DATA VALIDATION PROCEDURES USED FOR KOVACIC, MARSHALL, AND MEURER ARTICLE

VOLUME 10 EDITORIAL BOARD OF THE NYU JOURNAL OF INTELLECTUAL PROPERTY AND ENTERTAINMENT LAW (JIPEL)

As a policy of the journal, JIPEL provides readers with a short appendix that supplements authors' empirical analysis and attempts to validate a sample sets of findings, where possible. For a description of JIPEL's policy, please see the journal's Fall 2020 issue editorial on the subject.

In order to validate the authors' empirical analysis contained in this Article, journal staff reviewed the authors' patent tabulations for a subset of chemicals under the assumption that the accuracy of the coding of this subset is representative of the accuracy of the coding of all the chemicals.[1] Per the request of the JIPEL editors, the authors provided the journal a complete disaggregation of patent counts by chemical product. In its review, journal staff validated patent tabulations across all firms for three chemicals, Methacrylates, Polyethylene, and Polypropylene, which were associated with a total of 855 "results."[2] The total population of coded "results" numbered 6,121.[3] A "result" is defined as *one coded finding* for patenting by a firm on a chemical product in a single year, distinguished from "patent tabulation," which refers to the recorded *number of patents* sought for that firm / chemical / year. So, for example, BASF may have sought multiple patents related to a given chemical in a single year, but this would be considered one "result." JIPEL drew this distinction since it was interested in reviewing the potential error rate on the authors' findings by "result" as well as by patent tabulation, shown in Tables 1 and 2 below.

From this review, JIPEL staff did find slight discrepancies associated with approximately 31% of "results" across Methacrylates, Polyethylene, and Polypropylene, as shown in Table 2.[4] That said, these discrepancies tended to be in the amount of one to three patents greater or fewer than the authors' tabulated

---

[1] *See, e.g.*, *Sample Size Calculator*, CLINCALC, https://clincalc.com/stats/samplesize.aspx (last visited June 1, 2021) (describing a means to calculate minimum experiment sizes for a known population size). While JIPEL and the authors both followed the Article's Appendix B to architect their patent tabulations, it is possible that the errors that affected some or all of the three chemicals reviewed by JIPEL were dissimilar to errors that affected other studied chemicals.

[2] Methacrylates, Polyethylene, and Polypropylene were associated with 286, 261, and 308 "results," respectively.

[3] The "results" from the remaining chemicals totaled 5,292 "results."

[4] In total, JIPEL found discrepancies associated with 269 "results" across the three chemicals. Dividing 269 by 855 "results" gives a discrepancy rate of approximately 31%.

findings for patenting in a particular year. Thus, on net, JIPEL's total tabulated findings did not tend to be very different than the authors' findings. As shown in Table 1 below, in all periods, the authors' counts did not exceed the JIPEL's counts. And, the findings for the total number of patenting in the pre-plea, plea and post-plea periods tended to be very close.

TABLE 1: SUM OF PATENTING ACROSS FIRMS FOR A GIVEN CHEMICAL IN EACH PERIOD, SHOWING NET DIFFERENCE ("DIFF.") IN SUMMED TOTALS BETWEEN ARTICLE AUTHORS AND JIPEL

	Methacrylates			**Polyethylene**			**Polypropylene**		
	Authors	*JIPEL*	*Diff.*	*Authors*	*JIPEL*	*Diff.*	*Authors*	*JIPEL*	*Diff.*
Pre-plea	1688	1718	30 (1.78%)	353	362	9 (2.55%)	174	174	0 (0)
Plea	931	943	12 (1.29%)	934	973	39 (4.18%)	439	445	6 (1.37%)
Post-plea	1215	1292	77 (6.34%)	1774	1831	57 (3.21%)	1065	1084	19 (1.78%)

JIPEL also disaggregated its own tabulated errors on "results" by core versus non-core producers, as shown in Table 2, to determine if errors were any likelier for one set of firms versus the other.[5] JIPEL did observe greater errors in patenting "results" for core producers, but again, the magnitude of these errors remained very small, as seen in Table 1. JIPEL did not observe any greater magnitude of errors associated with "results" for core producers versus non-core producers.

[5] The authors explain their rationale for distinguishing between "core" and "non-core" producers in Section I of the main Article.

TABLE 2: JIPEL OBSERVED ERROR COUNTS FOR REVIEWED "RESULTS," SPLIT BETWEEN ERRORS ASSOCIATED WITH "RESULTS" FOR CORE AND NON-CORE PRODUCERS[6]

	Methacrylates		**Polyethylene**		**Polypropylene**	
	"Result" count (%)	*Error count (%)*	*"Result" count (%)*	*Error count (%)*	*"Result" count (%)*	*Error count (%)*
Core producer "results" and JIPEL observed errors	130 (45.45%)	80 (61.77%)	145 (55.56%)	49 (57.65%)	140 (45.45%)	28 (52.83%)
Non-core producer "results" and JIPEL observed errors	156 (54.55%)	51 (38.23%)	116 (44.44%)	36 (42.35%)	168 (54.55%)	25 (47.17%)
Total "results" and JIPEL observed errors	286	131	261	85	308	53

In sum, JIPEL finds that the aggregate differences in the number of patents recorded by the journal staff and the authors does not materially change the magnitude or direction of the findings for any of the three chemicals examined. Based on our assumption that discrepancies in the patents tabulated for these three chemicals by the authors and the JIPEL staff are representative of the magnitude of discrepancies for all the chemicals examined by the authors in this article, JIPEL data validation supports the authors' empirical analysis.

Some theories for why these errors persist include errors from human coding or errors in Google's automated document reading, which also automatically translates patent information across languages.[7] Errors might also be due to Google's "deduplication by family" option, which was turned on for the authors' and JIPEL's searches. This option is supposed to group together equivalent inventions and hide redundant patents from view.[8] It is possible that certain patents were hidden for the

[6] As noted above, JIPEL found discrepancies associated with 269 "results" across the three chemicals, the sum of 131, 85 and 53, shown in Table 2. Dividing 269 by 855 total "results" (the sum of 286, 261 and 308, shown in Table 2) gives a discrepancy rate of approximately 31%. In Table 2, JIPEL disaggregated "results" and its error rate on "results" by core and non-core producers. Percentages in Table 2, then, reflect the distribution of core versus non-core producer "results" and errors on "results" from JIPEL's analysis. The overall discrepancy rate remains 31%.

[7] *See About Google Patents: Coverage*, GOOGLE, https://support.google.com/faqs/answer/7049585 (last accessed June 1, 2021) (describing Google's process to upload and make available for digital searching 120 million global patents).

[8] *See About Google Patents: Search results page*, GOOGLE, https://support.google.com/faqs/answer/7049588/search-results-page?hl=en&ref_topic=6390989 (last accessed June 1, 2021). In its description of its deduplication by patent family option, Google

authors' searches that were visible to JIPEL, based on JIPEL performing its searches at a different time than the authors.

describes how similarly architected searches may nonetheless lead to slightly dissimilar conclusions. *Id.* The company observes how when using deduplication by family:

> Only the highest-ranking patent from the same "simple patent family" is displayed and the other family members are removed from the results list. The simple patent family is all of the patents that share the same set of priority claims. This is usually when the same or very similar patent is filed in more than one country.

Id. This grouping is done algorithmically using what Google describes as Cooperative Patent Classification (CPC) codes. *Id.* For further description of how patent families are created for global patents that seek protection for equivalent inventions, *see DOCDB Simple Patent Family*, EUR. PAT. OFF., https://www.epo.org/searching-for-patents/helpful-resources/first-time-here/patent-families/docdb.html (last accessed June 3, 2021).