14 Regulation of the biopharmaceutical industry

Patricia M. Danzon

1. INTRODUCTION

The pharmaceutical industry is of interest to the field of law and economics for three related reasons that modify the application of the standard analysis of structure, conduct and performance used in industrial organization. First, the pharmaceutical industry makes unusually high investment in Research and Development (R&D), which implies a high rate of technical change, high fixed costs relative to marginal cost, critical importance of patent protection, and novel price and product competitive strategies. This raises interesting positive and normative issues related to patent rules and generic entry, pricing and public policy.

Second, the industry is heavily regulated in all major functions. Early regulation focused on safety as a condition of market access for new drugs. Starting in the 1960s most countries have also required evidence of efficacy and manufacturing quality, and restrict promotion and advertising to both physicians and consumers. In the 1980s and 1990s, most countries added regulation to control drug prices, insurance reimbursement, and/or total expenditures. These policies arise out of concern that extensive insurance coverage makes consumer demand more price-inelastic, which tends to lead to increased utilization and higher prices, unless payers add constraints. Public insurers typically target controls at the supply side of the market, while private insurers rely more on consumer cost-sharing. Insurance and these control policies affect the nature of competition, returns to and incentives for R&D and ultimately social welfare. A growing literature examines both positive and normative issues raised by different forms of cost control regulation but many questions remain unanswered.

Third, most biopharmaceuticals are potentially global products that are diffused worldwide through local subsidiaries of multinational corporations and/or licensing to local firms. The profitability of pharmaceutical R&D thus depends on worldwide sales and on policies adopted in many national markets. Each country faces an incentive to adopt the regulatory policies that best control its pharmaceutical budget in the short run, free-riding on others to pay for the joint sunk costs of R&D. As countries increasingly adopt regulatory policies involving external referencing and parallel trade, the ability of industry to price
discriminate between countries is breaking down. However, theory suggests that price discrimination is superior to uniform pricing in achieving social goals of maximum access to existing drugs, while preserving incentives for R&D.

The structure of the industry has undergone dramatic change in the last two decades, with the emergence of biotechnology, genomics and related sciences that have transformed the nature of R&D and comparative advantage within the industry. Small firms play an increasingly important role in the discovery of new drugs and new R&D technologies. This continual flow of new entry makes the industry structurally competitive, provided that sources of financing are available either from private and public equity markets or through alliances with larger biopharmaceutical companies. The structuring of such alliances raises interesting questions related to agency and the nature of the firm.

The appropriate economic model of the pharmaceutical industry is either monopolistic competition or oligopoly with product differentiation. However, analysis must also take into account the roles of physician prescribing and third party payment as key factors in determining demand elasticities faced by producers. Moreover, pricing models must recognize the importance of R&D and fixed costs. Although the industry is characterized by high fixed costs, models in which firms endogenously choose sunk costs, either in the form of R&D or promotion, to retain competitive advantage and deter entry (Sutton, 1991) do not seem appropriate and are refuted by the evidence of entry over the last two decades by hundreds of small firms.

This chapter provides a review of basic issues and recent literature related to regulation and performance of this industry. Previous survey articles review earlier literature, in particular: Comanor (1986) examines political economy and the effects of US safety and efficacy regulation in the 1960s and 1970s; Scherer (1993, 2000) applies an industrial economics framework to discuss issues related to pricing, profits and technical progress; and Danzon and Keuffel (forthcoming) focus on regulation and its effects. All of these studies have a US focus, given the dominance of US-based literature and firms in this industry. Issues and evidence from other countries are included here where possible.

2. MARKET ACCESS REGULATION: SAFETY, EFFICACY AND QUALITY

2.1 Regulatory Structure

The US 1938 Food, Drug and Cosmetics Act (FDCA) required any firm seeking to market a new chemical entity (NCE) to file a new drug application (NDA) to demonstrate that the drug was safe for use as suggested by the proposed labeling. The Food and Drug Administration (FDA) had 180 days to...
reject the NDA. The FDCA established jurisdiction over drug advertising, but policing was left to the Federal Trade Commission (FTC) rather than the FDA. This Act also established the requirement that patients obtain a physician’s prescription to obtain retail drugs.

In the 1960s concern over the proliferation, pricing and advertising of drugs of dubious efficacy and concern to strengthen safety requirements, following the thalidomide tragedy that caused hundreds of birth defects in Europe while the drug was still under review in the US, led to the landmark 1962 Amendments to the FDCA. These Amendments strengthened safety requirements; added the requirement that drugs show proof of efficacy, usually by double blind, randomized controlled trials of the drug relative to placebo; removed the time limit within which the FDA could reject an NDA; extended FDA regulation to cover clinical testing and manufacturing; and restricted manufacturers’ promotion to approved indications. Basic requirements for promotional materials were defined, including that such materials cannot be false or misleading; they must provide a fair balance of risks and benefits; and they must provide a “brief summary” of contraindications, side-effects and effectiveness. Regulatory oversight of promotional material was ceded back to the FDA from the FTC.

From an economic perspective, the regulatory requirement for proof of safety and efficacy is plausibly an efficient mechanism to address the fact that physicians and consumers have imperfect and asymmetric information about new drugs. Imperfect information can lead to health hazard, wasted expenditures on ineffective drugs and possibly excessive product differentiation that undermines price competition. However, as regulators became more concerned with remote risks and risks associated with long-term use of chronic medications, the size and duration of required clinical trials have increased, leading to delay in the launch of new drugs, potential foregone benefits for consumers, and shorter effective patent life and foregone revenue for firms. High fixed costs of market entry may also lead to higher break-even prices, ceteris paribus, and fewer drugs, particularly drugs to treat rare diseases with small potential market size.

Subsequent legislation has addressed several of these cost-increasing effects of the 1962 Amendments. The Orphan Drug Act of 1983 (ODA) dramatically increased incentives to invest in orphan diseases (defined as conditions that affect fewer than 200,000 individuals in the US), granting market exclusivity for seven years and a 50% tax credit for R&D expenses. Following the ODA, the number of orphan drug approvals has increased significantly. By 1998, there were more than five times as many orphan drugs as in 1979, but fewer than twice as many non-orphan drugs, although approvals of these two drug types were similar prior to 1983 (Lichtenberg and Waldfogel, 2003).
To expedite regulatory review, the Prescription Drug User Fee Act (PDUFA) of 1993 and subsequent renewals requires firms to pay substantial submission fees that are used to hire more reviewers. PDUFA also created a system of priority vs. standard review, with target durations of 10 months for standard review and 6 months for priority review drugs. The 1997 FDA Modernization Act (FDAMA) created Fast Track status to potentially expedite the entire clinical trial process for novel drugs to address unmet medical needs for serious or life-threatening conditions. Fast track has reduced overall development times, but some have argued that fast track and priority review are associated with increased prevalence of post-approval adverse events (see below). In 2007 the FDA was authorized to approve new drugs conditional on a Risk Evaluation and Management Strategy (REMS), which can include restricted distribution, requirements to gather post-launch evidence on clinical outcomes and side-effects etc. This has reportedly made the FDA more willing to grant limited market authorization, pending collection of additional data (McCaughan, 2009).

Europe, Japan and other industrialized countries have similar market access regulation to the US. In 1995 the European Union established the European Medicines Evaluation Agency (EMEA) as a centralized approach to drug approval for EU member states. The centralized procedure provides simultaneous registration approval in all EU countries. The EMEA is the required approval route for biotech products and is optional for other new drugs. National systems remain for products that seek approval in only a few countries. Many other smaller countries rely largely on approval by the FDA, EMEA or other major agencies.

Since the 1990s, the US, the EU and Japan have worked through the International Commission on Harmonization to harmonize their regulatory requirements for safety, efficacy and manufacturing quality. Consequently, companies can compile a single dossier for submission to all major regulatory agencies, but with important local differences. For example, the EMEA typically requires trials of new drugs relative to current treatment, whereas the FDA more often uses a placebo comparator, except where use of placebo would imply unethical treatment of patients. Japan requires some trials on Japanese nationals. Most important, each agency still makes its own evaluation based on its own risk-benefit trade-off. Nevertheless, as a result of these harmonization procedures, differences in timing of drug launch in major markets more often reflect differences in reimbursement regulation, not of market access regulation (Danzon and Epstein, 2009).

2.2 Costs of Regulation

The 1960s and 1970s were characterized by rising costs of R&D, decline and
delays in the number of new drugs approved, and decline in the number of small firms. Much early economic analysis attempted to estimate the contribution of the 1962 FDCA Amendments to these trends (for example, Peltzman, 1973; Grabowski, Vernon and Thomas, 1978; Baily, 1972; Wiggins, 1981).

2.2.1 Number of new drug approvals
Grabowski, Vernon and Thomas (1978) report that the number of new chemical entities (NCEs) approved fell from 233 in the five-year period 1957–61 to 93 in 1962–6 and 76 in 1967–71. Some decline would be consistent with the intent of the legislation, if some of the prior introductions were ineffective. However, the percentage of total drug sales accounted for by new NCEs declined roughly in proportion to the number of drugs, from 20.0% in 1957–61 to 5.5% in 1967–71, which the authors contend is unlikely if only the less significant drugs were eliminated, assuming that these would have captured a proportionately smaller share of revenues.

Grabowski, Vernon and Thomas (1978) attempt to measure the marginal reduction plausibly attributed to the 1962 Amendments after controlling for other possible contributing factors, including the depletion of new product opportunities; the thalidomide tragedy that may have made manufacturers and physicians more risk averse, hence reduced demand for new drugs; and pharmacological advances that may have raised R&D costs independent of regulation. They conclude that research productivity, defined as number of NCEs per (lagged) R&D expenditure, declined sixfold between 1960–61 and 1966–70 in the US, compared to a threefold decline in the UK, and that the 1962 Amendments increased the cost per new NCE in the US by a factor of 2.3. They attribute these differentials to regulation, arguing that the UK would have been equally affected by exogenous changes in scientific opportunities or thalidomide-related change in demand.

2.2.2 R&D cost per NCE and declining R&D productivity
Determining the relative contribution of regulation vs. other factors to the increase in R&D cost per new drug approved is also problematic, due to other confounding trends. Baily (1972) and Wiggins (1981) concluded that the 1962 Amendments led to a large increase in the R&D cost per new drug approved, but with unexplained variation across therapeutic categories.

More recent evidence that the real cost per NCE approved continues to rise, despite no major change in explicit regulatory rules, suggests that other factors play a role. Specifically, DiMasi, Hansen and Grabowski (2003) and DiMasi and Grabowski (2007) found that capitalized cost per approved NCE, measured in present value at launch, grew from $138m. in the 1970s to $318m. in the 1980s, $802m. in the 1990s and over $1.2b. in 2009. Roughly
half of this total cost is actual out-of-pocket expense, including spending on drugs that fail but are an inevitable part of getting a drug approved. The remainder is foregone interest or opportunity cost of capital. Despite – or because of – the major advances in microbiology, combinatorial chemistry, high-throughput screening, robotics, bioinformatics, and genomics that revolutionized drug discovery in the 1990s, drug discovery costs have grown at a slower annual rate (2.3% in the 1990s) than the costs of clinical trials (11.8%), which reflect shifts in medical care technology and medical prices. The growth in clinical trial cost includes an increase in number of trial participants, more procedures and higher cost per participant. Besides changing regulatory requirements, other contributing factors include: rising clinical costs in step with advancing scientific and medical norms; change in types of drugs and diseases pursued, as R&D effort shifts towards more difficult diseases once the “low hanging” diseases have been addressed; increased focus on chronic diseases which require longer trials to detect cumulative effects; and collection of economic as well as clinical data, to satisfy growing payer demands for evidence of cost-effectiveness.

On the other hand, regulatory willingness to use biomarkers (such as tumor shrinkage) as surrogates for clinical endpoints (such as survival), priority review, and limited approvals with risk mitigation strategies etc. have contributed to the recent dramatic growth in number of biologics and other drugs under development or approved for cancer, inflammatory diseases etc. Scientific advance and relatively generous reimbursement have also no doubt played a role. However, it seems safe to conclude that, given recent regulatory measures to expedite approval of high priority drugs, the balance has shifted and there is now less concern over undue costs and delay at least for these high priority drugs, and perhaps more concern over adequate proof of safety and efficacy.

2.2.3 Lags in launch
Several analyses found that the 1962 Amendments delayed the approval of new drugs in the US relative to other countries (for example, Wardell, 1973; Wardell and Lasagna, 1975; Grabowski, 1976; Wiggins, 1981). Grabowski, Vernon and Thomas (1978) compare introduction dates in the US and the UK for drugs discovered in the US between 1960 and 1974. The proportion of drugs introduced first in the US declined significantly between 1960–62 and 1972–4, while the proportion introduced later in the US increasing significantly. The authors conclude that increased regulatory scrutiny in the US caused multinational companies to introduce new products abroad before their US launch. Dranove and Meltzer (1994) estimate that the average time from a drug’s first worldwide patent application to its approval by the FDA rose from 3.5 years in the 1950s to almost 6 years in the 1960s and 14 years in the mid
1980s. However, they also found that, starting in the 1950s, more important drugs – especially drugs that ultimately had larger sales – are developed and approved more rapidly than less important drugs. They attribute this differential to actions of drug companies as much as to regulatory priority setting.

However, since the 1990s the US no longer lags and may lead the major EU markets in number and timing of major new drug launches (Danzon, Wang and Wang 2005; Danzon and Epstein, 2009). Given the coordination of regulatory requirements between the EMEA and the FDA, differences in launch timing between the US and the EU appear to be driven less by differences in market approval requirements and more by price and reimbursement regulation in the EU, including the fact that price spillovers create incentives for manufacturers to intentionally delay launch in low-price markets. Japan is an exception, in that it has relatively high launch prices and long launch lags due to its unique requirement for country-specific clinical trials and long reimbursement process.

2.3 Benefits of Safety and Efficacy Regulation

Because the production of safety and efficacy information is a public good that may be suboptimally provided by private markets, regulation may benefit consumers, by reducing the risk that harmful or useless drugs are marketed. For safety, liability may provide an alternative corrective to regulation; however, liability is arguably less well-designed to weigh the detailed scientific evidence than regulatory review of control trial data by experts with training in the relevant science and statistics, before injuries occur. For efficacy, markets may acquire information over time, but learning by experience is slow and may entail real health and financial losses to patients and payers. The optimal regulatory policy would set safety and efficacy standards to achieve the optimal balance between costs and benefits of delayed access to new drugs. Unfortunately, estimates of the benefits of regulation are far less numerous than estimates of costs.

Peltzman (1973) attempts to weigh both the benefits and costs of the 1962 Amendments. He concludes that the benefits were minimal and were far outweighed by the costs, which he estimates as foregone consumer surplus due to the reduced flow of NCEs. These conclusions depend critically on the methods for estimating costs and benefits, which have been questioned (for example, Temin, 1979). In particular, benefits may be understated and costs may be overstated by ascribing the 1960s decline in NCE approvals solely to the Amendments. Nevertheless, this study is noteworthy in outlining a theoretical and empirical framework for evaluating the net benefits of the 1962 Amendments.

Several recent studies have examined the benefits and costs of the 1992
PDUFA priority review provisions. Between 1993 and 2003 the median time to approval for “priority” review products declined from 14.9 to 6.7 months, while review times for “standard” products only decreased from 27.2 to 23.1 months (Okie, 2005). Olson (2004) attempts to quantify the safety impact of PDUFA and compare the costs of faster approvals to the benefits. She finds that post-launch reports of adverse drug reactions (ADRs) are more likely for drugs that the FDA rates as “priority”, after controlling for drug utilization, disease characteristics, patient characteristics, drug review time and year-specific effects. Olson calculates the benefits from reduced delay using Lichtenberg’s (2002) estimate of increased life expectancy due to increased availability of new drugs. She finds a large benefit from the faster launch of new drugs with priority review status, which, under conservative assumptions (biasing against safety), is only reduced by 8% (measured in expected gain in life years) due to the higher ADRs. Subsequent research shows that the FDA’s post-marketing surveillance mechanism generally underreports ADRs, but the degree is not well established (Brewer and Colditz, 1999; Bennett et al., 2005). Whereas Olson finds significant negative safety effects of accelerated review, the General Accounting Office (US GAO, 2002) found that drug withdrawal rates differed insignificantly between the period before and after the PDUFA; however, this study did not control for other factors that may have influenced drug withdrawals rates.

None of these studies estimate the savings to firms from accelerating the R&D process, including lower capitalized costs of R&D and increased effective patent life. DiMasi (2002) estimates that a 25% reduction in phase length for all phases of clinical trials would reduce the average cost per NCE by $129m., or by 16.1% assuming a base cost of $802m. Since this estimate is based on a random sample of 68 drugs that entered clinical trials between 1983 and 1994, the absolute dollar savings may no longer be accurate. But if the percentage effect is roughly accurate, it suggests that measures to speed up regulatory review can have significant financial as well as health gains, provided quality is not adversely affected.

2.4 Discussion and Proposals for Reform of Safety and Efficacy Regulation

Some argue that drugs should be available for prescription after successful completion of phase II safety trials, with the stipulation that firms are mandated to continue with phase III and possibly post-launch efficacy trials, with results posted on the internet (Madden, 2004), such that patients and physicians could make their own evaluations as to whether expected benefits outweigh risks. The counterargument is that the limited safety and efficacy data available after phase II trials are seriously inadequate for informed
decision-making, which requires the more comprehensive data collected in phase III trials that are powered to provide statistically meaningful results. Moreover, FDA reviewers have specialized medical and statistical expertise and provide a public good in evaluating the evidence on safety and efficacy, including imposing minimum standards as a condition of launch. Such information would be underprovided in a free market regime and, even if mandated, is costly to assimilate for individual physicians, patients and payers.

Moreover, the social benefit of requiring minimum standards of safety and efficacy for marketed drugs has increased with the growth in number and complexity of drugs and with insurance coverage. In the 1960s far fewer drugs were available and virtually all consumers paid out-of-pocket. Hence the main potential benefit from a regulatory requirement for efficacy was to protect consumers from wasteful spending on useless drugs, including delayed recovery and other medical costs. Since then, the number, complexity and potency of drugs available have increased dramatically, with many consumers taking multiple prescriptions, especially seniors. Consequently, the potential frequency and severity of adverse drug reactions and interactions have increased, as has the information burden of staying informed and the potential cost from being misinformed. Moreover, the growth of insurance coverage has undermined individual consumers’ financial incentives to avoid ineffective drugs which could exacerbate wasteful spending on drugs with minimal benefit. Thus the case remains strong for a regulatory agency such as the FDA/EMEA to establish minimum standards of safety, efficacy and quality as a condition of market access.

The concern to reduce launch delay without sacrificing risk information is increasingly being addressed by supplementing pre-launch randomized controlled trials (RCTs) with post-launch observational evidence, from either controlled or uncontrolled studies. A growing proportion of drugs initially receive approval for a restricted subpopulation, with requirements for post-launch data collection, thereby permitting early access for patients with the greatest benefit/risk ratio, while protecting patients who have other alternatives until more information has been gathered. In addition, advances in data collection from routine care and in statistical methods for analyzing such data to adjust for possible non-random treatment assignments offer a potentially rich and relatively inexpensive source of information on larger populations, with detail on subpopulations and evidence on long term effects. Integrating such findings with FDA’s pre-launch data from RCTs could significantly enhance the information base available for post-launch decisions on labeling changes, treatment protocols and/or reimbursement decisions by payers.

A second unresolved regulatory issue is the optimal coordination of agency regulation and tort liability. The theory of optimal policy to control safety when markets are imperfectly informed generally views regulation and tort
liability as alternatives. In theory, since the FDA is an expert agency that employs specialists in clinical trial evaluation and is guided by advisory panels comprised of external medical and statistical experts that review and evaluate comprehensive RCT data on risks and benefits, the outcome of regulatory decisions should be better informed and more consistent across drugs than decisions of lay juries, made in the context of an adverse outcome to an individual patient who may have had many competing medical and life-style risk factors in addition to taking the drug at issue. The FDA approves drugs on the basis of population risks and benefits, which by definition are average effects, but it is intrinsically difficult to apply such trade-offs to individual patients in tort cases. For example, if the FDA decided that a 1% risk of an adverse outcome from a drug was acceptable in view of its benefits, how does a jury decide whether an individual patient’s adverse event is within this 1%, in which case the producer should not be found liable, or lies outside the 1%, in which case the drug may be less safe than expected and the firm should be liable? More generally, the concept of a “defective product”, which is the basis of product liability, is problematic when applied to drugs that necessarily entail risks and/or are ineffective for some patients. Unclear standards lead to erratic and unpredictable liability rulings, in which case incentives for safety are likely to be excessive (Craswell and Calfee, 1986). Moreover, tort decisions made ex post, after a drug has been on the market, are at risk of applying current information retroactively, that is, holding a firm liable for rare or cumulative adverse events that only emerge after widespread or long-term use, which the firm could not reasonably have foreseen and for which the FDA did not require testing. Given the extensive pre-market regulation of drugs, a possible compromise is that if a drug is in full compliance with FDA requirements, including full information disclosure by the company to the FDA, then FDA compliance should be a bar to tort claims except on grounds of gross negligence, or at least a bar to punitive damages.

A more extreme proposal would replace tort liability for negligence or product defect with a no-fault compensation fund, to provide compensation to patients injured by drugs without regard to producer negligence or product defect, funded by a tax on drugs. One model for this proposal is the Vaccine Injury Compensation Program (VICP), which was established in 1986 to provide compensation on a no-fault basis for injuries caused by vaccines, replacing tort liability on manufacturers, and funded by a tax on vaccines. However, the VICP model is relatively simple to administer because vaccine injuries are rare, they occur in otherwise healthy individuals and causation is usually clear. By contrast, patients take therapeutic drugs because they are sick; these drugs have probabilistic benefits but with no guarantee of benefit and some risk of side-effects. In these circumstances, if an individual patient does not benefit from a drug or suffers an adverse effect, determining whether
their condition is inappropriately caused by the drug or is simply the inevitable progression of their disease is problematic, both conceptually and empirically. For efficient deterrence, a no-fault compensation system must assign liability if and only if an adverse outcome is caused by a drug. Implementing such a system would be far more problematic for therapeutic drugs than for vaccines.

3. PATENTS AND REGULATION OF GENERIC ENTRY

3.1 Patents and Exclusivities for Originators

Patent protection bars generic entry for a limited period during which the originator firm may charge a price above marginal cost and thereby recoup the investment in R&D. In theory, the socially optimal patent term is defined by trading off the marginal utility gain from encouraging R&D against the loss from suboptimal consumption that may result if consumers face prices greater than marginal cost. Excessive investment to capture the rents may be an additional cost of patents (see, for example, Gilbert and Shapiro, 1990; Klemperer, 1990; Lerner, 1994; Levy, 1999). Under the World Trade Organization’s (WTO) requirements, all WTO members must adopt a patent regime with 20-year product patents from date of filing. This applies to all industries, including pharmaceuticals.

For pharmaceuticals, the nominal 20-year patent term is truncated by the R&D and regulatory process, which takes about 12 years on average, leaving only eight years of post-launch patent life. To compensate for this time lost, most countries provide some supplementary protection period. In the US, the 1984 Patent Term Restoration and Competition Act (Hatch-Waxman) grants originator drugs an extension of patent term of 0.5 and 1 year for each year lost due to clinical trials and regulatory review, respectively, up to a maximum 5 year extension and 14 years overall. Further extensions may result from filing of additional patents and/or new formulations.

In addition to patents, originator pharmaceuticals receive a “data exclusivity” period, during which the originator’s clinical trial data cannot be referenced by a generic company as part of its approval filing. In the US, new chemical entities (NCEs) receive 5 years of data exclusivity, whereas biologics will receive 12 years. In the EU, both NCEs and biologics receive 10 years data exclusivity (Kuhlik, 2004). Thus in practice, an originator’s effective patent life (time without generic competition) depends on the interplay of these regulatory protections and requirements and incentives for generic entry.
3.2 Regulation of Generic Entry

As a quid pro quo for patent restoration, the 1984 Hatch Waxman Act provides for expedited post-patent approval of generics, using an Abbreviated New Drug Application (ANDA) whereby the generic can simply show chemical and bioequivalence of their product to the originator product and reference the originator’s clinical trial data. The 1984 Act also permits generics to start work before the expiry on the originator’s patents (the so-called “Bolar Amendment”).

The Hatch-Waxman Act also provides an incentive for generic firms to challenge patents by granting 180 days of market exclusivity (other than the originator and its licensees) to the first generic firm to successfully challenge a patent (a Paragraph IV filing) and file a complete ANDA. Challengers must notify the originator firm, which can file for a legal stay that blocks generic entry for 30 months or until the case is resolved, whichever occurs first. This 30-month stay provision created incentives for originator firms to “evergreen” their patents, filing patents on ancillary features of a drug to extend potential patent life beyond the basic composition-of-matter patent term. Even if generics challenge and ultimately overturn these patents, each challenge is costly and could be countered by a 30-month stay, which could delay generic entry indefinitely. The number of Paragraph IV filings has increased from just 2% of expirations in the 1980s to 20% between 1998–2000, and higher for high-revenue products (Kuhlki, 2004). Whether this reflects increased filing of frivolous patents by originators and/or increasingly aggressive challenges by generic companies seeking payoffs in settlement is unclear. In some cases, the originator allegedly paid the generic challenger that received the 180-day exclusivity period to delay launch, thereby effectively delaying entry by other potential generic producers (FTC, 2002).

The 2003 Medicare Modernization Act reduced the incentives for frivolous patenting, by permitting only one 30-month stay per product; and reduced incentives for originator-generic collusion, by providing that the 180-day exclusivity period is forfeited if not used in a timely manner. The circumstances in which originators can legally settle with generic challengers remain somewhat unresolved because there are valid arguments on both sides. On the one hand, settlement may be an efficient means to resolve litigation and uncertainty about patent validity; on the other hand, the FTC tends to view such settlements as anti-competitive, which would be correct if the challenged patents are invalid and settlement delays competitive entry.

Litigation has also challenged the originator strategy of marketing an “authorized” generic to compete with the ANDA-approved generic during the Paragraph IV exclusivity period. Competition from an “authorized” generic reduces the price, quantity and profit earned by the exclusive ANDA generic
during the 180-day period. However, the evidence on Paragraph IV challenges suggests that generic firms still face strong incentives to challenge patents. Overall, this evidence on patent challenges, many of which are successful, suggests that a significant number of invalid patents are filed and pass preliminary review by the US Patent Office. Whether the 180-day exclusivity period provides excessive incentives for patent challenges is an important but unresolved question.

3.3 Non-patent Determinants of Generic Entry, Pricing and Diffusion

In the US the Hatch-Waxman Act laid the necessary statutory foundation for prompt generic entry after patent expiry. In 2009 generics accounted for over 60% of all prescriptions filled, and over 80% of prescriptions for post-patent compounds. But generics accounted for less than 20% of total drug expenditures because generic prices are low. Post-patent substitution of cheap generics thus permits great savings for US consumers and payers.

This rapidity of generic penetration and low generic prices reflect several other critical regulatory/reimbursement provisions which make the US generic market a pharmacy-driven market, with price-competitive, unbranded generics. First, during the 1970s–1980s, all states repealed anti-substitution dispensing laws and established default rules that allow pharmacists to substitute an FDA-approved bioequivalent generic for an originator drug unless the physician specifies that the originator brand is required. Second, since the 1990s all major payers (pharmacy benefit managers (PBMs), HMOs, Medicaid, and Medicare prescription drug plans) have established financial incentives for pharmacists to substitute generics, where available. These payers treat generics and brands as fully substitutable and reimburse pharmacies for multisource drugs using a form of generic reference pricing (see Section 4.4). Specifically, they typically pay pharmacies a Maximum Allowable Cost (MAC), which is based on the acquisition price of a low-cost generic, regardless of which generically equivalent product is dispensed. Since pharmacies capture the margin between the amount they are reimbursed (the MAC) and the price they pay to acquire a generic, they have strong incentives to buy the cheapest generics. Generic manufacturers therefore compete by discounting their prices to pharmacies and other large buyers. Generic competition in the US has been intensified by the consolidation of pharmacies into large pharmacy chains, mass merchandisers such as Walmart, and wholesalers that supply the independent pharmacies. These highly concentrated and price-sensitive buyers force intense generic price competition. Third, patients are incentivized to accept generics because if they want the brand, they must pay the difference between the MAC and the brand price (or a high, tiered co-payment).

In contrast to this pharmacy-driven generics market model in the US, where
generics are unbranded and compete on price, in many EU and Latin American countries generic markets have traditionally been physician-driven, with higher-priced, branded generics that compete on brand rather than price. In these countries, including France, Spain and Italy, generic substitution by pharmacies was not permitted unless the physician prescribed by generic name (which is rare, except in the UK). Moreover, because pharmacists were paid a percentage of the price of the dispensed drug, they had no incentive to prefer cheaper alternatives. In such contexts, generic producers market to physicians, who are typically price-insensitive and prescribe a branded generic or the originator by brand name. When generics are branded and compete on brand rather than price, generic market shares are usually lower and generic prices are higher than in a pharmacy-driven, unbranded generic market like the US (Danzon and Furukawa, 2003, 2011). Several EU countries have recently changed their regulation and reimbursement of generics, to encourage pharmacy substitution and generic price competition. Because most emerging markets have many copy products and do not require that generics be bioequivalent, generic quality is uncertain and products with a trusted brand can command high margins. Thus ironically, these relatively low-income countries have relatively high-priced generics, rather than the high-quality, low-price generics that deliver great savings to higher-income countries (Danzon, Mulcahy and Towse, 2011).

Empirical studies of generic entry have shown, not surprisingly, that generic prices are inversely related to the number of generic competitors (Grabowski and Vernon, 1992); generic entry is more likely for compounds with large markets (measured by pre-expiry brand revenue), chronic disease markets and oral-solid (pill) form (Scott Morton, 1999, 2000). Caves, Whinston and Hurwitz (1991) find that total volume does not increase after patent expiration, despite the significant drop in price due to generic entry, indicating that the price effect is offset by the negative promotion effect, because incentives for promotion cease after patent expiry. Similarly, Scott Morton (2000) finds no significant generic deterrent effect of incumbent advertising via detailing or journal advertising from 2–3 years prior to generic entry. This is unsurprising, given that the generic switching decision is made mainly by pharmacists and patients, in response to their financial incentives, not by physicians who are the target of detailing and journal advertising.

Originator firms in the US attempted to compete with generics in the 1990s, but most subsequently divested their generic activities, lacking the low-cost structures and broad product line needed to compete with generics in the US, except for “authorized” generics produced during the high-priced 180-day exclusivity period. Novartis was an exception in retaining and expanding its Sandoz generic division. More recently, several other large pharmaceutical companies have made alliances with branded generic producers in India and other emerging markets, to sell branded generics in these countries.
Depending on reimbursement rules, originators may respond to the rapid post-patent generic erosion of brand share by a range of strategies of limited value, including: raising price to maximize profit from the shrinking, relatively price-inelastic brand-loyal segment (Frank and Salkever, 1992) and/or shifting patients to newer formulations, such as a delayed release version or a follow-on product with a new patent. Such strategies may be less profitable in countries where payers apply reference pricing, reimbursing these new forms at the same price per dose as generic versions of the old formulation.

Post-patent switching to over-the-counter status may be profitable for the firm and valuable for consumers for certain drugs. If the OTC switch involves a change of formulation, strength or indication, the FDA requires additional clinical trials to show safety and efficacy under patient self-medication. To encourage these costly investments, the FDA grants three years of market exclusivity to a successful OTC switch, which delays entry of generic versions of the OTC formulation. OTC approval is traditionally only granted for drugs to treat conditions that are easily self-diagnosed; the potential for abuse or misuse is low; labeling can reasonably communicate any risks; and medical oversight is not required for effective and safe use. Prices of OTC products are usually lower than the prescription versions of the same medicines, because patients must generally pay out-of-pocket for OTC products and are therefore price-sensitive. Social welfare is likely to increase, unless the OTC entails significant patient risk or preempts a potentially cheaper generic prescription version (Temin, 1983). Keeler et al. (2002) estimate a demand function and medical and quality of life benefits for nicotine replacement drugs, which yielded a net social benefit of approximately $2b. per year from OTC conversion of these drugs.

3.4 Patents and “Access” in Developing Countries

Pharmaceutical patents can result in a welfare loss if prices exceed marginal cost, such that some consumers forego use, even though their marginal benefit would exceed marginal cost (suboptimal consumption). In most industrialized countries, comprehensive insurance coverage offsets this potential patent-induced tendency for underconsumption because patient co-payments are usually a small fraction of the full price.

However, suboptimal use of drugs due to patent-induced high prices remains a serious concern for middle and low-income countries (MLICs), where insurance is limited and most consumers pay out-of-pocket for drugs. Most of these countries now provide for a 20-year product patent, as required under the World Trade Organization (WTO)’s TRIPS (trade-related aspects of intellectual property) provisions. The proviso that governments may grant a compulsory licence to generic producers in the event of a “national emer-
has been invoked in some cases but is not extensively used, allegedly in part due to bilateral trade agreements initiated particularly by the US, that stipulate stricter patent provisions.

In practice, there is limited empirical evidence on whether product patents in MLICs would significantly raise prices and reduce consumption (see for example, Fink, 2001; Watal, 2000; Chaudhuri, Goldberg and Jia 2006). If demand facing a patent holder is highly price-elastic due to low willingness or ability to pay, then a firm’s profit-maximizing strategy may be to charge prices close to marginal cost, despite the patent. In fact, some companies did not bother to file patents in several African countries that (in theory at least) would enforce them (Attaran, 2004), suggesting that they perceived little value in patents due to some mix of highly elastic demand, costs of filing and weak enforcement. If demand is highly elastic such that, even with enforceable patents, profit-maximizing prices in low-income countries would be close to marginal cost, then the welfare loss due to patents is small, but so is the incentive to invest in R&D to treat diseases endemic to these countries. Chaudhuri, Goldberg and Jia (2006) estimate demand elasticities and supply parameters in the market for anti-infective drugs (quinolones) and conclude that patents would result in a welfare loss to consumers of $305m. per year, compared to a gain to patent holders of only $20m., and a reduction of “generic” firm profits of $35m. These welfare loss estimates are obviously sensitive to demand elasticities and might be reduced by within-country price discrimination. Danzon, Mulcahy and Towse (2011) estimate elasticity of prices with respect to mean per capita income across a broad range of MLICs countries and drugs at around 0.2, further confirming that prices are high relative to mean per capita income in low-income countries. However, they show that competition from other originators and branded generics is ineffective at reducing prices. This suggests that the problem is not so much patents per se as consumer uncertainty about quality, which leads originators and generics to use brand and price as proxies for quality.

The optimal policy approach to pharmaceuticals for less developed countries (LDCs) is radically different, depending on whether drugs are global or LDC-only. For global drugs that treat diseases such as diabetes, cardiovascular conditions etc. that are common in both developed and developing countries, patents combined with market segmentation and differential pricing can in principle enable firms to recoup their R&D investments by pricing above marginal cost in high-income countries while pricing close to marginal cost in LDCs. Thus differential pricing is key to reconciling patents to incentivize

R&D with affordability in LDCs. More formally, price discrimination across
countries can improve static efficiency by increasing consumption in LDCs,
while also enhancing dynamic efficiency, through quasi-Ramsey pricing of
the R&D joint assets (Danzon, 1997; Dumoulin, 2001; Jack and Lanjouw,
2005; Malueg and Schwartz, 1994; Maskus, 2001; Danzon and Tows, 2003,
2005). In practice, prices are high, relative to average per capita income, in
low-income countries, for many reasons, including: risks of external referenc-
ing and parallel trade (Danzon and Tows, 2003, 2005); income dispersion
(Flynn, Hollis and Palmedo, 2009) and quality uncertainty and branded gener-
ics (Danzon, Mulcahy and Tows 2011). Devising institutional frameworks
that support price discrimination both between and within countries, to
encourage lower prices for low-income populations in LDCs, remains an
important policy challenge.

However, for drugs to treat diseases that are endemic only in developing
countries, patents are ineffective at stimulating R&D, because there is insuffi-
cient high-income demand to pay the high prices necessary to recoup R&D
investments. Various “push” and “pull” subsidies have therefore been imple-
mented. “Push” subsidies focus on public private partnerships (PPPs), which
receive government and philanthropic funds plus private industry expertise
and resources, to address diseases such as malaria (Medicines for Malaria
Venture), tuberculosis (the Global Alliance for TB), an AIDS vaccine (the
International AIDS Vaccine Initiative, IAVI), and others. The basic issues are
outlined in Kremer (2002). “Pull” subsidies are designed to pay only for drugs
that meet target requirements, exemplified by the Advance Market
Commitment (AMC) approach, whereby international donors commit to
purchasing a specified number of doses at a specified price. The first AMC
was implemented in 2010 for the pneumococcal vaccine.

4. REGULATION OF PRICE AND REIMBURSEMENT
FOR ON-PATENT DRUGS

4.1 The Rationale for Price Regulation

Regulation of pharmaceutical prices is a priori anomalous because the phar-
maceutical industry is structurally competitive, with relatively low concen-
tration and easy entry, as evidenced by the continued entry of new start-up firms
and the growing share of new products discovered and launched by relatively
recent entrants. Competition occurs between therapeutic substitutes (different
compounds to treat the same condition) while a drug is on patent, in addition
to competition from generics after patent expiry. Acemoglu and Linn (2004)
show that entry of new drugs responds to expected market size. DiMasi and
Paquette (2004) find that follow-on compounds enter new classes rapidly, such that the period of market exclusivity of first entrants in new therapeutic classes declined from 10.2 years in the 1970s to 1.2 years in the late 1990s. Lichtenberg and Philipson (2002) show that therapeutic competition while a drug is on-patent reduces its net present value at launch by as much as post-patent generic entry.

However, despite this competitive market structure, price competition between on-patent drugs is weak due to two other institutional characteristics of pharmaceutical markets. First, in industrialized countries patients must obtain a physician’s prescription in order to get most drugs. If physicians are imperfect agents for patients and/or are not themselves at risk for drug spending, the separation of prescribing from consumption plausibly reduces demand elasticity.

Second, insurance coverage for pharmaceuticals reduces price-elasticity of patients, creating incentives for manufacturers to charge higher prices, in the absence of controls. Co-payments can mitigate this supplier moral hazard effect of insurance, but co-payments also reduce patient financial protection. In practice most public and private insurance plans include only very modest co-payments, often unrelated to the price of the drug and with an annual catastrophic cap on the patients’ total out-of-pocket spend. To counteract this price-increasing tendency of insurance, both private and public insurers limit the prices or reimbursement that they pay for all insured health services, including drugs.

4.2 Optimal Structure of Pharmaceutical Insurance and Price Controls

Standard models of optimal insurance contracts focus on the design of consumer co-payments to optimally balance their financial protection with incentives for overuse (consumer moral hazard) (for example, Pauly, 1968; Zeckhauser, 1971; Ma and Riordan, 2002). But for pharmaceuticals, the consumer co-payments necessary for optimal financial protection may be too low to constrain manufacturer pricing, especially for chronic and expensive drugs. Optimal provider cost-sharing has been analyzed for physician and hospital services (Ellis and McGuire, 1990), but such models do not apply to pharmaceuticals.

Optimal pricing and utilization of drugs in theory requires two conditions. First, static efficiency requires that existing drugs be available to patients whose marginal benefit exceeds the marginal cost of production. Second, dynamic efficiency requires that producers have optimal incentives to invest in R&D, which means that they should capture the full consumer surplus from new drugs. Insurance can reconcile these static and dynamic efficiency
requirements, if the consumers’ cost sharing is sufficiently low that it is not a barrier to appropriate use, and the top-up payment from the insurer to the manufacturer is sufficient for the producer to capture the full surplus (Garber, Jones and Romer, 2006; Jena and Philipson, 2008). However, given the inelastic demand that would result from such extensive insurance coverage, manufacturers would have incentives to set excessive prices in the absence of constraints. Danzon, Towse and Ferrandez (2011) show that optimal second-best pricing and utilization can be achieved if each payer defines a threshold willingness-to-pay of its citizens/enrollees for health gain in terms of an incremental cost effectiveness ratio (ICER) – for example, £30,000 per quality-adjusted life year (QALY) – and provides access to the drug only for patients for whom it is cost-effective at the price chosen by the manufacturer.

4.3 Price/Reimbursement Control in the US: Free Pricing with Negotiated and Regulated Discounts

In the US, manufacturers set list prices without constraints. Since the 1990s, private insurers and their pharmacy benefit managers (PBMs) have negotiated discounts off these list prices in return for preferred formulary placement. Because there are many actual and potential plans competing in the US private insurance market, no individual private plan has monopsony power. Drug manufacturers can and do choose not to supply a particular plan if it demands an unacceptably large discount off the list price for a particular drug.

The federal Medicare program for seniors and the disabled covers outpatient prescription drugs though the Medicare Part D drug benefit. Part D is delivered through private prescription drug plans (PDPs) that use negotiated formularies, similar to private PBMs. The federal government is specifically barred from negotiating drug prices, although this could be changed by future legislation if spending under the program far exceeds initial projections. Estimates for the Medicare drug benefit have already increased from $404b. for 2004–13 (CBO, 2004) to $724b. for 2006–15 (Kaiser Family Foundation, 2005).

The federal-state Medicaid program covers drugs for low-income families and the disabled. Medicaid requires that originator drugs give Medicaid the lower of (a) the “best price” offered to any non-federal purchaser or (b) a 15.1% discount off AMP (average manufacturer price), to increase to 23% in 2011. To discourage increases in private prices in response to the best price provision, an “excess-inflation” rebate is also required for price increases that exceed the CPI. For 2003, the combined effect of these mandatory discounts resulted in a 31.4% discount for Medicaid, relative to AMP (US CBO, 2005a). Similarly, the Big Four Federal programs (the Department of Defense, the Department of Veterans Affairs, the Public Health Service and the Coast
Guard) receive at minimum a mandatory 24% off the non-federal average manufacturing price, plus an excess inflation rebate. In 2003, the average Big Four price was roughly 38% below the AMP (US CBO, 2005b).

Thus public purchasers in the US have regulated discounts off private sector prices, but these private prices are freely set. Unsurprisingly, this regulatory structure that links mandatory discounts to public programs to the discounts drug firms grant to private plans has made firms less willing to grant discounts to private sector plans and may have contributed to increased list prices. GAO (1993) found that median best price discounts to HMOs declined from 24.4% before the Medicaid “best price” law went into effect in 1991, to 14.2% in 1993 (US GAO, 1993); CBO (1996) found similar evidence. Because discounts are confidential, academic studies have focused on the effects of the Medicaid best price provision on prices before buyer-specific discounts. Using these prices, Scott Morton (1997) found modest price increases in product categories with generic competition after the enactment of the Medicaid best price policy in 1991. Duggan and Scott Morton (2006) exploit the variation in the Medicaid market share for the top 200 selling products in the US and concluded that a 10% increase in Medicaid market share resulted in a 7–10% increase in average price. Widespread awareness that tying public prices to private prices leads to increases in private prices is one reason this approach was not adopted for Medicare Part D.

The early literature on competition in US pharmaceutical markets is interesting because it predates the advent of widespread insurance coverage, the associated price insensitivity of patients and tiered formulary approach used by payers. Opinion and evidence was divided on the extent and welfare effects of competition. Some viewed closely substitutable, patented products as wasteful “me-toos”, arguing that patent protection leads to excessive product differentiation and higher prices (for example, Comanor, 1986; Temin, 1979). Under this view, the 1962 Amendments’ requirement for proof of efficacy and restrictions on drug advertising restricted “excessive differentiation”. The alternative view was that the availability of more substitute products prior to 1962 increased price competition and benefited consumers (for example, Peltzman, 1973). Studies of launch prices and price trends over a drug’s lifecycle (e.g. Reekie, 1978; Lu and Comanor, 1998) found that new drugs that offered significant therapeutic advance were priced above existing drugs but tended to lower price over time, whereas imitators were priced lower initially but tended to increase prices. This evidence is consistent with some degree of competition but imperfectly informed buyers, such that sellers offer a low initial price to encourage use and build reputation or loyalty, then raise prices over time (Schmalensee, 1982).
The nature and extent of pharmaceutical competition changed in the 1990s as insurance insulated consumers from prices, but payers used formularies with tiered co-payments to encourage patients/physicians to use “preferred” drugs in each class. Because such strategies increase the cross-price elasticity of demand between therapeutic substitutes, manufacturers are willing to give discounts for preferred formulary status or actual gain in market share. These discounts are confidential, to encourage price competition, and are not available for academic research. Anecdotal evidence confirms the theoretical prediction that discounts are larger to purchasers that have tight control over drug use and in classes with several close substitute products.

However, this approach of negotiating discounts in return for preferred formulary status has not prevented high launch prices and annual prices increases that exceed the growth of the consumer price index (CPI) for several reasons. First, PBMs are incentivized by discounts but not by the absolute price of the drug, which is passed on to the final payer. Thus manufacturers may have incentives to compete by setting high list prices and give somewhat larger discounts than competitors. Second, incentives for discounts to private payers are muted by the requirement that Medicaid and federal payers receive the “best price” given to private payers or a minimum discount, whichever is lower, which acts as a tax on discounts greater than the required minimum. Third, payer leverage to negotiate discounts depends on availability of close substitute drugs. This approach may have worked for large, crowded therapeutic classes, such as anti-ulcerants, but it is less effective for specialty classes, such as cancer, which have fewer and less close substitutes. Payers increasingly place these expensive drugs on a specialty tier with a 20–30% coinsurance. However, this has proved ineffective at constraining the prices of specialty drugs, plausibly because most patients either have caps on their cost-sharing or receive cost-share rebates from manufacturers.

4.4 Price and Reimbursement Regulation

Most industrialized countries other than the US have either a universal national insurance scheme or a system of mandatory quasi-private social insurance funds that are regulated by the government. Governments or their surrogates therefore control prices of all reimbursed services, including drugs, as a way to control supplier moral hazard. If a firm does not seek reimbursement, it may usually sell at unregulated prices.

4.4.1 Direct price regulation

Under direct price regulation, as used in France, Italy, Spain, Japan, etc., a drug’s launch price and any price increases must be approved as a condition of reimbursement, and price decreases may be mandated. Price regulation
systems use two basic approaches to setting drug prices – internal and external referencing – and many countries use some variant of both.²

(1) “Internal referencing” benchmarks the new drug’s price to prices of other drugs in the same class. Mark-ups may be granted for improved outcomes, relative to comparator drugs and, if this comparison is rigorously done, this approach approximates value-based pricing and incremental cost-effectiveness review, but with important differences (see below).

“Reference price reimbursement” is an extreme form of internal benchmarking in which “similar” products are reimbursed the same price per daily dose – the reference price (RP). “Generic referencing” groups products with the same active ingredient and “therapeutic referencing” groups different compounds with similar mode of action and/or same indication. The RP is usually based on the price of the lowest or a low-priced product in the group. Manufacturers may charge prices above the RP, but patients must pay the excess. In practice, manufacturers typically drop their prices to the reference price to maintain market share, implying that demand is highly elastic at prices above the RP.

Although reference price reimbursement in theory limits reimbursement but not the manufacturer’s price, in practice RP reimbursement may be more constraining than internal benchmarking of prices for several reasons. First, whereas price regulation usually benchmarks to the most similar product and permits mark-ups for drugs with superior benefits, RP imposes the same reimbursement per daily dose for all products in a group. To obtain a higher reimbursement requires a separate classification, hence assignment of individual drugs is often litigated. Second, therapeutic RP systems cluster compounds without regard to patent status. Thus if generic entry occurs for one compound in a group and this generic price sets the RP, reimbursement for other on-patent products may drop to the generic price, thereby effectively truncating patent life of late entrants in a class unless patients are willing to pay surcharges. This may discourage R&D into follow-on products and new formulations of existing compounds, which may be welfare-enhancing or welfare-reducing, depending on the product and relative share of the country adopting such reimbursement within the global market. The experience so far in countries that have adopted therapeutic referencing (Germany, the Netherlands, New Zealand) is insufficient to predict the likely effects on R&D

² Although some countries, including Italy, have attempted to base prices on costs, this approach is not widely used because of the difficulty of obtaining accurate measurement of costs. Measuring R&D cost is particularly problematic, because it occurs over many years, includes the cost of failures and foregone interest, and is largely a joint cost that must be allocated across global markets. In practice, price regulation based on costs has relied on transfer pricing rules which were designed for tax purposes, not price regulation.
if the US, with its large share of global revenues and highly price-competitive generic market, were to adopt therapeutic RP (Danzon and Ketcham, 2004).

The early literature on RP is summarized in Lopez-Casasnovas and Puig-Junoy (2000); more recent studies are reviewed in Galizzi, Ghislandi and Miraldo (2011). Danzon and Ketcham (2004) provide empirical evidence on effects of RP in Germany, the Netherlands and New Zealand, the three most comprehensive RP systems. This evidence suggests that RP had little effect on average drug prices or drug availability in Germany or the Netherlands, but that effects on prices and availability were significant in New Zealand, which used broader classes and where the regulatory agency explicitly required RP-reducing price cuts as a condition for admitting new drugs to reimbursement.

The evidence on patient health outcomes under RP is mixed: some studies find no evidence of adverse effects, while others find an increase in adverse outcomes, possibly because patients switched to less appropriate drugs to avoid surcharges. The risks of such adverse effects depend on the substitutability of drugs within a group, which varies across therapeutic classes. For this reason, Australia and British Columbia only apply RP to therapeutic classes in which drugs are considered highly substitutable for most patients.

(2) “External referencing” benchmarks the price of a drug in country A to the mean, median or lowest price of the identical drug in specified other countries. For example, Italy uses an average European price, Canada uses the median of seven countries (five European countries plus the US and Japan), etc. External referencing thus constrains the manufacturer’s ability to price discriminate across countries. It creates incentives for manufacturers to seek a common launch price across linked markets and choose to delay or not launch in countries that do not pay that price, particularly those with small markets. Evidence from several studies supports these predictions (Danzon, Wang and Wang, 2005; Kyle, 2007; Lanjouw, 2005; Danzon and Epstein, 2009). Although parallel trade in principle creates similar incentive effects to external referencing, in practice manufacturers can usually constrain parallel trade flows by limited sales, dual pricing and other strategies in potential exporting countries.

External referencing by higher-income/higher-price countries to lower-income/lower-price countries imposes a welfare loss in the latter, if manufacturers seek to raise prices in these countries or delay launch, to avoid price erosion in higher-price markets. Evidence confirms that prices are generally higher, relative to mean per capita income, in relatively low-income, parallel exporting EU countries (Danzon and Epstein, 2009). More generally, regulatory strategies that lead to price convergence across countries at different income levels are likely to reduce overall welfare, compared with regulatory strategies that permit price discrimination, due to both output increases (static efficiency gain) and greater incentives for R&D (dynamic efficiency gain)
Moreover, differential pricing is also consistent with Ramsey pricing applied to paying for the joint costs of R&D (Ramsey, 1927; Baumol and Bradford, 1970; Danzon and Towse, 2003, 2005; Jack and Lanjouw, 2005). Traditionally, external referencing and parallel trade apply mostly between countries at similar levels of income within Europe. Welfare losses would likely be much larger if referencing or importation were extended to lower-income countries or if the US were to legalize drug importation from a broad group of countries, as has been proposed.

4.4.2 Drug budgets and expenditure controls
Many countries that regulate prices have added other measures to limit total drug spending. Such controls may be part of the initial price negotiation, which specifies the eligible population for which the drug is expected to be cost-effective and the target spend at the approved price. If actual spending exceeds this target, the price is cut proportionately such that spending remains within target. France applies such limits to some individual drugs, some therapeutic classes and each company’s total revenues, in order to constrain total drug spend within target levels and deter excessive promotion.

From 1993–2003 Germany implemented a budget for aggregate outpatient drug spending, with successive tiers of any overrun to be recouped from physicians and the pharmaceutical industry. Physicians responded by reducing the number of prescriptions and switching to cheaper drugs, leading to a 16% reduction in drug spending in the first year of the budget (Munnich and Sullivan, 1994). Schunenburg and Schoffski (1994) report that referrals to specialists and hospitals increased, because the drug budget excluded inpatient drugs. Germany’s aggregate drug budget was abolished in 2003, because enforcing the repayment of overruns was practically and politically problematic. Germany and the UK have also used physician-specific indicative budgets, with warnings but not financial penalties for overruns, possibly because such penalties could create incentives for physicians to avoid high-risk patients and/or make inappropriate drug choices if risk-adjustment tools are not sufficiently accurate.

4.4.3 Profit or rate-of-return controls
The UK’s Prescription Price Regulation Scheme (PPRS) is unique among industrialized countries in regulating the rate of return on capital, leaving manufacturers (relatively) free to set the price of individual drugs. The PPRS, which was renegotiated every five years between the pharmaceutical industry and the government, in theory provided an indirect control on drug prices by limiting each company’s revenues from sales to the UK National Health Service as a percentage of their capital, with specified limits on allowable
expenses and no post-launch price increases permitted. The allowed rate of return was around 17–21%; excesses could be repaid directly or through lower prices the following year. Companies with minimal capital in the UK could substitute a return-on-sales formula.

This system encouraged early launch of new drugs in the UK, by eliminating any need for prior price approval. Danzon and Percy (1996) found that although the rate of growth of capital and labor in the UK pharmaceutical industry has been high, relative to other UK industries and relative to pharmaceuticals in other countries, investment has not been biased towards capital relative to labor, as predicted by simple models of rate-of-return-on-capital regulation. Overall, the UK experienced relatively high total factor productivity growth, compared to other regulated and unregulated countries; however, whether this is due to the PPRS or to other policy or market characteristics of the UK cannot be determined. Although the UK’s drug prices have been higher than those in the price-regulated markets of France, Italy and Spain, they have not necessarily been out of line with income differences and precise differentials are sensitive to exchange rates (Danzon and Furukawa, 2003, 2008). The UK’s overall spending on drugs, either as a share of health spending or per capita, is not out of line with other EU countries, plausibly reflecting other characteristics of their health care system, including strong pharmacy incentives for generic substitution and physician training and incentives for cost-conscious prescribing. The UK pharmaceutical industry has also contributed more significantly to the flow of new medicines than most other countries of comparable size. Nevertheless, following a recent review and recommendation by the UK Office of Fair Trade that the UK move to a “value-based pricing” regulation, in place of profit regulation (UK Office of Fair Trade, 2007), in 2010 the UK government proposed such a shift. The details of this approach and its effects remain to be determined.

4.4.4 Cost-effectiveness requirements

A growing number of countries, including Australia, Canada, New Zealand and the UK, review the cost-effectiveness of new drugs as a condition of reimbursement. In other countries, such data are one input to price negotiations. The UK’s National Institute for Clinical Excellence (NICE) is a government-appointed but independent expert body that reviews the cost-effectiveness of technologies expected to have a major health or budgetary impact, including drugs, relative to current treatment, using specified methods to measure costs and quality-adjusted life years (QALYs). In 2004 Germany established a similar expert body (IQWIG) to review clinical effectiveness and now cost-effectiveness, and others are under debate in other countries.

The compelling rationale for cost-effectiveness review (CEA) as a condition of reimbursement is that CEA provides a systematic approach to evaluating
costs and outcomes of alternative uses of resources which is essential to achieve maximum health gain from fixed health budgets. Moreover, once a payer specifies a threshold incremental cost-effectiveness ratio (ICER) (for example, UK NICE uses roughly £30,000 per QALY), this provides an indirect control on the price a manufacturer can charge for a drug and still be cost-effective, given the drug’s incremental effectiveness relative to comparator products. For example, in the simplest case where the cost of treatment is simply the drug, with no other medical costs or cost offsets, the incremental price that a new drug can charge, relative to the current therapy, is simply the incremental QALYs gained, evaluated at the payer’s ICER threshold. If each payer/country chooses the ICER threshold that reflects its willingness to pay for health gain, this would achieve within-country price levels and between-country price differentials that are second-best optimal (Danzon, Towse and Ferrandez, 2011). The increasing use of some form of CEA and the growing literature on appropriate measurement of effectiveness and ICER thresholds offer hope that this approach can provide a more theoretically sound framework for drug price regulation.

5. INDUSTRY STRUCTURE, PRODUCTIVITY AND PROFITS

5.1 Barriers to Entry and Industry Structure

Several early studies argued that the 1960s increase in regulation of R&D created scale economies that favored large firms (Grabowski, 1976; Grabowski, Vernon and Thomas, 1978; Temin, 1979). Thomas (1990, 1996) showed that the post-1962 decline in NCE introductions in the US was concentrated in the smallest firms, many of which ceased R&D.

More recently, the biotechnology and genomics revolution appears to have eliminated the advantages of scale, at least for drug discovery, and this has dramatically changed the structure of the pharmaceutical industry. Previously, the chemistry basis of drug discovery implied an advantage for large firms that had large proprietary libraries of compounds, often created by their in-house chemists. Now, drug discovery is based on microbiology and associated sciences, with comparative advantage in smaller firms that are often spun out from academic research centers.

Large firms have adapted to this shift by acquiring biotechnology companies or in-licensing their compounds. Although some large firms have grown larger through horizontal mergers with peers and have often rationalized these horizontal mergers on grounds of economies of scale and scope in R&D, the empirical evidence does not support the claims (Danzon, Epstein and...
Nicholson, 2007). A growing share of new drug approvals is originated by smaller firms, including both biologics and some chemistry-based drugs. Initially these start-up small firms specialize in discovery research, sometimes forming alliances with larger firms that provide funding and expertise for late-stage clinical trials, regulatory approval and marketing, where experience and size play a greater role (Danzon, Nicholson and Pereira, 2005). The growth of contract research, sales and manufacturing organizations has increased the outsourcing opportunities for small firms and hence reduced their need to rely on larger, more experienced partners, and many hire experienced personnel from larger firms. Thus if the 1962 regulatory changes did disadvantage small firms and increase industry concentration, these effects have been dominated by technological change that has emerged from and benefited small firms.

It is possible that the high rate of new start-ups to some extent reflects excessive entry as firms compete for profits in a differentiated products oligopoly. However, it is also true that many new start-ups form around new technologies and that this scientific uncertainty can only be resolved by clinical testing that takes time. The rate of discovery of new technologies is driven in part by the extent of public funding of basic research and the incentives under the Bayh Dole Act (1980) to commercialize such research, and possibly by favorable tax treatment of R&D, especially for orphan drugs. Whether or not public funding for basic research is excessive or suboptimal is an important subject for research. Thus in the current environment it does not appear that regulation of market access or endogenous investments in sunk R&D costs are major contributors to excessive product differentiation or monopoly power, with the possible exception of orphan drugs that by regulatory design receive five years of market exclusivity and other subsidies. To the extent that incentives exist for excessive prices and excessive product differentiation, including extensions and new formulations, this is more plausibly due to the structure of insurance in the US than to market access regulation.

5.2 Profitability and its Measurement

The pharmaceutical industry is widely perceived to earn excessive profits. Accurate measurement of profits is particularly problematic for pharmaceuticals due to the length of product life and the importance of intangible investments in R&D and promotion. These are expensed under standard accounting practices, whereas from an economic perspective they are long-term investments to be recouped over the product life. Clarkson (1996) illustrates the effects of these adjustments for firms in 14 industries for the period 1980–93. Before adjustment, the average accounting rate of return on equity for the 14 industries was 12.3%; the pharmaceutical industry had the highest return of 24.4%. After adjustment for intangible capital, the average was 10.2%
compared to 13.3% for pharmaceuticals, which was less than the adjusted return for petroleum, computer software and foods.

A second approach to profit measurement uses the Lerner index of price relative to marginal production cost. Caves, Whinston and Hurwitz (1991) estimate this ratio at roughly 5, using the ratio of originator price relative to generic price several years after patent expiry. However, this post-patent price ratio overstates the average Lerner index over a product’s lifecycle in the US because originator prices rise and marginal costs decline with time since launch. More fundamentally, a one-year Lerner index based on short-run marginal production cost in one country is both theoretically and empirically inadequate as a measure of profit for global products with high and long-lived R&D investments.

A third – and conceptually more correct approach – measures the rate of return on investment in a cohort of drugs, using discounted cash flow estimates of costs and returns. Grabowski and Vernon (1990, 1996) and Grabowski, Vernon and DiMasi (2002) apply this approach to estimate the return on R&D for new drugs introduced in the 1970s, early 1980s and 1990s, respectively. Starting with market sales data for the US, they extrapolate to global sales and subtract estimated non-R&D costs to yield a discounted present value at launch of net sales. Comparing this NPV of net revenues to the estimated average capitalized R&D cost per new medical entity (NME) at launch, Grabowski and Vernon conclude that the 1970s drug cohort on average earned a return roughly equal to their cost of capital; the 1980s cohort on average yielded a positive net present value of $22.2m., or an internal rate of return of 11.1%, compared to the 10.5% cost of capital; and the 1990s cohort shows a similar, small but positive excess return. Given the large number of assumptions, confidence intervals are not reported and estimates are approximate.

More fundamentally, although this cohort rate-of-return approach in theory provides the most accurate measure of returns to R&D, it is arguably of limited policy relevance in an industry with low entry barriers but long lead times for product launch and high scientific and market risk. In particular, if the expected return on R&D exceeded the cost of capital, competitive entry would occur until any excess expected profit is eliminated. Such competitive adjustments are not instantaneous, due to time lags in R&D, and the actual realization of returns may differ radically from those anticipated due to regulatory, scientific and market risk. But if the assumption of dynamic competition with free entry is correct – and all the evidence suggests that it is – then if research were to estimate that returns either exceeded or fell short of the cost of capital over a particular time period, this would either reflect measurement error by the researchers or temporary market disequilibrium that will be corrected by competitive entry, such that the research estimate may already be obsolete for more recent cohorts.
Given that competitive entry to exploit R&D opportunities should reduce expected profits to competitive levels, the more important policy question is whether regulations governing pricing, reimbursement and generic entry yield an expected profit that attracts a socially appropriate level and mix of R&D and new drug flow. Alternatively stated, given low barriers to entry, changes in regulation and reimbursement may affect profitability in the short run. But in the long run, the level and mix of R&D readjust such that normal returns are realized on average. In the current context, if firms continue to be willing to invest despite rising costs per new medical entity approved, high failure rates and low marginal health gains from many new drugs, this suggests that prices are sufficiently high to make such R&D potentially profitable.

Viewing R&D as endogenous in this way emphasizes the societal importance of payers defining clearly what they are willing to pay for incremental health gain by setting ICER thresholds, possibly differentiated by disease type etc. The ex ante uncertainty as to the ultimate therapeutic value and timing of new drugs implies that ex post realizations will still yield some failures and some “me-too” drugs. Indeed, the optimal number of me-toos is uncertain, given their value as a competitive constraint and in improving therapies for some patients. Thus if product differentiation is perceived to be excessive in the pharmaceutical industry, this is more likely due to poorly structured generous insurance coverage and high reimbursed prices, rather than firm strategies to use endogenous investments in R&D or marketing as an (unsuccessful) entry barrier.

6. PROMOTION

6.1 Regulation of Promotion

Promotion of prescription drugs in the US has been regulated by the FDA since the 1962 Amendments. This statute restricts promotional claims to facts established in clinical trials; requires that risks as well as benefits be described in brief summary; and excludes promotion of unapproved indications. The FDA’s 1997 Guidance on direct-to-consumer advertising (DTCA) relaxed the requirement that the full product label, which includes all known risks, be displayed in broadcast ads. Rather, the requirement for a brief summary of risks and benefits could be provided by giving a website, a toll-free phone number, or reference to a print ad with the full label, in addition to advice to “see your physician.”

The US constitutional right to freedom of speech has been interpreted to include commercial speech and hence to constrain regulation of promotion. Promotion of information about off-label (unapproved) uses of drugs was not permitted until 1997, when companies were permitted to disseminate
peer-reviewed publications discussing off-label use. In its oversight of promo-
tion, as for its other activities, the FDA is required by statute to consider risks
and benefits; costs are not mentioned. Thus the FDA is concerned with the
effects of promotion on patients and physicians; whether or not it results in
unnecessary costs is beyond its purview.

Most countries other than the US restrict direct-to-consumer advertising
(DTCA) to “help seeking” ads, which inform consumers about a health condi-
tion and potential treatment for that condition without mentioning a specific
product. The only other country that permits product-specific DTCA is New
Zealand which, like the US, has a strict freedom of commercial speech
commitment. Several countries that regulate drug prices include features to
discourage promotion – for example, the UK PPRS limits the promotional
expenditure that can be deducted in calculating the net rate of return and
France penalizes “excessive” promotion through price penalties for over-
shooting sales targets. Whereas free samples are an important part of detailing
in the US, many other countries prohibit free samples.

Empirical evidence on cross-national differences in promotion and its
effects is more limited than for prices, in part because promotion content
varies across countries and data are limited. For example, counts of physician
detailing visits may be misleading if content differs, depending on time spent,
messaging allowed, whether sampling is permitted etc. Berndt, Danzon and
Kruse (2007) provide some evidence on cross-national differences in promo-
tion and in diffusion of new drugs. Most other studies are US-focused.

6.2 Information or Persuasion?

The pharmaceutical industry’s large expenditures on advertising and promo-
tion are controversial in both the economic literature and the policy debate,
with concern over both magnitude and form. Critics question the social value
of these large promotional expenditures and charge that they lead to increased
market power and higher prices. The alternative view is that promotion
provides useful information to physicians and consumers.

Considerable research has focused on estimating the competitive effect of
pharmaceutical promotional expenditures. Promotion is estimated at about
17.1% of sales – similar to spending on R&D and to promotion in several
other experience-good industries with product differentiation such as toys and
cosmetics (Berndt, 2005). This estimate is downward biased because it omits
the promotion-related components of clinical trials. On the other hand, the
estimate is upward biased because almost two-thirds reflects free samples
distributed to physicians for patient use (Berndt, 2005), and these samples are
valued at either a list price or a retail price that significantly exceeds the
economic cost to manufacturers. The next largest components of promotional
spending in the US are physician detailing, direct to consumer advertising, hospital detailing and medical journal advertising (Berndt, 2005). Direct to consumer advertising (DTCA) grew rapidly in the 1990s, and the broadcast share grew particularly rapidly after the 1997 FDA Guidance. DTCA is concentrated on the leading drugs in therapeutic categories that are particularly amenable to patient awareness and choice, such as antihistamines, antihyperlipidemics, anti-ulcerants etc. (Rosenthal et al., 2002; Berndt, 2005).

6.3 Evidence on Effects of Pharmaceutical Promotion

6.3.1 Promotion studies pre-1997
An early proponent of the anti-competitive hypothesis, Walker (1971) argues that large promotion expenditures raise entry barriers and increase market power, by requiring new entrants to make large outlays in order to attract attention to new products. The alternative view is that advertising may enhance competition by facilitating the introduction of new products and new firms (Schwartzman, 1976; Telser, 1975). Leffler (1981) found a positive relationship across therapeutic categories between selling effort as the dependent variable and number of new products introduced as the primary explanatory variable. He concludes that pharmaceutical advertising is at least partly informative, but also finds evidence that advertising of established products accomplishes “reminder” and “habit-formation” purposes. These results suggest, unsurprisingly, that advertising performs both functions and that its net effect on competition may differ, depending on the circumstances.

6.3.2 Effects of direct-to-consumer advertising (DTCA) post-1997
Several studies of DTCA have focused on its effects on drug sales in aggregate and on the share of the advertised brand. Although some of these studies use state-of-the-art methods applied to the best data available and provide valuable evidence, important positive and normative questions remain unresolved, due to both empirical challenges and the difficulty of drawing overall welfare conclusions.

The major empirical challenges are that DTCA is endogenously and simultaneously determined with other types of promotion, and may have long-term effects. Endogeneity can lead to biased estimates of the effect of promotion on use if market leaders have greater incentives to invest in promotion, because physicians tend to prescribe the best-in-class drug even without promotion. Simultaneous determination can lead to biased estimates if drugs that are more heavily promoted to consumers are also more heavily promoted to physicians and contracted for preferred formulary position. Estimating incremental effects of each type of promotion is problematic when all are highly correlated and data are often imperfect (Narayanan, Manchanda and Chintagunta, 2005).
Lagged effects of information stocks and habits of physicians and patients further compound estimation of marginal effects. Finally, the net effect of one firm’s promotion depends on competitors’ strategic responses.

Drawing welfare conclusions from the empirical evidence is problematic for pharmaceuticals. The economic/marketing literature generally views advertising that expands aggregate category sales as more likely to be informative and hence welfare-enhancing, whereas advertising that simply changes market shares without affecting aggregate use is more likely to be wasteful (Berndt, 2005; Kravitz et al., 2005). However, because consumers pay only a small fraction of the cost of pharmaceuticals out-of-pocket, even category-expanding effects could reflect unnecessary and/or unnecessarily costly use, even though such purchases are rational for individual consumers, given their insurance coverage. In general, studies suggest that DTCA has a greater effect on category sales than on individual brand sales (Rosenthal et al., 2002).

Wosinska (2002) finds that DTCA for cholesterol-reducing medications (statins) positively affects brand share only if the brand had preferred formulary status. She also finds that total DTCA spending had a minor impact on adherence to statin therapy, but current and lagged own DTCA has no affect on product adherence (Wosinska, 2004). Iizuka and Jin (2005b) find that DTCA increases total category sales, but brand-specific share is only significantly increased by physician promotion such as detailing and journal publications. They also find that DTCA increases the number of doctor visits at which a drug is prescribed (Iizuka and Jin, 2005a). Donohue and Berndt (2004) find that DTCA has no significant effect on choice of product, but that it does motivate individuals to visit the physician. Iizuka (2004) finds that high quality drugs, as defined by whether a drug had “priority” status for FDA approval, have significantly more DTCA spending.

6.3.3 Welfare conclusions
Some of the effects of DTCA appear consistent with welfare improvement, while other evidence suggests some inappropriate effects. None of the existing studies incorporate cost effects of DTCA, which would be relevant to an overall welfare evaluation. More fundamentally, the nature of all types of promotion is changing, as cost pressures, declining returns and the shift to specialty products have led to major cuts in sales force size. Safety issues with several widely advertised products have prompted both the FDA and industry to review their promotional policies (Dubois, 2003). Industry has issued voluntary guidelines for DTCA which reinforce the “fair balance” standard; stipulate that firms provide the FDA with copy of advertisements prior to, rather than concurrent with, planned public release (PhRMA, 2005); and call for firms to abstain from DTCA for several months after the launch of a new drug, to enable education of physicians in advance of DTCA release.
7. CONCLUDING COMMENTS

Economic analysis of the pharmaceutical industry has made many valuable contributions, in framing important issues and providing useful empirical evidence. Nevertheless, we still lack complete answers to some of the basic questions raised by policymakers and academic research. What is the optimal trade-off between access to new drugs and evidence on safety and efficacy, and how should pre-launch and post-launch data be combined? While the methodology of cost-effectiveness analysis has become increasingly sophisticated for use in reimbursement decisions, little progress has been made on the application of such concepts or other formal decision analytic tools to the weighing of risks and benefits in drug approval decisions, or determining optimal thresholds for safety and efficacy. While much progress has been made on streamlining the regulation of generics, to what extent should rules differ for approval and reimbursement of generic biologics/biosimilars? If the consumer and producer moral hazard effects of insurance justify some forms of control over prices, which forms provide the best trade-offs between reasonable control of costs, appropriate access for patients and incentives for innovative R&D? How should pharmaceutical promotion be regulated, to provide information and encourage appropriate use and adherence without stimulating inappropriate use? These are only some of the interesting questions that remain to be explored.

In summary, although there is a large and growing literature on the law and economics of the pharmaceutical industry that has produced valuable information and useful lessons learned, large and important issues remain for future research. Models from other industries are either not relevant or require significant adaptation and extension to fit this industry’s peculiar characteristics – in particular, high rates of R&D with life-or-death effects, patents, insurance, and physicians, consumers, payers and pharmacists as potential customers. This industry remains a fertile area for future research.

BIBLIOGRAPHY


Regulation of the biopharmaceutical industry


