Developing an empirical model of pricing and competition in pharmaceuticals *

(DISCUSSION PAPER)

Author: Rodrigo Refoios Camejo
Department of Health Policy & Management
Erasmus University
Rotterdam, The Netherlands

* Work in progress - please do not quote without author's permission

HESG June 2010
Cork

1. INTRODUCTION AND BACKGROUND

Competition in pharmaceuticals has been a much-debated issue in the health economics literature. It is often perceived that competition in the pharmaceutical market is weak because the presence of insurance and the agency physician/patient relation distort consumers' price elasticity. That is in fact the basis to support the reasonably strict regulation of pharmaceutical launch prices, which is formally widespread over Europe with the exceptions of Germany and the United Kingdom (UK).

Despite assuming that the pharmaceutical market is imperfectly competitive, drug prices are still likely to be negatively affected by the number of patented close substitutes or generic competitors. On the other hand, efficacy and safety measures as proven in clinical trials, promotional efforts or administration convenience are expected to concede a price premium over direct competitors. On top of these, other systemic factors as the implementation of newly negotiated pricing schemes or the development of clinical guidelines may act as shocks to the pricing landscape.

Although a patent-based reward system, which intentionally limits competition, being in place, the quasi-simultaneous launch of close therapeutic substitutes is frequent. In the literature, the pricing decision facing the developer in face of reference pricing has been theoretically analysed, but the subsequent stages of generalised market competition were not explored [1]. Real prices were found to decrease with time [2] but no actual attempt succeeded at clearly identifying the major price determinants through the molecules' life cycle.

Since the pharmaceutical industry's pricing strategy in the face of competition is not fully known, we intend to identify the nature of competition and estimate their role as determinants of the price of pharmaceuticals. We hope to play a part in understanding the dynamics of the pharmaceutical market, and in this discussion paper we approach the methodology used in our attempt at modeling those dynamics. We focus on a particular therapeutic area (anti-hypertensive drugs) in the United Kingdom market and use it as a case study for methods validation purposes.

This paper is structured as follows: section 2 introduces the methodology used, describes the data, and the puts forward a general model construction; section 3 reports on the results conferring methods applied and model appropriateness; and section 4 discusses the strengths and limitations of the model specification and assesses potential implications of those results in further analyses.

2. METHODS

The analysis assessed in this study aimed at empirically testing hypotheses on the impact of different factors on pricing of pharmaceuticals. We defined and included several indicators in the models and also intended to control for the introduction of wider systemic factors such as publication of clinical guidelines or the introduction of regulated pricing schemes. We applied different dynamic panel data methods and tested for the appropriateness of model specifications and indicators proposed through standard econometric tests.

2.1. Methodology Background

In economics there is often the need to capture the process of particular phenomena and the factors that contribute to their dynamics. Since cross sectional data sets do not allow testing any dynamic relationships, data sets with observations at different points in time must be used. Panel datasets, i.e. following the behavior of a number of individuals over a finite time period allow accounting for the simultaneous occurrence of dynamics and any unobserved individual heterogeneity in the phenomena of interest. This type of data structure is therefore seen to represent an advantage over aggregated time series since no microeconomic dynamics will be masked by the bias inherent to data aggregation.

In econometrics, the nature of the data is of considerable importance since it determines to a great extent the particular type of model to use. Autocorrelation and partial autocorrelation coefficients can be estimated directly from the raw data and used to guide this decision. Once models are estimated they can be compared with alternative specifications and their appropriateness can be assessed. One critical factor with panel data models

is the number of time periods T included, since normal autoregressive models were shown not to be consistent for finite T.[3]

To overcome this issue, a number of consistent estimators have been proposed. One of the most used is one of two variations (first-difference transform or orthogonal deviations) of a Generalised Method of Moments (GMM) estimator developed by Arellano and Bond, and Arrellano and Bover.[4] [5] However, Blundell and Bond [6] observed that with highly persistent data first-differenced estimators may suffer of a severe small sample bias due to weak instruments, and suggest a system GMM estimator. Both these estimators are expected to hold for a relatively large N and small T, but they can be biased and inefficient in the case of small N.[7]

In short, these estimators are devised to be used where there is a dynamic process in place with current realisations influenced by past values of the dependent variable. They allow for arbitrarily distributed fixed individual effects using those variations over time to identify parameters. The non-fixed effects-related disturbances may be serially correlated and have patterns of heteroskedasticity within individuals, but must be uncorrelated across individuals.[8] Additionally, variables do not need to be strictly exogenous and some may be pre-determined, i.e. with current disturbances eventually influenced by past ones, and still be used in the regression equation (although we can only include the levels which are unrelated to the error term).

2.2. Data

Products included in the dataset comprised all generic and non-generic antihypertensive drugs (i.e. recommended for first, second and third line treatment of hypertension in the UK [9]) with reported sales in the UK market between June 1998 and June 2008 as identified from the Intercontinental Medical Statistics (IMS) Health database. Original sample consisted of 1017 presentations, which were spread over 214 products and 50 molecules. Retail pricing, launch and market quarterly data were available for all presentations. In order to address the dimensions of pricing, several indicators were developed for price, competition, quantity and quality from the measurements available in the dataset.

Price

Price was given by the price per usual daily maintenance dosage. This was computed for each presentation using the retail list price and the dosage suggested in the products' most recent Summary of Product Characteristics (SPC). The price per daily maintenance dose was also computed at molecule and product level using the volume-weighted average of all presentations available in each time period.

Molecule launch price per daily maintenance dose and product launch price per maintenance dose were available for only a few individual products. The price per daily maintenance dose of molecule, subclass, and disease area market leaders were recorded for all products in each period. It should be noted that as available pricing data did not reflect discounts practiced at the distribution level, price variation may be distorted, especially when volume competition from generic products is in place.

Competition

A number of indicators were developed to use as explanatory variables trying to capture competition within molecules, within subclasses and within the disease area as a whole. These comprised the number of generic and non-generic products available within each molecule and subclass in each time period as well as, if available, at time of products' launch.

Lagged market share (defined by relative number of packs) of molecule, subclass and disease area market leaders were computed. Time of entry of first generic in molecule relative to originator; and time of entry of second molecule relative to subclass originator were also though as a possible indicator of the amount of competition faced.

In developing these indicators, it was assumed price does not affect the number of generic or non-generic close substitutes entering the market in the same molecule and subclass. This may be a limitation as the price is expected to influence profitability, which in theory is itself a determinant of the number of competitors. As a consequence, the effect of competition determinants on pricing may be underestimated.

Quantity

Data on the quantity dimension were available at presentation, product, molecule and subclass levels. Indicators calculated included the number of packs of product sold relative to the respective molecule and subclass sales in each quarterly period.

Quality

Quality indicators were present as measures of clinical efficacy and safety at the subclass level. These included average decrease of systolic blood pressure and incidence of adverse and serious adverse events. Ideally, measures of clinical effectiveness at molecule level should have been used but these were not available for all molecules and therefore could not be included. To address this issue, the lag to the first molecule in subclass was included under the assumption that more recent compounds within a subclass are generally more effective.

Product life cycle indicators were introduced though measures of product age including variables as the lag to product and molecule launch. The products' market entry lag relative to: the first product in subclass; the first product in molecule; and to the first generic in molecule, were also raised. Additionally, dummy variables were included to identify therapeutic subclass and control for products being innovators, generics or licensed copies.

2.3. The Model(s)

As discussed above, in order to take account of the dynamic elements of the pricing process, we developed different panel data models and autoregressive distributed lag models. The proposed general model is of the form:

$$y_{i_t} = \alpha + \beta_1 y_{i_{t-1}} + \beta_2 x_{i_t} + \beta_3 x_{i_{t-1}} + e_{i_t} + u_{i_t}$$

The analyses were conducted at the level of the product as defined by subclass, molecule and product name. Although data were available at the level of the individual presentation as defined by pack size and dosage, pricing strategies are expected to be run at a product level and using weighted price indicators was thought to be more appropriate.

Since products' list price is not expected to vary continuously due to the cost

and logistic implications such modifications entail, observations were defined in yearly terms as given by data points relative to third quarter of each study year.

Variables

Taking in consideration collinearity and overidentification issues, a subset of theoretically pertinent variables were selected from the available data. Those included the price per daily maintenance dose (price_MD) as dependent variable and up to 4 of its lags (L1.price_MD, L2.price_MD, L3.price_MD, L4.price_MD).

Product-specific characteristics as the entry lag of product molecule relative to first in subclass (lag_m2_1st_sc) and time elapsed since molecule launch (d_lag2_mol) were included to control for the products' place in the subclass launch sequence and stage in the products' lifecycle, respectively. Dummy variables regarding product innovative status (original, generic) and subclass dummy variables (scC3A, scC8A, scC9A, scC9B, scC9C, scC9D, scC9X) were also included.

In what competition is concerned, the indicators selected were the price of the market leader in class (pMD_ml_c) and subclass (pMD_ml_sc); the number of products in the subclass (n_prod_sc); and the number of generics in the molecule (n_gen_m). A lag structure for competition indicators was tested under the assumption that these may be expected to impact on price mostly in the following period. Quantity indicators regarding product sold relative to molecule (q_prodRmo) and molecule sold relative to subclass (q_mol_sc) were assumed to be pre-determined, whilst price of molecule market leader (pMD_ml_mol) was assumed to be endogenous. As suggested by Roodman [8], time dummies corresponding to the observation year were incorporated in an attempt to minimise any correlation across individuals in the idiosyncratic disturbances. Please see Appendix 1 for the complete set of variables used and respective labels.

Analysis

We adopted a general to simple model strategy when specifying the model and subsequently compared the appropriateness of the restricted models. An

80% confidence interval was used as a decision rule to drop coefficients based on significance levels. We consider this level to be appropriate due to the riskless exploratory nature of the analysis and based on the knowledge that other factors not present in our data are expected to be of influence at an uncontrollable systemic level.

We started by applying Ordinary Least Squares (OLS) and Least Squares Dummy Variables (LSDV) to the model. This was intended to be used later to assess the consistency of other estimators as a suggested by Bond.[10] We specified the model using the first-differences method suggested by Arellano and Bond (using the `xtabond` command). The alternative improved system GMM estimator proposed by [6], which is expected to perform better in the presence of highly persistent series, was subsequently applied through the command `xtdpdsys`. When employing these methods we explored alternative specification choices (e.g. variables included and lags used; one or two-step estimation; use of robust errors, etc) and assessed their implications. All analyses were run using the statistical software package Stata 11.

3. RESULTS

It should be noted that the purpose of this paper is solely to discuss the methodology applied in defining a model for pharmaceutical pricing with our particular set of data. Hence, no considerations are made regarding the estimated coefficients and the practical meaning of those findings.

Data structure

Panel is unbalanced in the sense that data is not available for all products at every time period. This is expected since a number of products were launched in the market during the study period and therefore were not available in the early years of the analysis.

Preliminary Analyses

We started by simply regressing the selected variables (Figure 1). Several variables appear to be significant at the selected 80% confidence level and an adjusted R^2 of approximately 0.97 would look promising. However, the estimation suffers from what is called the dynamic panel bias, due to the correlation between the lag value of the dependent variable and the error

term. Although the endogeneity problem might have been negligible in case T was large, with small T it violates a necessary assumption for the OLS estimates consistency. As explained before, these OLS point estimates can however be useful is assessing the consistency of the transformed estimators.

```
. regress price_MD L1.price_MD L2.price_MD L3.price_MD generic original  d_lag2_mol lag_m2_1st_sc lag_m2_1st_sc
>  lag_prod_sc  scC3A scC8A scC9A scC9B scC9C scC9D scC9X n_prod_sc n_mol_sc n_gen_m L1.n_prod_sc L1.n_mol_sc L
> 1.n_gen_m  y1998 y1999 y2000 y2001 y2002 y2003 y2004 y2005 y2006 y2007 y2008 q_prodRmol q_prodRsc q_molRsc  p
> MD_ml_mol pMD_ml_sc pMD_ml_c if prod_level==1, level (80)
      Source |       SS       df       MS              Number of obs =     941
-------------+------------------------------           F( 26,   914) = 1079.70
       Model |  18.3028928    26  .703957417           Prob > F      =  0.0000
    Residual |  .595922225   914  .000651994           R-squared     =  0.9685
-------------+------------------------------           Adj R-squared =  0.9676
       Total |  18.8988151   940  .020105122           Root MSE      =  .02553

------------------------------------------------------------------------------
    price_MD |      Coef.   Std. Err.      t    P>|t|     [80% Conf. Interval]
-------------+----------------------------------------------------------------
    price_MD |
         L1. |    1.005323   .0522156    19.25   0.000     .9383581    1.072289
         L2. |  (omitted)
         L3. |   -.0346358    .052671    -0.66   0.511    -.1021852    .0329137
             |
     generic |    .0014506   .0026406     0.55   0.583    -.0019359    .0048371
    original |     .001579   .0026939     0.59   0.558    -.0018758    .0050339
  d_lag2_mol |    .0002798   .0000973     2.88   0.004     .000155    .0004047
lag_m2_1st~c |    .0002537   .0000981     2.59   0.010     .000128    .0003795
lag_m2_1st~c |  (omitted)
 lag_prod_sc |   -.000704   .0002643    -2.66   0.008    -.001043    -.000365
        scC3A |  (omitted)
        scC8A |  (omitted)
        scC9A |  -.0438103   .0177688    -2.47   0.014    -.0665984   -.0210223
        scC9B |  (omitted)
        scC9C |  (omitted)
        scC9D |  -.1381229   .0400436    -3.45   0.001    -.1894779   -.0867678
        scC9X |  (omitted)
   n_prod_sc |  -.0096844   .0025597    -3.78   0.000    -.0129672   -.0064016
    n_mol_sc |  -.0218347   .0059038    -3.70   0.000    -.0294062   -.0142633
      n_gen_m |  -.0197142   .0044554    -4.42   0.000    -.0254282   -.0140001
              |
   n_prod_sc |
         L1. |    .0094223   .0026076     3.61   0.000     .0060782    .0127665
             |
    n_mol_sc |
         L1. |  (omitted)
             |
      n_gen_m |
         L1. |    .0191201   .0043713     4.37   0.000      .013514    .0247261
             |
       y1998 |  (omitted)
       y1999 |  (omitted)
       y2000 |  (omitted)
       y2001 |    .0261453   .0100499     2.60   0.009     .0132565    .0390342
       y2002 |    .0229188   .0091517     2.50   0.012      .011182    .0346556
       y2003 |    .0193062   .0082938     2.33   0.020     .0086696    .0299428
       y2004 |    .0160441   .0075828     2.12   0.035     .0063194    .0257688
       y2005 |    .0130498   .0070026     1.86   0.063     .0040691    .0220305
       y2006 |    .0093933   .0065482     1.43   0.152     .0009954    .0177912
       y2007 |    .0062345    .006337     0.98   0.325    -.0018925    .0143616
       y2008 |  (omitted)
  q_prodRmol |  -.0024313   .0039681    -0.61   0.540    -.0075203    .0026576
   q_prodRsc |   .0024913   .0140822     0.18   0.860    -.0155688    .0205514
    q_molRsc |   .0001621   .0011725     0.14   0.890    -.0013416    .0016659
  pMD_ml_mol |   .0044237   .0074891     0.59   0.555     -.005181    .0140283
   pMD_ml_sc |  -2.052451   .6212192    -3.30   0.001    -2.849151   -1.25575
   pMD_ml_c  |  (omitted)
       _cons |   .3189816   .0904235     3.53   0.000     .2030154    .4349478
------------------------------------------------------------------------------
```

Figure 1 – Simple regression using OLS

In the next step, by applying the LSDV estimator, we try to deal with the endogeneity problem and expunge the fixed effects from the error term by entering dummies for each product (Figure 2). However, the dynamic panel bias is not completely eliminated with the transformation. In this case, contrarily to the initial OLS regression, the coefficients are negatively correlated originating a downward bias.[8] It can be noted the point estimate for the lag of the dependent variable is lower (0.8358) than the OLS estimate (1.005), and as suggested by Bond [10] these can be useful bounds for a consistency check on the transformed estimators.

```
. xi: regress price_MD L1.price_MD L2.price_MD L3.price_MD generic original  d_lag2_mol lag_m2_1st_sc lag_m2_1s
> t_sc lag_prod_sc  scC3A scC8A scC9A scC9B scC9C scC9D scC9X n_prod_sc n_mol_sc n_gen_m L1.n_prod_sc L1.n_mol_
> sc L1.n_gen_m  y1998 y1999 y2000 y2001 y2002 y2003 y2004 y2005 y2006 y2007 y2008 q_prodRmol q_prodRsc q_molRs
> c  pMD_ml_mol pMD_ml_sc pMD_ml_c i.id if prod_level==1, level (80)
      Source |       SS       df       MS              Number of obs =     941
-------------+------------------------------           F(173,   767) =  234.35
       Model |  18.5479194   173  .107213407           Prob > F      =  0.0000
    Residual |  .350895685   767  .000457491           R-squared     =  0.9814
-------------+------------------------------           Adj R-squared =  0.9772
       Total |  18.8988151   940  .020105122           Root MSE      =  .02139

------------------------------------------------------------------------------
    price_MD |      Coef.   Std. Err.      t    P>|t|     [80% Conf. Interval]
-------------+----------------------------------------------------------------
    price_MD |
         L1. |   .8358259   .0983927     8.49   0.000     .7096219      .96203
         L2. |  (omitted)
         L3. |  -.0317529   .0481356    -0.66   0.510    -.0934943    .0299886
             |
     generic |  -.2225125   .0832653    -2.67   0.008    -.3293132   -.1157118
    original |  -.0550813    .111754    -0.49   0.622    -.1984233    .0882606
  d_lag2_mol |   .0016586   .0016516     1.00   0.316    -.0004599     .003777
lag_m2_1st~c |   .0008136   .0019974     0.41   0.684    -.0017484    .0033756
lag_m2_1st~c |  (omitted)
 lag_prod_sc |  -.0011368   .0060161    -0.19   0.850    -.0088535    .0065798
       scC3A |  (omitted)
       scC8A |  (omitted)
       scC9A |  -.1236349   .0413244    -2.99   0.003     -.17664   -.0706299
       scC9B |  (omitted)
       scC9C |  (omitted)
       scC9D |  -.3813091   .1942969    -1.96   0.050    -.6305252    -.132093
       scC9X |  (omitted)
  n_prod_sc |  -.0119163   .0023839    -5.00   0.000    -.0149741   -.0088585
   n_mol_sc |  -.0259505   .0989271    -0.26   0.793    -.1528399    .100939
    n_gen_m |  -.0195316   .0040045    -4.88   0.000     -.024668   -.0143952
             |
  n_prod_sc |
         L1. |   .0068864   .0050857     1.35   0.176     .0003631    .0134096
             |
   n_mol_sc |
         L1. |  (omitted)
             |
    n_gen_m |
         L1. |  -.0004946   .0133043    -0.04   0.970    -.0175594    .0165702
             |
       y1998 |  (omitted)
       y1999 |  (omitted)
       y2000 |  (omitted)
       y2001 |   .1268818   .1341878     0.95   0.345    -.0452351    .2989986
       y2002 |   .1073474   .1144281     0.94   0.348    -.0394245    .2541193
       y2003 |   .0874962   .0946062     0.92   0.355     -.033851    .2088435
       y2004 |   .0678282   .0748057     0.91   0.365    -.0281217    .1637782
       y2005 |   .0479125    .055059     0.87   0.384    -.0227092    .1185343
       y2006 |   .0259439   .0354071     0.73   0.464    -.0194712     .071359
       y2007 |   .0057577   .0161428     0.36   0.721     -.014948    .0264634
       y2008 |  (omitted)
  q_prodRmol |  -.0736441    .180678    -0.41   0.684    -.3053918    .1581036
   q_prodRsc |   .4880795   .3960904     1.23   0.218    -.0199684    .9961274
   q_molRsc  |  -.0029383   .0139484    -0.21   0.833    -.0208292    .0149527
  pMD_ml_mol |   .1943597   .0714716     2.72   0.007     .1026861    .2860332
   pMD_ml_sc |  -1.369643   .5848027    -2.34   0.019    -2.119743   -.6195418
    pMD_ml_c |  (omitted)
      _Iid_13 |  (omitted)
```

Figure 2 – Regression using the LSDV estimator

First Difference GMM (Arellano and Bond)

We then applied the first difference transform run through the `xtabond` command, which is expected to perform better than both OLS and LSDV with less bias and smaller variances:

```
. xtabond price_MD generic original  d_lag2_mol lag_m2_1st_sc lag_prod_sc  scC3A scC8A scC9A scC9B scC9C scC9D
> scC9X n_prod_sc n_mol_sc n_gen_m L1.n_prod_sc L1.n_mol_sc L1.n_gen_m  y1998 y1999 y2000 y2001 y2002 y2003 y20
> 04 y2005 y2006 y2007 y2008 if prod_level==1, noconstant lags(3) maxldep(3) pre(q_prodRmol q_prodRsc q_molRsc
>  pMD_ml_mol pMD_ml_sc pMD_ml_c) level(80) artests(2)
```

This estimation however may not be robust to heteroskedasticity or serial correlation in the errors and a two-step estimation implying a reweighing based on the second moments should be undertaken if sample is not expected to be homoskedastic (Figure 3).

```
. xtabond price_MD generic original  d_lag2_mol lag_m2_1st_sc lag_prod_sc  scC3A scC8A scC9A scC9B scC9C scC9D
> scC9X n_prod_sc n_mol_sc n_gen_m L1.n_prod_sc L1.n_mol_sc L1.n_gen_m  y1998 y1999 y2000 y2001 y2002 y2003 y20
> 04 y2005 y2006 y2007 y2008 if prod_level==1, noconstant lags(3) maxldep(3) twostep pre(q_prodRmol q_prodRsc q
> _molRsc  pMD_ml_mol pMD_ml_sc pMD_ml_c) level(80) artests(2)

Arellano-Bond dynamic panel-data estimation  Number of obs         =       782
Group variable: id                           Number of groups      =       145
Time variable: year
                                             Obs per group:    min =         1
                                                               avg =  5.393103
                                                               max =         7

Number of instruments =    186               Wald chi2(11)         =  10232.97
                                             Prob > chi2           =    0.0000
Two-step results
```

price_MD	Coef.	Std. Err.	z	P>|z|	[80% Conf.	Interval]
price_MD						
L1.	(omitted)					
L3.	-.0150194	.0010636	-14.12	0.000	-.0163825	-.0136563
q_prodRmol	.0674473	.048516	1.39	0.164	.0052715	.1296231
q_prodRsc	.5888794	.0486637	12.10	0.000	.5265143	.6512445
q_molRsc	-.0014601	.0013771	-1.06	0.289	-.003225	.0003047
pMD_ml_mol	.1990753	.0189812	10.49	0.000	.1747499	.2234007
pMD_ml_sc	-1.292163	.0979626	-13.19	0.000	-1.417707	-1.166619
generic	(omitted)					
original	(omitted)					
d_lag2_mol	-.000029	.0248857	-0.00	0.999	-.0319214	.0318634
lag_m2_1st~c	-.0007328	.0005274	-1.39	0.165	-.0014087	-.0000568
lag_prod_sc	(omitted)					
scC9A	(omitted)					
scC9B	(omitted)					
scC9C	(omitted)					
n_prod_sc	-.0146765	.0007136	-20.57	0.000	-.015591	-.013762
n_mol_sc	(omitted)					
n_gen_m	-.0228006	.0014422	-15.81	0.000	-.0246488	-.0209523
n_prod_sc						
L1.	(omitted)					
n_gen_m						
L1.	(omitted)					
y2002	.0004749	.2986289	0.00	0.999	-.3822334	.3831833
y2003	.0009343	.5972578	0.00	0.999	-.7644824	.766351
y2004	.0014926	.8958868	0.00	0.999	-1.146632	1.149618
y2005	.0018013	1.194516	0.00	0.999	-1.529032	1.532635
y2006	.0003334	1.493145	0.00	1.000	-1.913208	1.913875
y2007	.0004819	1.791774	0.00	1.000	-2.295768	2.296732
y2008	.01358	2.015746	0.01	0.995	-2.569702	2.596862

```
Warning: gmm two-step standard errors are biased; robust standard
         errors are recommended.
Instruments for differenced equation
        GMM-type: L(2/4).price_MD L(1/.).q_prodRmol L(1/.).q_prodRsc L(1/.).q_molRsc L(1/.).pMD_ml_mol
                  L(1/.).pMD_ml_sc L(1/.).pMD_ml_c
        Standard: D.d_lag2_mol D.lag_m2_1st_sc D.n_prod_sc D.n_gen_m D.y2001 D.y2002 D.y2003 D.y2004
                  D.y2005 D.y2006 D.y2007
```

Figure 3 – Estimation using two-step first-difference GMM

However, if the number of instruments is large and the sample is small, standard errors can be downward biased and suggest spurious precision. As Arrelano and Bond discuss [4], in small samples reweighing using the moments' own estimated variances and co-variances can result in overvaluing observations that fit the model. Although coefficients are still expected to consistent, a correction is suggested to produce standard errors robust to heteroskedasticity.[11] We apply this correction to the two-step by using the option `vce(robust)` (Figure 4). We can observe that after the standard error correction is applied only 3 theoretically meaningful variables fall within the 20% significance level.

```
. xtabond price_MD generic original  d_lag2_mol lag_m2_1st_sc lag_prod_sc  scC3A scC8A scC9A scC9B scC9C scC9D
> scC9X n_prod_sc n_mol_sc n_gen_m L1.n_prod_sc L1.n_mol_sc L1.n_gen_m  y1998 y1999 y2000 y2001 y2002 y2003 y20
> 04 y2005 y2006 y2007 y2008 if prod_level==1, noconstant lags(3) maxldep(3) pre(q_prodRmol q_prodRsc q_molRsc
>  pMD_ml_mol pMD_ml_sc pMD_ml_c) vce(robust) level(80) artests(2)
Arellano-Bond dynamic panel-data estimation  Number of obs         =       782
Group variable: id                           Number of groups      =       145
Time variable: year
                                             Obs per group:    min =         1
                                                               avg =  5.393103
                                                               max =         7

Number of instruments =    186               Wald chi2(16)         =    186.37
                                             Prob > chi2           =    0.0000
One-step results
                                        (Std. Err. adjusted for clustering on id)
------------------------------------------------------------------------------
             |               Robust
    price_MD |      Coef.   Std. Err.      z    P>|z|     [80% Conf. Interval]
-------------+----------------------------------------------------------------
    price_MD |
         L1. |  (omitted)
         L3. |  -.0161001   .0059176    -2.72   0.007    -.0236838   -.0085163
             |
  q_prodRmol |   .0329358   .1643729     0.20   0.841    -.1777166    .2435881
   q_prodRsc |   .5867788   .5212634     1.13   0.260    -.0812471    1.254805
   q_molRsc  |  -.0003426   .0190344    -0.02   0.986    -.0247361    .0240509
  pMD_ml_mol |   .1690325   .1301224     1.30   0.194     .0022739    .3357911
   pMD_ml_sc |  -1.426177   1.510479    -0.94   0.345    -3.361934    .5095804
     generic |  (omitted)
    original |  (omitted)
  d_lag2_mol |   .0001679   .0001435     1.17   0.242     -.000016    .0003518
lag_m2_1st~c |  -.0003458    .001789    -0.19   0.847    -.0026385     .001947
 lag_prod_sc |  (omitted)
       scC9A |  (omitted)
       scC9B |  (omitted)
       scC9C |  (omitted)
   n_prod_sc |  -.0160238   .0092511    -1.73   0.083    -.0278796    -.004168
    n_mol_sc |  (omitted)
     n_gen_m |  -.0255328   .0185518    -1.38   0.169    -.0493078   -.0017577
             |
   n_prod_sc |
         L1. |  (omitted)
             |
     n_gen_m |
         L1. |  (omitted)
             |
       y2002 |  -.0018763   .0017193    -1.09   0.275    -.0040796    .0003271
       y2003 |  -.0037692   .0034388    -1.10   0.273    -.0081762    .0006378
       y2004 |   -.005554   .0050791    -1.09   0.274    -.0120632    .0009551
       y2005 |  -.0076126   .0068644    -1.11   0.267    -.0164098    .0011845
       y2006 |  -.0116105   .0097451    -1.19   0.233    -.0240994    .0008784
       y2007 |  -.0138411   .0113368    -1.22   0.222    -.0283699    .0006876
       y2008 |  (omitted)
------------------------------------------------------------------------------
Instruments for differenced equation
        GMM-type: L(2/4).price_MD L(1/.).q_prodRmol L(1/.).q_prodRsc L(1/.).q_molRsc L(1/.).pMD_ml_mol
                  L(1/.).pMD_ml_sc L(1/.).pMD_ml_c
        Standard: D.d_lag2_mol D.lag_m2_1st_sc D.n_prod_sc D.n_gen_m D.y2001 D.y2002 D.y2003 D.y2004
                  D.y2005 D.y2006 D.y2007
```

Figure 4 – Estimation using two-step first difference GMM with correction for robust standard errors

System GMM (Blundell and Bond)

A major concern with the first-difference GMM estimator explored above is that in the case of unbalanced panels, any gaps in the data will be amplified. To overcome this, Blundell and Bond [6] propose transforming the instruments (instead of the regressors) and develop a system GMM estimator. Noting that non-deflated pharmaceutical prices are expected to be non-stationary, this estimator appears to accommodate the concerns that could spring from the use of our dataset (Figure 5).

```
. xtdpdsys price_MD generic  lag2_mol  scC3A scC8A scC9A scC9B scC9C scC9D scC9X n_prod_sc n_mol_sc n_gen_m L1.
> n_prod_sc L1.n_mol_sc L1.n_gen_m  y1998 y1999 y2000 y2001 y2002 y2003 y2004 y2005 y2006 y2007 y2008 if prod_l
> evel==1, noconstant lags(3) maxldep(3) twostep pre(pMD_ml_mol pMD_ml_sc pMD_ml_c, lagstruct(0,.)) vce(robust)
>  level(80) artests(2)

System dynamic panel-data estimation         Number of obs         =       941
Group variable: id                           Number of groups      =       158
Time variable: year
                                             Obs per group:    min =         1
                                                               avg =  5.955696
                                                               max =         8

Number of instruments =     99               Wald chi2(15)         =     87.21
                                             Prob > chi2           =    0.0000
Two-step results
```

price_MD	Coef.	WC-Robust Std. Err.	z	P>|z|	[80% Conf.	Interval]
price_MD						
L1.	(omitted)					
L3.	-.0086646	.0050567	-1.71	0.087	-.0151449	-.0021842
pMD_ml_mol	.2669394	.1002868	2.66	0.008	.1384167	.3954622
pMD_ml_sc	-1.336358	1.300945	-1.03	0.304	-3.003586	.3308705
generic	(omitted)					
lag2_mol	.0045493	.0165677	0.27	0.784	-.0166831	.0257817
scC9A	(omitted)					
scC9B	(omitted)					
scC9C	(omitted)					
scC9D	(omitted)					
n_prod_sc	-.0153773	.0062644	-2.45	0.014	-.0234054	-.0073492
n_mol_sc	.0078373	.0936218	0.08	0.933	-.1121439	.1278186
n_gen_m	-.0141476	.0064608	-2.19	0.029	-.0224275	-.0058677
n_prod_sc						
L1.	.0043047	.0640614	0.07	0.946	-.0777933	.0864028
n_gen_m						
L1.	(omitted)					
y2001	-.0140877	.0066132	-2.13	0.033	-.0225628	-.0056125
y2002	-.0140153	.0066339	-2.11	0.035	-.022517	-.0055135
y2003	-.0139477	.0066417	-2.10	0.036	-.0224594	-.005436
y2004	-.0138714	.0066231	-2.09	0.036	-.0223592	-.0053835
y2005	-.013925	.0066448	-2.10	0.036	-.0224407	-.0054094
y2006	-.0149432	.0070767	-2.11	0.035	-.0240123	-.0058741
y2007	-.0151517	.0069707	-2.17	0.030	-.024085	-.0062184

```
Instruments for differenced equation
        GMM-type: L(2/4).price_MD L(1/.).pMD_ml_mol L(1/.).pMD_ml_sc L(1/.).pMD_ml_c
        Standard: D.n_prod_sc D.n_gen_m D.y2000 D.y2001 D.y2002 D.y2003 D.y2004 D.y2005 D.y2006 D.y2007
Instruments for level equation
        GMM-type: LD.price_MD D.pMD_ml_mol D.pMD_ml_sc D.pMD_ml_c
```

Figure 5 – Estimation using system GMM with correction for robust standard errors

If we decided to accept this specification, we would need to run the Arellano-Bond test for auto-correlation. This was not performed due to a software fault. If serial correlation of order one was found we would need to restrict the use of lags to the second and subsequent lags.

4. DISCUSSION

In our attempt to develop an econometric model accounting for the dynamics of pharmaceutical pricing we were confronted with the specificities of the data available. The very nature of the market we wanted to study with new entrants arriving and other products leaving during the study period caused the panel to be unbalanced and requested the use of more elaborated estimators.

Initial conditions are usually fundamental in model outcomes as the impact on subsequent observations cannot be ignored. In the case of our study this may be of particular importance since the study timeline captures different

products from various therapeutic subclasses at different times of their product cycle. In the absence of data on launch prices, this cannot be safely controlled for despite our attempt to do so by including lifecycle-related variables as the time elapsed since molecule launch and molecules' launch lag to first drug in respective subclass

As stated before, the validity of the additional instruments in the system GMM estimator depends on the assumption that changes in variables are uncorrelated with the fixed effects. To account for these, we included time dummies as these tend to minimise any existent correlation across individuals in the disturbances, which is essential to achieve robust estimates of the coefficient standard errors. It should therefore be noted such dummy variables only have a structural role and their coefficients are not empirically meaningful.

In the same way, the value of the coefficients obtained for the lags of the dependent variables are not of much interest in themselves. However, they provide more information and can still be of use in achieving consistent estimates for the other equation parameters. A decision on the lag structure is therefore important since, due to the nature of the dataset, an increase or decrease in the number of lags of the dependent variable will necessarily affect sample size. In particular, any marked increase in the number of lags of the dependent variable in the model results in excluding more recent products for which not many lagged observations are available.

In system GMM, contrarily to first-difference GMM, time-invariant variables e.g. `original`, `subclass` or `lag_m2_1st_sc`) can be included as regressors. These are not expected to affect the estimation of other coefficients, but with small T the use of a Within Groups-like variable (i.e. that can be 0 or 1 for almost all products) may result in a downward bias as discussed for LSDV. In our dataset there are subclasses with small representation and technically the respective subclass dummies should be dropped from the regression equation. In testing for that effect in our system GMM analysis, no changes were noted in the precision of the variables of interest if such subclass dummies were dropped.

We were not able to run the standard Arrelano-Bond autocorrelation and Sargan/Hansen overfitting tests which are essential is defining model validity. If regarding the first the concern is not high since software tends to resolve the problem automatically by dropping the autocorrelated lags, the absence of an overidentification test may be worrisome. The problem of over fitting is related to the fact that too many instruments can result in including endogenous variables which will then undermine consistency of the estimates. Since these tests are prone to be weak anyway, the number of instruments used is reported. In case the count of instruments was too high, we would test for robustness to reducing it by limiting the lags used and/or collapsing instruments.

.As reported in the methods section, the present study was conducted at a product level under the assumption that pricing strategies are usually designed at product level. A presentation or molecule-level analysis was in the process of being conducted but could not be included in this paper due to lack of time. Further analyses may include alternative model specifications such as using only products for which the launch price was available, i.e. products that were launched in the UK at some point between June 1998 and June 2008.

Points for discussion / further exploration

In the absence of clinical and safety data at the molecule level we failed to include the quality dimension in the analysis. We were also unable to control for systemic changes (e.g. change in target sBP in guidelines across time or the implementation of pricing schemes). How can we overcome this in order to make the model more explanatory?

Would a straightforward fixed effects autoregressive model using quarterly data suffice, taking into consideration that if T is large dynamic panel bias becomes insignificant?

There appears to be a marked difference on the price of products from different subclasses. Do the dummies in place adequately control for this and would a log-log model, in which coefficients could be read as elasticities, be more adequate?

Considering we are trying to capture the reactive nature of pharmaceutical pricing, is the use of autoregressive distributed lag models, i.e. models where the past values of variables also play a role as regressors, a necessary approach?

Noting that the panel is unbalanced because of the nature of the pharmaceutical market which implies new products coming in and out during the study, how can we include more information and use deeper lags without reducing the sample and eventually restricting analyses to older products?

REFERENCES

1. Miraldo, M., *Reference pricing and firms' pricing strategies.* Journal of Health Economics, 2009. **28**: p. 176-197.
2. Hoyle, M., *Future drug prices and cost-effectiveness analyses.* Pharmacoeconomics, 2008. **26**(7): p. 589-602.
3. Nickell, S., *Biases in dynamic models with fixed effects.* Econometrica, 1981. **49**(6): p. 1417-1426.
4. Arellano, M. and S. Bond, *Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations.* Review of Economic Studies 1991. **58**: p. 277-297.
5. Arellano, M. and O. Bover, *Another look at the instrumental variables estimation of error-component models.* Journal of Econometrics, 1995. **68**: p. 29-51.
6. Blundell, R. and S. Bond, *Initial conditions and moment restrictions in dynamic Panel Data models.* Journal of Econometrics, 1998. **87**: p. 115-143.
7. Bruno, G., *Approximating the bias of the LSDV estimator for dynamic unbalanced panel data models.* Economics Letters, 2005. **87**(3): p. 361-366.
8. Roodman, D., *How to do xtabond2: an introduction to "difference" and "system" GMM in Stata*, in *Center for Global Development*. 2006.
9. NCCCC, *Hypertension: management of hypertension in adults in primary care: partial update*. 2006, National Collaborating Centre for Chronic Conditions, Royal College of Physicians: London.
10. Bond, S., *Dynamic Panel Data models: a guide to micro data methods and practice*, in *Cemmap working paper* 2002.
11. Windmeijer, F., *A finite sample correction for the variance of linear efficient two-step GMM estimators.* Journal of Econometrics, 2005. **126**: p. 25-51.

APPENDIX 1

Table I – Set of variables used and respective labels

Variable	Label
price_MD	Product's weighted price per daily maintenance dose
L1.price_MD	1st lag of price per daily maintenance dose
L2.price_MD	2nd lag of price per daily maintenance dose
L3.price_MD	3rd lag of price per daily maintenance dose
L4.price_MD	4th lag of price per daily maintenance dose
L5.price_MD	5th lag of price per daily maintenance dose
generic	Innovative status dummy (==1 if product is generic)
original	Innovative status dummy (==1 if product is originator)
d_lag2_mol	Time elapsed since molecule launch
lag_m2_1st_sc	Entry lag of molecule relative to 1st molecule in subclass
lag_prod_sc	Entry lag of 2nd product in respective subclass
n_prod_sc	Number of products in subclass
n_mol_sc	Number of molecules in subclass
n_gen_m	Number of generics in molecule
L1.n_prod_sc	Number of products in subclass in previous period
L1.n_mol_sc	Number of molecules in subclass in previous period
L1.n_gen_m	Number of generics in molecule in previous period
q_prodRmol	Quantity of product sold relative to molecule
q_prodRsc	Quantity of product sold relative to subclass
q_molRsc	Quantity of molecule sold relative to subclass
pMD_ml_mol	Price of molecule market leader
pMD_ml_sc	Price of subclass market leader
pMD_ml_c	Price per daily maintenance dose of class market leader
scC3A	Subclass dummy (==1 if diuretics)
scC8A	Subclass dummy (==1 if calcium antagonists plain)
scC9A	Subclass dummy (==1 if ACE inhibitors plain)
scC9B	Subclass dummy (==1 if ACE inhibitors comb)
scC9C	Subclass dummy (==1 if angiotensin II antagonist plain)
scC9D	Subclass dummy (==1 if angiotensin II antagonist comb)
scC9X	Subclass dummy (==1 if other rennin-angiotensin agents)
y1998	Time dummy (==1 if t==1998)
y1999	Time dummy (==1 if t==1999)
y2000	Time dummy (==1 if t==2000)
y2001	Time dummy (==1 if t==2001)
y2002	Time dummy (==1 if t==2002)
y2003	Time dummy (==1 if t==2003)
y2004	Time dummy (==1 if t==2004)
y2005	Time dummy (==1 if t==2005)
y2006	Time dummy (==1 if t==2006)
y2007	Time dummy (==1 if t==2007)
y2008	Time dummy (==1 if t==2008)