_c$ already reaches 0.96 of the baseline productivity.

Table 1. Proximate and productivity effects of disease.

<table>
<thead>
<tr>
<th></th>
<th>(a) $H = 1$</th>
<th>(b) $H' = 2.5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h$</td>
<td>1.20</td>
<td>3.00</td>
</tr>
<tr>
<td>$d$</td>
<td>0.20</td>
<td>0.50</td>
</tr>
<tr>
<td>$A_d^{1/(1-\alpha)}$</td>
<td>0.28</td>
<td>0.28</td>
</tr>
<tr>
<td>$A_t^{1/(1-\alpha)}$</td>
<td>0.78</td>
<td>0.78</td>
</tr>
<tr>
<td>$A_s^{1/(1-\alpha)}$</td>
<td>0.95</td>
<td>0.95</td>
</tr>
<tr>
<td>$A_y^{1/(1-\alpha)}$</td>
<td>0.98</td>
<td>0.98</td>
</tr>
<tr>
<td>$A_{12}^{1/(1-\alpha)}$</td>
<td>0.98</td>
<td>0.98</td>
</tr>
<tr>
<td>$A_{13}^{1/(1-\alpha)}$</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

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

Mean health human capital has increased over time and health disparities have narrowed across individuals and countries. For instance, 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 the extremes for centralized markets $A_c$ under $(h,d) = (1.2, 0.2)$ and $A_n$ under $(h',d') = (2.6, 2.0)$ is perhaps the most representative change during modern times. In such comparison,

$$Y/N = 0.28 \times 1.0 \rightarrow 1.0 \times 2.5 = (Y/N)' ,$$

with output per capita increasing by a factor of 8.9 as $H' \rightarrow 2.5 \times H$ and $A_n^{1/(1-\alpha)} \rightarrow 3.6 \times A_c^{1/(1-\alpha)}$. An increase in output by a factor of 8.9 goes a long way in accounting for the historical development of rich countries or the current differences between poor and rich countries.\(^9\) Even under $(h,d) = (1.15, 0.4)$, the productivity gains are closer to the proximate gains, i.e., $A_n^{1/(1-\alpha)} \rightarrow 1.78 \times A_c^{1/(1-\alpha)}$. The previous estimates hence suggest large productivity gains associated with health human capital investments.

**Some remarks.** A focus on $A_c$ and $A_n$ can be justified by considering the productivity of cities. In the past, cities played a central role in production and exchange, but also in disease transmission. Until the twentieth century, disease spillovers and the close proximity of individuals induced a considerable “urban mortality penalty;” see, e.g., Birchenall (2007). It is perhaps obvious but the scale and current high concentration of workers in cities, where knowledge externalities take place, would not be possible with the disease spillovers of the past.

Large aggregate efficiency gains when disease spillovers are eliminated are reasonable. They may even be a lower bound for income changes. I have ignored complementarities between health human capital and other forms of human or physical capital, and technology. Physical capital, education, and technological changes are likely to lead to much higher income levels as $A$ and $H$ increase. Also, the model is static but health exhibits intertemporal and intergenerational complementarities associated with gains that originate as early as during fetal development; Fogel (1994). The main offsetting effect not considered here is the demographic response to disease. Increases in population would counter the productivity and proximate effects of disease on aggregate income. Population growth in response

\(^9\)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).
to health improvements, however, is likely transitory as during the demographic transition when fertility adjusted too slowly to mortality reductions; see, e.g., Acemoglu and Johnson (2007).

It is possible to consider a different approach to calibrate $\rho$. The consensus on the elasticity of substitution between skilled and unskilled labor, as measured by educational attainment, is a range between 1 and 2 with a common estimate of 1.4; see, e.g., Jones (2014) and Caselli and Ciccone (2019). The value $\rho = 1 - 1/1.4 = 0.28$ is consistent with Assumption 2. Under $(h, d) = (1.2, 0.2)$ in part (a), $A^1/1(1-\alpha) = 0.90$ so disease spillovers would be less significant. The value $\rho = 0.28$, however, is a conservative measure for disease spillovers. The empirical literature on education consider values of elasticities of substitution below 1 as implausible given the evolution of wage inequality; see Jones (2014). Elasticities above one imply $\rho > 0$. The difference between positive and negative values of $\rho$ is important because positive values give a greater weight to healthy individuals in $\mu_\rho(h, d)$ whereas negative values of $\rho$ give a greater weight to diseased individuals. (This can be seen easily for $\mu_\infty(h, d) = \max\{h, d\}$ and $\mu_{-\infty}(h, d) = \min\{h, d\}$.) Assuming $\rho > 0$ therefore limits the scope of disease spillovers.

4 Some Historical Perspectives

Having discussed the potential role of disease in the social organization of production and exchange, I next discuss briefly some historical accounts that showcase the role of segregation by disease in production and exchange. Some of the channels discussed here are not tacitly recognized in economics but are often discussed by historians. A detailed study of the instances discussed here would require, at a minimum, a series of dedicated papers. The goal of this section is more in line with creating a catalog of past responses to disease that aligns with this paper’s thesis and might be helpful in organizing future work.

**Caste.** The canonical example of social segregation is the caste system. Caste divides 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,
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 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 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. Untouchability arises from performing ‘dirty’ production tasks.

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.”

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.
tasks associated with “taboo infections” such as animal skinning, leather-work, scavenging, and removing sewage and corpses; see, e.g., Hutton (1969, p. 89). Moreover, the separation underlying the ‘practice of untouchability’ restricts direct physical contact but also other indirect forms of 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.” Although caste serves many purposes, a central function of the Indian caste system appears to be the regulation of direct and indirect physical contact as a disease-avoidance mechanism.

Accounts of the Indian caste system date back to the Vedic period somewhere around 1500-1000 BC. 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. 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 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.

The role of disease as contributing factor in the caste system was also noted by Singh

---

12 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.
(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 based on social norms and a social consensus. As in the lifetime isolation of lepers, discussed many times in the *Old Testament*, the enforcement of physical separation between diseased and healthy individuals is based on social and not legal sanctions. Quarantines are the earliest example of segregation by disease enforced by the authority of the state using principles of coercion.

Quarantines are preventive measures that apply to a general population presumed to be diseased. That is, quarantines represent the physical separation of all individuals who have the potential of transmitting infections, i.e., even asymptomatic individuals who may have been exposed to disease. (Semantically, isolation represents the physical separation between symptomatic individuals or diseased individuals.) Quarantine measures are relatively modern likely because they require a centralized authority. They began as a strategy by Mediterranean city-state governments to minimize the risk of infection during the Black Death. The 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 and it spread along trade routes (Biraben, 1975). 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.” Trade and factor mobility restrictions were 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.\footnote{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.}

Social segregation 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. 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 \textit{cordon sanitaire} between the Habsburg and Ottoman empires established by Austria in the 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 cities.

The principles of segregation by disease, even prior to the advent of the \textit{germ theory} of disease, were applied to many additional diseases such as tuberculosis and leprosy. The segregation of infected individuals into dedicated hospitals (i.e., sanatoria and preventoria for tuberculosis) followed the same principles of physical separation but, due to the chronic nature of these diseases, the separation was permanent. In Birchenall (2021), I present a more detailed discussion of the use of quarantines to counter Yellow Fever and plague pandemics in the West.

\textbf{Zoning.} Since antiquity, cities have featured a number of positive and negative agglomeration externalities due to the close proximity between individuals as well as between individuals and firms. Until the early twentieth century, cities had much higher infectious disease loads than rural areas with a large “urban mortality penalty” reversed only after
widespread land-use regulation and major urban infrastructure projects such as sewers and water sanitation. Mortality rates before the age of 15 in rapidly industrializing European cities, for example, exceeded mortality in rural areas by at least a factor of 6; 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 transmission and facilitated the spread of communicable diseases such as tuberculosis, typhoid fever, and cholera (Birchenall, 2007).

Even during the age of miasmas, public health officials considered ways to improve city sanitation using principles based on contagion. As in the response of Mediterranean cities 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” and appropriate public health responses were put in place; see, e.g., Melosi (2008, pp. 41-43). Indeed, “[d]isease was the raison d’être for the advent of urban planning in Europe and the USA” according to Pinter-Wollman et al. (2018, p. 3). The regulation of urban spaces collectively known as zoning was one of the central public health strategies to counter the harmful effects of urbanization. (Zoning is defined as the spatial segregation of residential, commercial, and industrial land-uses.)

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. Zoning served to create distance between residential areas and polluting firms, noxious fumes, and animal waste; see, e.g., Schilling and Linton (2005, p. 98).

‘Silent trade.’ Tropical parasitic and infectious diseases have been heavily concentrated
in sub-Saharan Africa. An illustration of the effect of disease on market exchange is the tendency for trade in pre-colonial Africa to be conducted at the border of territories, rather than at centralized markets. This 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. 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 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.”

Exchange based on iterative transactions that minimize direct physical contact are likely the result of severe disease spillovers. (Language barriers and general concerns for safety are also seen as reasons for the silent trade.) In the context of sub-Saharan Africa, Hartwig and Patterson (1978, pp. 6-7) note that “some long-distance trade involved relatively little human movement because specific communities prohibited transit traffic. Instead, these

---

14 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).”

15 Grierson (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.
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 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).

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 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-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 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 restrictions prevented “upland Indians” from going to “coastal plains” and vice versa. In their military campaigns at lower altitudes, Inca soldiers were replaced “every two months since the coastal lands are unhealthy for the

\[^{17}\text{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. Historically, African empires also had administrative centers but of a much smaller scale compared to the empires of the New World. In fact, “some African monarchies deliberately avoided urbanization.” For example, “[i]n 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). Notice, nonetheless, that the majority of the Ethiopian population still resides in the northern highlands in the interior of the country.}]
During colonial times, European settlers in the Caribbean established ‘hill stations’ for military personnel as a way to avoid tropical diseases. 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.

5 Conclusion

This paper examined how disease spillovers (i.e., contagion) influence the social organization of production and exchange. I sketched out an analytical framework whose inferences follow from simple concavity principles and Jensen’s inequality. Its main purpose was to characterize aggregate efficiency across a number of centralized and decentralized production and exchange protocols with differential exposure to disease. In the calibrated numerical example, disease spillovers had a large effect on aggregate efficiency and output. In a development accounting framework, a three-fold difference in measured productivity 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. Having discussed the theoretical and practical implications of segregation by disease, it is useful to summarize some general principles in the theory and the perspectives offered by historians.

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 demography of the economy; they are not about the size of the different types of agents in the economy or about the total 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 measuring 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 international 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 are still relevant for understanding long-run economic growth. As the theory makes clear, all production and exchange protocols are less efficient than technologically possible. These 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 ‘untouchability’ and other discriminatory practices. High altitude cities in Latin America similarly served to consolidate pre-colonial empires. The increased distance
from coastal areas and rugged terrain, however, has led to much higher transportation and integration costs in today's globalized world. 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). Difficulties in centralized state-building appear important in sub-Saharan Africa today where severe, and often fatal, infectious diseases (i.e., Ebola) are recurrent phenomena. But not all historical legacies are negative. The segregation of land-use functions is the most prevalent land use planning tool in modern cities and towns. In the absence of vaccines, quarantines and isolation remain one of the most important and effective preventive public health measures against communicable disease outbreaks. Clearly, not all past social responses to disease have become maladaptive.

References


37


6 Appendix: Some Extensions

The paper considers a highly stylized model. This Appendix provides some brief remarks and extensions for a two-class society. The insights also apply to a multi-class society, presented later in this Appendix.

Inequality and welfare. I focused on the social organization of production and exchange. Segregation, however, might improve welfare if disease spillovers are sufficiently negative relative to the degree of inequality that can be socially tolerated.

Suppose that production is equally consumed within each pair of workers and let the 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 ex post utility of \((h, d)\) and \((d, d)\) meetings, respectively. Let \( u(n(e)) \) denote community \( e \)'s utility, i.e.,

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

and define the economy's social welfare function as the CES aggregate of expected utilities across communities, i.e.,

\[
U(u) \equiv \frac{1}{1 - \varphi} \int_0^1 u(n(e))^{1-\varphi} \text{d}e. \tag{18}
\]

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 presence of \( \epsilon \) and \( \varphi \) reduce the incentives of social segregation, so Assumption 2 must be strengthened. 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 7** 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 \( (18) \).

**Proof.** The proof of the first case is standard because the objective function \( (18) \) has the same form as \( (8) \). Thus, as in \( (7) \) and Proposition 1, Proposition 7 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 \( n(e) \).\(^{18}\) ■

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

\[\text{First derivative of } U(u) \text{ with respect to } n(e) = [n(e)^2\phi(h, h) + (1 - n(e))^2\phi(d, d)]^{-\varphi} \text{ for } \rho < (1 - \alpha)(1 - \epsilon).\]

\(^{18}\)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) -

\[\text{The first derivative of } U(u) \text{ with respect to } n(e) = [n(e)^2\phi(h, h) + (1 - n(e))^2\phi(d, d)]^{-\varphi} \text{ for } \rho < (1 - \alpha)(1 - \epsilon).\]

\(^{18}\)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) -

40
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.

**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).\(^{19}\) 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.\(^{20}\)

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 (19).)

Let \( 0 < \beta < 1 \) denote the discount factor. The social planner allocates healthy workers \( \phi(d, d) \). After some manipulations, the second derivative can be written as

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

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.

\(^{19}\)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.

\(^{20}\)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.
across communities to maximize the discounted present value of aggregate output,

\[ v(\gamma) = \max \left\{ \int_0^1 q(n(e))de + \beta v(\gamma') \right\}, \text{ s.t., (1) and (19)}. \]

The following proposition characterizes the dynamic assignment:

**Proposition 8** 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 (5), the assignment problem now contains a Lagrange multiplier \( \lambda' \) on the constraint (19). The second-order condition for \( n(e) \) depends on

\[ [f(h, h) - 2f(h, d) + f(d, d)] + \lambda' 2\tau. \tag{20} \]

The first term in (20) captures the static spillovers of disease (7) and the second the dynamic spillovers. This second term is positive since \( \lambda' > 0 \). The added constraint (19) thus increases the social planner’s incentives for segregation. The transmission rate \( \pi \) does not influence (20) because the vector transmission is uniform across locations. The human transmission rate \( \tau \), however, induce segregation in (20) even if \( q(n(e)) \) is linear, i.e., \( 1 - \alpha = \rho \). Under segregation, (19) 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 (19) or if there is segregation, disease evolves exogenously from the perspective of a social planner.

In Proposition 8, the fraction of healthy individuals is endogenous in a dynamic environment. Under segregation, (19) becomes \( \gamma' - \gamma = \sigma(1 - \gamma) - \pi m \gamma \), which converges monotonically to \( \gamma_s = \sigma/(\sigma + \pi m) \). 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 \( m \). 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 $h_1 < h_2 < \ldots < h_J$. 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., $\phi_i = \phi$. 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 $F(i, j) = f(h_i, h_j)$. 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 $n(e) \equiv (n_1(e), \ldots, n_J(e))$, which is consistent with the baseline model. Since workers meet randomly, expected output in community $e$ is given by a quadratic form $2q(n(e)) \equiv n(e)^T Fn(e)$. 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 9** 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 $\mu_p(h, d)$, as in the benchmark model. Since the entries $F(i, j)$ are real, the matrix $F$ is also Hermitian. Therefore, if pairwise comparisons satisfy

$$F(i, j) < \frac{1}{J-1} \sqrt{F(i, i)F(j, j)},$$

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., $\sqrt{F(i, i)F(j, j)} \leq (F(i, i) + F(j, j))/2$,
imply that
\[ F_{(i,j)} < \frac{1}{J-1} \frac{F_{(i,i)} + F_{(j,j)}}{2}, \] (22)
is a sufficient condition for \( F \) to be positive definite. Since \( F_{(i,i)} = z h_i^{1-\alpha} k_i^{\alpha} \), \( F_{(j,j)} = z h_j^{1-\alpha} k_j^{\alpha} \), and \( F_{(i,j)} = z [(h_i^c + h_j^c)/2]^{(1-\alpha)/\rho} k^{\alpha} \), some simple algebra shows that (22) can be written as (21). 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.

When \( J = 2 \), expression (21) 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 (21) 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 9 leads to generalizations of the baseline model. Under segregation, the aggregate production function is still Cobb-Douglas \( Y_s = A_s K^{\alpha} (HN)^{1-\alpha} \), and aggregate efficiency is given by
\[ A_s \equiv z \left[ \sum_{j=1}^{J} \gamma_j \left( \frac{h_j}{H} \right)^{1-\alpha} \right], \text{ with } H \equiv \frac{1}{N} \sum_{j=1}^{J} \gamma_j h_j. \] (23)
Expression (23) extends (9) 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
\[ A_c \equiv z \left[ \sum_{j=1}^{J} \gamma_j \left( \frac{h_j}{H} \right)^{\rho \gamma_j h_j^{1-\alpha} k_j^{\alpha}}, \text{ s.t., } \sum_{j=1}^{J} \gamma_j k_j = K, \right. \]
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
\[ \max_{k_j} \sum_{j=1}^{J} z \gamma_j h_j^{1-\alpha} k_j^{\alpha}, \text{ s.t., } \sum_{j=1}^{J} \gamma_j k_j = K, \]
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 9 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 9, 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.