European Economic Review 83 (2016) 165–184

Contents lists available at ScienceDirect

European Economic Review

journal homepage: www.elsevier.com/locate/eer

Who was colonized and when? A cross-country analysis of determinants[☆]

Arhan Ertan[a], Martin Fiszbein[b], Louis Putterman[c,*]

[a] Kadir Has University, Turkey
[b] Boston University, United States
[c] Brown University, Box B, Providence, RI 02912, USA

ARTICLE INFO

Article history:
Received 12 August 2014
Accepted 11 October 2015
Available online 26 November 2015

JEL classification:
N00
N90
O43
O11

Keywords:
Colonization
Early Development
State History
Neolithic revolution

ABSTRACT

The process of colonization has shaped the economic and demographic contours of the modern world. In this paper, we study the determinants of the occurrence and timing of colonization of non-European countries by Western European powers. Of particular interest is the role of early development measures that are known to be strong correlates of present-day levels of income. We show that non-European societies with longer histories of agriculture and statehood and higher levels of technology adoption in 1500 were less likely to be colonized, and tended to be colonized later if at all. We also find that proximity to the colonizing powers, disease environment, and latitude are significant predictors of the occurrence and timing of colonization, although their impacts are less robust to choice of country sample. Our models have high explanatory power, and their support for the significance of early development is robust to the use of alternative indicators of early development and disease, to the use of instruments to focus on the exogenous component of early development, and to the joint estimation of the colonization and timing equations to correct for potential selection bias.

1. Introduction

There is little disagreement among historians that the process by which Western European nations set sail into the Atlantic and Indian Oceans, began the conquest of their islands and coastlines, and eventually came to control vast swaths of territory in the Americas, Africa, Asia and the Pacific, is one of the most important factors that shaped the economic and demographic contours of the modern world. The age of colonialism began with the European discoveries of sea routes around Africa's southern coast (1488) and to the Americas (1492), or perhaps a bit earlier with the settlement of previously uninhabited Atlantic islands like Cape Verde in 1462 (Landes, 1998). Thereafter, by discovery, conquest, and settlement, the emerging nation-states of Portugal, Spain, the Dutch Republic, France, and England expanded their reach, spreading

[☆] Financial support from the William R. Rhodes Center for International Economics and Finance is gratefully acknowledged. We wish to thank Frank Kleibergen and Philipp Ketz for econometric advice, Yunan Ji for her help with estimation work, Dina Akbar, Jack Du, Brian Hanrahan, and Tanya Nguyen for helpful research assistance, and participants at presentations at Brown University, Bogazici University, City University of New York, and EH Clio Lab for their helpful comments. We are grateful to two anonymous referees for their helpful suggestions. We thank Douglas Hibbs and Ola Olsson for providing their data on biogeography and climate, and Raphael Auer, who responded to inquiries regarding his disease measure (EDE) by calculating values for those countries in which our respective samples fail to overlap.

* Corresponding author. Tel.: +1 508 517 6976.
E-mail address: Louis_Putterman@Brown.Edu (L. Putterman).

http://dx.doi.org/10.1016/j.euroecorev.2015.10.012
0014-2921/

European institutions, culture, and genes, and forcing or inducing massive cross-continental movements of Africans and others. By the time that the era of colonization ended in the decades after World War II, the populations of countries in the Americas, Australia, New Zealand, and elsewhere had been radically transformed, and new nation-states had been brought into being on four continents—North and South America, Africa, and Australia—with borders bearing no relation to pre-colonial precedents.

On the eve of World War II, two-fifths of the world's land area and a third of its population were in colonies, dependencies, or dominions of Western European colonizing powers. A further third of world territory had been colonized by these European powers sometime between the 15th and 19th centuries and had already emerged as independent nations. In many of the latter cases, however, it was not the once-colonized peoples that became independent, but rather the descendants of the colonizers, so that the process of colonization was never truly reversed. In other cases, post-colonial populations were mainly descended from people that the colonizers had imported as slave or indentured laborers, or by admixtures of indigenous, “imported” and colonizing populations. What is called “the Third World” or “the developing world” consists overwhelmingly of ex-colonies, including both ones that underwent dramatic changes in source population of the kinds just described (such as those in the Americas) and ones that did not (such as most in Africa, India, and ex-colonial Asia—see Putterman and Weil, 2010; Chanda et al., 2014).

Yet not all of the non-European world was colonized by Western European maritime powers. Turkey, Iran, China and Japan are among the Eurasian countries not colonized by Western Europeans, while parts of Central Asia that became independent states when the Soviet Union dissolved in 1991 had been absorbed into the land based empire of Russia and were never ruled by Western European colonizers (Landes, 1998; Acemoglu et al., 2001, 2002, 2005, likewise distinguish between European maritime colonization and Eurasian land-based empires). Furthermore, places that were colonized by Western Europeans came under their rule at very different times: for example, the late 15th and early 16th century for the Americas, but the late 19th and early 20th century for most of sub-Saharan Africa—a difference of four hundred years. The Philippines was under Spanish rule by the early 17th century, whereas Australia, New Zealand, New Guinea and Vietnam were not colonized until the 19th century, and countries including Syria and Jordan experienced Western European rule only after World War I.

The impact of the colonial era is recognized in some of the most influential papers on long run economic growth. But none of them, to our knowledge, attempt to explain why some countries were colonized and others not, or why some were colonies as early as the 15th century while many others became colonies only in the late 19th or early 20th centuries.

Our attempt to explain the occurrence and timing of colonization extends the literature on the persistence of early developmental advantages, which was recently surveyed by Spolaore and Wacziarg (2013). Outside of economics, the most influential work in this literature is Diamond (1997), which places the question of colonization front and center, emphasizing the asymmetric character of the colonization process. “The modern United States,” Diamond writes, “is a European-molded society, occupying lands conquered from Native Americans and incorporating the descendants of millions of sub-Saharan black Africans brought to America as slaves. Modern Europe is not a society molded by sub-Saharan black Africans who brought millions of Native Americans as slaves. … The whole modern world has been shaped by lopsided outcomes (Diamond, pp. 24–25).” How, Diamond asks, can this lopsidedness be explained?

Diamond's analysis, with its emphasis on the geographic distribution of the precursors of major domesticated plants and animals, has been much discussed by economists. But while several studies (beginning with Hibbs and Olsson, 2004) have found support for his thesis about the impact of early agriculture on subsequent economic development, we are the first to statistically examine his related idea regarding the impact of early agriculture on colonization. Diamond used a broad set of descriptive case studies to build an explanation of why European powers colonized (most of) the Americas, Africa and Oceania, and not the other way around. In this paper, we take the general idea that early development contributes to the explanation of colonization patterns and provide a statistical assessment by directing our attention to the cross-sectional variation in the occurrence and the timing of colonization in the non-European world.

While testing the impact of early agrarian civilizations on colonization provides the initial impetus to our study, we also bring additional geographic and disease considerations to bear in our analysis. We find that both nautical distance from Western Europe, and the distance to be traversed overland in the cases of landlocked and semi-landlocked countries (explained below), play roles in both the occurrence and timing of colonization. We find the presence of disease environments deadly to Europeans to be a major delayer, but not preventer, of colonization.

A common criticism of Diamond's discussion concerns its relative silence on the divergence between European and other Eurasian civilizations (Morris, 2010; Acemoglu and Robinson, 2012). Explaining why Atlantic-facing rather than other Eurasian states began the colonization of the Americas and Oceania is beyond the scope of our paper. However, we do show that the relative lateness of European colonial acquisitions in North Africa, the Middle East and Asia is consistent with the role of relative technological and organizational leads in explaining colonization's timing. That is, Western Europeans tended to colonize earlier the non-Eurasian areas with substantially lower levels of technology, state experience, and duration of reliance on agriculture, and most of their colonization of regions in or near non-European Old World core civilizations occurred only after the technological gap between Western Europe, on the one hand, and North Africa and Asia, on the other, had grown much larger, partly through colonial acquisition and intra-European competition. The importance of Europe's growing technological lead for explaining later European conquests within Eurasia fits a more general analytical rubric connecting colonization with technological and organizational advantages. That rubric adds, to Diamond's focus on

Eurasian versus non-Eurasian differences, complementary attention to early modern gaps that may to some degree reflect advantages Western Europe enjoyed thanks to its initial post-15th century colonization lead.

Our paper considers only colonization by European maritime powers in a particular era of world history, not colonization as a broader phenomenon. In ancient times, the Assyrian, Persian, Hellenistic, Roman, and other empires conquered large parts of the Near East and the Mediterranean basin. The Mongol conquests of the 13th century, followed by the fall of the Byzantine Empire, the rise of the Ottoman and Mughal empires, and the ultimate division of large parts of Eurasia between Russian and Chinese empires, reshaped Old World history through transmission of technology, genes, and disease (McNeill, 1998). Modern usage and the recent scholarship mentioned earlier distinguish these land-based empires from the overseas empires established by Portugal, Spain, the Netherlands, France, and Britain beginning in the 15th century. It is the latter epoch of colonization of non-European regions by Western European powers, not the earlier or contemporaneous land-based empires, on which we focus. Not only did the conquest of distant lands by the countries mentioned (and by later, minor emulators Belgium, Germany and Italy) play a key role in determining the borders of the nation-states of the contemporary Americas, sub-Saharan Africa, and Oceania, but it may well have contributed directly to the emergence of industrial capitalism and thereby of the modern global economy (Acemoglu et al., 2005).

Note that our focus is on the occurrence and the timing of colonization, not on which of the Western European countries colonized a given territory or how long colonization lasted. Although those questions are also important ones, we do not expect the factors that are our focus, especially economic and social development circa 1500, to have the same bearing on whether, say, France versus Portugal was the colonizer, or when independence occurred. Our interest is mainly in systematically testing the hypothesis that development circa 1500 helps to explain who got colonized and when, and in determining what additional roles disease and geography played. This leaves exploration of other questions for future research.

Consistent with our initial conjecture, we find that early development, which is captured in our regression analysis by three different measures –agricultural history, state history, and technology in 1500– both decreases the probability of being colonized and delays the date of colonization. Geographic proximity to Europe increases the likelihood and hastens the occurrence of colonization, while distance from the equator has the opposite effects. The role played by the disease environment is more complicated: a less favorable disease environment causes colonization to occur later in time, but its effect on the probability of being colonized is not significant, perhaps because initial disease barriers to colonization had been sufficiently lowered by medical advances before the age of colonization ended. The early development variables are jointly significant determinants of both the occurrence and timing of colonization, and the full model including also the geographic variables and disease environment has high explanatory power for both outcome variables.

The remainder of the paper proceeds as follows: Section 2 briefly discusses relevant literature. Section 3 sets out our hypotheses, empirical strategy, and the data to be used. In Section 4, we present the results of the Probit and Ordinary Least Squares (OLS) regressions that predict the occurrence and timing of colonization, and explore robustness to sample changes, IV estimation, and Heckman selection models. Section 5 concludes by summarizing our main findings.

2. Literature

Several influential papers on long run economic growth consider the impact of the colonial era. La Porta et al. (1999) emphasize the importance of the European origins of legal systems. Hall and Jones (1999) attribute large cross-country differences in productivity to differences in "social infrastructure," instrumented by the proportion speaking European languages. Sokoloff and Engermann (2000) argue that factor endowments were important in explaining long-run economic success in the Americas partly by determining the type of settlers and labor force drawn to different regions, a view that finds statistical support in Easterly and Levine (2003). Acemoglu et al. (2001, 2002) argue that differing types of institutions dating back to differing modes of European colonization account for much of the cross-country divergence in current incomes. Putterman and Weil (2010), Ashraf and Galor (2013), and Chanda et al. (2014) emphasize the impact of colonization on comparative development through the movement of people and resulting changes in population composition, while Easterly and Levine (2012) focus on the presence of Europeans in countries during their years as colonies. None of these papers attempt to explain why colonization occurred in some but not other non-European countries, or why its timing varied so widely.

As mentioned earlier, our approach to explaining colonization's differential occurrence and timing builds on ideas in the literature on the deep roots of comparative development. That literature, which includes Bockstette et al. (2002), Hibbs and Olsson (2004), Olsson and Hibbs (2005), Comin et al. (2010), Putterman and Weil (2010), Ang (2013), and Spolaore and Wacziarg (2013, 2014), provides evidence that differences in economic outcomes observed in the late 20th century are strongly predicted by differences in time of adoption of agriculture, early presence of macro-level polities, and levels of technology at least half a millennium ago.[1]

Diamond (1997) argues that it was differences in technological and organizational development attributable to the disparate timing and dissemination of the agricultural and subsequent urban revolutions on the different continents, along

[1] See also Alesina et al. (2013) and Nunn (2014).

with the susceptibility of indigenous people elsewhere to Old World diseases, that account for European colonization and the massive population shifts of the colonial era and its aftermath. The crux of his argument that differences in the distributions of wild plants and animals suitable for domestication, and other climatic and geographic factors, account for differences in the timing of adoption of agriculture, has been tested by Hibbs and Olsson (2004). The idea that the timing of adoption of agriculture accounts for much of the difference in level of economic development in the modern world has also been tested with supportive results by Olsson and Hibbs (2005), Putterman (2008), and Putterman and Weil (2010). But as noted in the introduction, we are aware of no research that statistically tests Diamond's explanation of colonization patterns across the globe.

A number of papers, including Grier (1999) and Feyrer and Sacerdote (2009), have investigated the impact of the duration of colonization on subsequent economic outcomes. In a clever design, Feyrer and Sacerdote use patterns of wind speed and direction to instrument for duration of colonization in 81 island nations. However, they do not attempt to explain colonization's occurrence (they study only colonies) or its timing (they study rather its duration), nor do they relate the outcomes studied to the historical factors on which we focus. Also, the set of countries studied by them accounts only for a small part of the colonized world: while we consider 111 countries that together account for 95.4% of the world's population outside of Europe, Feyrer and Sacerdote's island nations account for only 1.5% of that population.

We are aware of only one published study that includes a statistical analysis of which non-European countries were colonized. In a paper focusing on the effects of colonization on democracy, Hariri (2012) posits that for non-European countries with past histories of autocratic rule, colonization had a positive effect on current levels of political democracy thanks to its role in breaking up old autocratic social formations (see also Olsson, 2009). Hariri's hypothesis that pre-existing states are bad for current democracy because they made colonial conquest less likely is at one point tested using the Bockstette et al. state history measure which is also one of our three main proxies of early development.[2] However, Hariri interprets state history as an indicator of governmental capacity to mount an organized military defense, and thus focuses specifically on political history, whereas we view state history mainly as one among several plausible indicators of broad social and technological development. Moreover, his analysis does not consider the timing of colonization, and it leaves out the remaining determinants of colonization and of its timing that are included in our models, perhaps in part because the main dependent variable in his study is democracy, not colonization.

The goal of our paper is to test the hypothesis that pre-modern economic and technological development is an important determinant of which non-European countries were colonized by Western Europeans and when, while accounting also for how disease environment and other geographic considerations impacted colonization. Diamond (1997) argues that differences in levels of technological and social development associated with the timing of agricultural revolutions, as well as the diffusion of technical knowledge across landmasses at similar latitudes versus obstruction of such diffusion by latitudinal differences, deserts, and oceans, are the main factors explaining who was in a position to conquer whom beginning in the 15th century. The observation (supported by Maddison, 2001) that advanced Eurasian societies, including Europe but also China, Korea and Japan, Ottoman Turkey, Persia, and Moghul India, enjoyed similar levels of development around 1500, leads to our conjecture that those areas would have been much less likely to be colonized by Europeans before Europe obtained a decisive technological advantage over them.

Apart from differences in levels of development prior to 1500, other major factors that we hypothesize affect the occurrence and timing of colonization include distance from Western Europe, presence of land barriers, and disease burdens. Landes (1998) discusses the role of geographic proximity and accessibility in the colonization first of islands off of West Africa, then those in the Caribbean, then the parts of the American mainland and stopping points on the ocean route from Western Europe to India, and so forth. Relative proximity seems likely to help explain why the American east coast, and parts of Pacific-facing South America near the Isthmus of Panama, were colonized before most of the American west and interior, and likewise before Australia and New Zealand. Once coastal areas were reached, overland movement of men and arms constituted a substantial additional cost, and we take into account that accessing, e.g., Spain's Pacific-facing American colonies, while possible entirely by sea, was usually facilitated by a land crossing in Central America. We also control for latitude, both because currents favored European crossings to the Caribbean (Landes, 1998), and because semi-tropical areas were favored for their plantation potential (Sokoloff and Engermann, 2000).

The role of malaria and yellow fever in impeding European penetration of Africa and parts of Southeast Asia is also noted by Landes as well as Acemoglu et al. (2001). Although the disease factor cuts both ways in that the indigenous population's susceptibility to European diseases made conquest easier in many instances, we are unaware of a reliable index of the degree of such susceptibility, and we therefore abstract from the indigenous susceptibility side of the disease question in our analysis.

3. Hypotheses and data

Our starting point is the desire to test the conjecture that those non-European regions that were closer to Europe with respect to technology, agrarian history and history of political organization in the 15th century, were the ones least likely to

[2] The data from Bockstette et al. are also used in a precursor to our paper, Ertan (2007).

fall to European colonial expansion, and tended to do so later in the colonial era, if at all. Societies with large scale states having armies using technologies close to the Eurasian technological frontier of the early modern period, for instance the Ottoman, Safavid, Mughal, and Ming empires, were not ones that Europeans could easily dominate in 1500.[3] Some of these would fall later, however, as a growing technological gap with Europe opened. States alone were not enough, since state-level societies which lacked steel weapons, armor and gun powder, e.g. the Incas and Aztecs, readily fell to European conquerors. The territories of stateless societies in other parts of the Americas and Oceania were still easier for Europeans to acquire, given both technological gaps and small or absent macro polities at the time of European contact.

Relative levels of development in 1500 can be described in terms of a rough continuum. On one end are societies that relied on hunting and gathering, lacked state-level polities, and exhibited low population densities and the absence of cities (Australia, parts of Southeast Asia, parts of southern Africa, and parts of the upper Amazon River basin). On the other were ones practicing intensive agriculture and animal husbandry and having higher population densities, degrees of urbanization, and technological capacities (much of Europe, North Africa, the Middle East, Iran, South Asia, and East Asia). In between were ones having intermediate population density, state presence and technological sophistication (e.g., parts of West Africa, Mexico and the Andes, Indonesia and the Philippines).

As indicators of developmental status, we use three previously-studied measures, which are strongly correlated with one another (see Section 4.2) but may capture somewhat different aspects of what we call "early development." First, time elapsed since the transition to agriculture is used to capture the extent to which the practice of settled agriculture may have effected changes in organization, technology, and outlook. Rather than Hibbs and Olsson's measure, which classifies countries into nine global regions of agricultural spread and assigns to all within a given region the same origin date for agriculture—for example, Iraq and Ireland share a common transition year since Ireland was an ultimate recipient of the Fertile Crescent "agricultural package"—we use Putterman and Trainor's (2006) country-specific estimate of the number of centuries since transition to reliance on agriculture. In this data set, Ireland's transition is identified as occurring 5000 years before the present, versus Iraq's at 10,000 years BP. However, to capture level of development as of 1500 CE, we redefine *agyears* as years before 1500, measured in hundreds (hence Ireland has value 45, Iraq 95). We assign the value 0 to the two cases in which agricultural transition is dated to later than 1500 CE.[4]

Our second indicator of early development is the index of state presence, scale, and home-based character since 1 CE—dubbed *statehist*—originally compiled by Bockstette et al. (2002). It assigns to a present-day country a positive value for a given past half-century if its territory contained a polity at the macro or supra-tribal level, with a higher value if the polity was locally based rather than externally imposed, and if more of the present territory was under unified rule. As in Putterman and Weil (2010), we use the version of the index that covers the years 1–1500 CE with each previous half century discounted by an additional 5%. An extended version of the index, constructed by Borcan et al. (2014), covers the years 3500 BCE – 1500 CE. We have also considered this version, using a 1% backward time discount rate to ensure that states in the earlier millennia receive non-negligible weight. Since the basic analysis reported in Section 4.1 is very similar with both versions, and they have a very high correlation (0.86), we decided to use the version described at the beginning of this paragraph, which is already well-established in the literature.[5]

Our third and last measure of early development is the composite index of technologies including writing, plough use, firearms and steel in use by the population in 1500, assembled by Comin et al. (2010). The authors developed such indices for three years (1000 BCE, 1 CE and 1500 CE) and found that each of the three is correlated with year 2002 per capita income. The year 1500 index is a highly significant predictor of recent income even after addition of numerous controls including region dummies, and it is this index, referred to as *tech1500*, that is our direct technology measure.[6]

Geography and disease. Although the idea that societies less technologically advanced than Europe was in the 15th and ensuing centuries were more easily subdued by Europeans is our central focus, other factors also appear to have influenced who was colonized and when. The preoccupation of the early explorers with the spices of the semi-tropical "indies," the high potential for growing plantation crops like sugar in subtropical and tropical latitudes (see again Sokoloff and Engermann, 2000 and Easterly and Levine, 2003), and the wind- and current-influenced navigation routes crucial to sailing, caused some lands to be explored and colonized earlier than others (compare, e.g., Cape Verde [1461], Cuba [1511] and Mexico [1521] with New Zealand [1840], Fiji [1874] and Papua New Guinea [1884]). We control for tropical climate using

[3] Indeed, would-be colonizers were in a number of cases expelled by Asian powers; examples include the reclaiming of Oman and Zanzibar from the Portuguese by the Sultan of Oman in 1650, and the expulsion of Spanish and Dutch forces from Taiwan by a Chinese general in 1661.

[4] Those countries are Australia (1600 CE) and Mauritius (1638 CE). We checked all estimates using years before 2000 CE as the transition year in these and all other countries, finding no qualitative change on any estimate. Note that our *agyears* data includes one country for which the Putterman and Trainor data has no value: Fiji, which we thought useful to include given under-representation of Oceania in the sample and Fiji's relatively large population among island nations of that region. We assume agriculture arrived with Austronesian settlers around 1500 BCE, since Encyclopedia Britannica says human habitation dates "at least" to that year and it is consistent with the estimate in Diamond (1997) that the Austronesian settlement wave reached Fiji from the Solomon Archipelago (estimated arrival date 1600 BCE) and Santa Cruz, and reached Samoa from Fiji (estimated arrival date 1200 BCE).

[5] Although we construct *statehist* in the same way as Putterman and Weil, i.e. covering the period 1–1500 CE and using a 5% discounting rate, we use as the underlying data the updated information of Borcan et al., which includes small changes for some early CE period years for a small subset of countries.

[6] Two other measures of pre-1500 development used in influential studies are the urbanization rate (see, e.g., Acemoglu et al., 2002) and the estimated level of population density in 1500 (see, for instance, Ashraf and Galor, 2013, and for all measures mentioned in the text and this note, Chanda et al., 2014). We eschewed using urbanization because of its availability for too few countries. We chose not to focus on population density due to measurement and other issues, but do include it in robustness tests (see below).

Table 1
Descriptive statistics.

Variable	Mean	Standard deviation	Minimum	Maximum
col	0.8288	0.3784	0	1
colyr	1777.39	152.29	1462	1936
Ln*colyr*	7.4791	0.0890	7.2876	7.5684
agyears	37.0405	25.1452	0	100
statehist	0.2870	0.3230	0	1
tech1500	0.3802	0.2709	0	0.883
latitude	19.9016	12.6755	0.228	48.19
navdist	6.5289	3.7316	0.965	14.054
landdist	0.1363	0.2702	0	1.209
malaria ecology	5.5609	8.1888	0	32.203
EDE	0.2867	0.8973	−2.1843	2.6054
Ln*pd1500*	0.5967	1.4123	−3.8170	3.8424
landlocked	0.2432	0.4310	0	1
biogeography	36.1155	34.5690	6.4706	100

latitude. For distance from Europe, we use navigation distance (*navdist*) from a port centrally located among those used by the main colonizing powers—Camaret-sur-mer, located at the northwestern tip of France—to control for effective distance from Western Europe at the time of colonization. Distances are calculated using routes appropriate to the era prior to opening of the Suez and Panama canals.

Areas deep in the hinterlands of continents, for instance Afghanistan or Mongolia, were not directly encountered by naval exploration and were far more costly to reach with armed personnel and equipment, given the greater cost of overland travel. We create a variable, *landdist*, to which we assign the value 0 for coastally accessed countries like Haiti and India. For landlocked countries like Mongolia, *landdist* is the number of kilometers of overland travel from the nearest port to the county's significant urban center nearest the sea. *landdist* also takes nonzero values for some countries that have sea coasts but that were accessed by colonizing powers mainly via land or river routes—e.g., Jordan, which has a Red Sea port but which from a European standpoint during the relevant period was seen more as a landlocked region reachable from the Mediterranean. Other examples are Sudan, usually reached via the Nile rather than the Red Sea, and El Salvador, Ecuador, Peru, and Chile, which were more often reached by routes that included a land crossing around Panama than by circumnavigating South America's southern tip.[7]

That the “scramble for Africa” did not take place until nearly four centuries after Spain began colonizing the New World is often attributed to the hazards posed by Africa's disease environment. Malaria and yellow fever have also been credited with discouraging European settlement elsewhere, such as New Guinea.[8] When controlling for disease, it is important to avoid endogeneity—seeing fewer deaths or cases of disease due to prevention and care made possible by higher income and more advanced technology. Two measures seem suitable, by this criterion. The first is the now widely used *malaria ecology* variable, constructed to indicate the climatic potential for malaria. The second is the more general Early Disease Environment (*EDE*) measure developed by Auer (2013) using data on colonial era non-combat soldier mortality and several dimensions of climate to predict “the logarithm of the annualized probability of death for European males in the age cohort of soldiers.” Auer argues that *EDE* is the better of the two measures when considering both colonized and non-colonized countries because *malaria ecology* displays little variation among the non-colonies. This, plus its coverage in principle of diseases other than malaria, makes it preferable for our purposes, although we report on some robustness tests which use the *malaria ecology* variable and obtain similar results.

Regarding which countries should be treated as having been colonized and when, judgments are unavoidable due to the existence of gray areas such as whether being indirectly ruled or being deemed a protectorate constitutes colonization, and whether the country is a colony as soon as the eventual colonizer has a coastal toehold. We consider colonization to have begun once 20% or more of a country's territory (using year 2000 boundaries) is deemed by sources to have been largely under the control of the colonizing power, provided that the majority of the territory would eventually be controlled by that or by a subsequent European colonizer. We developed our own data for both having been colonized ($col=1$) and year of colonization (*colyr*) from various sources. Part I of the Supplementary Online Appendix lists the countries in our sample by year of colonization or never-colonized status, and explains the basis on which our decisions were made in cases in which some judgment is required. Table 1 provides descriptive statistics for all of the variables, and the Data Appendix gives brief

[7] We apply the same treatment to Bolivia, which early in its history included a small coastline, and similar treatment to Georgia, which while in principle accessible through the Mediterranean and Black Seas, was in practice reachable by Europeans only over land during most of the age of colonization, due to Ottoman control of the passage between the two seas. For the Pacific-facing cases, *landdist* is the land distance crossed on the Isthmus of Panama, in km., while the distances for Georgia, Jordan and Sudan are calculated, like those for landlocked countries, by measuring the land distance from a relevant port to the country's nearest significant urban center.

[8] Marcus (2009, p. 41) writes, “Malaria … interfered with European colonization in parts of Southeast Asia. For example, malaria was well established in New Guinea, especially in the lowland areas. It inhibited European settlement there.” New Guinea resembles Africa in that it took centuries after landings on its coasts before Europeans saw areas further inland. For a broader discussion of the role of disease in world history, see McNeill (1998).

Table 2
Averages of early development indicators for groups of non-European and colonizing countries.

	Colonized before 1842	Colonized in 1842 or later	Never colonized	Colonizing countries
agyears	24.96	36.11	63.87	67.75
statehist	0.143	0.297	0.546	0.683
tech1500	0.217	0.415	0.610	0.938
number of countries	38	54	19	8

Note: the colonizing countries are Britain, Belgium, France, Germany, Italy, Portugal, Spain, and the Netherlands. The information in Part I of the Supplementary Online Appendix guides the classification of countries into the other categories. Both *statehist* and *tech1500* are normalized to take values in the [0, 1] interval. *agyears* is the number of years, in hundreds, since the first reliance on cultivated food for subsistence within a country's territory.

descriptions and source notes on each variable. Several of our robustness tests consider sensitivity of our results to our main case treatments of difficult-to-decide cases, including Ethiopia, Taiwan, and the Levant countries.[9]

We assembled data for all non-European countries with populations of over one-half million for which information on the variables of interest is available. In our sample, we have a total of 111 countries, consisting of 92 non-European countries which were colonized for some time by Western European countries and 19 that were not colonized by those countries (Afghanistan, Armenia, Azerbaijan, China, Georgia, Iran, Japan, Kazakhstan, South Korea, Kyrgyzstan, Liberia, Mongolia, Nepal, Saudi Arabia, Taiwan, Thailand, Turkey, Turkmenistan, and Uzbekistan).[10]

Note that countries in Central Asia and the Caucasus that were colonized by Russia and later incorporated into the Soviet Union are not European colonies in the sense of this paper, and we likewise treat countries that emerged from the Ottoman Empire as if they were not colonies until ruled by Britain, France or Italy. Since we cannot rule out the possibility that some of these countries might have been colonized (or colonized earlier) by the European powers had the Ottoman and Russian empires not existed, we perform the robustness checks of estimating our models on restricted samples: without the non-European countries of the former Soviet Union (FSU), namely Armenia, Azerbaijan, Georgia, Kazakhstan, Kyrgyzstan, Turkmenistan, and Uzbekistan; without the "Levant" countries Lebanon, Syria, Israel, Jordan and Iraq, which were among the last countries to pass into Western European hands; and without countries from either group. We also experiment both with dropping and with treating as never colonized Ethiopia, a country colonized exceptionally late (1936) and for an exceptionally brief period (five years). And we experiment both with dropping and with treating as colonized Taiwan, an island that had limited Spanish and Dutch settlements in the mid-1600s before the arrival of most of its current population's ancestors from mainland China, and that we accordingly consider never colonized in our main analysis.[11]

4. Colonization and its timing: results

4.1. Single variable comparisons

A simple exposition of our idea that early development affected colonization's occurrence and timing is facilitated by Table 2, which compares the average levels of development in 1500, according to our different indicators, for four groups of countries: the non-European countries in our sample that were colonized before 1842, those that were colonized thereafter, those that were never colonized, and, finally, the 8 European colonizing powers. The division into "early" and "late" colonized (made for this exercise only) is marked by the British takeover of Hong Kong in 1842, which puts 38 countries colonized over 380 years into the "early colonized" set and 54 countries colonized over 94 years (from 1842 to the final case we consider, Ethiopia's 1936 takeover by Italy) into the "late colonized" set. We chose that division line because it roughly divides the colonial era into epochs before and after the qualitative transformation of the West's technological lead by the industrialization process, symbolically coinciding with the wresting of a territory's control from once mighty China by a much smaller European power, Britain.

[9] An interesting suggestion by one of the reviewers is that colonization might be treated as a continuous rather than dichotomous variable. However, many alternative approaches to ordering such a continuum can be imagined, so much judgment would doubtless still be required.

[10] Our sample of colonized countries includes one, Cape Verde, the colonization of which began prior to 1500 (specifically, in 1462 as mentioned above), raising the issue of a potentially reverse temporal relationship from colonization to indicators of early development measured as of 1500. We believe those impacts—arrival of agriculture with the colonizers in 1462 and state presence in the last 38 of the 1500 years covered by *statehist*—to be too minor to justify dropping Cape Verde from our sample. Cape Verde's *statehist* value is 0.03 due to its pre-1500 colonization, whereas it would otherwise have been 0.00; 0.03 is the third lowest positive value observed, little different from the 43 zero values and far below the 0.47 average among the 68 sample countries having positive *statehist* values. In any case, our results are robust to dropping Cape Verde from the sample (see footnote 17).

[11] Taiwan's colonization by Japan from 1895 to 1945 is treated in the same fashion as that of South Korea, namely as not pertinent, given our paper's focus on colonization by Western European powers. We do not perform the same tests for South Korea and Taiwan as a group that we perform for the Levant and formerly Soviet-ruled countries because, in comparison to the Ottoman and Russian empire cases, it seems less likely that Taiwan and South Korea were delayed in or prevented from becoming Western European colonies mainly due to Japan's influence (see further discussion of Taiwan in the text, below). Relatedly, we do not report tests dropping also the North African countries that were for a time controlled by the Ottoman Empire, due to weaker Ottoman control and earlier European colonization of these than of the Levant.

For the non-European countries, the average values for all three indicators of pre-1500 development line up precisely as expected: late-colonized countries have longer agricultural histories, more state history, and higher indices of technology, than do early-colonized ones, and never-colonized countries in turn exceed late-colonized ones on the same three measures. Corresponding *t*-tests find all paired differences between late- and early-colonized and between never- and late-colonized countries to be significant at the 5% level, the majority being significant at the 1% level as well. More conservatively, we can apply a non-parametric test, the Mann–Whitney (hereafter, MW) *U* test, and this confirms the same result with the exception of the late- versus early-colonized countries difference for years of agriculture (p-value $\approx$ 0.102 in two-tailed test; p-values for most of the tests mentioned here are reported in Table S.1 in the Supplementary Online Appendix). When early and late colonized countries are combined into a single group, the differences in agricultural history, state history, and technology in 1500 between colonized and never colonized non-European countries are all significant at the 1% level in both *t*-tests and MW tests.

Colonizers themselves score higher on average for the three measures than do both early and late colonized countries. All differences between colonizers and colonized countries are significant at the 1% level by both *t* and MW tests, and the same holds for the difference between colonizers and all colonized countries (pooling early and late colonized ones).[12] The differences between colonizers and never-colonized countries are interesting, since as pointed out some non-colonized countries including China, Japan and Turkey are believed to have scored similarly to many European countries on indicators of development not long before the colonial era (Maddison, 2001; Morris, 2010). This is consistent with the fact that the differences between the 18 never colonized countries and the 8 colonizing countries are not statistically significant for *agyears* (according to both *t* and MW tests) nor for *statehist* according to the MW test (though the *t*-test indicates the difference is significant with p-value $\approx$ 0.07). Among the three early development indicators, *tech1500* is the only one for which differences are significant according to both tests. Even with respect to that measure, differences between colonizers and some never-colonized countries are small: China's *tech1500* index of 0.883, for instance, is close to that of the Netherlands, Belgium, Germany and Italy, which share value 0.900.

The idea that the European countries first colonized the weakest non-European areas, then somewhat stronger ones, and did not colonize at all the non-European countries with the highest levels of pre-modern development, thus finds substantial support for our individual measures of pre-1500 development. Our idea that an increase in the colonizing powers' lead over once-similar Old World countries facilitated the colonization of more of the latter at the end than at the beginning of the colonial era is also consistent with the available data on the evolution of comparative development over the colonial period from Maddison (2001). While too limited to be considered in our regression analysis, those data do permit a basic illustration. The ratio of average income per capita in the eight colonizing powers to income per capita in non-European countries rose from 1.43 in 1500 to 1.74 in 1600, to 2.06 in 1700, to 2.17 in 1820, and to 2.99 in 1870. A marked upward trend also appears when we consider specifically the three late-colonized countries for which Maddison provides data (Egypt, Iraq, and Morocco): income per capita in the colonizing powers was 1.58 times as high as in those three countries in 1500, then 2.11 times as high in 1600, 2.35 in 1700, 2.54 in 1820, and 3.11 in 1870.[13] As suggested by the work of Acemoglu et al. (2005), the widening of the lead of colonizing powers within the Old World may well have been aided by the resource extraction and market expansion occasioned by their early colonizing activities.

4.2. Baseline multivariate regressions

In Table 3, we begin to report our multivariate analysis by looking at the determinants of which of the 111 non-European countries in our sample were colonized. For this binomial dependent variable, we considered Probit and Logit regressions as well as a Linear Probability Model (LPM), choosing to display Probit results for purposes of consistency with the Heckman selection model discussed later; Logit and LPM specifications yield qualitatively similar results (see Tables S.2 and S.3 in the Supplementary Online Appendix). We report six specifications, all but one of which include the disease measure *EDE*, *navdist*, *landdist*, *latitude*, and at least one measure of pre-modern development. The regressions of columns (1)–(3) each contain only one such development indicator, while those of columns (4) and (6) include three of those indicators (respectively) without and with the accompanying variables, and column (5) includes only those other explanatory variables. Including the three indicators in (4) provides a sense of their explanatory power relative to that of the variables other than the early development measures included in (5) and relative to that of the full model in (6). But since the three indicators are very highly correlated,[14] results from specifications (4) and (6) may not provide reliable estimates of the effects of each particular indicator nor do they conclusively indicate which one was most relevant.

[12] The results are qualitatively the same when we use the index of state history extended by Borcan et al. (2014) to cover the years 3500 BCE – 1500 CE (with a 1% discount rate), the sole exception being that the difference between colonizers and late colonized countries is significant by both tests at the 10% level only.

[13] Income per capita for these three late-colonized countries as well as for the eight colonizing powers are computed as unweighted averages of country-level income per capita figures from Maddison (2001). Income per capita for non-European countries is calculated after subtracting total income corresponding to 30 Western European countries and the 7 Eastern European countries located outside of the former Soviet Union from total world income (and likewise for population). Thus, the figures for non-European income per capita are effectively a population-weighted cross-country average that includes the European portion of the former Soviet Union, which cannot be separated from the Caucasian and Central Asian countries in Maddison's data for the period of interest. The figures are very similar if total income and population for all former Soviet Union countries are also netted out.

[14] The correlation of *agyears* and *statehist* is 0.643, that of *agyears* and *tech1500* is 0.743, and that of *statehist* and *tech1500* is 0.792.

Table 3
Determinants of colonization. Dependent variable: *col.*

	(1)	(2)	(3)	(4)	(5)	(6)
statehist	−1.168**			−0.716		−0.824
	(0.498)			(0.697)		(0.646)
tech1500		−1.204**		0.0183		0.661
		(0.607)		(1.006)		(1.012)
agyears			−0.0202***	−0.0250***		−0.0195**
			(0.00597)	(0.00919)		(0.00825)
EDE	0.154	0.168	0.289		0.0689	0.284
	(0.206)	(0.213)	(0.228)		(0.223)	(0.233)
navdist	−0.116**	−0.118**	−0.129**		−0.138**	−0.124**
	(0.0584)	(0.0591)	(0.0623)		(0.0603)	(0.0595)
landdist	−1.953***	−1.994***	−2.021***		−1.935***	−1.998***
	(0.637)	(0.672)	(0.696)		(0.689)	(0.672)
latitude	−0.0471**	−0.0451**	−0.0280**		−0.0590**	−0.0284*
	(0.0205)	(0.0194)	(0.0142)		(0.0229)	(0.0146)
constant	3.726***	3.784***	3.852***	2.368***	3.704***	3.819***
	(1.095)	(1.104)	(1.078)	(0.321)	(1.070)	(1.069)
Observations	111	111	111	111	111	111
McFadden's pseudo R^2	0.499	0.490	0.524	0.266	0.459	0.531

Probit regressions, robust standard errors in parentheses.
*** $p < 0.01$.
** $p < 0.05$.
* $p < 0.1$.

When entered individually in columns (1)–(3), each pre-modern development indicator obtains a negative and significant coefficient, supporting our hypothesis that having experienced greater pre-modern development tended to ward off colonization. The coefficients for *statehist* and *tech1500* are significant at the 5% level, that for *agyears* at 1%. The marginal effects of early development measures computed at the means of regressors provide some sense of the magnitude of the effects: an increase of one standard deviation in early development is associated, for specifications (1), (2), and (3), respectively, with a reduction in the probability of being colonized of 5.1%, 4.7%, and 6.4%. When all three measures are included, in columns (4) and (6), only *agyears* is individually significant, at the 1% and 5% levels.

Turning to the other variables, across specifications the coefficients on *landdist* are negative and significant at the 1% level, those on *navdist* and *latitude* are negative and significant mainly at the 5% level (with one exception, significant at 10%). *EDE* obtains consistently positive but insignificant coefficients. The McFadden Pseudo R^2 is 0.266 for the model including only early development variables, while for the model with the other regressors it is 0.459, and for the full model in column (6) it reaches 0.531. The full model has high explanatory power, yielding a Count R^2 of 0.937 and an Adjusted Count R^2 of 0.632.[15]

In Table 4, we use the same sets of explanatory variables to explain not whether but when colonization occurred. Our sample thus includes observations only for the 92 countries of our sample that were subjected to Western European colonization at some time between 1463 and 1936. We use the natural logarithm of colonization year as our dependent variable to reduce the influence of extreme values on our estimates, but the results are qualitatively the same if we use the year of colonization without any transformation (see Table S.4 in the Supplementary Online Appendix).

Looking first at the measures of early development that are our main focus, columns (1)–(3) show that each of the three when entered along with the geographic and disease controls obtains a positive coefficient, as predicted. Significance ranges from 10% for *statehist* to 5% for *agyears* and 1% for *tech1500*. For the specifications in which all three variables are included, only *tech1500* obtains a statistically significant coefficient, with its significance remaining at the 1% level in both columns (4) and (6). The estimated coefficients in specifications (1)–(3) imply that a one-standard-deviation increase in early development is associated with an increase of 0.17–0.35 standard deviations in the outcome variable, depending on the measure of early development (the standardized coefficient is highest for *tech1500*, the indicator that remains significant in columns (4) and (6)). The R^2 of the column (4) model suggests that the combination of early development measures can explain about 22% of the variation in colonization's timing.

Regarding the geographic and disease controls, we find that *navdist* appears to significantly delay colonization in some estimates, consistent with the intuitions based on examples mentioned in Section 2. However, *navdist's* coefficient becomes insignificant, though remaining positive and similar in magnitude, when *tech1500* is added in specifications (2) and (6). In all columns, we find positive effects of *landdist* and *latitude* on timing, significant at the 5% or 1% level. These variables, which tend to make colonization less likely, also tend to delay its occurrence. One variable not significant in explaining

[15] The Count R^2 is the fraction of correctly predicted outcomes considering the predicting outcome to be 1 (0) when the predicted probability is above (below) 0.5. The Adjusted Count R^2 is the fraction of correctly predicted outcomes beyond a baseline that predicts for every observation the most frequently observed outcome.

Table 4
Determinants of the timing of colonization. Dependent variable: Ln*colyr*.

	(1)	(2)	(3)	(4)	(5)	(6)
statehist	0.0464*			−0.0699		−0.0525
	(0.0275)			(0.0494)		(0.0446)
tech1500		0.116***		0.249***		0.176***
		(0.0315)		(0.0597)		(0.0559)
agyears			0.000772**	−0.000470		−0.000161
			(0.000317)	(0.000395)		(0.000428)
EDE	0.0578***	0.0480***	0.0577***		0.0611***	0.0455***
	(0.0116)	(0.0107)	(0.0113)		(0.0109)	(0.0105)
navdist	0.00596**	0.00391	0.00616***		0.00698***	0.00362
	(0.00232)	(0.00245)	(0.00225)		(0.00223)	(0.00256)
landdist	0.0949**	0.103***	0.0971***		0.0934**	0.106***
	(0.0363)	(0.0334)	(0.0368)		(0.0357)	(0.0321)
latitude	0.00253**	0.00194**	0.00239**		0.00299***	0.00203**
	(0.000980)	(0.000875)	(0.000972)		(0.000920)	(0.000830)
constant	7.357***	7.355***	7.345***	7.427***	7.353***	7.353***
	(0.0273)	(0.0256)	(0.0273)	(0.0160)	(0.0263)	(0.0271)
Observations	92	92	92	92	92	92
R^2	0.326	0.391	0.338	0.221	0.303	0.404

OLS regressions, robust standard errors in parentheses.
*** $p < 0.01$.
** $p < 0.05$.
* $p < 0.1$.

colonization's occurrence has a significant impact on timing: higher disease mortality as measured by *EDE* is significantly associated with later colonization, as anticipated. Overall, the regressions appear to explain somewhere under half of the variance in colonization’s timing, with a maximum R^2 of 0.404.

4.3. Robustness to alternative sample compositions, country classifications, and measures

In this section, we discuss estimates of the specifications of Tables 3 and 4 for samples excluding the late-colonized Levant countries, the never-colonized FSU countries that emerged from the U.S.S.R. in 1991, or both, as well as inquiring into whether our results hold if we make alternative determinations about the colonization status and timing of Ethiopia and Taiwan, and whether they hold if Africa and Asia (the Old World) only are considered. To conserve space, the regression results discussed here are shown in the Supplementary Online Appendix.

4.3.1. Estimated effects of early development with alternative samples and classifications

We test the effect of removing either our five Levant countries, seven FSU countries, or both, because it is plausible that each might have experienced different colonization outcomes but for historical circumstances associated with the Ottoman and Russian empires and the successor Soviet Union. Some studies of colonization do not treat the Levant countries as late colonies of Britain and France because their period of colonial rule was short and occurred formally under the rubric of League of Nations mandates. The ex-Soviet states were routinely left out of comparative development studies until the past decade or so because of incomparability between Soviet-era statistical systems and non-Communist international measures.

We can think *a priori* of several possible ways in which conclusions we draw from our full sample might be importantly influenced by the way we treat these two sets of countries. The Levant countries have the world's longest experience of agriculture as well as of states, although *statehist* values per se (which give greater weight to experience under home-based states) are lowered by lengthy periods under Hellenistic, Roman, Byzantine, Ottoman and other rulers. We want to check whether inclusion of these countries is driving significance of *agyears* and *statehist* in Tables 3 and 4. Most of the FSU countries also obtained domesticated crops and animals from the Middle East fairly early and many experienced some early state presence. Conceivably, our inclusion of the Levant countries as late colonized countries works to undermine predictions that early development of agriculture and states reduced the likelihood of colonization, while strengthening predictions that it made colonization later if it occurred. Inclusion of the FSU countries as never-colonized conceivably strengthens the case that early agriculture and states deters colonization, although neither *agyears* nor *statehist* is exceedingly high in them. Because many FSU countries are distant from ocean outlets, their failure to have been colonized could also be impacting the coefficient on *landdist*.

Tables S.5–S.7 and S.12 show that our qualitative results regarding the effects of early development on both colonization and its timing are not especially sensitive to whether either set of countries is included in our sample. As anticipated, indications of the impact of early development on the likelihood of being colonized are strengthened by dropping the Levant countries (the Pseudo R^2 of specification (4) rises from 0.266 to 0.455), whereas indications of the impact of early development on colonization's timing are marginally weakened (the R^2 of the corresponding specification falls from 0.221 to 0.209).[16] Dropping the FSU countries impacts only the *col* regressions, somewhat lowering the values of the Pseudo R^2, especially for specification (5), which includes only variables other than the early development measures. Both *statehist* and *agyears* are significant at the 5% level, rather than *agyears* alone being significant at 1% level, in specification (4) without FSU. While the dropping of the Levant and FSU observations pull in opposite directions with respect to the *col* regressions, the result of dropping all twelve countries is, like that of dropping Levant alone, basically favorable to the hypothesis of early development deterring colonization, with the same significance cut-offs and higher Pseudo R^2 (0.367 rather than the base sample's 0.266) in specification (4).

There are also some *a priori* reasons for checking whether our decisions on how to treat Ethiopia and Taiwan affect our results on early development and colonization. Ethiopia has the world's longest record of unbroken state history during 1–1500 C.E., according to our data. Thus, our decision to treat the country as having been colonized in 1936 is likely to be unfavorable to the early development conjecture, at least with respect to the *statehist* measure, while being favorable (through that channel) to early development's tendency to retard colonization. As for Taiwan, if we were to apply to that country the high *statehist* and *agyears* values of mainland China, our decision to treat it as not colonized in 1624, but rather as never colonized (Japanese colonization not fitting our definition) could be suspected of aiding both the conjecture on deterring and that on delaying colonization. However, Taiwan history before 1500 is quite unlike China's: it had no supra-tribal government, and practice of agriculture on the island has not much more than half the history of agriculture in China. Taiwan also has the low *tech1500* value of other "Austronesian" islands like Fiji, since Han Chinese migration to the island was still in its early stages in 1500. What impact to expect from our coding of Taiwan for colonization is thus somewhat unclear.

In the Supplementary Online Appendix, we report alternative estimates of our full model, corresponding to column (6) of Tables 3 and 4, in which we drop Ethiopia, Taiwan, or both, or apply an alternative assumption regarding their colonization (see Tables S.8 and S.13). We find that treating Ethiopia as never colonized would slightly strengthen our findings, causing the coefficient on *statehist*, not only that on *agyears*, to be significant at the 5% level. Simply dropping Ethiopia from the sample has little impact in either the *col* or the Ln*colyr* regressions. Excluding Taiwan from the sample or treating Taiwan as having been colonized in 1624 has no effect on the early development impacts on Ln*colyr*, but slightly weakens those for *col*, lowering the significance of the coefficient on *agyears* to only the 10% level.[17]

Testing our conjecture when leaving out the "New World," defined here as the Americas and Oceania, strikes us as a demanding challenge for our main thesis about early developmental disadvantages and colonization. Developmental gaps between Europe and the New World regions tended to be large because of the millennia-long absence of technological diffusion, whereas the Old World, especially Asia and North Africa, include many societies sharing similar technological levels with Europe around 1500. Our non-European sample contains 82 Old World and 29 New World countries, of which 63 Old World countries and all 29 New World countries were colonized. The fact that 100% of the inhabited regions that had lacked contact with Eurasia for millennia versus only 77% of regions which had had such contact were ultimately colonized, is in itself *prima facie* evidence for our theme. But is inclusion of New World countries in our sample critical to our conclusions?

The results in Tables S.11 and S.16 answer mostly in the negative. Table S.11 shows that the power of our early development variables as a group is little different for explaining which countries were colonized when we restrict our attention to the Old World only. While columns (1) and (2), in which only *statehist* or only *tech1500* enter, show different estimated coefficients on those variables, overall explanatory power and the pattern of coefficient signs and their significance levels are the same for specifications (4) and (6). Models for timing of colonization in Table S.16 perform rather differently in the Old World sample, in that it is *agyears* that strongly predicts timing of colonization, a job performed instead by *tech1500* in the full sample. Evidently, the difference in timing of adoption of agriculture in many African as opposed to most Asian countries predicts better the later colonization process in Africa than do differences in technologies recorded by Comin *et al.*'s sources. And while the coefficient on *agyears* displays significance at the 1% level in specification (4) for the Old World sample, much less of the overall variance ($R^2 = 0.0497$) is explained by the model. Still, both the *col* and the Ln*colyr* results in the Old World only sample can be viewed as supportive of our early development hypothesis, especially a simple version that emphasizes the timing of transition to agriculture: within Africa and Eurasia, the earlier was the transition to

[16] That small overall weakening is accompanied by a rise of the coefficient on *agyears* to be significant at the 10% level with "wrong" sign, but the coefficient on *tech1500* barely changes and remains significant at the 1% level.

[17] Another country the coding of which for colonization has been debated is Liberia, which we treat as never colonized but which Auer (2013) considers to have been a U.S. colony from 1820 to 1847. As explained in Part I of our Supplementary Online Appendix, Liberia fails to meet our criteria for colonization because the American Colonization Society that organized the settlement in Liberia of freed African slaves (few of whose ancestors were local to Liberia) controlled too little of the territory of present-day Liberia during those years when it came closest to having formal U.S. sponsorship. We nevertheless checked whether treating Liberia as having been colonized in 1820 significantly alters any of our results, and confirmed that it does not do so, both when only Liberia is added to the list of colonized countries and when Liberia and Taiwan are simultaneously added to that list. A minor exception is that when only Liberia is added to the list of colonized countries (Table S.9), *EDE* obtains a positive coefficient that is significant at the 10% level in two of the regressions for *col*. See also Tables S.10, S.14 and S.15. Finally, we checked that inclusion of Cape Verde, the earliest colonized country in our sample, is not crucial for any results. We estimated variants of Tables 3 and 4 regressions in which we exclude Cape Verde and found no qualitative changes except that in model (1) of Table 4, the estimate of the coefficient on *statehist* has $p = 0.125$, short of its significance at the 10% level in the baseline specification.

agriculture on the territory of what is a country today, the less likely was it to be colonized by Europeans, and the later it was colonized if colonization occurred.[18]

4.3.2. Estimated effects of geographic and disease variables with alternative samples and classifications

In discussing how the other independent variables of our model perform in the alternative samples, we begin with the disease measure *EDE*, which lacks significant effect on colonization but significantly delays colonization's timing in our main sample. *EDE*'s lack of effect on colonization is unchanged when we drop the Levant and/or FSU countries, and is also unaffected by how we treat Ethiopia and Taiwan. The positive coefficient on *EDE* in the *col* regression becomes significant at the 5% level, however, in full specification (6) for the Old World only sample. Compared to other countries in Africa and Asia, those with the worst disease environments were more likely to be colonized eventually, which seems intuitive given that the 19 never-colonized countries include none of those beset by tropical and sub-tropical diseases like malaria and yellow fever, and that sub-Saharan and southeast Asian countries had both lower levels of population density, urbanization, and state development, and harsher disease environments. Thus, the delaying effect on colonization due to a harsh disease environment is present only when the Americas and Oceania cases are included.

Consider next the importance of distance from Western Europe. The variable *navdist* has a robustly negative and significant effect on the likelihood of being colonized in almost every specification and sample, including Old World only. Its effect on the timing of colonization is more sensitive to sample composition, however. In the full sample, the effect of *navdist* on Ln*colyr* is consistently positive but becomes insignificant in full model (6). Table S.12 shows that the delaying effect of *navdist* on colonization is significant with a higher confidence level when the Levant countries—late to colonization but not so far from Western Europe—are left out. When the sample is restricted to Old World countries (including the Levant), however, the coefficient on *navdist* changes sign and is significant in some specifications, although not (6). Considering Africa and Asia only, being closer to Europe did not make colonization earlier, controlling for other factors. This conclusion is consistent with the fact that South Africa, Mozambique, Indonesia, India and Sri Lanka are all relatively further from Western Europe than are parts of North Africa, the Middle East, and West Africa that were colonized later. Both Ottoman influence in the Middle East and the discouraging disease environment of West (as with most of sub-Saharan) Africa may have contributed to this outcome.

In the main sample, the overland distance required to access a landlocked country or one reached in practice by crossing land (*landdist*) shows a consistently significant negative impact on the likelihood of being colonized and a positive effect on the timing of colonization, if it occurred. The main news about this variable from the sample robustness tests is that the significance of the effect of land distance on whether colonization occurred at all is quite sensitive to whether the FSU countries are in the sample, as expected. The coefficients on *landdist* remain negative but are insignificant without FSU. An alternative interpretation might therefore be suggested: perhaps only the power of the land-based Russian empire, not being landlocked or requiring land passage *per se*, discouraged colonization. However, we cannot rule out that the logistical and transport considerations that led us to include *landdist* in our models also played a part in the non-acquisition of the landlocked FSU countries by Western European colonizers. That is, even if Russian rule was a major factor discouraging their colonization by powers such as Britain, the more complicated logistics of a campaign to wrest control of places like Turkmenistan and Tajikistan from the Russian sphere due to their inland locations (and perhaps also their lower value to colonizers due to their landlocked geography[19]) may have contributed to their not being colonized by Western countries. Apart from this, *landdist*'s delaying of the *timing* of colonization remains significant in all other samples, including the Old World only one.

Turning finally to *latitude*, its coefficient remains consistently negative in all estimates of effects on *col*, but tends to lose significance when the FSU countries are excluded, and also is insignificant in full model (6) for the Old World only sample. The effect might still be considered fairly robust, since the coefficient's *p*-value is not so different (although above the 10% threshold). But *latitude*'s significance in (6) is in any case always marginal in the *col* regressions; moreover, the estimated coefficient is not significant at conventional levels if Ethiopia is treated as never colonized or if Ethiopia and Taiwan are dropped from the sample.

The apparent effect of *latitude* on colonization's timing is even more sensitive to the sample used. The significant positive effects seen in Table 4 are robust to dropping the Levant countries and to different treatments of Ethiopia and Taiwan, but when attention is restricted to the Old World, there is a sign change, with the now negative coefficient being significant in all specifications but (6). Thus, while greater latitude delayed colonization in the non-European world as a whole, looking at

[18] We have also considered regressions with continental fixed effects. The main result about the relevance of early development continues to hold, although not as robustly across specifications: for both outcomes under consideration, we find that at least one of the three early development indicators is significant in all specifications where it is included, although when included individually not all early development indicators appear significant, in contrast to our baseline specifications. It is important to note that within-continent variation is limited, particularly for colonization status. All of our sample countries in the Americas and Oceania were colonized, and thus those 29 observations are dropped from the *col* equation when we include continent fixed effects. Thus, this specification is similar to that of the Old World sample, except that with continent dummies the results are not affected by differences between Asia and Africa but rather depend exclusively on variation within Asia and within Africa.

[19] Historians treat the "closing of the steppe" by Russian and Chinese conquests in the centuries following the Mongol expansion as being more defensively than acquisitively motivated. Control over regions such as Xinjiang and Kazakhstan probably had far less value to Western European colonizers than they did to these regional empires which had suffered from steppe nomadic invasions for centuries.

Africa and Asia only it is associated with *earlier* colonization—a result likely to be attributable to late colonization of most sub-Saharan countries.

4.3.3. Estimation with alternative measures of early development, geography and disease environment

We also check the robustness of our core results to small differences in the choice of early development, geographic and disease measures. First, we estimated models identical to our main model except for the use of the log of estimated population density in 1500 (Ln*pd1500*) as a measure of early development. Models of development in the very long run such as Galor and Weil (2000) predict that prior to the industrial revolution, population growth offset technological improvements, so technological progress is better measured by population density than by per capita income; Ashraf and Galor (2011) provide supportive evidence.

While we chose not to use population density as one of our main measures for development prior to 1500 for various reasons, including the facts that *tech1500* can be considered a more direct measure of the level of technology and that the widely used population estimates based on McEvedy and Jones (1978) are imprecise and are in many cases region-based so that assumptions are needed to disaggregate to country level, we nonetheless judge using estimated population density as an alternative or additional measure of early development to be a sensible robustness check. We tried using Ln*pd1500*, as calculated by Ashraf and Galor, as sole measure of early development, paralleling specifications (1), (2) and (3) of Tables 3 and 4, and also adding Ln*pd1500* in addition to *statehist*, *agyears*, and *tech1500* in specifications paralleling those of columns (4) and (6) of those tables. In the specifications in which it appears alone, Ln*pd1500* obtains a negative coefficient significant at the 1% level, in the counterpart of Table 3 (predicting *col*), and a positive coefficient that falls short of significance at the 10% level in the counterpart of Table 4 (predicting Ln*colyr*), thus strongly supporting our hypothesis on colonization's occurrence and only marginally supporting our hypothesis on colonization's timing. When it is added to the other three early development measures, Ln*pd1500* raises the values of R^2 and Pseudo R^2 slightly in both specifications for both dependent variables, and while statistically insignificant in most of these estimates, it actually displaces *agyears* as the sole individually significant variable among the four early development measures in what corresponds to the full specification (6) for Ln*colyr*, being significant at the 5% level and having the expected negative sign. These results, shown in Tables S.17 and S.18, thus generally support our core hypothesis with respect to both outcomes.[20]

We have also considered combining our three main measures of early development into a single composite index of early development—the first principal component of our three indicators. The composite index has pairwise correlations with *statehist*, *tech1500*, and *agyears* of 0.898, 0.938 and 0.876, respectively. The counterparts of Tables 3 and 4 are shown in Tables S.19 and S.20. The composite index of early development has the expected sign and is significant at the 1% level in all the specifications where it is included.

Next, as a robustness test of our disease control measure, we estimated regressions identical to those of Tables 3 and 4 apart from substituting *malaria ecology* for *EDE*. The results, shown in Tables S.21 and S.22, show little sensitivity of the estimated effects of the early development indicators to which of the two disease measures is used. We find slightly more sensitivity of significance levels of the coefficients on some of the geography measures, and a considerable qualitative change in the significance of the disease measure's effect on colonization, but not that for its effect on colonization's timing. In the colonization regressions, the significance of the measured impact of *tech1500* falls from the 5% to the 10% level in the column (2) estimate, while in the columns (1) and (3) estimates of the Ln*colyr* regressions, the impacts of *statehist* increase in significance (to 5% and 1% respectively) when *malaria ecology* is used. We leave it to the interested reader to inspect the effects on the geography measures, which include a drop in significance for the coefficients on *landdist* in the regressions for Ln*colyr*. Regarding the impact of disease itself, if measured by *malaria ecology* it not only delayed colonization—a conclusion that holds with either disease measure—but it also made being colonized somewhat less likely. This result is neither very robust nor very easy to support with historical intuition, and given its secondary importance to our concerns, we think it best to place little weight on it.[21]

Our main robustness tests on the geography side probe the manner of controlling for distance overland. We checked whether a simpler dummy variable which takes value 1 when *landdist* >0, or a more conventional *landlocked* variable that is 1 only for truly landlocked countries (and is thus 0 for Jordan, Sudan, El Salvador, etc.), or the combination of the *landlocked* dummy and our *landdist*, yield substantially different results. We find that our results are not highly sensitive to the manner of controlling for land barriers or being landlocked. Tables S.23 and S.24 show the variants of Tables 3 and 4 in which

[20] Galor and Ashraf's data, which are ultimately based on McEvedy and Jones (see above), lack population density estimates for three countries in our sample: Hong Kong, Mauritius, and Taiwan. Tables S.17 and S.18 report estimates using year 1500 population density estimates for Hong Kong and Taiwan, as described in the table notes. Since Mauritius is believed to have been unpopulated in 1500, it does not appear in estimates using natural log of population density, but is included in further robustness tests that use level of population density or ln(1+popden1500). In still other robustness tests mentioned in those tables' notes, we experimented with dropping each possible subset of the three countries for which our own population density estimates are required.

[21] It is easy to think of cases, especially the majority of countries in sub-Saharan Africa, which support the expectation that the disease environment and malaria in particular delayed the year of colonization but failed to prevent countries from being colonized by the early 20th Century. In contrast, countries hospitable to malaria that were never colonized are hard to identify. Among countries that our main case coding treat as never colonized that also had a particularly high disease burden, we are able to think of Liberia, only. Since Liberia's coding is debatable (see footnote 17), we re-ran the regressions using malaria ecology under the alternative assumption that Liberia was colonized and found that in the regressions for colonization, the coefficient on malaria becomes mainly positive and always insignificant.

landlocked is substituted for *landdist*. The fact that *landdist* is significant with higher confidence levels than the *landlocked* dummy in several specifications of the model for *col* speaks to the advantage of using of *landdist* in our main specification.[22]

4.4. Estimation by instrumental variable and Heckman selection models

In this section, we consider two possible concerns about our baseline models that lead us to investigate sensitivity of our results to alternative estimating methods.

The first concern is that our indicators of early economic, political and technological development might be correlated with the error terms in our regressions and not have any causal significance in their own right. The observed correlations of colonization's occurrence and timing with early development indicators could be driven by omitted variables, such as the degree of previous contact or cultural proximity with European powers, or some dimension of early development not captured by our indicators.

The second concern is the possibility of sample selection bias in our estimations regarding the determinants of colonization’s timing. Naturally, our sample for these estimations consists only of countries that were colonized, and these may be different from non-colonized countries in important dimensions that are not captured by our controls.

To address the first concern, we exploit a source of exogenous variation in early development, drawing again on insights from Diamond (1997). The advent of agriculture across world regions, Diamond argues, was determined by the presence of wild precursors of the planet’s main grain crops and large domesticated animals. Hibbs and Olsson (2004) assemble the data on the presence of these plants and animals, map them into an index of “biogeography,” and show the ability of that index to predict the dates of macro-regional agricultural revolutions.[23]

Building on that work, we can use the biogeography index as an instrumental variable for *agyears*, under the assumption that the presence of plants and animals affected colonization by affecting the timing of the agricultural revolution but not through any other channels; in other words, our IV strategy is defensible insofar as the diversity of plants and animals suitable for domestication had a significant bearing on development paths in the distant past but had no direct effects on the occurrence and timing of colonization other than through the impact of agrarian histories on levels of development circa 1500.

We can also estimate IV versions of the models that include *statehist* or *tech1500* instead of *agyears*. The *biogeography* index is highly correlated with both *statehist* and *tech1500*, presumably because the emergence of states and the development of technology were tightly connected with the emergence of agriculture.[24] In these cases, the identifying assumption is that *biogeography* affected colonization’s occurrence and timing exclusively by affecting either *statehist* or *tech1500,* presumably through the timing of the agricultural revolution.

Our intention in these IV estimations is simply to assess the robustness of the observed impact of early development on colonization and its timing. Note that while our baseline estimations included specifications with all three indicators of early development, we omit such specifications here because we only have one instrumental variable. In any case, recall that we view the three indicators of early development as tightly connected dimensions of the same phenomenon, and when using such specifications elsewhere in the paper we do not intend to provide a conclusive assessment regarding which dimension of early development has a stronger impact on colonization patterns.

The results from our two-stage least squares estimations paralleling those of the first three specifications of Tables 3 and 4, with the early development variables instrumented by *biogeography*, are displayed in Tables 5 and 6. First stage results indicate that *biogeography* had a strong positive effect on early development in all specifications, and the Kleibergen-Paap tests reject the weak instrument null hypothesis with very high levels of confidence (the *p*-values are below 0.001 in all cases). Thus, while the *biogeography* index does not display as much variation across countries as might be desired (in our sample it has distinct values only for 8 macro-regions), it appears to be a strong predictor of early development.

In the IV Probit estimates of Table 5 (obtained with the maximum likelihood method), the effect of each early development variable has the expected sign and is statistically significant at the 1% level. Calculations of the marginal effects at the means indicate that a one standard deviation increase in early development reduces the probability of being colonized by (depending on which early developmen**t** measure is considered) between 9% and 12.6%. In the IV models for Ln*colyr*, shown in Table 6, each instrumented early development variable obtains a positive coefficient which is significant at the 1% level; a one-standard-deviation increase in early development is associated with an increase of 0.42–0.50 standard deviations in the outcome variable, depending on the measure of early development. Thus, for all three individual proxies of early development, IV estimation supports our proposition that greater early development makes colonization less likely to occur and delays its occurrence, consistent with the baseline estimations reported in Section 4.2. Moreover, for both the

[22] An additional robustness check that we have considered is replacing absolute latitude, which is meant to capture climatic features, by the Köppen-based climate index used by Olsson and Hibbs (2005). That variable (which is highly correlated with latitude) does not obtain significant coefficients when we use it in addition to latitude, and when substituted for latitude it performs similarly but with less robustness across specifications. The results for other variables are qualitatively the same.

[23] We assigned values of *biogeography* to a few countries missing from Hibbs and Olsson’s sample, as detailed in the Data Appendix.

[24] Borcan et al. show that time of transition to agriculture is a highly significant predictor of both time of first state emergence, and of the *statehist* index as of 1500 CE, even when numerous controls and macro region fixed effects are included, for their sample of 151 countries.

Table 5
Determinants of colonization, IV probit regressions (maximum likelihood estimation). Dependent variable: *col. (statehist*, *agyears*, *tech1500* are instrumented by *biogeography*)

	(1)	(2)	(3)
statehist	−2.376***		
	(0.574)		
tech1500		−2.361***	
		(0.658)	
agyears			−0.0264***
			(0.00803)
EDE	0.305	0.394*	0.310
	(0.203)	(0.225)	(0.220)
navdist	−0.0783*	−0.0847*	−0.118**
	(0.0453)	(0.0484)	(0.0558)
landdist	−2.193***	−2.157***	−2.038***
	(0.738)	(0.718)	(0.711)
latitude	−0.0189	−0.0212	−0.0213
	(0.0121)	(0.0148)	(0.0132)
constant	3.114***	3.407***	3.835***
	(0.805)	(0.966)	(1.047)
Observations	111	111	111
First stage			
dep. var.:	*statehist*	*tech1500*	*agyears*
EDE	−0.00219	0.0200	−1.847
	(0.0356)	(0.0230)	(1.948)
navdist	0.0162**	0.0152***	0.425
	(0.00691)	(0.00531)	(0.433)
landdist	−0.153	−0.0842	−0.159
	(0.125)	(0.0675)	(5.599)
latitude	−0.00299	−0.00403**	−0.357**
	(0.00287)	(0.00159)	(0.154)
biogeography	0.00683***	0.00702***	0.669***
	(0.000643)	(0.000550)	(0.0657)
constant	0.0158	0.113***	17.76***
	(0.0756)	(0.0423)	(3.586)
athrho	0.521**	0.460**	0.246
	(0.221)	(0.204)	(0.160)
Observations	111	111	111

Robust standard errors in parentheses.
*** $p < 0.01$.
** $p < 0.05$.
* $p < 0.1$.

occurrence and the timing of colonization, the IV estimates of the effects of early development are larger in magnitude than the baseline estimates.

A second potential issue with our OLS models estimating the date of colonization is the possibility of sample selection bias affecting the estimated effect of early development. Places with high levels of early development indicators were less likely to be colonized. Some of those places, though, were colonized, presumably because they had low values of some unmeasured variable that deterred colonization elsewhere. Thus, places with high early development tend to have low values of this unmeasured variable in the sample of colonized countries, while places with low early development have a more even distribution of values for that variable in this sample. If this unmeasured variable also delays colonization timing, then the estimated negative effect of early development on timing will underestimate its absolute magnitude.

Fortunately, the correction for sample selection bias proposed by Heckman (1979) has a very natural application in our context. We have studied the determinants of colonization, and the *col* equation can be used as the "selection equation" to properly take into account selection into colonization when estimating the determinants of Ln*colyr*.

We provide full information maximum likelihood (FIML) estimates of the Heckman model obtained from iterative joint estimation of the *col* and Ln*colyr* equations.[25] Table 7 shows our results. The null hypothesis of no selection bias is rejected at the 95% confidence level in specifications 3, 4, and 6, but it is not rejected for the other specifications.[26] In all specifications,

[25] We report the FIML estimates because they are more efficient than those from Heckman's original two-step procedure (see Nawata, 1994; Puhani, 2000; Greene, 2012); in any case, for our analysis both sets of estimates are very similar.

[26] This is indicated by the significance of *athrho*, the estimated inverse hyperbolic tangent of *rho*, which measures the correlation between the error of the *col* equation and the error of the Ln*colyr* equation.

Table 6
Determinants of timing of colonization, IV regressions. Dependent variable: Ln*colyr*. (*statehist*, *agyears*, *tech1500* are instrumented by *biogeography*)

	(1)	(2)	(3)
statehist	0.138***		
	(0.0375)		
tech1500		0.139***	
		(0.0326)	
agyears			0.00159***
			(0.000461)
EDE	0.0513***	0.0454***	0.0541***
	(0.0125)	(0.0103)	(0.0115)
navdist	0.00395	0.00330	0.00529**
	(0.00245)	(0.00258)	(0.00255)
landdist	0.0981**	0.105***	0.101***
	(0.0417)	(0.0323)	(0.0373)
latitude	0.00164	0.00173**	0.00176*
	(0.00108)	(0.000859)	(0.00104)
Constant	7.365***	7.356***	7.336***
	(0.0290)	(0.0250)	(0.0272)
Observations	92	92	92
R^2	0.237	0.387	0.299
First stage			
dep. var.:	*statehist*	*tech1500*	*agyears*
EDE	0.1470	0.0574***	−0.4996
	(0.3548)	(0.0208)	(2.2298)
navdist	0.1572**	0.0204***	0.5218
	(0.0073)	(0.0043)	(0.4595)
landdist	0.1340	−0.08164	9.6774
	(0.1310)	(0.0769)	(8.2331)
latitude	−0.0024	−0.0031*	−0.2865
	(0.0030)	(0.0018)	(0.1895)
biogeography	0.0077***	0.0077***	0.6667***
	(0.0009)	(0.0006)	0.0596
Constant	−0.0492	0.0212	14.2278***
	(0.0812)	(0.0477)	(5.1065)
K–P LM test p-value	0.000	0.000	0.000
Observations	92	92	92
R^2	0.499	0.7473	0.6377

Robust standard errors in parentheses.
*** $p < 0.01$.
** $p < 0.05$.
* $p < 0.1$.

even in those where the null of no selection bias is rejected, the estimates of the determinants of Ln*colyr* shown in the upper panel of Table 7 are broadly similar to those in Table 4. Results for *EDE*, *navdist*, *landdist* and *latitude* are all quite similar to those obtained before. Most importantly, regarding the effects of early development on the timing of colonization, the coefficient on *tech1500* retains its positive sign and 1% significance in both models (4) and (6). Thus, our proposition that higher early development delayed colonization also receives support from the FIML selection model.

Other coefficients remain similar in value and significance level, with that on *agyears* strengthening from 5% to 1% significance level in model (3). A couple of differences with the OLS results deserve to be mentioned. The positive coefficient on *statehist* in model (1) becomes slightly smaller and loses its already marginal significance, while the “wrongly” signed negative coefficient on *agyears* obtained in model (4) has a larger absolute value and is significant in the FIML estimation.

It is important to note that, contrary to what is often recommended in this context, in our specifications none of the variables in the selection equation are excluded from the second equation, as we are unable to make a convincing case that any variable in our model for *col* is a valid candidate for exclusion. In such cases, identification is achieved only through the nonlinearity of the inverse Mills ratio, and the validity of the specified model relies on the assumption that the error terms in the two equations are jointly normally distributed. Although this distributional assumption is certainly stringent, we think that even without exclusion restrictions, this framework provides an informative check for our estimates of the determinants of the timing of colonization.

Overall, while the estimation of the selection model does not produce a conclusive assessment of whether sample selection bias is present in the single equation analysis of the determinant of colonization's timing, it shows that the main results hold when explicitly taking selection into account.

Table 7
Determinants of the timing and the occurrence of colonization. Heckman selection model/maximum likelihood estimation.

Equation		(1)	(2)	(3)	(4)	(5)	(6)
Ln*colyr*	*statehist*	0.0323			-0.0786		−0.0662
		(0.0273)			(0.0480)		(0.0416)
	tech1500		0.109***		0.230***		0.226***
			(0.0396)		(0.0566)		(0.0567)
	agyears			0.00114***	-0.00126**		-0.000771
				(0.000283)	(0.000596)		(0.000528)
	EDE	0.0574***	0.0456***	0.0528***		0.0613***	0.0395***
		(0.0112)	(0.0128)	(0.0107)		(0.0107)	(0.00968)
	navdist	0.00431*	0.00125	0.00817***		0.00592***	0.000515
		(0.00249)	(0.00457)	(0.00216)		(0.00223)	(0.00268)
	landdist	0.0734*	0.0833**	0.118***		0.0824***	0.0742***
		(0.0394)	(0.0424)	(0.0360)		(0.0309)	(0.0263)
	latitude	0.00198*	0.00120	0.00284***		0.00257***	0.00133*
		(0.00114)	(0.00131)	(0.000915)		(0.000945)	(0.000749)
	constant	7.372***	7.378***	7.323***	7.440***	7.362***	7.384***
		(0.0288)	(0.0375)	(0.0278)	(0.0173)	(0.0254)	(0.0234)
col	*statehist*	−1.380*			−1.030*		−1.310**
		(0.807)			(0.541)		(0.589)
	tech1500		-1.556		1.132*		−0.135
			(1.622)		(0.643)		(1.367)
	agyears			−0.0170***	−0.0346***		−0.0112
				(0.00531)	(0.00736)		(0.0255)
	EDE	0.222	0.384	-0.264		0.137	0.595
		(0.254)	(0.244)	(0.322)		(0.244)	(0.592)
	navdist	−0.104	−0.0596	−0.0984***		−0.136**	−0.0499
		(0.0713)	(0.137)	(0.0306)		(0.0584)	(0.0899)
	landdist	−1.524***	−1.632	−1.816***		−1.911***	−1.231
		(0.533)	(0.994)	(0.463)		(0.672)	(0.770)
	latitude	−0.0554***	−0.0432**	−0.0338*		−0.0566**	-0.0302
		(0.0198)	(0.0198)	(0.0177)		(0.0227)	(0.0509)
	constant	3.890***	3.382***	3.138***	2.447***	3.613***	3.264***
		(1.268)	(0.978)	(0.418)	(0.225)	(0.961)	(0.585)
	athrho	1.235	1.495	−16.13***	16.63***	0.487	16.07***
		(1.316)	(2.031)	(0.0641)	(0.0582)	(0.390)	(0.0694)
	Observations	111	111	111	111	111	111

Robust standard errors in parentheses.
Note: lower panel shows Probit regressions for *col*; upper panel shows maximum likelihood estimates of the regression for Ln*colyr*.
*** $p < 0.01$.
** $p < 0.05$.
* $p < 0.1$.

5. Conclusion

We used country-level data on a sample of countries that together account for more than 95% of the world's population outside of Europe to investigate what factors account for which countries were and were not colonized, and for when colonization occurred during the course of the almost five centuries of the epoch of European overseas colonization. Our main interest lay in testing the proposition that the level of economic, political, and technological development in the 15th century was a major determinant of who was colonized and when, with local disease environment playing an important role in timing, and with auxiliary roles for a small number of geographic factors.

Our most important result was that we find support for the conjecture that both the occurrence and the timing of colonization are to a significant degree explained by the level of development of the potential targets of colonization on the eve of the colonial era, as measured by years elapsed since transition to agriculture, experience of indigenous state-level polities, and level of technology adoption as of 1500 CE. In brief, countries with longer histories or higher levels of pre-1500 development tend to have been colonized later, or not at all, by Western European colonizing powers. We also find navigation distance from northwest Europe, overland distance for countries that are landlocked or otherwise required a land passage to reach, latitude, and disease environment to be significant predictors of both the occurrence and timing of colonization across countries.

We checked sensitivity of our results to changes in sample and to a few potentially controversial decisions on coding colonization date and status, finding the main results, especially regarding impact of pre-1500 development on colonization, to be robust. We likewise confirmed robustness of these qualitative results when we control for endogeneity of our early development indicators by IV estimation, and when our equations for colonization and its timing are jointly estimated in a

Heckman selection model. Many results hold even when Old World countries alone are considered, although which early development indicator appears as the most robust predictor of colonization patterns, and the role of disease in delaying colonization, are affected by dropping the Americas and Oceania from consideration.

The most important implication of our research lies in demonstrating, for the first time in a statistically rigorous way, that the large differences in levels of technological and social development which marked the world on the eve of European exploration and colonization, differences that have been shown to predict much of the variation in levels and rates of development to the present day, played an important role also in determining both the occurrence and the timing of colonization. European colonization was not an all-at-once event but rather a process unfolding over five centuries, one that began with the conquest of lands technologically well behind the panoply of advanced Eurasian civilizations and that ended with a relatively short era of European dominance over most of the globe including Eurasia. European powers were unable to dominate near equals in North Africa and Asia at the beginning of their colonial expansion, but had colonized most by the end. The facts that more technologically advanced non-European countries including Turkey, Iran, China, Japan and Korea were never colonized by Europeans, that the Central Asian countries fell under land-based Russian dominance rather than that of European maritime powers, and that the countries of the Levant were only briefly European-ruled after the collapse of the Ottoman empire, also appear to be attributable to a substantial degree to relative pre-modern development, with an added role for landlocked status in the case of Central Asia.

A large number of key features of today's global mosaic are closely bound up with the colonization process that we've explored. One important impact of the interaction of relative developmental differences circa 1500 with the process of European colonization is that many of the regions that were least technologically sophisticated and thus least densely settled in 1500, including countries like Argentina, Uruguay, the U.S., Canada, Australia, and New Zealand, were put on the path to becoming mainly European-settled societies. Equally important, many Latin American countries and South Africa became cauldrons of social and political conflict thanks to temporary domination and significant settlement by Europeans, while other regions, especially in the Caribbean but also in parts of the U.S., Brazil, and elsewhere, became home to a vast African diaspora. Each of these transformations, their contrast with the far greater continuities of population in most of Africa and Asia, and the much shorter duration of colonization in most of Africa than in the Western Hemisphere, are important for understanding the social and economic problems of today's world. For these reasons, a systematic understanding of where and when European colonization took place demands an important place in the study of economic and social history.

Finally, scholars studying the impact of colonial rule on specific outcomes such as per capita income, inequality, rate of economic growth and level of income, have rarely incorporated in their work systematic consideration of the ways in which colonization and the manner of colonization may be endogenous to conditioning variables including those studied here. By demonstrating that the occurrence and timing of colonization can be statistically predicted, our study will hopefully contribute to raising the standards to be expected of future studies of the colonial era's impact.

Data Appendix (description and sources of all variables)

col Dummy variable set to 1 if most of the country's territory was colonized by Belgium, England, France, Germany, Italy, the Netherlands, Portugal or Spain during the period between 1462 and 1945, otherwise 0. Judgment on whether foreign involvement meets the standard of colonization is our own and is explained for each country in the Supplementary Online Appendix. Colonies include cases of indirect rule as well as League of Nations protectorates but exclude cases where sources speak merely of a foreign "sphere of influence."

colyr First year in which colonial rule by one of the powers mentioned is considered to have been effective over 20% or more of the present-day country's territory. Determination of colonization year is our own and is explained for each country in the Supplementary Online Appendix. Ln*colyr* is the natural logarithm of *colyr*.

agyears Number of years before year 2000, in hundreds, that a substantial population living within what are the present country's borders began to obtain most of their calories from agriculture. Data are from Putterman with Trainor (2006), which in turn details its sources.

statehist Discounted and normalized value of index for presence of supra-tribal government on territory constituting the present-day country, covering years 1 CE–1500 CE In a given year, the index value is the product of three indices covering the unit interval: (1) an index for existence of a state, (2) an index for that state being domestically based (=1 if so, 0.5 if imposed by an external power), and (3) an index for territorial extent and unity (states ruling small shares of the country's current territory and multiple simultaneously extant states get lower values). Values are aggregated into 50-year periods, the period x half centuries prior to 1500 is discounted by $(1.05)^x$, the resulting numbers are summed, and the sum is normalized to the [0, 1] interval by dividing by the hypothetical maximum value. Data are from Borcan et al. (2014).

tech1500 An index of the adoption of agricultural, military, communications, and other technologies, from Comin et al. (2010).

latitude Absolute value of latitude. Source: Weil (2009). Unit of measure: degrees.

navdist Distance between Camaret-sur-mer and the closest port of historical significance (usually the main port) in each country, without considering routes going through the Suez or Panama canals. Source: AXSMarine distance table (www.axsmarine.com). Unit of measure: thousands of nautical miles.

landdist The distance that the colonizers had to traverse through ground transportation. For landlocked countries, it reflects the distance from the country's historically most important city (usually but not always the current capital) to the closest oceanic port. For El Salvador, Ecuador, Peru, and Chile, *landdist* is the distance between Panama City and Balboa—the Atlantic and Pacific ports that are now joined by the Panama Canal. Source: Google Earth. Unit of measure: thousands of nautical miles.

malaria The malaria ecology index measures the suitability of a country's climate to mosquito breeding as well as the prevalence of mosquito species that feed only on humans. The source is Kiszewski et al. (2004).

EDE European disease environment, measured as the logarithm of the annualized probability of death for European males in the age cohort of soldiers, as calculated by Raphael Auer for Auer (2013). Values for countries not in the sample used in that paper were also calculated by Auer and provided to the authors in personal communication.

Ln*pd1500* Log of population density in 1500 CE, from Ashraf and Galor (2011).

landlocked Dummy variable that indicates if a country has direct access to the ocean—if it does not, *landlocked* takes a value of 1.

biogeography Based on the numbers of large-seeded grasses and numbers of large animals suitable for domestication, from background data of Olsson and Hibbs (2005). A larger number indicates a richer set of potential domesticates among naturally occurring species. Country values are shared within world regions of agricultural spread. For countries in our data set not included in that of Olsson and Hibbs, we adopt values based on the following regional assignments: Afghanistan, Armenia, Azerbaijan, Kazakhstan, Kyrgyzstan, Lebanon, Turkmenistan, Uzbekistan, assigned to Near East, N. Africa and Europe region; Liberia assigned to sub-Saharan African region; Myanmar[27] and Vietnam, to E. Asia region.

Supplementary material

The Supplementary Online Appendix and data and replication files can be found in the online version at http://dx.doi.org/10.1016/j.euroecorev.2015.10.012.

References

Acemoglu, Daron, Johnson, Simon, Robinson, James, 2001. The colonial origins of comparative development: an empirical investigation. Am. Econ. Rev. 91 (5), 1369–1401.

Acemoglu, Daron, Johnson, Simon, Robinson, James, 2002. Reversal of fortune: geography and institutions in the making of the modern world income distribution. Q. J. Econ. 117 (4), 1231–1294.

Acemoglu, Daron, Johnson, Simon, Robinson, James, 2005. The rise of Europe: Atlantic trade, institutional change, and economic growth. Am. Econ. Rev. 95 (3), 546–579.

Acemoglu, Daron, Robinson, James, 2012. Why Nations Fail: The Origins of Power, Prosperity and Poverty, Crown Publishers, New York.

Alesina, Alberto, Giuliano, Paola, Nunn, Nathan, 2013. On the origins of gender roles: women and the plough. Q. J. Econ. 128 (2), 469–530.

Ang, James, 2013. Institutions and the long-run impact of early development. J. Dev. Econ. 108, 1–18.

Ashraf, Quamrul, Galor, Oded, 2011. Dynamics and stagnation in the Malthusian epoch. Am. Econ. Rev. 101, 2003–2041.

Ashraf, Quamrul, Galor, Oded, 2013. The 'Out of Africa' hypothesis, human genetic diversity, and comparative economic development. Am. Econ. Rev. 103 (1), 1–46.

Auer, Raphael, 2013. Geography, institutions, and the making of comparative development. J. Econ. Growth 18, 179–215.

Bockstette, Valerie, Chanda, Areendam, Putterman, Louis, 2002. State and markets: the advantage of an early start. J. Econ. Growth 7, 347–369.

Borcan, Oana, Olsson, Ola, Putterman, Louis, 2014. State History and Economic Development: Evidence from Six Millennia. Brown University Department of Economics, Working Paper 2014–18.

Chanda, Areendam, Cook, Justin, Putterman, Louis, 2014. Persistence of fortune: accounting for population movements, there was no post-colombian reversal. Am. Econ. Rev. – Macroecon. 6 (3), 1–28.

Comin, Diego, Easterly, William, Gong, Erick, 2010. Was the wealth of nations determined in 1000 BC? Am. Econ. J.: Macroecon. 2 (3), 65–97.

Diamond, Jared, 1997. Guns, Germs and Steel: The Fate of Human Societies. Norton & Co, New York.

Easterly, William, Levine, Ross, 2003. Tropics, germs, and crops: how endowments influence economic development. J. Monet. Econ. 50 (1), 3–39.

Easterly, William, Levine, Ross, 2012. The European Origins of Economic Development. NBER Working Paper 18162.

Ertan, Arhan, 2007. Determinants and Economic Consequences of Institutions: Analyses of Colonization, Trade Policy and a Public Goods Experiment (Doctoral Dissertation). Brown University, Providence, R.I..

Feyrer, James, Sacerdote, Bruce, 2009. Colonialism and modern income: islands and natural experiments. Rev. Econ. Stat. 91 (2), 245–262.

Galor, Oded, Weil, David N., 2000. Population, technology, and growth: from Malthusian stagnation to the demographic transition and beyond. Am. Econ. Rev. 90 (4), 806–828.

Greene, William H., 2012. Econometric Analysis, 7th edition. Prentice Hall, Upper Saddle River, NJ.

Grier, Robin, 1999. Colonial legacies and economic growth. Public Choice 98 (3–4), 317–335.

Hall, Robert, Jones, Charles, 1999. Why do some countries produce so much more output than others? Q. J. Econ. 114 (1), 83–116.

Hariri, Jacob, 2012. The autocratic legacy of early statehood. Am. Polit. Sci. Rev. 106 (3), 471–494.

Heckman, James, 1979. Sample selection bias as a specification error. Econometrica 47 (1), 153–161.

Hibbs, Douglas, Olsson, Ola, 2004. Geography, biogeography, and why some countries are rich and others are poor. Proc,. Natl. Acad. Sci. 101 (10), 3715–3720.

[27] Hibbs and Olsson's background data file lists Near East plant and animal numbers for Myanmar, to which we have reassigned the values of Asian countries including Cambodia, Laos, India and Bangladesh.

Kiszewski, Anothony, Mellinger, Andrew, Spielman, Andrew, Malaney, Pia, Sachs, Sonia Ehrlich, Sachs, Jeffrey, 2004. A global index representing the stability of malaria transmission. Am. J. Trop. Med. Hyg. 70 (5), 486–498.

La Porta, Rafael, Lopez-de-Silanes, Florencio, Shleifer, Andrei, Vishny, Robert, 1999. The quality of government. J. Law Econ. Organ. 15 (1), 222–279.

Landes, David S., 1998. The Wealth and Poverty of Nations: Why Some are so Rich and Some so Poor. Norton & Co, New York.

Maddison, Angus, 2001. The World Economy: A Millennial Perspective. Development Center of the OECD, Paris.

Marcus, Bernard, 2009. Deadly Diseases and Epidemics: Malaria, 2nd edition. Chelsea House, New York.

McEvedy, Colin, Jones, Richard, 1978. Atlas of World Population History. Facts on File, New York.

McNeill, William, 1998. Plagues and Peoples. Anchor Books, New York.

Morris, Ian, 2010. Why the West Rules—For Now: The Patterns of History, and What They Reveal About the Future. Profile Books, London.

Nawata, Kazumitsu, 1994. Estimation of sample selection bias models by the maximum likelihood estimator and Heckman's two step estimator. Econ. Lett. 45, 33–40.

Nunn, Nathan, 2014. Historical development. In: Aghion, P., Durlauf, S. (Eds.), Handbook of Economic Growth, North-Holland, Amsterdam.

Olsson, Ola, 2009. On the democratic legacy of colonialism. J. Comp. Econ. 37 (4), 534–551.

Olsson, Ola, Hibbs, Douglas, 2005. Biogeography and long-run economic development. Eur. Econ. Rev. 49 (4), 909–938.

Puhani, Patrick, 2000. The Heckman correction for sample selection and its critique. J. Econ. Surv. 14 (1), 53–68.

Putterman, Louis, Cary Anne Trainor, 2006. Agricultural Transition Year Country Data Set. Data set posted at 〈http://www.econ.brown.edu/fac/Louis_Putterman/agricultural%20data%20page.htm〉.

Putterman, Louis, 2008. Agriculture, diffusion and development: ripple effects of the neolithic revolution. Economica 75, 729–748.

Putterman, Louis, Weil, David, 2010. Post-1500 population flows and the long run determinants of economic growth and inequality. Q. J. Econ. 125 (4), 1627–1682.

Sokoloff, Kenneth, Engermann, Stanley, 2000. Institutions, factor endowments and paths to development in the new world. J. Econ. Perspect. 14 (3), 217–232.

Spolaore, Enrico, Wacziarg, Romain, 2013. How deep are the roots of economic development? J. Econ. Lit. 51 (2), 325–369.

Spolaore, Enrico, Wacziarg, Romain, 2014. Long-term barriers to economic development. In: Aghion, P., Durlauf, S. (Eds.), Handbook of Economic Growth, North-Holland, Amsterdam.

Weil, David, 2009. Economic Growth, 2nd edition. Addison Wesley, Boston.