International Journal of Pharmaceutics, 51 (1989) 115–127
Elsevier

IJP 01724

Influence of crystallization conditions on the physical properties of acetaminophen crystals: evaluation by multiple linear regression

Albert H.-L. Chow * and David J.W. Grant

Faculty of Pharmacy, University of Toronto, Toronto, Ont. (Canada)

(Received 7 June 1988)
(Modified version received 20 September 1988)
(Accepted 22 September 1988)

Key words: *p*-Acetoxyacetanilide; Water content; Length-to-width ratio; Entropy of fusion; Multiple linear regression

Summary

The comparative influence of the following 3 aqueous crystallization variables on various physical properties of acetaminophen (paracetamol, **P**) has been appraised statistically by multiple linear regression analysis: the concentration of the additive, *p*-acetoxyacetanilide, **A**, in the solution (PAA), the stirring speed (SPEED) and the initial supersaturation of **P** (SIGMA) in the solution. The following properties of the crystals have been studied: the uptake of **A** and water during crystallization, the length-to-width ratio and the entropy of fusion (ΔS^f). The uptake of **A** depends almost exclusively on PAA, and much less so, on SPEED and/or on SIGMA. The length-to-width ratio also depends chiefly on PAA and, to some extent, on SIGMA but only in the presence of **A**, i.e. SIGMA only plays a mediating role. The uptake of water, however, is determined, in descending order of significance, by SIGMA, SPEED, and PAA. Comparison of the multiple r^2 and residual mean-square values suggests that both the additive uptake and the length-to-width ratio are governed mainly by the crystallization conditions while the water content is less so. On the other hand, the entropy of fusion is partially linked to the crystallization conditions possibly via the density of growth defects.

Introduction

Our previous studies have established that the uptake of *p*-acetoxyacetanilide (**A**), water content, length-to-width ratio and the entropy of fusion of acetaminophen (**P**) crystals depend to various extents on the concentration of **A**, the stirring speed and the initial supersaturation during solution-phase crystallization (Chow et al., 1985; Chow and Grant, 1988a and b). To describe the various relationships quantitatively, the data have been treated statistically by multiple linear regression analysis. The objectives of the present analysis are two-fold: (1) to develop empirical and statistically adequate equations to express the dependence of the various physical properties on the conditions of crystallization; and
(2) to quantify and to compare statistically the influence of the various crystallization variables on each of the measured physical properties.

* *Present address:* Faculty of Pharmaceutical Sciences, University of British Columbia, 2146 East Mall, Vancouver, B.C. V6T 1W5, Canada.
Correspondence: D.J.W. Grant, College of Pharmacy, University of Minnesota, Health Sciences Center Unit F, 308 Harvard Street S.E., Minneapolis, MN 55455, USA.

It is hoped that this general approach will find further applications in pharmaceutics.

Experimental details and data collection

The conditions of crystallization, methods of analysis and experimental details have been reported by Chow et al. (1985) and by Chow and Grant (1988a and b).

Depending on the grade or purity, all chemicals and reagents were used either as received or after purification by crystallization twice from the previously stated organic solvents (either glass-distilled or HPLC-grade). Acetaminophen (**P**) was crystallized from distilled water in a 3-necked round-bottom flask (500 ml) in the presence and absence of the stated concentration of *p*-acetoxyacetanilide (**A**) using one of the following 3 procedures:

(1) **P** (9 g) was dissolved in water (390 ml) containing a defined concentration of **A** at 55 °C. The supersaturation was created by cooling the solution to 30 °C. While cooling and being stirred at 240 rpm, the solution was seeded at 42 °C with 1 mg of crystals (400–200 mesh; 30–75 μm) to initiate crystallization, which was allowed to continue for 2 h. The crystals were then harvested, air-dried overnight, further dried under vacuum for 2 days and sieved into various size fractions before use. This procedure corresponds to crystallization in the presence of various concentrations of **A**, while the stirring speed and the initial supersaturation were held constant.

(2) Essentially the same procedure as (1) was followed with the following modifications. **P** (9 g) was crystallized from water (390 ml) containing zero or 500 mg $\cdot$ dm^{-3} **A** at various stirring speeds (200–400 rpm). This procedure corresponds to crystallization at various stirring speeds, while the concentration of **A** is either zero or 500 mg $\cdot$ dm^{-3} and the initial supersaturation was maintained constant as stated above.

(3) The crystals were prepared using the procedure detailed in (1) with the following modifications. **P** in various amounts (8.5–13 g) was crystallized from water (390 ml) containing zero or 500 mg $\cdot$ dm^{-3} **A** at a stirring speed of 240 rpm. The initial supersaturation is defined here as the initial concentration of **P** minus the solubility of **P** at 30 °C (16.45 g $\cdot$ dm^{-3}). This procedure corresponds to crystallization under various initial supersaturations, while the concentration of **A** was either zero or 500 mg $\cdot$ dm^{-3} and the stirring speed was held constant at 240 rpm.

After harvesting and drying the **P** crystals, a defined sieve fraction (250–355 μm, 60–45 mesh; or 355–500 μm, 45–35 mesh) was used in subsequent characterization. To obtain sufficient crystals for all the measurements, crystals of the same sieve size from 6–8 separate batches were pooled together and mixed to form a larger batch. Such experimental constraint precludes the assessment of inter-batch variation. To minimize batch differences due to such minor variables as day of the week, time of the day and order of crystallization, the various batches of "doped" and "undoped" crystals were prepared in a fully randomized fashion.

The amount of **A** incorporated into the crystals was measured (Chow et al., 1985; Chow and Grant, 1988a) in triplicate by high-performance liquid chromatography (HPLC). The water content of the crystals was determined (Chow et al., 1985) in duplicate by Karl-Fischer titration. Quadruplicate determination of the enthalpy of fusion, melting point and entropy of fusion was performed (Chow et al., 1985; Chow and Grant, 1988a) using a differential scanning calorimeter (Perkin-Elmer 2C). Size measurement of 20 randomly sampled crystals from each pooled batch employed an optical microscope (Karl Zeiss with Hughes-Owens objectives) equipped with a calibrated eyepiece. The ideal molar entropy of mixing of the various components of the crystals was calculated from analytical data using the classical equation employed previously (Chow et al., 1985; Chow and Grant, 1988a).

The mean values of replicate measurements rather than of individual determinations were used in the present statistical analysis primarily for the following reasons.

(a) The replicate measurements here actually are "pseudo-replicates"; such repeated measurements afford valuable information on the variation within sample (e.g. instrumental errors),

but not on the variation between samples, and thus are incapable of improving the estimation of the population statistics.

(b) The observed variation is small relative to the variation across the independent variables (< 5%). The exception is the length-to-width ratio, the variability of which depends on crystal shape (~ 10% for prismatic habit and ~ 40% for acicular crystals).

Underlying approach

In regression analysis, selection of appropriate variables for describing a given set of data depends on the purpose for which the equation is constructed (Chatterjie and Price, 1977). If the prime objective is description, one chooses the minimum number of independent variables which account for the most substantial part of the variation in the dependent variable. This choice is a compromise of the following two conflicting criteria:

(a) to explain as much of the variation as possible, which suggests the inclusion of a large number of variables; and

(b) to adhere to the principle of parsimony, which favours the use of as few variables as possible.

If the aim is to predict a future observation or to estimate the mean response corresponding to a given observation, the variables are selected with a view to minimizing the variance of prediction, or more directly, the residual mean square (RMS) value. If the purpose is to determine the magnitude by which the value of an independent variable must be altered to obtain a specified value of the dependent variable, the coefficients of the variables in the equation should be measured precisely (i.e. the standard errors of the regression coefficients should be small).

In practice, it is extremely difficult, if not impossible, to find a set of variables that would satisfy all of the above criteria. The purpose for which the regression equation is developed determines the criteria that are to be optimized in its formulation. The statistical arguments presented in this report are based upon the objectives and criteria defined above.

Data analysis by multiple linear regression

The abbreviations of the variables used in computation are listed below:

PAA, $mg \cdot dm^{-3}$ or $g \cdot dm^{-3} \times 10^{-3}$	concentration of **A** in solution
SIGMA, $g \cdot dm^{-3}$	initial supersaturation of **P** in solution, i.e. initial concentration of **P** minus solubility of **P**
SPEED	stirring speed of the solution
PAASIG, $g^2 \cdot dm^{-6} \times 10^{-3}$	interaction term for PAA and SIGMA, i.e. PAA × SIGMA
PAASPD, $g \cdot dm^{-3} \times 10^{-3} \times rpm$	interaction term for PAA and SPEED, i.e. PAA × SPEED
SIGSPD, $g \cdot dm^{-3} \times rpm$	interaction term for SIGMA and SPEED, i.e. SIGMA × SPEED
PASISP, $g^2 \cdot dm^{-6} \times 10^{-3} \times rpm$	interaction term for PAA, SIGMA and SPEED, i.e. PAA × SIGMA × SPEED
PAAUPT, mole fraction	uptake of A by the crystals
WATER, mole fraction	water content of the crystals
LWR	length-to-width ratio of the crystals
ENTFUS, $J \cdot K^{-1} \cdot mol^{-1}$	entropy of fusion of the crystals
ENTMIX, $J \cdot K^{-1} \cdot mol^{-1}$	ideal molar entropy of mixing for the various components in the crystals

PAA, SIGMA and SPEED are the independent variables, i.e. the predictor variables, which correspond to the 3 crystallization conditions in the solution, their composite effects being represented by the interaction terms, PAASIG, PAASPD, SIGSPD and PASISP. The 4 measured physical properties of the crystals are represented by the 4 dependent variables, PAAUPT, WATER, LWR and ENTFUS. The variable, ENTMIX, is only employed for the calculation of the disruption index (York and Grant, 1985; Grant and York, 1986).

The data were analyzed by multiple linear regression using the BMDP2R (forward selection,

FS, and backward elimination, BE, procedure; minimum acceptable F to enter = 0.1 and maximum acceptable F to remove = 4) and the BMDP9R (all possible subsets regression; selection criteria based on Mallows' C_p and r^2) statistical computer programmes (BMDP manual, 1983).

Employing the FS-BE procedure, preliminary regression analysis was performed to determine the relative significance of the predictor variables, namely, PAA, SIGMA and SPEED, in explaining the variation in each of the response variables, namely, PAAUPT, WATER, LWR and ENTFUS. For reasons stated in the Experimental Details and Data Collection section, mean values of replicate measurements of each of the response variables were used in the analysis. Depending on the outcome of the preliminary analysis and on the circumstance, the regression analysis was repeated with the inclusion of the interaction variables. Normally, single-component predictor variables are preferable to mixed or interaction variables primarily because of the ease of interpretation. The interaction variables do, however, assist in "smoothing out" any existing non-linearity in the relationship.

Uptake of additive

For PAAUPT, only PAA, which has the strongest correlation with PAAUPT (i.e. highest F value) is entered into the equation, the remaining variables, SPEED and SIGMA, being insignificant in the presence of PAA. This correlation arises because the uptake of A can only occur in the presence of A. Since the influences of stirring rate and initial supersaturation on the uptake of A have been clearly demonstrated, the interaction terms, particularly those containing PAA are expected to account for part of the variation in PAAUPT. Thus, the regression analysis was repeated using PAA, PAASIG, SIGSPD, PAASPD and PASISP as regressors in the selection procedures. Only PAA and PASISP were retained in the equation after the FS-BE procedure. The selection criterion based on C_p statistic also indicates that these two variables are the best. As mentioned before, the variables should be selected with a view to minimizing the RMS, which provides a measure of the adequacy of the model. As shown in Table 1, the equation containing the PASISP term has a considerably lower RMS than that without, strongly suggesting that the PASISP term is required for describing the data. Another criterion that is used here to evaluate the sufficiency of the model equation is the Mallows' C_p statistic (see Appendix for definition). The desirable subsets of variables produce values of C_p that are close to p (i.e. the number of parameters including the intercept). However, this outcome depends very much on the estimate of the variance of the residuals, σ^2 (or RMS). Usually, the estimate of σ^2 is obtained from the residual sum of squares (RSS) for the full model. If the full model has a large number of variables with no explanatory power (i.e. population regression coefficients are zero or almost zero), the estimate of σ^2 from the full model would be large. The loss of degrees of freedom for the devisor would not be compensated by a reduction in the RSS. If σ^2 is large, then the value of C_p will be small. For C_p to provide useful information, σ^2 must be estimated reliably. As can be seen in Table 1, an increase in p decreases the RMS at first but later increases it. Furthermore, the RMS for the full model with 5

TABLE 1

Variable selection by the forward-selection procedure

Variable	p	C_p	RMS	r	r^2	adjusted r^2
PAA	2	37.4	2.766×10^{-8}	0.98725	0.974670	0.973732
+ PASISP	3	0.15	1.066×10^{-8}	0.99529	0.974671	0.989875
+ SIGSPD	4	2.04	1.103×10^{-8}	0.99531	0.990643	0.989521
+ PAASPD	5	4.02	1.149×10^{-8}	0.99531	0.990648	0.989090
+ SPEED	6	6.00	1.198×10^{-8}	0.99532	0.990658	0.988627

TABLE 2

Parameters and statistics for the linear regression of PAAUPT on PAA and PASISP

Variable	Regression coefficient	Standard error	Standardized coefficient	*t*-Statistic	Tolerance
Intercept	-1.31×10^{-5}	2.76×10^{-5}		−0.48	
PAA	4.02×10^{-6}	1.35×10^{-7}	1.230	29.9 *	0.2133
PASISP	-4.24×10^{-10}	6.39×10^{-11}	−0.273	−6.64 *	0.2132

RMS = 1.07×10^{-8}; $r^2 = 0.991$; d.f. = $n - 3$, where $n = 29$.

* Significant at the 5% level.

$$\text{PAAUPT} = 4.02 \times 10^{-6}(\text{PAA}) - 4.24 \times 10^{-10}(\text{PASISP}) - 1.31 \times 10^{-5} \quad (1)$$

Outliers: PAAUPT [a]	SPEED [b]	SIGMA [b]	PAA [c]	PASISP [c]	ϵ [d]
0.00199	240	5.34	500	640,800	2.73
0.00130	240	9.19	500	1,102,799	−2.27

[a] Dependent variable.

[b] Conditions not explicitly in Eqn. 1, although implicit in the interaction term, PASISP.

[c] Independent variable.

[d] Standardized residual defined in the Appendix.

variables is larger than that for the model with 2 variables (PAA and PASISP). Consequently, the C_p is distorted and is not useful here in variable selection. The plot of C_p against p (Fig. 1), however, shows that the sets of variables corresponding to points close to the line, $C_p = p$, invariably contain both terms, PAA and PASISP.

A summary of the regression statistics for the equation with PAA and PASISP as the predictor variables is given in Table 2. According to Eqn. 1 in Table 2, the uptake of A increases with increasing concentration of A in solution, but decreases with increasing stirring speed and/or initial supersaturation. These predictions agree with the experimental observations (Chow et al., 1985; Chow and Grant, 1988a and b). In Table 2, all the regression coefficients except the intercept are

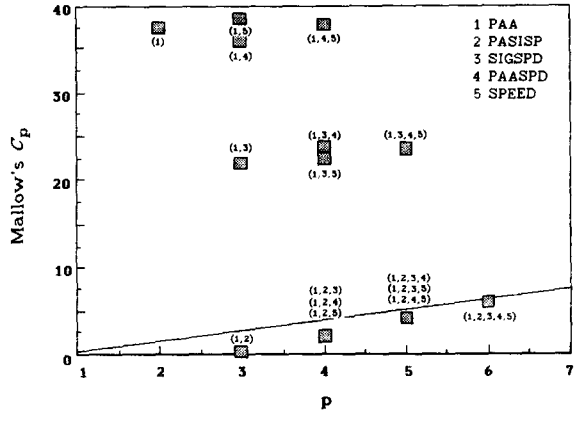

Fig. 1. Mallows' C_p plot against p, the number of parameters, including the intercept. This plot illustrates the best subsets of variables among PAA, PASISP, SIGSPD, PAASPD and SPEED for defining PAAUPT.

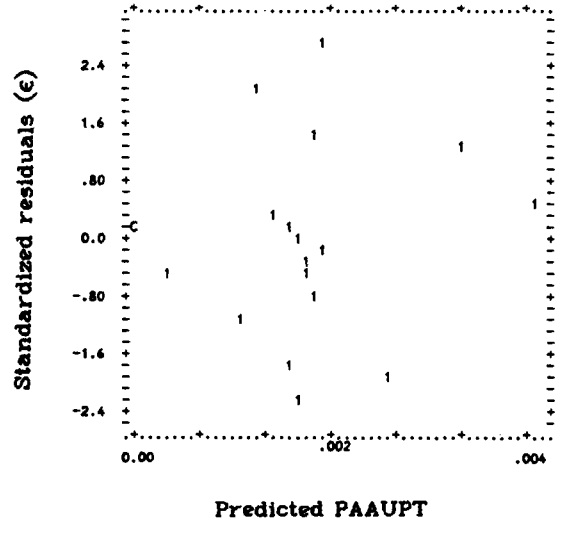

Fig. 2. Standardized residuals plotted against the predicted PAAUPT for the linear regression of PAAUPT on PAA and PASISP (Eqn. 1, Table 2). The symbol C represents 12 virtually coincident points.

TABLE 3

Parameters and statistics for the linear regression of WATER on PAA, SIGMA, SPEED and PASISP

Variable	Regression coefficient	Standard error	Standardized coefficient	*t*-Statistic	Tolerance
Intercept	-5.47×10^{-2}	1.42×10^{-2}		-3.86 *	
PAA	5.48×10^{-5}	1.69×10^{-5}	0.980	3.24 *	0.1083
SIGMA	4.74×10^{-3}	7.52×10^{-4}	0.907	6.30 *	0.4773
SPEED	1.87×10^{-4}	3.71×10^{-5}	0.575	5.04 *	0.7616
PASISP	-4.14×10^{-8}	8.07×10^{-8}	-1.560	-5.13 *	0.1070

RMS = 8.50×10^{-5}; $r^2 = 0.763$; d.f. = $n - 5$, where $n = 29$.

* significant at the 5% level

$$\text{WATER} = 5.48 \times 10^{-5}(\text{PAA}) + 4.74 \times 10^{-3}(\text{SIGMA}) + 1.87 \times 10^{-4}(\text{SPEED}) - 4.14 \times 10^{-8}(\text{PASISP}) - 0.0547 \quad (2)$$

Outlier [a]: WATER [b]	PAA [c]	SIGMA [c]	SPEED [c]	PASISP [c]	ϵ [d]
0.0414	0	6.63	240	0	2.28

[a] Shown also in Table 4, Figs. 3 and 4.

[b] Dependent variable.

[c] Independent variable.

[d] Standardized residual defined in the Appendix.

statistically significant, i.e. different from zero ($P < 0.05$). An intercept value which is close to zero is anticipated since there can be no uptake of **A** in the absence of **A**. Although the two independent variables are highly correlated with each other ($r = 0.887$; $P < 0.0001$), the parameter estimates are acceptable on account of their reasonably low standard errors.

The residual plots are also satisfactory (Fig. 2). Although 3 of the residuals seem to spread out more at the mid-range of the predicted values of PAAUPT, they do not necessarily constitute a typical case of non-constant error variances since they represent only a small fraction of all the residuals. The observed spread of the residuals is more likely to be related to the experimental design or to the sampling technique. In addition, 2 of these 3 residuals have statistically significant standardized values, $\epsilon (P < 0.05)$, indicating that the corresponding observations are outliers (Table 2). However, in the present case, these observations are unlikely to influence the overall statistics to any significant extent since they are located in the mid-range of all the observations (i.e. low leverage) and have relatively small Cook's distances [$D = 0.09 < F_{3,26} = 0.1$; $D = 0.26 < F_{3,26} = 0.25$] (see Appendix for definition), and since their removal from the analysis does not significantly improve r^2 over the existing value of 0.991, which is considered to be excellent. The presence of these outliers may be largely the result of the greater precision of the other observations. The absence of systematic pattern among the residuals and the insignificant results from Durbin-Watson test suggest that the residuals do not exhibit any significant serial correlation. Thus, it can be concluded that Eqn. 1 in Table 2 is adequate for describing the data. An additional useful piece of information that can be derived from multiple regression analysis is the sensitivity of the response variable to each predictor variable. This value can be determined by the regression coefficients of the predictor variables. However, since the predictor variables are expressed in different units and on different scales, their regression coefficients cannot be directly compared without standardization. This procedure is usually accomplished by transforming the values for the response observations (y) and for each predictor variable (x) to quantities akin to the normal standard deviates, and performing the regression analysis on these transformed values, which are characterized by zero mean and unit variance (see Appendix for definition) (Gunst and Mason, 1980).

The standardized regression coefficients thus obtained are 1.23 and −0.273 for PAA and PASISP, respectively, meaning that PAAUPT increases by 1.23 for every unit increase in PAA, but decreases by 0.273 for every unit increase in PASISP.

Incorporation of water

Preliminary analysis indicates that PAA, SIGMA and SPEED are all important predictors for WATER, as shown by their statistically significant regression coefficients at the 5% level. Since only 50% of the total variation are accounted for by the 3 variables, the analysis was extended to include the various interaction terms. For ease of interpretation, restrictions are placed upon selecting only those interaction variables that are significant in the presence of all the 3 single-component variables. Only PASISP has been shown to satisfy this criterion. Table 3 presents the final form of the equation (Eqn. 2) and the related statistics. Eqn. 2 predicts that:

(a) in the absence of **A**, an increase in the stirring speed or in the initial supersaturation will cause an increase in water content;

(b) an increase in the concentration of **A** in solution will lead to a decrease and then an increase in water content (since the water content is more sensitive to the PASISP term than to the PAA term, as reflected by their standardized regression coefficients); and

(c) in the presence of **A**, an increase in the stirring rate or in the initial supersaturation will similarly decrease and then increase the water content.

These predictions are consistent with the experimental observations (Chow et al., 1985; Chow and Grant, 1988a and b).

The regression parameters are statistically significant at the 5% level and have reasonably low standard errors. Except for the correlation between PAA and PASISP ($r = 0.887$) mentioned earlier, PAA, SIGMA and SPEED are essentially independent of each other, as designed a priori for the experimental studies. Based on the *t*-statistics, the relative importance of the predictor variables in influencing the water content of the crystals follows the order: SIGMA > PASISP > SPEED > PAA. On the other hand, the sensitivity of water content to the various crystallization variables, as measured by the absolute magnitudes of the standardized regression coefficients, decreases in the order: PASISP > PAA > SIGMA > SPEED.

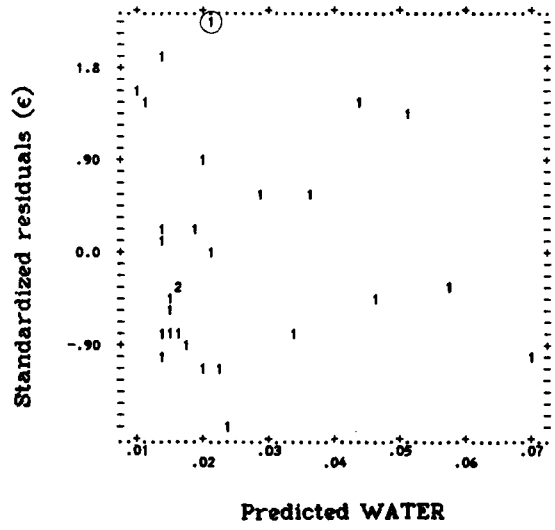

Fig. 3. Standardized residuals plotted against the predicted WATER for the linear regression of WATER on PAA, SIGMA, SPEED and PASISP (Eqn. 2, Table 3). The number 2 represents two virtually coincident points. The outlier is circled.

The residual analysis yields generally satisfactory results. The standardized residuals, ϵ, are randomly distributed about the line, $\epsilon = 0$, forming an approximately horizontal band (Fig. 3). There is no distinct systematic trend in the residuals and so no problems of autocorrelation between residuals and/or of missing variables (e.g. quadratic and interaction terms) need be resolved. However, the circled data point is an outlier (Table 3), as shown by its statistically significant standardized residual ($\epsilon = 2.28$; $P < 0.05$). Whilst atypical, this outlier was not removed from the analysis since it may contain useful information not shared by other observations (Montgomery and Peck, 1982). In addition, its removal led to the appearance of other outliers, which would suggest that some of the data points still show relatively large deviation from the fitted curve. This point has been substantiated by the fact that 24% of the total variation in WATER remain unaccounted for after the regression, which would further indicate that the water content of the crystals is not very well controlled by the conditions of crystalli-

TABLE 4

Parameters and statistics for the linear regression of WATER on PAA, SIGMA, PAASPD, SIGSPD and PASISP (with the outlier)

Variable	Regression coefficient	Standard error	Standardized coefficient	*t*-Statistic	Tolerance
Intercept	-8.66×10^{-3}	5.72×10^{-3}		−1.51	
PAA	1.44×10^{-4}	2.64×10^{-5}	2.58	5.45 *	0.0277
SIGMA	-5.74×10^{-3}	1.34×10^{-3}	−1.10	−4.28 *	0.0937
PAASPD	-4.04×10^{-7}	1.03×10^{-7}	−1.88	−3.91 *	0.0269
SIGSPD	4.21×10^{-5}	5.67×10^{-6}	1.92	7.42 *	0.0928
PASISP	-3.41×10^{-8}	6.66×10^{-9}	−1.28	−5.11 *	0.0984

RMS = 5.30×10^{-5}; $r^2 = 0.858$; d.f. = $n - 6$, where $n = 29$

* Significant at the 5% level.

$$\text{WATER} = 1.44 \times 10^{-4}(\text{PAA}) - 5.74 \times 10^{-3}(\text{SIGMA}) - 4.04 \times 10^{-7}(\text{PAASPD}) + 4.21 \times 10^{-5}(\text{SIGSPD}) - 3.41 \times 10^{-8}(\text{PASISP}) - 0.00866 \quad (3)$$

zation. Although the fitted equation is probably adequate from the standpoint of description, it is not satisfactory enough for prediction of future responses, for which a high multiple r^2 value is desired.

In order to develop an equation which would afford much better predictability, the regression analysis was repeated with no restrictions imposed on the order in which the variables should be selected. All the single-component and interaction variables were included in the analysis with the exception of SPEED and PAASIG, both of which have been found, in this case, to be below the tolerance limits required for accurate computation of the inverse of the covariance matrix. The selection of best subsets of variables is guided by Mallows' C_p and r^2. All the regressors appear to provide essential information about the water content of the crystals, as suggested by their statistically significant regression coefficients ($P < 0.05$). A summary of the statistics for the final equation (Eqn. 3) is given in Table 4.

Eqn. 3 yields a multiple $r^2 = 0.86$, a significant improvement over Eqn. 2 in terms of predictability. However, Eqn. 3 is more difficult to interpret since it involves 3 interaction terms. Also, the negative sign of the regression coefficient for SIGMA would imply that the water content decreases with increasing initial supersaturation, which seems to contradict the observed increases in water content. The tolerance of the various predictor variables is generally low, which is indicative of high correlation among the variables. The residual plots are satisfactory despite the fact that the majority of the residuals tend to concentrate at the low end of the predicted values of WATER (Fig. 4), which is possibly linked to the experimental design or to the sampling technique. As before, the same circled data point manifests itself as an outlier, having a standardized residual

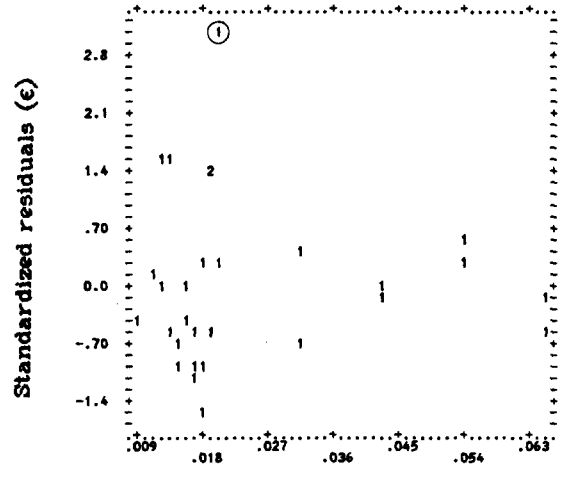

Fig. 4. Standardized residuals plotted against the predicted WATER for the linear regression of WATER on PAA, SIGMA, PAASPD, SIGSPD and PASISP, with circled outlier (Eqn. 3, Table 4). The number 2 represents two virtually coincident points.

TABLE 5

Parameters and statistics for the linear regression of WATER on PAA, SIGMA, PAASPD, SIGSPD and PASISP (without the outlier)

Variable	Regression coefficient	Standard error	Standardized coefficient	*t*-statistic	Tolerance
Intercept	-1.39×10^{-2}	4.66×10^{-3}		-2.97 *	
PAA	1.56×10^{-4}	2.09×10^{-5}	2.78	7.46 *	0.0282
SIGMA	-6.15×10^{-3}	1.05×10^{-3}	-1.19	-5.83 *	0.0937
PAASPD	-4.21×10^{-7}	8.08×10^{-8}	-1.95	-5.20 *	0.0279
SIGSPD	4.52×10^{-5}	4.50×10^{-6}	2.08	10.0 *	0.0916
PASISP	-3.63×10^{-8}	5.23×10^{-9}	-1.36	-6.93 *	0.1008

RMS = 3.30×10^{-5}; $r^2 = 0.914$; d.f. = n-6, where $n = 28$.
* Significant at the 5% level.

$$\text{WATER} = 1.56 \times 10^{-4}(\text{PAA}) - 6.15 \times 10^{-3}(\text{SIGMA}) - 4.21 \times 10^{-7}(\text{PAASPD}) + 4.52 \times 10^{-5}(\text{SIGSPD}) - 3.63 \times 10^{-8}(\text{PASISP}) - 0.0139 \quad (4)$$

of $\epsilon = 3.09$, which is much higher than that based on Eqn. 2 (see Table 3). The removal of this data point from the analysis affords an $r^2 = 0.91$, an increase of 5%. The modified regression equation (Eqn. 4) and related statistics are shown in Table 5, while the residuals are plotted in Fig. 5. Since the regression statistics and analysis of the residuals yield satisfactory results, no further analysis is required. For simplicity and for the ease of explanation, Eqn. 2 is the equation of choice while Eqn. 4 would be preferable for the sole purpose of prediction.

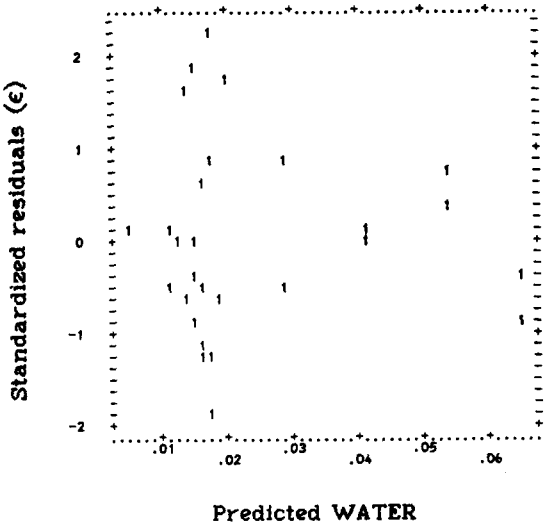

Fig. 5. Standardized residuals plotted against the predicted WATER for the linear regression of WATER on PAA, SIGMA, PAASPD, SIGSPD and PASISP, without outlier (Eqn. 4, Table 5).

Length-to-width ratio

Regression of LWR against PAA, SIGMA and SPEED using the FS-BE procedure results in the retention of PAA and SIGMA in the final equation, SPEED being insignificant in the presence of PAA and SIGMA. Since the LWR depends strongly on PAA, and since the effects of SPEED and SIGMA on the LWR are chiefly mediated through PAA, the regression analysis was repeated with the inclusion of the various interaction variables. Only PAA and PAASIG remained in the equation (Eqn. 5 in Table 6) after the FS-BE procedure. The statistical results in Table 6 agree closely with the experimental observations that the length-to-width ratio of the crystals increases with increasing concentration of **A** in the crystallization solution, but decreases with increasing initial supersaturation (Chow and Grant, 1988b) at a fixed concentration of **A** (500 mg · dm^{-3}). The use of C_p statistic as a selection criterion also yields the same results. (However, for the same reasons indicated earlier, C_p is not a very useful indicator of the best subset of variables in the present analysis.)

The parameter estimates are sufficiently precise, as shown by their low standard errors. The residual plots appear to give some indication of systematic pattern (Fig. 6). However, the Durbin-Watson test for serial correlation in this case is

TABLE 6

Parameters and statistics for the linear regression of LWR on PAA and PAASIG

Variable	Regression coefficient	Standard error	Standardized coefficient	*t*-Statistic	Tolerance
Intercept	1.17	0.472		2.49 *	
PAA	0.0324	0.00229	1.62	14.2 *	0.2182
PAASIG	−0.00199	0.000271	−0.840	−7.35 *	0.2182

RMS = 3.14; $r^2 = 0.926$; d.f. = $n - 3$, where $n = 29$.
* Significant at the 5% level.

$$LWR = 0.0324(PAA) - 0.00199(PAASIG) + 1.17 \quad (5)$$

Outliers: LWR [a]	SPEED [b]	SIGMA [b]	PAA [c]	PAASIG [c]	ϵ [d]
6.95	200	6.63	500	3315	−2.22
8.43	240	5.34	500	2670	−2.14
18.83	240	6.63	700	4641	2.51

[a] Dependent variable.
[b] Conditions not explicitly in Eqn. 5, although SIGMA is implicit in the interaction term, PAASIG.
[c] Independent variable.
[d] Standardized residual defined in the Appendix.

inconclusive. Since there are only a few data points at the upper end of the predicted LWR values, the suggestion of systematic trend in the residuals cannot be unequivocally established. To examine this point further, various powers of PAA (i.e. PAA^2, PAA^3,...) were introduced into the equation, but produced only small changes in the RMS, r^2 or the existing residual pattern. Thus, Eqn. 5 was considered adequate for explaining the data. Although 3 outliers (Table 6) were also identified, they were not deleted from the analysis since they may control some of the key properties of the model equation (Montgomery and Peck, 1982). Judging from the *t*-statistic in Table 6, PAA is more important than PAASIG in determining the LWR. The relatively high absolute value of the standardized regression coefficient for PAA also suggests that LWR is relatively sensitive to changes in PAA.

Entropy of fusion

The analysis in the present case is fairly straightforward. As established in previous reports (Chow et al., 1985; Chow and Grant, 1988a and b), the entropy of fusion of the crystals depends largely on the ideal molar entropy of mixing and,

TABLE 7

Parameters and statistics for the linear regression of ENTFUS on ENTMIX, PAA, SIGMA and SPEED

Variable	Regression coefficient	Standard error	Standardized coefficient	*t*-Statistic	Tolerance
Intercept	68.5	1.27		54.0 *	
ENTMIX	−4.49	0.523	−0.695	−8.58 *	0.5929
PAA	0.00164	0.000693	0.167	2.37 *	0.7873
SIGMA	−0.229	0.0668	−0.249	−3.43 *	0.7395
SPEED	−0.0111	0.00402	−0.194	−2.77 *	0.7935

RMS = 1.04; $r^2 = 0.907$; d.f. = $n - 5$, where $n = 29$.
* significant at the 5% level.

$$ENTFUS = 68.5 - 4.49(ENTMIX) + 0.00164(PAA) - 0.229(SIGMA) - 0.0111(SPEED) \quad (6)$$

to a much lesser extent, on the conditions of crystallization, suggesting that the extent of lattice disruption in the **P** crystals cannot be solely explained by the presence of **A** and water in the crystals. The defects arising from crystallization alone (i.e. growth defects) may not be accounted for (Mullin, 1972). Assuming that a proportionate relationship exists between the conditions of crystallization and the concentrations of growth defects generated, regression analysis, as shown in Table 7, was carried out with ENTMIX, PAA, SIGMA and SPEED as the regressors. All the predictor variables were shown to be important determinants of ENTFUS, as evidenced by their statistically significant regression coefficients ($P < 0.05$). The use of interaction variables in place of single-component variables exerted little influence on r^2, RMS or the pattern of residuals. Thus, Eqn. 6 in Table 7 with the ENTMIX, PAA, SIGMA and SPEED terms is deemed the best for explaining the variation in ENTFUS. Eqn. 6 indicates that an increase in ENTMIX, SIGMA and/or SPEED would result in a decrease in ENTFUS, while an increase in PAA would give rise to an increase in ENTFUS; these conclusions

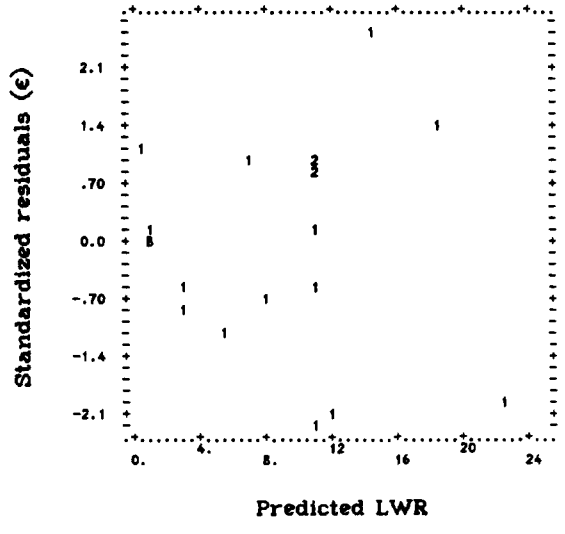

Fig. 6. Standardized residuals plotted against the predicted LWR for the linear regression of LWR on PAA and PAASIG (Eqn. 5, Table 6). The symbol 2 and B represent 2 and 11 virtually coincident points, respectively.

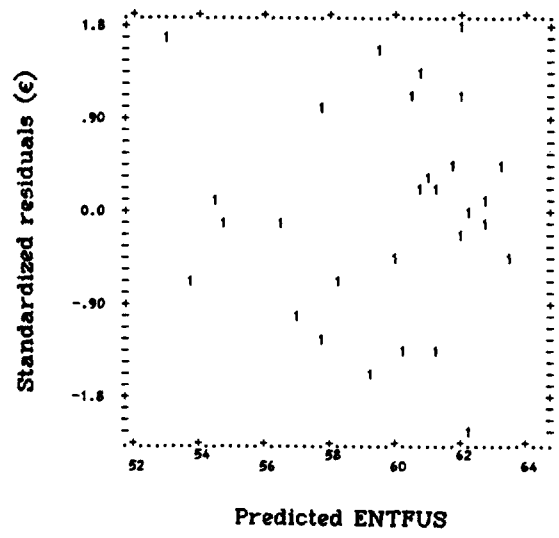

Fig. 7. Standardized residuals plotted against the predicted ENTFUS for the linear regression of ENTFUS on ENTMIX, PAA, SIGMA and SPEED (Eqn. 6, Table 7).

are in close conformity with the experimental findings detailed in previous reports (Chow et al., 1985; Chow and Grant, 1988a and b). As shown in Table 7, the regression coefficients have reasonably good precision, as suggested by their small standard errors. The residual plots are quite satisfactory (Fig. 7). Ordered arrangement of the standardized residuals is not observed; they are randomly scattered about the line, $\epsilon = 0$, and are contained within a narrow horizontal band. There is, however, one outlier, which has a standardized residual of $\epsilon = 2.08$. Being a borderline case and for the reasons given earlier, the outlier was retained in the present analysis. The relative importance of the predictor variables in controlling ENTFUS, as delineated by the t-statistics, follows the order: ENTMIX > SIGMA > SPEED > PAA. The sensitivity of ENTFUS to the various predictor variables also conforms to the same decreasing order, as determined by the absolute values of the standardized regression coefficients. The regression coefficient for ENTMIX (i.e. the negative value of the disruption index) based on the pooled data in the present analysis is obviously lower than those calculated for individual crystallization cases (Chow et al., 1985; Chow and Grant, 1988a

and b), suggesting that the disruption indices (York and Grant, 1985; Grant and York, 1986) in the latter cases measure not only the crystal defects induced by incorporated **A** and water, but also those emanating from growth (Mullin, 1972).

Conclusions

The influence of the various conditions of crystallization on the uptake of **A** and water, length-to-width ratio (extent of habit elongation) and the entropy of fusion of the crystals has been evaluated statistically using multiple linear regression analysis. The uptake of **A** is determined largely by the concentration of **A** in the solution, and much less so, by the stirring rate and/or by the initial supersaturation. The extent of habit elongation also depends strongly on the concentration of **A** in the solution, and to a lesser extent, on the initial supersaturation but only in the presence of **A**, i.e. the latter only plays a mediating role. On the other hand, the uptake of water appears to depend, in ascending order of importance, on the additive concentration, stirring rate and initial supersaturation. On the basis of the multiple r^2 and residual mean square, both the additive uptake and the length-to-width ratio appear to be determined mainly by the conditions of crystallization, while the water content seems to be less so. The entropy of fusion is governed by the ideal molar entropy of mixing, and considerably less so, in descending order of significance, by the initial supersaturation, stirring speed and concentration of **A**. The sensitivity of the entropy of fusion to these various regressors also follows the same order of decrease, as reflected by the absolute magnitudes of the standardized regression coefficients. The regression coefficient for the ideal molar entropy of mixing term derived from pooled data is significantly lower than those determined for separate crystallization cases reported previously (Chow et al., 1985; Chow and Grant, 1988a and b), suggesting that the disruption indices (York and Grant, 1985) in the latter cases reflect not only the impurity defects, but possibly also the growth defects (Mullin, 1972).

Appendix

Definitions of some statistical terms used in linear regression analysis.

Multiple r^2

$$r^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

where RSS is the residual sum of squares after regression and TSS is the total sum of squares without regression.

Adjusted r^2

$$r^2_{\mathrm{adj}} = 1 - \frac{s^2_{\mathrm{y.x}}}{s^2_{\mathrm{y}}} = r^2 - \frac{p'(1-r^2)}{(n-p'-1)}$$

where $s^2_{\mathrm{y.x}}$ is the variance after regression (i.e. residual mean square, RMS), s^2_{y} is the variance without regression (or variance of observations), n is the number of observations and p' is the number of parameters (excluding intercept).

Mallows' C_p

$$C_{\mathrm{p}} = \frac{\mathrm{RSS}}{s^2} - (n - 2p)$$

where RSS is the residual sum of squares for the best subset being tested, p is the number of parameters (including the intercept) and s^2 is the RMS based on the regression using all independent variables. If the selected subset of variables truly affords the "best" estimate of the population residuals variance, RSS will be equal to the product of s^2 and p, and C will therefore be equal to p.

Standardized (Studentized) residual

$$\epsilon = e_{\mathrm{i}}/s_{\mathrm{ei}}$$

where e_{i} is the raw residual for the i^{th} observation and s_{ei} is the standard error of the residual for the i^{th} observation.

Cook's distance

$$D = \left(\frac{\epsilon^2}{p}\right)\left(\frac{h_i}{1-h_i}\right)$$

where h_i is the leverage for the i^{th} observation and ϵ and p are defined above.

Standardized regression coefficient

$$\beta' = \beta\sqrt{\left(\frac{\sum(x_i - x)^2}{\sum(y_i - y)^2}\right)}$$

where β is the ordinary (unscaled) regression coefficient.

Tolerance

$$T_j = 1 - R_j^2$$

where R_j^2 is the squared multiple correlation of x_j variable with the other independent variables.

Durbin-Watson statistic

$$d = \frac{\sum(e_i - e_{i-1})^2}{\sum e_i^2}$$

where e_i is the residual for the i^{th} observation and e_{i-1} is the residual for the $(i-1)^{th}$ observation.

Acknowledgements

We thank Frank W. Horner Ltd., Montréal, Québec and McNeil Consumer Products Co., Guelph, Ont., for gifts of materials. We also thank the Medical Research Council of Canada for an operating grant for D.J.W.G. (MT-7835) and the Ontario Ministry of Colleges and Universities for an Ontario Graduate Scholarship for A.H.-L.C.

References

BMDP Statistical Software Manual, University of California Press, Berkeley, CA, U.S.A. 1983.

Chatterjee, S. and Price, B., *Regression Analysis by Example*, John Wiley, NY, 1977, pp. 193–214.

Chow, A.H.-L., Chow, P.K.K., Wang Zhongshan and Grant, D.J.W., Modification of acetaminophen crystals: influence of growth in aqueous solutions containing *p*-acetoxyacetanilide on crystal properties. *Int. J. Pharm.*, 24 (1985) 239–258.

Chow, A.H.-L. and Grant, D.J.W., Modification of acetaminophen crystals. II. Influence of stirring rate during solution-phase growth on crystal properties in the presence and absence of *p*-acetoxyacetanilide. *Int. J. Pharm.*, 41 (1988a) 29–39.

Chow, A.H.-L. and Grant, D.J.W., Modification of acetaminophen crystals. III. Influence of initial supersaturation during solution-phase growth on crystal properties in the presence and absence of *p*-acetoxyacetanilide. *Int. J. Pharm.*, 42 (1988b) 123–133.

Grant, D.J.W. and York, P., A disruption index for quantifying the solid state disorder induced by additives or impurities. II. Evaluation from heat of solution. *Int. J. Pharm.*, 28 (1986) 103–112.

Gunst, R.F. and Mason, R.L., *Regression Analysis and its Applications*, Marcel Dekker, New York, N.Y., 1980, pp. 73–77.

Montgomery, D.C. and Peck, E.A., *Introduction to Linear Regression Analysis*, Wiley, New York/Toronto, 1982, pp. 57–59.

Mullin, J.W., *Crystallisation*, 2nd edn., Butterworth, London, 1972, pp. 136–173.

York, P. and Grant, D.J.W., A disruption index for quantifying the solid state disorder induced by additives or impurities. I. Definition and evaluation from heat of fusion. *Int. J. Pharm.*, 25 (1985) 57–72.